Perception and production of English vowels by Chilean learners of English: effect of auditory and visual modalities on phonetic training

Yasna Pereira

A thesis submitted in fulfilment of requirements for the degree of
Doctor of Philosophy

to

Department of Speech, Hearing and Phonetic Sciences
University College London (UCL), London, UK

2013

Declaration

I, Yasna Pereira, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Yasna Pereira

Abstract

The aim of this thesis was to examine the perception of English vowels by L2 learners of English with Spanish as L1 (Chilean-Spanish), and more specifically the degree to which they are able to take advantage of visual cues to vowel distinctions. Two main studies were conducted for this thesis. In study 1, data was collected from L2 beginners, L2 advanced learners and native speakers of Southern British English (ENS). Participants were tested on their perception of 11 English vowels in audio (A), audiovisual (AV) and video-alone (V) mode. ENS participants were tested to investigate whether visual cues are available to distinguish English vowels, while L2 participants were tested to see how sensitive they were to acoustic and visual cues for English vowels.

Study 2 reports the outcome of a vowel training study. To compare the effect of different training modalities, three groups of L2 learners (beginner level) were given five sessions of high-variability vowel training in either A, AV or V mode. Perception and production of English vowels in isolated words and sentences was tested pre/post training, and the participants' auditory frequency discrimination and visual bias was also evaluated. To examine the impact of perceptual training on L2 learners' vowel production, recordings of key words embedded in read sentences were made pre and post-training. Acoustic-phonetic analyses were carried out on the vowels in the keywords. Additionally, the vowels were presented to native listeners in a rating test to judge whether the perceptual training resulted in significant improvement in intelligibility.

In summary, the study with native English listeners showed that there was visual information available to distinguish at least certain English vowel contrasts. L2 learners showed low sensitivity to visual information. Their vowel perception improved after training, regardless of the training mode used, and perceptual training also led to improved vowel production. However, no improvement was found in their overall sensitivity to visual information.

Acknowledgements

I would like to express my deepest gratitude to my main supervisor, Professor Valerie Hazan, for her guidance, support and enlightening teaching all along my PhD years. I am grateful for her contribution to my research experience; she always showed appreciation and respect for my ideas, helped me become more critical when needed and encouraged me to take new challenges. All of this has made my coming to UCL a great personal and professional experience.

I am also thankful to a number of other people who have helped me with different aspects of my research. First, Paul Iverson, my secondary supervisor, who kindly adapted his Vowel Trainer software to meet the requirements of my research in the main vowel training study. I am also thankful for all his help and support regarding the statistical analysis of the data presented in this thesis. While preparing my experiments, I received help from many other kind souls. I owe my gratitude to Andrew Faulkner for his support with the processing of audio-visual files and for allowing the use of material developed for previous studies (BKB sentences). I was also lucky that Stuart Rosen and Mike Coleman had developed a frequency discrimination test and kindly allowed me to use it for my study. On using the Lab and gadgets, Steve Nevard was the best teacher ever. He kindly put up with my non-techy experience and helped me on countless occasions when everything seemed to be going wrong. Finally, I would also like to thank Mike Coleman for developing a platform for the True-or-False test used in this research. His help came when we were desperate to find a way to present this new material.

This research project would not have been possible without the generous help of some of my colleagues in Concepción who persuaded their students into taking part in the experiments that make up this thesis. My eternal gratitude to Paola Fanta, Juan Pablo Cerda, Javiera Otárola, Marcela E. Cabrera, Mirsa Beltrán and Olga Roca. There were around 115 participants whose names need to remain anonymous to follow the ethics code of practice from UCL; without the generous and motivated participants my research idea would not have made any progress.

Along this PhD journey, the personal and professional learning also came from my PhD colleagues who made this time an unforgettable experience. I will take with me countless memories of support, learning and sharing; to all of them my eternal gratitude. Special thanks for their technical support to José Joaquín Atria (Praat scripts), Mark Wibrow (adding noise to files), Sonia Granlund (academic and personal advice), to my "R" guru Yasuaki Shinohara and to Dorothea Hackman (for English language advice). And to all the guys who make 326 a wonderful environment to work: Tim Schoof, Young-Shin Kim, Georgina Oliver, Wafa Alshangiti, Katharine Mair, Lucy Carrol, Emma Brint, Nada Al-Sari, Louise Stringer, Mauricio Figueroa, Csaba Redey-Nagy, Albert Lee, Dong-Jin Shin and Kurt Steinmetzger.

Miles away across the Atlantic, there is also a number of people in Chile who supported my dream of coming to UCL to do my PhD. I will always be thankful for the support and trust of Maria Edith Larenas and Marcela Cabrera A., Head of the Foreign Languages Department at Universidad de Concepción (previous and current Head, respectively). I should also thank my friends Hernán Pérez and Jaime Soto who encouraged me to pursue further studies in my MA and later in this PhD; they are and will always be my best inspiration.

The funding for this PhD programme came from BecasChile, a Chilean government funding body, and also from Universidad de Concepción. The latter granted my study leave from work. I am grateful for their financial support without which my PhD studies would not have been possible.

Finally, my personal and emotional strength comes from my family. Their support has been essential during these years. I am grateful to have their encouragement and trust in every personal and professional project I have started.

Contents

List of Figures

List of Tables

Thesis overview

The main aim of this thesis was to examine the perception of English vowels by L2 learners of English with Chilean-Spanish as native language (L1), and specifically to what extent they are able to use visual cues to perceive vowel distinctions. In this section, the overall structure of this thesis will be briefly described to anticipate what will be explained in more detail in each chapter later. Tests, aims and participants who took part in the different studies will be presented.

Summary of chapters

This thesis includes five chapters. In Chapter 1, research studies relevant to the main aim of this thesis will be discussed. The general topics which will be covered are L2 speech perception, use of visual information in L2 speech perception, L2 vowel training and L2 speech perception and production change after training.

Chapter 2 describes results from a study which examined the extent to which L2 learners with L1-Spanish are sensitive to visual cues for English vowel contrasts. Additionally, the contribution of visual information to vowel distinction for English native speakers was also explored. A vowel perception test was given three groups of participants: two groups of L2 learners of English (Chilean-Spanish L1) with different levels of proficiency (beginner, advanced) and a group of native speakers of English (ENS). This vowel perception test was presented to the L2 learners in three modes: audio (A), audio-visual (AV) and video-alone (V) modes in clear. The same test was presented in noise (A, AV) to a group of native speakers of English (ENS) to find if visual information improved vowel identification. Further tests were presented to gain information about the participants' auditory and visual abilities. To measure visual bias in L2 learners, a McGurk test was used. A frequency discrimination test was used to measure L2 learners' auditory perception of small differences in frequency. A BKB-sentence test was used to measure keyword intelligibility in sentence material in A and AV mode. A summary of tests, aims and participants is presented below in Table A.

Table A Summary of tests, aims per test and group of participants which appear in Chapter 2.

Test	**Aim**	**Participants**
Frequency discrimination test	Measure of auditory frequency discrimination capacity	47 L2 beginner learners 37 L2 advanced learners
McGurk test	Obtain a measure of Visual bias	47 L2 beginner learners
BKB sentence test	Measure capacity to perceive key-words in sentence material in A and AV mode.	47 L2 beginner learners
Vowel test (In Clear for L2 learners In noise for ENS)	Measure perception of English vowels in A, AV and V mode in CVC words.	47 L2 beginner learners 37 L2 advanced learners 20 ENS

Chapter 3 presents the results of a study which compared three vowel training modalities and their impact in the perception of English vowels by L2 learners. The three modalities used were: auditory (AT), audio-visual (AVT) and video-only (VT) training. Three groups of L2 learners of English (total: 47 participants) with beginner-proficiency level (Chilean Spanish L1) participated in the training (Table B). Most of the beginner learners had participated in the experiment described in Chapter 2; however, they were tested again for this study which was conducted 6 months later. The test battery used in Chapter 2 was used again as pre and post training measures. A new sentence test, True-or-False sentence test, was added to this test battery to assess the effect of training on the perception of English vowels beyond the isolated-word level. This test presented minimal-pair words in sentences and relied on the accurate identification of vowels for a correct answer as to whether the sentence was meaningful ('true') or semantically-unpredictable ('False'). A summary of tests, aims and participants is presented in Table C.

Table B Summary of the vowel training programmes and material. The list of words are the same for all three training programmes.

Training modality	**Material**	**Participants**
Audio training (AT) (5 sessions)	Audio recordings of: 140 words (fixed tokens, 14 vowels x10 words) 85 words (adaptive procedure)	17 L2 beginner learners
Audio-visual training (AVT) (5 sessions)	Video recordings of: 140 words (fixed tokens, 14 vowels x10 words) 85 words (adaptive procedure)	14 L2 beginner learners
Video-only training (VT) (5 sessions)	Video (no sound) recordings of: 140 words (fixed tokens, 14 vowels x10 words) 85 words (adaptive procedure)	16 L2 beginner learners

Table C Summary of tests, aims per test and group of participants which appear in Chapter 3. These tests were used as pre and post tests for the L2 beginner group who took the Vowel training sessions. The tests are listed in the order they were presented to participants.

Test	**Aim**	**Participants**
Frequency discrimination test	Same as in Table A	47 L2 beginner learners
McGurk test	Same as in Table A	47 L2 beginner learners
Vowel test In Clear for L2 learners	Same as in Table A	47 L2 beginner learners (37 had participated in the study reported in Chapter 2)
BKB sentence test	Same as in Table A	47 L2 beginner learners
True-or-False (TF) sentence test Vocabulary test **	Measure perception of English vowels in sentence material. Check knowledge of the vocabulary used in the TF test	47 L2 beginner learners ** 13 Control (beginner learners) ** 37 L2 advanced learners 20 ENS

Chapter 4 addresses the issue of the impact of the perceptual training on L2 vowel production and the perception-production link. Participants in the training study described in Chapter 3 were recorded reading key-words in carrier sentences before and after they were given one type of vowel training programme. These materials were then used in a goodness-rating study presented to English native speakers (ENS) to judge whether the keywords produced post-training were perceived as more native-like than those produced before training. Additionally, spectral and duration measures are reported together with the relation between vowel identification and vowel production measures (Table D).

Table D Summary of information of measures (F1, F2 and duration of vowels) and goodness rating test presented in Chapter 4.

Test/measure	**Material**	**Participants**
Measure: F1, F2 of vowels Aim: find whether spectral change has taken place after training	vowel tokens recorded at pre and post test (3x11x2) (keywords recorded in read sentences)	47 L2 beginners
Measure: Duration of vowels Aim: find whether duration change has taken place after training	vowel tokens at pre and post test (3x11x2) (same used for F1, F2) from Pre and post test	47 L2 beginners
Goodness rating test Aim: find how English native speakers (ENS) rated L2 vowel production before and after training	2068 tokens (2 tokens per vowel at pre and post test for the 47 participants (2x11x2x47). Tokens were the same for all participants.	11 ENS

Finally, Chapter 5 presents the general discussion based on the findings of chapters 2, 3, and 4 of this thesis. Limitations of the current study and suggestions for future research will also be addressed in this final chapter.

Chapter 1

Introduction

The aim of this thesis is to investigate the perception of English vowels by learners of English (L2 learners) with Spanish as their native language (L1). It has been suggested that English vowels may be misperceived by non-native speakers due to the assimilation of two English phonemes to a single-phoneme category in the learners L1 (Best, 1995); this may cause misunderstandings and communication breakdown. Research in second language (L2) speech perception has found that visual information can generally aid L2 speech perception, particularly when the contrasts are visually salient (Hardison, 1999; Hazan, Sennema, Faulkner, Ortega-Llebaria, Iba & Chung, 2006). However, the extent to which visual cues may help L2 learners in the perception of English vowels has not been yet established; as most of the studies using visual information have focused on consonant contrasts. One of the main objectives of this study is to establish whether L2 learners are sensitive to visual cues to English vowel contrasts and if not, whether it is possible to train them in the use of the visual information available for English vowel perception.

1.1 Sources of information in speech

Understanding the information present in speech that may cue the identity of a speech sound is a complex process. The various sources of information that are used to decode speech may be grouped into "bottom-up information" that is present in the incoming signal

and “top-down information” which involves the perceiver’s stored linguistic knowledge (Cook & Ellis, 2001). These two concepts were presented in the context of auditory perception, but an expansion of the content of the input to the visual domain is needed in the context of this study. The speech signal not only comprises the acoustic information which carries linguistic and indexical properties (cueing phoneme and speaker identity) but also visual information (e.g. lip and jaw movements) of speech (McGurk & MacDonald, 1976; Rosenblum, 2005; 2008). On the other hand, the top-down information domain includes phonetic, phonological, lexical, syntactic and semantic knowledge together with the cognitive processes needed to decode speech. These sources of information and the cognitive processes involved interact in complex ways which make establishing clear boundaries a hard task (Cook & Ellis, 2001).

Native speakers integrate linguistic and non-linguistic information when perceiving speech (Nygaard, 2005) and make use of the range of visual information available (Reisberg, McLean & Goldfield, 1987). However, studies on non-native speakers’ speech perception have found that integrating the information available may not take place in exactly the same fashion as in the L1, leading to problematic issues which concern L2 speech perception.

1.2 L2 speech perception

It is widely accepted that L2 speech perception presents various challenges to non-native speakers. It has been suggested that speech perception in an L2 is somehow similar to hearing in an adverse condition in the L1 as learners have to face the task of dealing with an “imperfect signal and imperfect knowledge” of the L2 (García-Lecumberri, Cook & Cutler, 2010, p. 864). It is an imperfect signal in the sense that the sounds in the input do not necessarily match phoneme categories that L2 perceivers have as mental representations of sounds in their native language. Together with this poor signal and incomplete knowledge of the target language, there exist a number of other factors that interact and can make L2 speech perception more problematic.

1.2.1 Factors affecting L2 speech perception

A wide number of factors and interactions affecting L2 speech perception have been found in L2 studies. Strange (1992) suggested that the way these factors interact may explain the great variability in the results of L2 speech perception studies with adult learners. The author proposed an adapted version of Jenkins' (1979) model, the "Tetrahedral model for cross-language perception". In this adapted version, Strange presented four groups of factors that interact in L2 speech perception. a) *Subject variables* which refer to L1 experience, L2 experience, starting age and aptitude for learning a second language. b) *Training task variables* which include L2 learning instruction experience, types of tasks used in laboratory procedures (e.g. identification and discrimination tasks), types of input and speakers' variability, set of stimulus, speakers' number and presentation criteria. c) *Criterial task variables* which refer to laboratory procedures (as in 2) but now related to tasks that target cognitive load and task which aim at the transfer of perceptual learning. Finally, the last group of variables concerns d) the *Stimulus variables.* In this group the author included variables as the type of contrast, relation between the L1 and L2 categories, phonetic and phonotactic context of the contrasts and the nature of the stimulus (natural vs. synthetic stimulus, single vs. multiple cues). These variable interactions would predict complex outcomes and may help to explain the large individual variability in the results of L2 speech perception and production studies (Strange, 1992, pp. 200-201). In this section of Chapter 1, L2 perceptual and training studies will be presented under each of these four grouping factors (Strange, 1992) to illustrate some of the issues that have been the focus of L2 speech perception research.

a. Subject's variables

Among the "subject" factors that have been extensively studied lies "age", as in the age when someone started learning the L2. Many times this relates to the time when the person immigrated to the L2-speaking country as well. It may also refer to the length of time someone has been living in the L2-speaking environment. Flege introduced the concepts of "age of arrival" (AOA) and "length of residence" (LOR) to refer to these two factors mentioned above. Flege, Bohn and Jang (1997) tested adult German, Spanish, Mandarin, and Korean L2 speakers of English (20 in each group, mean age 25) who had spent

between 0.7 and 7.4 years (LOR) in the United States and had received intensive English instruction on arrival. Participants were tested in the perception and production of the English vowels /iː/–/æ/. The results showed that the perception and production of English vowels was more accurate for the more experienced participants, though there was variability in performance depending on the L1 background. Yet, the authors suggested that these findings confirmed that the amount of L2 language experience plays a role in L2 learning as had also been found in a previous study by Munro, Flege and MacKay (1996).

Age of arrival (AOA) and length of residence (LOR) have also been found to interact with the amount of use of L1 and L2. Piske, Flege, MacKay and Meador (2002) tested a group of Italian immigrants in Canada on their production of English vowels (11 vowels). The participants had learnt English as children or adult, so the group was divided into early (E) or late (L) learners and subdivided according to the amount of Italian L1 used on daily basis (H: high, L: low use). They worked with three groups: early-high (E-H), early-low (E-L) and late-high (L-H). Native speakers of Canadian English rated the intelligibility of the L2 learners' vowel production. The vowels produced by the E-L group was rated as more native-like, whereas comparisons between the E-L and E-H scores were significantly different for 7 vowels, with lower ratings given to E-H's productions. The L-H group obtained the lowest scores. These results indicated that the use of the native language (Italian) may interfere with establishing L2 categories, regardless of how early learner may have started their experience with the language. These findings are in line with results in Flege and Mackay (2004); early Italian L2 learners, who differed in the amount of use of their L1, had different levels of performance in their perception of Canadian English vowels. Learners who reported lower use of L1 were better at discriminating English vowels. These findings may imply that the mere fact of an early start of learning a second language does not guarantee accurate perceptual and production competence in the L2 per se; they also reinforce the view that difficulties in learning may be due to L1 interference.

The idea that "early is better" refers to the advantage that early L2 learners have over adult L2 learners due to the quantity and quality of input in the L2. Early learners are more likely to have a larger amount of interaction with native speakers. Therefore, early age constitutes

an advantage in the sense that it provides more opportunities for L2 language categories to be established through contact with native speakers rather than due to exclusively greater plasticity in younger learners (Højen & Flege, 2006). However, this age factor may interact with use of L1 as discussed above. Late adult L2 learners' constrains may also derive from the lack of opportunities to experience the L2 in contact with native speakers, making their L2 input poor in quantity and quality (Flege, 1997, 1999). Therefore, constrains are less likely to be due to a maturational and cognitive effect of aging as it used to be interpreted from Lenneberg's (1967) proposal in the "Critical period" for language learning.

b. Training task variables

The two topics covered in this section will be the Type of task and speaker-related variability. Although the original component was called training variables, Strange (1992) actually referred to testing and training factors under this heading. Thus, L2 perceptual and training studies included in this section will relate to these aspects.

The two most typically used tasks for perceptual experiments are identification (ID) and discrimination (DIS) of new phonemic contrast. Discrimination tasks focus on lower-level phonetic information (e.g. spot the odd-one out token in a set of three contrasts) whereas identification training focuses on higher-level processing skills (e.g. assign a label to a specific token). It has been suggested that discrimination training could contribute to improve learners' perception of the English consonants /l/-/r/ contrasts and could also transfers to identification tasks (Strange & Dittmann, 1984). However, other authors have argued that identification training may be more appropriate because it helps to change the perceptual space. ID tasks would facilitate establishing new categories and generalisation of the learning to new tokens, while allowing more attention to between-category distinctions (Lambacher, Martens, Kakehi, Marasinghe & Molholt, 2005; Logan, Lively & Pisoni, 1991; Lively, Logan & Pisoni, 1993; Bradlow et al., 1997). More recently, Shinohara and Iverson (2012) used a combined identification and discrimination training to Japanese subjects learning English /r/−/l/. They gave participants 10 training sessions (5 ID sessions, 5 DIS sessions) and controlled for the order of the training approaches (ID-DIS, DIS-ID). Their results showed that both modalities improved perception and production of

the contrasts to similar amounts, but no extra effect of combing the two approaches was found.

Another source of difficulty in L2 perception is the large amount of speaker-related variability that is naturally occurring in speech. Speakers vary in the average speaking rate that they may use, in their regional accent, in how intelligible their speech is (Bradlow, Torretta & Pisoni, 1996) and there is also variability in the perceived information about the talker's identity (Nygaard & Pisoni, 1998; Remez, Fellowes & Nagel, 2007). In spite of all this variability, L1 listeners manage to decode the message contained in the signal. In spite of the difficulty that this variability may pose on L2 learners, studies have shown they are able to learn to attend to this variability through training which introduces multiple speakers as training material. Logan, Lively and Pisoni (1991) devised a multi-talker-natural-token training programme to improve perception of English /r/ and /l/ in six Japanese learners of English. They used minimal pairs in an identification task for training. Their high-variability training method using natural speech proved to be effective in improving participants' perception of the contrasts. Lively et al. (1993) compared high variability training versus training with one speaker only for English /r/-/l/ and found that generalisation to new tokens only occurred in the group trained with a larger number of speakers, though both groups showed improvement after training. However, some studies have also shown this variability may become detrimental. Perrachione, Lee, Ha and Wong (2011) trained learners on L2 phonological contrasts based on pitch contrasts using a high-variability training method. In the study, learner's perceptual abilities were measured and then correlated with their improvement after training. Although introducing talker variability in training has been suggested as beneficial for L2 learners in previous research (Logan et al. 1991; Iverson & Evans, 2009), this variability only helped those learners that had high perceptual scores before training and was detrimental for the weaker perceivers in this study. Consequently, it seems important to consider initial perceptual skills and how they relate to the amount of improvement obtained after training to account for individual differences. It may also be informative as to who may need a different type of traininig.

The benefit of the use of multiple speakers and multiple tokens in training has been found in a number of other studies showing greater perceptual improvement of the contrasts trained and greater generalisation to new speakers and tokens (Iverson & Evans, 2009; Logan, Lively & Pisoni, 1991; Lively, Logan & Pisoni, 1993). However, most learners in the English as a foreign language (EFL) context usually have reduced access to multiple speakers in the classroom and exposure to this natural variability outside the classroom is scarce. This issue is of relevance for the current study given that the participants tested were all EFL learners. Thus some of the factors affecting their perceptual learning may be due to the EFL context constrains.

c. Criterial tasks variables

Under this topic, Strange (1992) refers to aspects concerning cognitive processes involved in the execution and aim of the measure obtained with a given task used in the test batteries.

The amount of cognitive effort demanded by the tasks used to measure L2 speech perception may contribute to variability in participants' performance. At the level of cognitive demands, Díaz, Mitterer, Broersma and Sebastián-Gallés (2012) tested 55 late Dutch-English bilinguals on the English /æ/–/ɛ/ contrast using a categorization, word identification and lexical decision task. The results showed that late bilinguals performed significantly better in a lower-level phonetic identification task than in a lexical decision task. This would indicate that performance may be lower when more cognitive resources are needed as in the case of lexical processing.

On another cognitive aspect, Aliaga-García, Mora and Cerviño-Povedano (2010) found that phonological short-term memory was related to L2 perceptual accuracy performance when categorising and discriminating English vowels. Learners with higher short-term memory showed higher perceptual scores and benefited more from English vowel training. On the same aspect, MacKay, Meador and Flege (2001) found that the Italian participants' phonological short-term memory could account for between 8 to 15% of the variance in the identification of English consonants in their study. Thus, it seems reasonable to consider these cognitive aspects when designing and analysing the learners' performance on the

tasks used in L2 speech perception tests or training programmes. In addition, they could also provide a source to explain individual differences in L2 speech performance.

Another factor that brings in more complexity to the outcome of perceptual studies is the type of learning which is being targeted by a given technique in L2 speech testing or training. Training studies are better to illustrate this issue. Some training studies focus on promoting perception of either a small or larger set of contrasts, some studies measure improvement in perception of the same type of trained contrasts, whereas other studies assess the impact of training on the generalisation of the learning to new tokens and talkers (Iverson, Hazan & Bannister, 2005; Logan et al., 1991; Lively et al., 1993; Nishi & Kewley-Port, 2007; Strange & Dittmann, 1984; Wang & Munro, 2004). Some training methods aim at changing the attention of the learner to a particular critical cue by using different methodologies to improve perception of a given contrast (Kondaurova & Francis, 2010). While other studies have also tried to find a link between perception and production by measuring the impact of training on perception as well as production of the trained contrasts (Bradlow, Akahane-Yamada & Tohkura, 1999; Hazan, Sennema, Iba & Faulkner, 2005; Lambacher, Martens, Kakehi, Marasinghe & Molholt, 2005).

Most of the training studies report some degree of improvement after training but a smaller number of studies have been able to test whether the learning is retained after some months of completing the training. Wang & Munro (2004) used computer-training for three English vowel contrasts with Mandarin and Cantonese L2 participants and retested them 3 months after completing training. Transfer and retention was found in participants, though with slightly lower scores than in the post test. It is enormously valuable to test whether learning has had any long-term impact on L2 learners taking part in perceptual training studies; however, it is probably due to practical reasons that researchers find it difficult to test participants some months after they have been trained. In the current study, a subset of 34 participants was tested three times in a year but testing their retention after training would have been difficult. Besides, there was the issue of to what extent their vowel perception continued improving as a factor of their intensive classroom instruction at university.

An aspect that is central in the way transfer or generalisation of the learning is measured is the nature of the material used. Typically, pre and post test material uses new tokens produced by new speakers. These tokens are mostly short words (monosyllabic material, e.g. “hVd” words) presented in isolation. To our knowledge, it has not been shown whether this learning through exposure to high variability material during training can generalise to perception of the contrasts in more complex materials. It remains to be seeing whether learners can actually use this learning at word level for more naturalistic situations in which they have to perceive, for example, vowel or consonant contrasts in sentences, or longer chunks of spoken language. This is most likely what an L2 learner wants to successfully achieve.

d. The stimulus

Success in the perception of L2 sounds may also be affected by factors that relate to the input used when testing or training L2 learners. Some of the factors that will be considered here are: the type of contrast, the type of stimulus, the relation between the L1 and L2 sound system and the use of different types of cues to the L2 contrast.

The perception of vowels or consonants, that is the type of contrast used as stimulus, presents different degrees of difficulty for L2 learners. Consonants tend to be more stable than vowels in their acoustical properties (Liberman, Cooper, Shankweiler & Studdert-Kennedy, 1967) and therefore they may be easier to perceive. Vowels tend to have higher intensity and longer duration than consonants, less vocal tract constriction and they are all voiced sounds (Best, Halle, Bohn & Faber, 2003; Knight, 2011). Vowels may present more within-stimulus variability as they may differ in pitch, loudness and quality (Ladefoged, 2006) and due to their central place in the syllable, they tend be articulated between consonants and influenced by these neighbouring sounds.

The nature of the stimulus used may promote different types of learning. Generalisation to perception of natural tokens after training seems to occur only if the training tokens have also been natural speech. The use of synthetic input in training material has shown to improve perception of novel contrasts only for the same type of input (synthetic) but has failed to generalise to new natural tokens. Strange & Dittmann (1984) tested eight Japanese

female students living in USA on their perception of English /r/–/l/ contrast. They gave the participants 18 discrimination training sessions. Comparisons of the pre and post training measures showed there was significant improvement and transfer to identification tasks but only when using synthetic tokens. They failed to transfer their learning to natural speech perception. These findings revealed the impact of the type of stimulus used and confirmed that natural input is better to improve natural L2 speech perception.

Another factor affecting L2 speech perception is the L2 sounds relation to a native category in the L1. Models for L2 speech perception like the Perceptual Assimilation Model (Best, 1995, Best & Tyler, 2007) and the Speech Learning Model (1995) predict that the perceptual distance or similarity of an L2 contrast with an L1 category would determine the degree of success in accurately perceiving the contrast and in establishing new categories (models will be discussed in section 1.3). For example, studies on vowel contrasts have shown that Spanish learners initially perceive English /iː/–/ɪ/ contrast as two instances of their L1 Spanish /i/. The English /iː/ is perceived as closer to the Spanish /i/ category, whereas English /ɪ/ is perceived as a poor example of Spanish /i/ (Flege & MacKay, 2004; Flege, Bohn & Jang, 1997; Fox, Flege & Munro, 1995; Morrison, 2008).

Concerning the relation between the L1 and L2 system, the size of the phoneme inventories may also play a role. Iverson & Evans (2007) compared the perception of English vowels in two groups of L2 learners with different L1 vowel size inventories. They found that learners of English with smaller vowel inventories (Spanish, French) showed poorer identification and learning performance than L2 learners with larger vowel inventories (German, Norwegian). The results from the Spanish and French L2 learners would contradict assertions made by the SLM which predicts that L2 learners with a smaller vowel inventory would have more room to establish new categories in their perceptual space (Flege, 1995).

Attending to critical cues to accurately perceive an L2 contrast seems to be problematic for L2 learners. Studies on the auditory perception of the English contrast /r/–/l/ by Japanese

learners show that these learners weigh more information from Formant 2 (F2) which is less relevant to perceive /r/, instead of using Formant 3 (F3) as native speakers do (Aoyama, Flege, Guion, Akahane-Yamada & Yamada, 2004; Bradlow, Torreta & Pisoni, 1996; Bradlow, Pisoni, Akahane-Yamada & Tohkura, 1997; Bradlow, Akahane-Yamada, Pisoni & Tohkura, 1999; Iverson, Kuhl, Akahane-Yamada, Diesch, Kettermmann & Siebert, 2003; Iverson, Hazan & Bannister, 2005; Yamada & Tohkura, 1992). The same lack of weighting of critical information has been found in L2 learners who rely more on duration rather than on spectral cues for English vowel contrasts, unlike English native speakers (Cebrian, 2006; Flege, Bohn & Jang, 1997; Iverson et al., 2003; Kondaurova & Francis, 2008; Morrison, 2002).

There seems to be no consensus yet concerning the nature of the use of temporal cues for English vowel perception by L2 learners. Different accounts have been offered for this problem. Bohn (1995) explained the reliance on duration cues with his "Desensitisation Hypothesis". He suggested that the use of duration cues in L2 vowel perception is the strategy L2 learners would most likely use when the spectral information is not enough to perceive the acoustic difference between vowels. That is, when learners lack previous linguistic experience with the use of spectral cues to differentiate vowel contrasts (desensitisation). In this situation, the use of duration cues seems to be the easiest resource available. He tested Spanish, Mandarin and German learners of English on the perception of English contrasts /e/–/æ/and /iː/–/ɪ/. He found that all three groups of learners used duration as the main cue to perceive the contrasts. These L2 learners had either no experience with duration in their L1 for vowel contrasts (Spanish), some experience in the distinction of two tones length (Mandarin) or experience with using duration for vowel contrasts (German). Spanish and Mandarin speakers were expected to only use spectral information for vowel distinction, based on their sensitisation given by experience with these cues, whereas only the German speakers were predicted to use spectral and durational information. These results suggested that duration cues are more accessible to L2 learners when spectral cues are insufficient to perceive the L2 vowel contrasts, regardless of the status of durational cues in the learner's L1.

An alternative explanation on the use of duration by L2 learners was presented by Escudero & Boersma (2004). They suggested that L2 learners who have never used temporal cues to distinguish vowel contrasts in their L1 have an empty space in the duration dimension that would allow them to easily create this category. This account hypothesized that L2 learners have access to "L1-like acquisition strategies" that allow them to move boundaries between categories. L2 learners would be able to create new categories for the L2 sounds by using principles from distributional learning (Boersma, Escudero & Hayes, 2003). This distributional learning perspective would pose that learners can use one cue at a time for any phonological contrast, so duration is the first cue for L2 learners perceiving a vowel contrast. Once the categories for short and long duration are established, the learner can have access to spectral cues and integrate the temporal information for the contrast. This model considers the use of L1 categories to perceive the L2 sounds as a starting point, and then moves on to using the learner's perceptual space available to create new categories. These two accounts (Bohn, 1995; Escudero & Boersma, 2004) challenge the traditional idea of "L1-Transfer".

A different approach to the use of duration cues in L2 speech perception is supported by studies which use "L1-Transfer" as the explanation to this phenomenon. Morrison (2008, 2009) suggested that the use of duration cues in L2 vowel perception acts as a secondary stage in the perceptual process of English vowel contrasts. L2 learners would first have access to a category-goodness-assimilation (CGA) process which is multidimensional and allows the learners to transfer an L1 category to perceive an L2 sound. Morrison tested Spanish L2 learners on the English /iː/-/ɪ/ contrast and found that after using the CGA strategy, L2 learners used duration at the initial stage. After exposure to English, learners reversed their duration strategy if they were given negative reinforcement on duration and positive reinforcement on the use of spectral cues. Participants were able to achieve English native-like perception for the vowel contrast. On a somehow similar view on duration as an L1 transfer strategy, Kondaurova and Francis (2008) suggested that even when an L2 learner does not have L1 durational contrasts to distinguish phonemes, they may have experience with listening to allophonic use of duration. That is, making any phoneme longer as a result of stress, or voicing of coda stop consonants, for example. This

experience would be enough to transfer the temporal cue strategy to L2 vowel perception. Though the authors do not make clear what processes would be involved in the transfer of such a strategy to the perception of vowel contrasts.

Another important question in L2 vowel perception is whether the use of cues is similar across learners with different L1 background. Iverson & Evans (2007), mentioned previously (part c), found that L2 learners with larger vowel inventories (German and Norwegian learners) were more accurate at perceiving English vowels. However, there were no fundamental differences in what German, Norwegian, Spanish and French individuals learned; they all used spectral and temporal cues for English vowel perception. The authors suggested that learners with different L1 background may learn to perceive L2 vowels in a very similar way. This use of cues would contradict most of the vowel studies that have suggested that most L2 learners would use mainly durational cues to perceive the difference between tense-lax vowel contrasts (Bohn, 1995; Cebrian, 2006; Escudero & Boersma, 2004; Flege et al., 1997; Kondaurova & Francis, 2008; Morrison, 2008; Wang & Munro, 1999).

When referring to cues for speech perception, most of the studies have focused on auditory cues mainly, although speech is bimodal (Rosenblum, 2005). However, there is a growing body of research on L2 speech perception which has shown that visual information can aid L2 learners in perceiving some English contrasts when the visual information is salient (Hardison, 1996; Hazan et al., 2005; Wang, Behne & Jiang, 2008). L2 training studies have also found that perception of L2 contrasts can be enhanced if visual information is used (Hardison, 2003; Hazan, et al., 2005; Hazan & Sennema, 2007; Sennema, Hazan & Faulkner, 2003). Most of these studies have used visual information for English consonants perception and very few have explored visual information for English vowel perception (Ortega-Llebaria, Faulkner & Hazan, 2001).

In this section of chapter one, factors and their interactions affecting L2 speech perception have been presented to illustrate the complexity of the phenomenon. In the following section, some theoretical models that have been used to explain L2 speech perception, and some of them production, will be presented.

1.3 L2 Speech perception models

One of the models that has largely influenced research in L2 speech perception is the Perceptual Assimilation Model (PAM) by Best (1995) that was initially proposed as an explanation for the perception of non-native sounds by naïve monolingual "mature listeners" with no experience with the non-native language. The model was subsequently used by L2 researchers to explain second language speech perception problems although not initially meant to be used with this population. PAM suggested that perception of non-native phonemes is regulated by an "assimilation process" to an L1 category. An L2 phoneme may be a good or poor example of an L1 sound, or unlike any L1 phoneme. This model is based on the "Direct-realist approach" to perception (Gibson & Gibson, 1995) in which a central tenet is that perceivers attend to speech gestures.

Best & Tyler (2007) reviewed PAM in the light of monolingual and late-adult bilingual or multilingual speech perception to find evidence of common processes involved when perceiving non-native speech in this population and, thus, extended PAM to L2 speech perception (PAM-L2). The authors maintained all the assimilation categories and predictions in the original PAM and made clear that these predictions would only apply to monolinguals and L2 learners, the latter living in a bilingual environment or immersed in the L2 language culture. They made an observation about L2 learners who are learning the second language in a classroom environment where there is neither need nor chance to be exposed to the L2 outside the educational context. These learners are usually referred to as EFL learners (EFL: English as a foreign language). Given that experience with the L2 would have a fundamental place in the acquisition of the L2 inventory, the adequate quantity and quality of input must be provided through genuine interaction with native speakers of the L2 and in a natural occurring environment in the country where target language is spoken. The authors also clarified their agreements and discrepancies with the L2 speech perception competing model, the Speech Learning Model (Flege, 1995).

The PAM model predicted six different assimilation types, varying in degrees of discrimination of the non-native phonemes (Best, 1993, p.56). First, a "*two-category contrast*" in which two non-native sounds would be assimilated to two different L1 phonemes, allowing easy discrimination of the L2 sounds. A "*category-goodness difference*" in which two sounds are perceived as similar to one in the L1, one phoneme being a good and the other less good assimilation to the L1 phoneme. In this latter case, discrimination would be very good to moderate. Another type of contrast suggested was a "*Single-category assimilation"* when two L2 contrasts are assimilated to only one L1 category but both phonemes are considered good exemplars of the L1 sound. This would make differentiation of the L2 sounds very hard. On the other hand, if the listener cannot find any similar sound for the two L2 phonemes, they fall into *"both-uncategorised"* type and their discrimination would be predicted to be poor to moderate. Very good discrimination is predicted for two phonemes when only one of them is categorised as familiar to another L1-counterpart but the other L2 sound is left uncategorised, *"uncategorised vs. categorised"* contrast type. Finally, very good to good discrimination is predicted in the case of two non-native categories that cannot be assimilated to any L1 sound, because they are considered non-speech categories. In this case, the *"non-assimilable"* label is used for these non-speech gestures.

The Speech Learning Model (SLM) proposed by Flege (1995) is the other model which is usually contrasted with PAM's predictions. This model aimed at explaining patterns in the perception and production of L2 sounds by experienced L2 learners who were immersed in the culture of the target language and have used the L2 for some time. This model assumes a direct perception-production relation and is based on an auditory perceptual account for speech; failure in accurately producing an L2 sound is due to inability to correctly perceive the acoustic characteristics of such L2 phoneme, though not the only source of perceptual problems. Four postulates (P) were presented in this model. In the first one (P1), the author states that the capacity of the mechanisms and processes involved in the perception and production of the L1 sounds is never lost as human beings grow older. It is this ability to process native sounds which can be used while acquiring new sound categories of an L2. In Best & Tyler's (2007) view, PAM and SLM agree in that the speech perception

mechanisms are always in an on-going process of refinement but the way in which children and adults do this when they learn an L2 is different. However, there is strong disagreement in terms of the central idea of category formation of sounds in SLM, given that according to PAM speech perception is realised by "extracting invariants of articulatory gestures" and not auditory cues.

In the second postulate (P2), Flege posits that phonetic categories are long-term memory representations that contain the language-specific information of speech sounds needed to process speech. In P3, SLM claims that phonetic categories established for the native-language experience change over the lifespan and those changes reflect influences of both the L1 and L2 sound system characteristics. The idea of mental representations presented in P2 and P3 is another difference with Best's model. This is due to PAM's assumption that it is through the perception of articulatory gestures and perceptual learning that the perceiver gets tuned into the sounds of the L1 and L2. Other than that, there is no major disagreement with the idea presented in P3 of a refinement of the perceptual system through experience.

Finally, in P4 Flege suggests that when an L1 and L2 sound occupy the same phonological space, bilinguals would struggle to perceive the difference between them. In PAM, it is generally accepted that L1 and L2 phonological categories may co-exist in a common phonological space. Moreover, there would be no problem as long as the listener is able to perceive them as two phonetically distinct realizations of one phonological category.

SLM makes predictions about L2 perception and production patterns through seven hypotheses. These hypotheses (H) refer to the allophonic-level relation between sounds in the L1 and L2 (H1), the establishing of new categories for an L2 sound when it is perceived as different from any L1 sound (H2), and the advantage of the distance in similarity between an L1 and L2 category that would contribute to their perceived difference (H3). However, if an L2 phoneme is perceived as equal to an L1, the category formation for the L2 sound may be difficult to achieve (H5). In SLM, the age at which learners start learning the L2 plays an important role in perceiving the phonetic difference between L1 and L2 sounds (H4), the younger a learner starts acquiring the L2 phoneme inventory, the better. H6 states that monolinguals and bilinguals differ in the way they form an L2 category.

Bilinguals need their categories to be distinguished from the L1 sounds or weigh cues for the L2 sounds differently as they do for the sounds in the L1.

Both models, PAM and SLM, are attempts to explain the L2 speech perception process in terms of similarity or distance of the L2 sounds to a phoneme in the L1. As mentioned earlier, PAM-L2 (Best & Tyler, 2007) examines the problem from a "direct-realist approach" to perception in which perceivers attend to gestures in the L1 or L2 which may represent realisations of the same L1 phonological unit. Whereas SLM considers that adult learners attend to cues in the L1 and L2 sounds to create categories which share a phonological space where they interact and may eventually form new categories or not. However, Bohn & Best (2007) found that none of these two models could fully explain the results in their study for German participants achieving similar scores to American English native speakers in perceiving the approximants /w, j/, even though they lack this contrast in their L1 (/w/ does not exist in German). So, could there be any other explanation?

Iverson et al. (2003) offered another possible account for the difficulties L2 adult learners encounter when listening to L2 Speech. This proposal takes into account that early L1 language experience may prevent L2 sound category formation. The authors adhere to the idea that from childhood to adolescence, individuals go through the process of tuning into their L1 sound system and establishing their L1 perceptual space. This specialisation mechanism would result in reduced perceptual sensitivity within the L1 phoneme inventory (Kuhl, 1992) and would become an obstacle when adults try to learn non-native sounds. Iverson et al., (2003) tested Japanese, German and Americans in the perception of English /r/ and /l/ and found that Japanese are more sensitive to second formant (F2) variation in the stimulus, cue that is not critical for native speakers to distinguish between the two phonemes. Germans listeners showed a near similar sensitivity to Formant 3 (F3) as native speakers did. The Japanese perceptual sensitivity to the irrelevant F2 cue would contribute to create the wrong category representation for /ɹ/ and confuse it with /l/. On the other hand, the German listeners showed better sensitivity to the critical F3 cue and could distinguish /ɹ/ and /l/ at a near-native level.

These findings illustrate how L1 language experience and the representations in the perceptual space for L1 may affect the perception of L2 sounds. The perceptual interference account agrees with the idea in PAM and SLM that higher-level linguistic processes are changed when adults learn an L2 but claims that processes at lower-level in perception interfere with higher-level processes due to a "loss of perceptual sensitivity to non-native phoneme contrasts" (Iverson et al., 2003:B54).

PAM/PAM-L2, SLM and the perceptual interference (PI) account focus on different population. PAM/PAM-L2 targets late bilinguals immersed in the country of the L2 and also naïve monolingual listeners, while SLM makes predictions about experienced L2 adult-learners who also live in the target language culture. Studies using the PI account have usually focused on L2 learners who have learnt the L2 in a classroom environment (EFL context) and do not necessarily have experience of living in the country where the target L2 language is spoken. Although these models were conceived with a different population in mind, they have usually been addressed to account for the results in L2 speech perception studies, regardless of the type of participants tested (bilinguals, L2 learners or EFL learners).

Finally, it seems interesting that even when the bimodality of speech in the native language has been well established for more than three decades (McGurk & MacDonald, 1976; Rosenblum, 2005; 2008) and there is a growing body of research on the effect of visual cues in L2 speech (Hardison, 1999; 2003; Hazan et al., 2005; Hazan et al., 2006; Wang, Behne & Jiang, 2008; among others), a model that may account for the use of visual cues in L2 speech perception has yet to be developed. Moreover, the focus of studies on L2 speech perception remains mainly in the auditory perceptual realm. In the next section, visual information and its possible contribution to L2 speech perception will be presented.

1.4 Use of visual cues in speech perception

It is has been well established that during speech perception both the auditory and visual information are available and influence speech perception in face to face communication. Besides, the weighting of these two sources of information may be affected by talker's variability and even noise in the environment (McGurk & Macdonald, 1976; Massaro, 1987; Massaro, Cohen, Gesi, Heredia & Tsuzaki, 1993; Rosenblum, 2005).

McGurk & MacDonald (1976) proposed that there is a fast integration of audio and visual information during face to face speech perception and that the use of visual cues is not restricted to adverse conditions (noisy contexts or hearing impairment). To test the bimodality of speech information, they used discrepant auditory (A) and visual (V) one-syllable stimulus ("A-ba", "V-ga") with children and adult English native speakers. Results showed that incongruent audio and visual input triggered responses like "da" that had not been presented to participants (aka the *McGurk effect*); with adults being more affected by incongruent A and V information. They suggested that when there is no dominant modality, as speech is bimodal, making a decision between the two sources of conflicting input (A and V) is difficult for the listeners and this would trigger the "illusion" ("da") type of responses. They also concluded that adults were more influenced by visual cues than children because the use of visual information is developed through language experience as humans grow older.

1.4.1 When does the integration of visual cues take place?

The importance of visual information has been approached from different theoretical perspectives suggesting that the integration of visual cues is automatic (e.g. motor or ecological theories) or it is a separate supportive component of the auditory signal (auditory theories). Concerning when the integration takes place, two broad proposals have been suggested depending on whether the integration takes place at an early or late stage in the perception process (Green, 1998). Early integration of auditory (A) and visual (V) cues models propose that there is interaction of A and V information at a pre-phonetic level. Integration takes place before the extraction of phonetic prototypes. Auditory and visual information are perceived and integrated before a phonetic category is assigned to that A+V input (Braida, 1991; Green, 1998; Rosenblum & Gordon, 2001; Rosenblum, 2005;

Schwartz, Berthommier & Savariaux, 2004; Summerfield, 1987). The other perspective supports the late integration of A and V cues after feature extraction (Massaro, 1987; Bernstein et al., 2004).

In an early integration model supported by Summerfield (1987; 1992), the audio-visual stimulus is perceived, A and V features are extracted and combined. After that, a decision about the phonetic categorisation is made. The counter approach would be Massaro's (Massaro & Cohen, 1983; Massaro, 1987) Fuzzy Logical Model of Perception (FLMP) which suggests three stages in the perceptual process. In the first stage, continuous perceptual characteristics are extracted from the stimulus separately. In the second stage, the features (A and V) are integrated and this information is matched against a perceptual unit, a "potential prototype". In the final stage, each potential prototype is assessed and a decision is made. The best pattern is chosen as a response. In this model, one piece of information is more informative if the other cue is ambiguous. The main difference between these two approaches is that the sources of information in the L1 (A, V) are evaluated together (A+V) by the early integration models and separately (A, V) by the late integration models and then integrated before final perceptual recognition takes place.

Some studies have addressed the issue whether the integration of auditory and visual information in speech perception is "universal" or whether it might be affected by language-specific factors such as the characteristics of the phoneme inventory or by cultural factors. Massaro et al. (1993) studied how language and culture may play a role in speech perception in face-to-face communication. They tested native speakers of Japanese, Spanish, American English and Dutch using the McGurk effect as a paradigm. They concluded that there was no dominance of the auditory modality across the different language participants, suggesting similarity in the speech perception processes across language and culture. Variation in the way visual cues are used may derive from differences in the phonetic realisation of syllables or from language phonotactic constraints.

The relative influence of visual information derived from the McGurk effect was found to be stronger in some speakers of Spanish (Fuster-Duran, 1996) and Italians (Bovo et al., 2009) and weaker on Japanese (Sekiyama & Tohkura, 1991; 1993) and Chinese speakers

(Sekiyama, 1997). These results gave rise to the suggestion that the processing of visual information may have a cultural component that affects speakers of cultures which do not encourage listeners to look at their interlocutor (Sekiyama &Tohkura, 1991; 1993). From a developmental perspective, Sekiyama & Burnham (2008) established that young Japanese and English children (6 years old) made similar use of visual cues for speech perception. However, a difference was found when comparing older children of the same language groups (from 6 to 11 years old). Japanese adult speakers seem to need less visual information in their L1 speech perception than English adults. This would explain the findings with Japanese learners of English who seem to be less affected by the McGurk effect. The authors are in favour of the idea that visual information would only be helpful when the auditory information is ambiguous.

In line with Sekiyama and colleagues, other researchers have also have argued that the use of visual cues becomes relevant only when the acoustic signal is degraded. Benoit & Le Goff (1998) studied the benefit of speech-reading in French. They compared the intelligibility of audio (A) and audio-visual (AV) speech under distortion conditions using material produced by a text-to-speech synthesizer and by a human speaker. Their results suggested that the benefit of visual cues was relevant only when audio was degraded. As has been shown in other studies on the bimodality of speech, it is true that speakers attend to visual cues when speech is affected by noise or the perceiver has some hearing-impairment. But it has been proved that visual information is quickly integrated during normal speech perception condition and not only when the signal is degraded (Alsius, Navarra, Campbell, & Soto-Faraco, 2005; McGurk & MacDonald, 1976; Massaro & Cohen, 1983; Massaro, Cohen & Smeele, 1995; Reisberg, McLean & Goldfield, 1987).

All these experiments have tested participants on the use of visual cues for one-syllable words using material which is language-neutral, the consonants chosen are found in most of the perceivers L1. It remains to be seen whether visual cues may also aid perception of longer input; for example, with real words in isolated contexts or in more naturalistic environments as in sentence-length material. We will now turn our attention to where the

visual information in speech comes from and what processes are involved in the integration of visual and auditory cues.

1.4.2 Sources of visual information during speech perception

It has been established that the visual information for speech comes from lips and jaw movements, as well as seeing the whole face. Summerfield (1979; 1992) tested listeners with no hearing impairment on sentences in noise spoken by British English native speakers under two audiovisual (AV) conditions: showing speaker's face or only the speaker' lips. Both modalities seemed to contribute to the number of words participants identified correctly, but higher scores were obtained when the whole face of the speaker could be seen (whole-face: 43%, lips-only: 31%). These findings indicated that in L1 speech perception, visual information of speech not only comes from lips but also from seeing the speaker's whole face.

Fisher (1968) suggested that visual phonemes, "visemes", could be used to describe the units that stand for any contrastive visual segment in speech perception. Many researchers (Fisher, 1968; Binnie, Montgomery & Jackson, 1974; Walden, Erdman, Montgomery, Schwartz & Prosek, 1981; Kricos & Lesner, 1982; Owens & Blazek, 1985) set out the task of establishing a standardised list of visemes for consonants varying in the number and type of talkers chosen for the visual input, the population tested (normal hearing and/or hearing-impaired), vowel context for the consonants tested and the stringency of the criterion (70-75% recognition). These variables have made researchers realise that it seems difficult to establish a unique list of visemes for a particular language, given that on top of the variables mentioned above there are other factors that affect the perception of the visemes as well. These factors relate to the talker's visual intelligibility, the perceiver's visual speech-reading capacity, phonemic context (CV/VCV-words), length of the input and amount of light and viewing angle (Walden et al., 1981; Owen & Blazek, 1985).

Most of the studies in the field suggest that groups of sounds that share visual information would also share viseme category. For example, there would be only one viseme for the phonemes in /p, b, m/, /f, v/, /θ, ð/or /ʧ, ʤ, ʃʒ/ as each group shares or has a near place of articulation. This would make any list of visemes look shorter than the list of

phonemes. Lesner & Kricos (1981) investigated whether visemes for vowels and diphthongs would differ in their visual perception. They assumed that perceiving visual cues for vowels would be easier than for diphthongs given than the former require "unique articulatory movements" for their production. Their results revealed that diphthongs were easier to be visually perceived than vowels, perhaps due to their co-articulation movements. However, they found that the degree of visibility of the visemes varied across speakers. This variability would be the consequence of some speakers being more visually intelligible than others.

If one viseme may stand for more than one phoneme, these findings may bear implications for speech perception in L1 and L2 concerning the integration of visual and auditory information.

1.5 Use of visual cues in L2 speech perception

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 the visual cues add information to the contrast. Its use may also vary across L2 learners as a function of their L1 perceptual categories. Hardison (1996) used the McGurk effect to test the influence of visual input on advanced learners of English with Japanese, Korean, Spanish or Malay L1 and a control group of native speakers of English (NS). The stimulus consisted of syllables with /p, f, w, r, t, k / in audio (congruent & incongruent) and audio-visual mode (in clear & noise). The authors suggested that there was an effect of visual cues for non-native (NN) as well as for native perceivers, but its use varied across language group and phoneme contrast and context. For example, Japanese and Koreans improved their perception of /f/ and /r/ when visual cues were added in congruent context. However, their scores were significantly lower than the other speaker groups in auditory (A) condition. This is interesting because neither Japanese nor Korean have these two sounds in their L1; yet, visual cues helped in the distinction. With

incongruent input, visual /t, k/ on auditory /p/, there was a strong visual effect for all three NN groups and the NS as well. The author argued that differences in L2 perceptual processing may be influenced by the L1 perceptual categories of the learners.

The degree of salience of the visual cues is another factor that needs to be considered. The visual information for bilabial or labio-dental contrasts is highly salient and should be easier to perceive than for alveolar sounds. Hazan et al. (2006) tested Spanish and Japanese L2 learners on /p, b, v/ in audio (A), audio-visual (AV) and video only (V) conditions (experiment 1) and found that both groups of learners obtained higher scores in AV condition, though the Japanese learners seemed to benefit less from visual cues than the Spanish group.

Poorer use of visual cues has been reported in other studies for Japanese learners in Sekiyama & Tohkura (1993) but Hardison (1996) found that Japanese learners were affected by the McGurk effect (visual benefit). Hazan et al. (2006) also tested Japanese and Korean learners of English on a less visually salient contrast, /l/–/r/ (experiment 2). Neither of the two groups showed visual benefit in general though Koreans' A and AV scores were better than V scores. Japanese showed poor scores in all three conditions (A, AV, V). All the contrasts tested in this study presented some level of difficulty to the L2 learners because they do not exist or they are realised in a slightly different way in their L1. The authors suggested that the results reflect the effect of the learners' native language background and the salience of visual cues for L2 contrasts.

Although visual cues have been shown to help the perception of difficult L2 contrasts, the weight of a cue for a type of contrast may also determine the amount of benefit the L2 learner may obtain from the visual information available. Ortega-Llebaria, Faulkner and Hazan (2001) examined the contribution of visual cues for English consonant (16) and vowel (9) contrasts in Spanish learners of English and English native speakers. The study focused on speech contrasts for English consonants and vowels commonly confused by this group of L2 learners. Consonant identification scores improved for both groups in AV condition compared to the A, but confusions that were language dependant (voicing or

manner) did not improve with the addition of visual cues. Besides, Spanish participants did not benefit from visual information to distinguish contrasts that are allophonic in Spanish but phonemic in English (e.g. English /s/-/z/). These findings suggested that the weight of visual information may be different depending on whether visual cues are being used for a phonemic or allophonic contrast.

The use of visual cues in L2 speech perception also seems to vary as a function of L1 and L2 language experience. Wang, Behne and Jiang (2008) tested the use of visual cues in L2 learners with Mandarin as L1 and Canadian-English speakers on dental (/ð,ɵ/) and labiodental and alveolar (/f, v , s, z/) English contrasts. They presented the stimuli in quiet and café-noise in A, V, AV (AV, congruent and incongruent). Visual information seemed most useful in AV quiet congruent condition for both groups. Linguistic experience with English showed to positively influence the use of visual cues in Mandarin speakers with longer residence in Canada. This group was close in overall performance to the native speakers. On the use of visual cues and L1 experience, Wang, Behne and Jiang (2009) compared non-native English speakers (Korean & Mandarin) and Canadian-English speakers in their use of visual cues on the same set of contrasts as in their 2008 study (dental, labiodental & alveolar contrasts). Their results showed that Koreans were less able to use visual cues in Visual-only condition for the labiodental contrasts, which do not occur in Korean, than the English and Mandarin speakers. However, Koreans were able to achieve native-like scores in A and AV. With dental contrasts that are non-existent in Korean and Mandarin (/ð, ɵ/), both groups of L2 learners showed benefit of visual cues in AV condition. Lower scores were found in A and AV modes for the Mandarin group. These findings confirm that L2 learners can benefit from visual cues when the visual information is non-existent in the L1. The authors suggested that learners may differ in the way they use visual information available for speech perception as a function of their L1 background.

Different levels of proficiency in L2 learners have been found to interact with the amount of visual information used in general comprehension measures. A study comparing the

impact of two AV modes and A mode of a lecture given to learners of English (low-intermediate & advanced proficiency) was conducted by Sueyoshi & Hardison (2005). They used three presentation modes: AV-gesture-face (AV including gestures and face), AV-face (only face presented, no gestures) and A-only condition. Participants' comprehension level was measured from answers to questions about the lecture using a multiple choice questionnaire. The results were significantly better in both AV conditions for the two participant groups. However, the higher proficiency group obtained higher scores in the AV-face condition whereas the lower proficiency participants showed higher preference for AV-gesture-face with higher scores in this latter modality than in any other condition. This poses the question whether the use of gestures and visual cues might be related to the learner's amount of experience with the L2 language (English), as lower proficiency learners seemed to need more visual information to achieve better comprehension of the message.

An account which considers how L2 visual cues may be perceived in relation to the L1 has been suggested by Hazan et al. (2006). The authors presented three types of possible scenarios for the use of visual cues (VC) in L2 speech perception/acquisition: a) relatively similar visual cues for a viseme exist in the L1 and L2; b) the visual cues for a viseme exist only in the L2 but not in the L1, and c) the visual cues for a viseme exist in the L1 and L2 but are used to mark different phonetic distinctions. This proposal takes into consideration some of the aspects included in Flege's SLM (1995) for phonemes.

In the first scenario (a), where the L2 viseme contrast has a similar counterpart in the L1, it is expected that, visually, this contrast will be assimilated to the L1 viseme category, and no new category will be formed. For instance, the English /iː/−/ɪ/ contrast would be assimilated to Chilean-Spanish /i/. Even though they are spectrally and visually different in English, their spectral and visual realizations conform to the range of naturally acceptable allophonic variability for the Chilean-Spanish (Ch-Spanish) /i/. In the second scenario (b), if the viseme does not exist in the L1, this would facilitate its perception or acquisition as a new viseme. For example, the English dental fricative /θ/ may be more

easily perceived by Ch-Spanish learners, as no viseme or phoneme resembles the English viseme. Finally, in the third case (c) if the viseme exists in both the L2 and L1 but it is used to mark a different contrast in the L1, there would be no need to establish a new viseme category. However, a new association of the viseme with the corresponding phoneme will be needed. Hazan et al. (2006) found that the existence of a labio-dental fricative viseme /f/ aided Spanish learners of English to perceive the contrast between English bilabials /b/–/p/ and the dental fricative /v/, even though the latter is not a phoneme category in Ch-Spanish.

Different factors affecting the use of visual cues by L2 learners have been discussed here. Additional contexts in which visual information aids L2 speech perception have also been found as the effect of attending more to visual cues when the speaker is perceived as foreign (Chen & Hazan, 2007), the use of visual cues to discriminate tones (Burnham, Lau, Tam & Schoknecht, 2001; Chen & Massaro, 2008), the familiarity with the talker that may influence the amount of visual information used (Hardison, 2006). Factors discussed earlier as the cultural aspects may also determine the extent to which some L2 learners can perceive visual cues in speech (Sekiyama, 1997; Massaro, Cohen & Smeele, 1995). All the factors discussed so far provide a wider picture of the complexity of the process of integrating 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. In the next part of this chapter, the impact and benefit of L2 perceptual auditory and audio-visual training for L2 speech perception will be discussed.

1.6 English vowel training

The state-of-the-art in L2 perceptual training for English vowels has shown that the most effective training method seems to be one that uses high-variability of speakers, natural speech, feedback and which trains a larger set of vowels (Lambacher et al., 2005; Iverson &

Evans, 2009; Nishi & Kewley-Port, 2007). Generalisation to new tokens and new speakers, as well as retention after some months has also been found (Iverson & Evans, 2009). Most of the training studies have been given to L2 learners who have had little experience with the L2 and can benefit fully from access to intensive native speakers' input. A recent study has also shown that even more experienced learners can also benefit from this type of high variability training (Iverson, Pinet & Evans, 2012). Yet, most of the training studies have been conducted using auditory input only.

Although research using visual information for training has mainly focused on English consonants (Hardison, 2003; Hazan et al., 2005; Hazan & Sennema, 2007; Sennema et al., 2003), visual cues could also contribute for L2 vowel perception. Aliaga-García (2010) compared two audio-visual techniques for L2 vowel training: "audio-visual-identification" (AV-ID) versus "audio-visual-articulatory" (AV-ART) training. In the first technique AV-ID, participants were presented with tokens in video material for an identification task. In the AV-ART technique, participants were presented with audio-visual tokens and had to imitate their pronunciation and record themselves saying the tokens. This study tested 64 Catalan-Spanish learners of English in their perception of eight English vowels (/iː, ɪ, e, ɜː, æ, ʌ, ɑː, ʊ, uː/). Participants were given 10 sessions of either AV-ID or AV-ART training and were tested before and after training on the perception of natural and synthesised vowels. Overall, results showed that both techniques contributed to improve vowel perception. There was no effect of training technique; this means that participants improved in similar amounts regardless of the training type they took. Trainees also improved in their cue weighting, shifting their reliance on duration to more use of spectral information for tense-lax vowel distinction. The author suggested that identification tasks in AV-ID training can be better to establish long-term phonetic representations of the L2 phoneme categories. However, due to the experimental design of the study which lacks an audio-only training group, it is not possible to determine whether the benefit found was due to AV material only. This is because participants might have been relying mainly on the audio of the audio-visual input and may have ignored the visual cues.

Also on the contribution of visual information for speech perception, research on virtual tutors (VT) developed by Wik (2011), Wik and Hjalmarsson (2009) has explored the benefit of seeing a virtual tutor (a.k.a. embodied conversational agent, ECA; or talking heads) in perceiving and producing Swedish vowels and the impact of the feedback delivered by ECAs. This type of training has concentrated on improving perception and production of certain aspects like Swedish vowel duration, detecting insertion and deletion of vowels at the segmental level. It also provides practice and feedback on lexical stress. The virtual tutor software has been developed on expected learner's errors which are part of the data base used by the software to provide learners feedback. 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, providing appropriate feedback.

Summarising, L2 speech training has been shown to benefit the perception of L2 contrasts. Additionally, some studies have reported long-term retention of this perceptual learning; this might suggest that either L2 categories have been established in the L2 learner's perceptual space (Flege, 1995) or that learners have become more efficient at using the L2 categories they already had prior to training (Iverson & Evans, 2009). Fewer studies have been able to test the impact of perceptual training on the learners' production of the sounds trained. In the next section, the link between perception and production will be discussed briefly in the light of some findings of studies using L2 perceptual training.

1.7 The L2 speech perception and production link

In general, it has been claimed that by using training the learners' perceptual space is modified and new phonemic categories may be created (Lively et al., 1993; Logan et al., 1991). It has also been suggested that errors in the production of L2 contrasts may reflect difficulties in the perception of those phonemes based on a direct link between perception and production (Flege, 1995). However, the fact that training would help create new categories has been brought into question in studies that find no change in the best

exemplars for English vowels, though learners become more efficient at applying existing L1 and L2 categories (Iverson & Evans, 2009). Based on results on a production training study which showed improvement only in the production of contrasts but not in perception, Hattori and Iverson (2010) provided evidence as to question the direct link between perception and production. They tested Japanese L2 learners of English in their perception and production of the English /r/–/l/ contrasts and found that poor identification accuracy was not related to the participants' poor production accuracy. Furthermore, they found that production training improved Japanese production of the English /r/–/l/ but did not improve their perception of the contrast. These results suggested that these processes may run independently and learners may use different strategies to process the L2 sounds.

So, if perceptual training does not change the internal representation of the categories and production training only brings production improvement, one may argue that this is evidence that the perception-production relation may need to be revisited. Thus, models like the SLM (Flege, 1995) which aim at accounting for the link between perception and production would necessarily need an alternative account for this link.

Typically, the impact of perceptual training on production has been measured by exploring the relation between improvement in perception and production, assuming that some improvement in perception has been transferred to production. Research supporting this point of view would also claim that generalisation from perception improvement to production would confirm the relation between these two speech processes (Akahane-Yamada, Tohkura, Bradlow & Pisoni, 1996; Bradlow et al., 1997; Hazan et al., 2005; Lengeris & Hazan, 2010; Lambacher et al., 2005). Interestingly, there seems to be conflicting evidence as for this direct relation between perception and production. Bradlow et al., (1997) found a greater effect of training on perception than on production. Such differences were expected, given that learners had been trained on perception but not on production. However, the amount of individual variability found in the perception as well as in the production improvement of the contrasts was not necessarily related. That is, the participants that improved more in perception were not necessarily the ones who obtained higher scores in their production. Similar results have been reported by Iverson et al.

(2012). They found improvement in perception and production of English vowels after training. They also found that individual differences in production were significantly correlated with identification performance at pre and post test but the production improvement was not correlated to identification accuracy before or after training.

In a recent study, criticism has been presented to previous research as for the impact of production training. Herd, Jongman and Sereno (2013) questioned the procedure used in production training studies which reported improvement in perception as well as production. The usual procedure exposes learners not only to watching some kind of visual input (wave forms, spectrograms or articulatory movements) but also to listening to the stimuli (Hirata, 2004). In their study, Herd and colleagues compared perceptual and production training of /d, ɾ, r/ in learners of Spanish with American English as L1. They carefully controlled the stimuli in each training paradigm so that participants could hear the sounds in the perceptual training but not in the production training. Instead, they presented the written form, the waveform and spectrogram of the word and were encouraged to practise and record the words presented. They also included a third combined "perception-production" training mode. The results showed similar results for the perception and production training, both groups improved their perception mainly and production to a lesser extent. However, they found that perceptual improvement differed depending on the contrast trained. Perception training showed more improvement in the /d, ɾ/ contrast and production training showed more improvement in the perception of the /ɾ, r/ contrast. While the combined perception-production group improved more than the other two training groups in production, they made no improvement in perception. These results would suggest that the perceptual sessions may need to be increased for this combined group to find gains in perception. The authors claimed that the number of sessions per group played a role in the results. They also suggested that the effectiveness of a given training modality depends on the contrast being trained. In their study, perception trained was more effective to improve the perception of two allophonic realisation of the same phoneme; however, production training was better to train learners on new phoneme

categories. These results contribute to question the traditional view of a direct relation between perception and production.

Elicitation techniques for production studies rely mainly on written material which may have an impact on the production data collected. Piske et al. (2010) compared errors produced by learners in continuous speech (recorded task) with reported errors for the same profile of participants in a previous study (Piske et al., 2002) when they had read words and non-words aloud. In the 2010 study, responses to auditory input were recorded and analysed. They found that the errors reported for early bilinguals when using written-input task were not present in the sentences produced by the same kind of learners when using auditory input and continuous speech in a recorded task. These results suggested that elicitation techniques for L2 production need to be taken into account when designing experiments and analysing the data, as they may interfere with participants' production inaccuracies.

In summary, research on the relation between L2 speech perception and production seems to have been dominated by the idea of a direct link between these two speech processes. More recently, a few studies have contributed to bring this link into question. In the future, more research needs to be conducted to further explore the relation between these two speech mechanisms in L2 learners.

One of the most studied problems in the perception of an L2 has been the influence of the L1 system and the interactions between the L1 and L2 sound inventory when perceiving the L2 sounds. Because the vowel inventory for Spanish and English differs in quantity and quality a brief reference to the L1 Spanish vowel system and the typical assimilation patterns reported for Spanish learners of English will be presented in the next section.

1.8 Spanish and English vowels

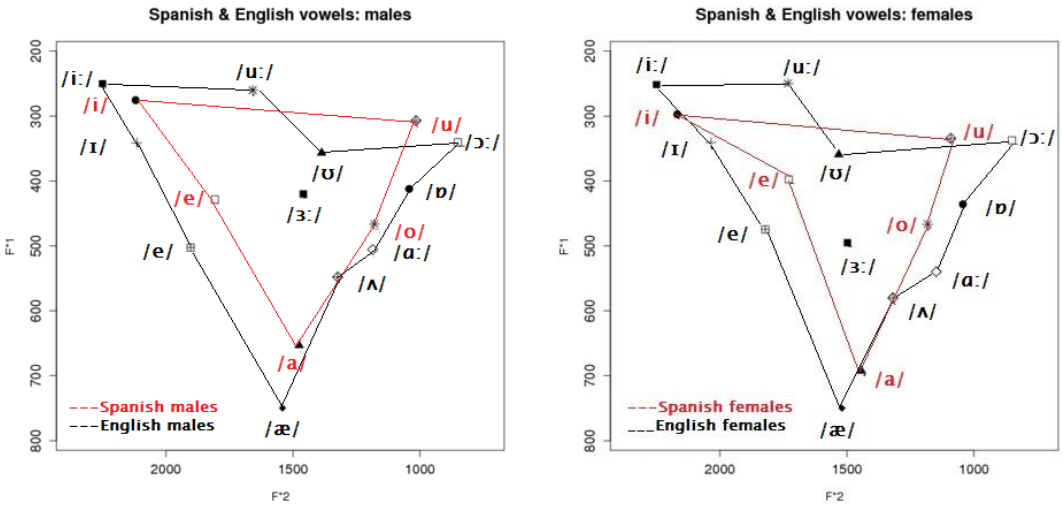


Figure 1.1 Vowel plots for English native speakers (5 males, 5 females) and Chilean-Spanish speakers (31 males, 30 females). Values for English vowels were taken from Hawkins & Midgley (2005) and Moreiras (2006) for male and females respectively. Data for Chilean Spanish speakers was taken from Sadowsky (2012). Formant values were normalised (Lobanov's method).

Research on the perception of non-native sounds has suggested that learners use their L1 phoneme categories to perceive L2 novel contrasts. This may result in, for example, new L2 sounds being assimilated to different phonemes in the L1 (Best & Tyler, 2007) or identified as two possible realizations of the same L1 phoneme (Flege 1995). The participants in this study have Chilean-Spanish as their native language, a five-vowel system (/i/, /e/, /a/, /o/, /u/) with well distinguished phonemes along the F1/F2 plane. These five vowels are not contrasted in terms of duration. The English vowel system is made up of twelve monophthongs (/iː/, /ɪ/, /e/, /ɜː/, /æ/, /ʌ/, /ɑː/, /ɒ/, /ɔː/, /ə/, /uː/, /ʊ/) which are spectrally distinct. The differences between the vowel inventories are not only in the number of vowels in each language but also in the spectral values of their formants as can be observed in the vowel plots for Standard Southern British English and Chilean-Spanish in Fig. 1.1.

In terms of spectral information, a more expanded space for English vowels can be noticed and the distance between English and Spanish vowels reveals very few English sounds are near a Spanish counterpart. For example, English vowel /iː/ is lower in F1 (i.e. more open) than Spanish /i/. On the other hand, English /ɪ/ seems closer to the Spanish counterpart but a bit higher in F1 and lower in F2, more back and a bit closer. In terms of similarities, neither vowel system uses lip-rounding, length or nasalization contrastively for vowels (Bradlow, 1995); although the English tense vowels have been described as having longer duration than the lax vowels (Ladefoged, 2006). Different studies have described the perception of English vowels by L2 learners, most of them describing pairs of vowels (Bent, Bradlow & Smith, 2008; Bohn & Flege, 1990; Flege, MacKay & Meador, 1999; Morrison, 2009; among others). However, Iverson and Evans (2007) presented in section 1.2.1 provided information on the whole vowel system assimilation patterns for four groups of L2 learners (German, Norwegian, Spanish and French). The learners heard a word (hVd) containing an English vowel and identified which "hVd" Spanish word sounded closest to the English "hVd" word. This study revealed assimilation patters; English /ɪ/, /iː/ were rated as the closest to Spanish /i/; English /e/ was assimilated to Spanish /e/; English /ɑː/ was assimilated to Spanish /a/; English /ɜː/, /ʌ/, /ɒ/ and /ɔː/ were assimilated to Spanish /o/ and English /ʊ/, /uː/ were assimilated to Spanish /u/. The Spanish and French learners of English initially assimilated more than one English vowel to a single L1 vowel whereas the German and Norwegian showed a pattern of one English vowel assimilated to one L1 category. The results suggested that the assimilation patterns cannot predict what L2 learners can learn with perceptual training, as all of them showed learning of spectral and durational cues for the perception of English vowels. Assimilation patterns will not be examined in the current study; however, confusions in the perception of English vowels may show some relation to the assimilation problems mentioned above.

The aim of this introductory chapter was to review and discuss different aspects that have been found to be relevant in the L2 speech perception and production field. In the next section, the main research question for the current thesis will be presented.

1.9 The current study

In the light of the issues presented in the literature review, the main research questions addressed in this thesis were:

a. To what extent are L2 learners with L1-Spanish sensitive to visual cues for English vowel contrasts?

-Are there visual cues available to English vowels (EV) to native speakers of English?

-How does the use of visual cues for English vowel identification compare between native speakers of English (ENS) and L2 learners?

b. Can L2 learners be trained to attend to visual cues in the perception of EV contrasts?

c. What is the effect of vowel training modality on the perception of EV contrasts?

-To what extent does perceptual training lead to improvement in the perception of EV in isolated words?

-To what extent does perceptual training lead to improvement in the perception of EV in more naturalistic speech at sentence-level with new tokens and talkers?

d. What is the effect of modality of perceptual vowel training on the production of EV?

-If change is found in the production of EV after training, Is this change due to enhancement of spectral or temporal cues for the tense-lax vowel contrasts?

e. Is L2 learners' perception of English vowels related to production?

Chapter 2

Use of visual cues in the perception of English vowels by L2 learners

Studies in L2 speech perception have found that visual cues aid perception of L2 sounds when the contrasts are salient (Hardison, 1999; Hazan et al., 2006). Most of these studies have focused on English consonants, using monosyllabic words (non-words) rather than real words (Alm, Behne & Wang, 2009; Hardison, 1999; Hazan et al., 2006; Schwartz et al., 2004). However, there is lack of research on the extent to which visual information may contribute to English vowel perception in L2 learners. The aim of this study was to investigate whether L2 learners with Chilean-Spanish (Ch-Spanish) as L1 can use visual information to improve their perception of English vowels.

Research on the perception of English vowels by native speakers using visual information only has found a lot of variability in the perception of viseme categories. This variability depends on the talker's visual intelligibility, the perceiver's speech-reading capacity, phonemic context (CV/VCV-words), length of the input and amount of light and viewing angle (Walden et al., 1981; Owen & Blazek, 1985). Lesner and Kricos (1981) compared the perception of 10 vowels and five diphthongs (American English) and found that diphthongs were more easily perceived than vowels. They also found patterns which showed a viseme could represent more than one phoneme. For example, the viseme {a} would stand for the lower back vowels /ɑ/, /ɔ:/, /ʌ/ and confusions for /ɑ/, /ʌ/ with /ɛ/, /æ/, /eɪ/ were also found. Another issue to consider is the relative dependence of the neighbouring sound which may facilitate visual perception. The visual perception of consonants has been found to depend on the vowel which follows. Owens and Blazek (1985) conducted a study on consonant viseme identification (VCV words) and found that vowel /ɑ/ contributed more

than /ʊ/ in consonant perception; the latter caused the identification scores to lower. Cohen, Walker & Massaro (1996) described viseme categories for American English vowels in five groups: high-front vowels (/iː/, /ɪ/), non-high front vowels (/eɪ/, /e/, /æ/, /aɪ/), lower-back and central vowels (/a/, /ɔː/, /ʌ/), mid-back rounded (/oʊ/) and high-back rounded (/ʊ/, /ʊ/). Though, these viseme categories are for American English; something similar would be expected for British English vowels in SSBE accent, bearing in mind the differences in the two vowel systems.

Regarding confusions between vowels from a speech-reading perspective; Kaplan, Bally and Garretson (1985) suggested that vowels are confused with the vowel just above or below in the vowel triangle. They also claimed that the most visually distinctive vowels are the ones in the extreme corner of the vowel triangle (/iː/, /a/, /ʊ/).

Research on the perception of English vowels by Spanish speakers has suggested that Spanish learners tend to perceptually assimilate the tense-lax English vowels to the nearest Spanish category (Flege et al., 1997, Fox et al., 1994; Iverson & Evans, 2007; Morrison, 2008). Thus, auditory confusions for the English tense-lax contrasts by the participants in the current study are expected. In the absence of research concerning the use of visual cues for English vowel perception by L2 learners, it could be hypothesized that beginner learners may show visual confusions for the English tense-lax vowels contrasts and for the /æ/–/ʌ/–/ɑː/ contrasts. For instance, English /iː/–/ɪ/ would be visually confused with Spanish /i/; English /ʊ/,/uː/ confused with Spanish /u/; and English /ɒ/,/ɔː/ with Spanish /o/; the English /æ/–/ʌ/–/ɑː/ would be confused with Spanish /a/. Based on findings that suggest that the use of visual cues increases with more language experience (Wang et al., 2008), the more experienced learners in this study (L2 advanced) would be expected to benefit from visual cues to improve their vowel identification in AV mode and to reduce the amount of confusions. However, it remains to be seeing how ENS make use of visual cues for English vowel perception.

2.1 Aims

The aim of this study was to find out whether L2 learners can use visual cues in the perception of English vowels. To explore this, the following research questions were used to guide this study:

a) Which vowels can be reliably identified by native English speakers on the basis of visual information only.

b) To what extent L2 learners are sensitive to visual cues in the perception of English vowels.

c) If they are, to what extent do L2 learners integrate visual cues for the perception of English vowels in audio-visual mode?

d) Are individual differences in English vowel perception related to L2 learners' auditory frequency discrimination and degree of visual bias?

e) Is the L2 learners' capacity to perceive key-words in sentences related to vowel identification in isolated words?

f) Does the use of visual cues in L2 speech perception vary as a function of L2 proficiency level?

2.2 Methods

2.2.1 Participants

a. L2 beginner group: 47 university students (32 female, 15 male) were recruited and tested in Chile at Universidad de Concepcion. They were all native speakers of Spanish and were in their first year of the Teacher of English Training Programme and the Translation Programme. They had a "beginners-level" of English (Common European Framework for Languages (CEF) A2-B1). Participants' age ranged from 18 to 23 years old (M: 19.6, SD:

1.1) and all had had previous experience with English at High school. Students typically start receiving English instruction at school from the age of twelve.

All participants had had a month of intensive instruction in English at university. A measure of the participants' knowledge of English was obtained from the overall results of the standardized Cambridge Preliminary English Test (PET) that targets the Common European Framework B1 level (beginners). This exam includes listening comprehension, reading, writing, grammar and vocabulary assessment. All correct answers from the Reading (25 points), Writing (25 points) and Listening test (25 points) were transformed into percentage. A total of 75 points equalled 100%. Participants self-reported not having any hearing impairment and agreed to be volunteers in this study.

b. L2 advanced group: 37 university students, aged between 22 and 27 years old (M:23.5, SD: 1.2) there were 34 female and 3 male participants. This group of L2 learners of English were recruited at Universidad San Sebastian in Concepcion, Chile. They were students of the teacher training programme in their fourth year of the programme; the undergraduate programme length is five years. They were all proficient L2 speakers of English (CEF B2-C1). To establish a measure of their proficiency level, the final mark from their just completed term of their English module was transformed into percentage. Marks are given on a 7 point scale, 7 equals 100%. This is a comprehensive measure of proficiency as it covers the assessment of reading, writing, listening and speaking. All participants self-reported not having any hearing impairment and agreed to be volunteers in this study.

c. English native speakers (ENS): 20 native speakers of English were recruited in London and tested at UCL in the Chandler House Speech Sciences Laboratory. Participants were university students with Standard Southern British English (SSBE) accent and their age ranged between 23 and 28 years old (M: 25.1; SD: .9). There were 16 female and 4 male native speakers. Participants received a small payment for their collaboration. All participants self-reported not having any hearing impairment.

2.2.2 Test battery

The test battery included tests of vowel identification in three modes (audio, audio-visual and video-only), as well as further tests to gain more information about the participants' auditory acuity for frequency discrimination, degree of visual bias in speech perception and comprehension for sentence level materials.

The main test used to measure vowel identification was a *Vowel Test* which presented stimuli in three modes: audio (A), audio-visual (AV) and video-only (V). This vowel identification test was given to the L2 beginners and L2 advanced learners in clear. Noise was added to the vowel identification test given to the ENS to make results comparable. A frequency discrimination test was included to measure auditory acuity of the L2 beginners and L2 advanced learners. This measure was included based on recent findings which suggested that good auditory acuity was related to more sensitivity to acoustic cues in a new vowel contrast (Lengeris & Hazan, 2010). In the current study, L2 beginners' degree of visual bias was measured with a McGurk effect test as in Sekiyama, Burnham, Tam and Erdener (2003). This measure was found to relate to speech perception in AV mode in Chen and Hazan (2009). Finally, to test the degree to which intelligibility increased as a result of adding visual cues in the perception of sentence-length material, a BKB-sentence test was given to L2 beginner participants in audio (BKB-A) and audio-visual (BKB-AV) mode. It was not possible to give the McGurk test and the BKB-sentence test to the L2 advanced group given to the participants' availability. All tests, except for the frequency discrimination test (non-speech), involved speech stimuli produced by Southern British English (SSBE) native speakers.

Test battery materials

a. Vowel test

Audio-visual stimuli of 11 English vowels (/æ/, /ʌ/, /ɑ:/, /ɪ/, /i:/, /e/, /ɜ:/, /ɒ/, /ɔ:/, /ʊ/, /u:/) embedded in /bVt/ and /hVd/ words were used. Four native speakers of Southern British English (2 males & 2 females) recorded a list of 61 randomized words

containing English vowels in the two contexts described above (e.g. “bat”, “had”, “bet”, “head”). Three repetitions per word were included to make a selection later. The laptop screen was located below the camera and the speaker was asked to read the word, then look up and repeat it to the camera. Words that constitute non-words were excluded from the list (no word for the pronunciation /bʊt/ was found). Another list of words containing the same 11 English vowels was filmed using a different SSBE male speaker reading a randomized list of 33 frequent words (11x3); these were used as examples before the vowel test started (“*cat, cup, card, sit, pet, feel, word, pot, caught, full & food)* and they were also used as the “response button word”.

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 (Fig. 2.1). In order to get a high-quality audio recording, a Bruel & Kjaer 2231 microphone was connected to a DAT recorder and recordings were made at a sampling rate of 48 kHz. The video material was digitally transferred to a PC and time-aligned with the DAT audio recording; the original audio on the video was then replaced by the audio on the DAT tape. 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 (11) in two contexts (“hVd”, “bVt”-words), except for /ʊ/ (10 vowels x 2 contexts and 1 vowel x 1 context). The video material was edited using Adobe premier pro and was then compressed to Microsoft Video1 format using Virtualdub software. The same procedure was used for all the video material included in the test battery.

b. Frequency discrimination test

To measure frequency discrimination acuity, a frequency discrimination test which used a non-speech single-formant continuum was used (as in Lengeris & Hazan, 2010). The synthesised stimuli were 150 ms in duration and included a single formant with a 100 Hz bandwidth which varied in frequency from 1250 to 1500 Hz, similar to a vowel second formant (F2). Fundamental frequency (F0) was constant at 120 Hz (male-speaker like). There were 51 tokens in the continuum which varied in equal formant frequency steps.

c. McGurk test

To collect information about the participants' visual bias, the weighting of auditory and visual information was measured by using a McGurk test (McGurk & McDonald, 1976). This test measures the influence of visual speech on perception when the auditory component of the stimulus is discrepant with the visual component. This test was comprised of short recordings and video clips of two English speakers (1 male & 1female) producing CV syllables (ba, ga, da) which were a subset of the stimuli also used in Chen and Hazan (2009). A set of 'incongruent' stimuli created by cross-splicing the audio channel from one syllable with the video from another (e.g. auditory /ba/+ visual /ga/) was included. These incongruent stimuli were used to investigate whether listeners gave greater weighting to the auditory or visual information when these were put in conflict. The stimuli (Table 2.1) used in this test consisted of /ba, da, ga/ syllables presented audio-visually in clear (AVcl) and in noise (AVn) using congruent and incongruent tokens, in audio condition in clear (Acl) and in noise (An) and in video-only condition with no sound (V). A signal-to-noise-ratio (SNR) of -12 dB was used for the noise conditions (A, AV). Further details about stimulus construction can be found in Chen and Hazan (2009).

Table 2.1 Stimuli used in the McGurk test.

Modality	Token	Number of tokens	Total
A clear	"ba, "da", "ga"	6 each	18
A noise	"ba, "da", "ga"	6 each	18
AV clear congruent	"ba, "da", "ga"	6 each	18
AV clear incongruent	"ba_ga", "ga_ba"	6 each	12
AV noise congruent	"ba, "da", "ga"	6 each	18
AV noise incongruent	"ba_ga"	6 each	6
Video	"ba, "da", "ga"	6 each	18

d. BKB-sentence test

To have a sense of the participants' perception of sentence-length material, two lists of BKB sentences (Bench, Kowal & Bamford, 1979) were used. Video recordings of these lists by a 21-year-old female native speaker of English (SSBE) were made at UCL

(Faulkner & Rosen, 1999). Each set includes 16 sentences containing simple lexical and syntactical items that would be appropriate for children native speakers. Some examples of BKB sentences are: "*the clown had a funny face", "the car engine's running", "the angry man shouted"* and *"the dog sleeps in a basket".* Each set of sentences contains 50 key words.

Procedure

The tests were presented to participants individually. L2 Beginner learners had two sessions of 60 minutes for session 1 and 20 minutes for session 2, L2 Advanced learners and English native speakers had only a 60-minute session. The testing of the L2 Beginner group was conducted in the Phonetics Lab of the Spanish department at Universidad de Concepcion, Chile. The L2 Advanced group was tested at the English Department PC Lab at Universidad San Sebastian Concepcion-Chile. Native speakers of English were tested in London in the Research Lab at Chandler House, University College London. All tests were presented using headphones (Genius HS-04SU) connected to a laptop. Participants were allowed to adjust the volume of their headphones at a comfortable listening level when doing the tests. Participants were allowed short breaks in-between tests. First, participants were given written information about the purpose of the study and the test battery they would be given in the session. They were allowed to ask questions prior to signing a consent form. Tests were presented in the following order from non-speech to speech stimuli:

<u>L2 beginners: session 1</u>

- frequency discrimination test
- McGurk test
- Vowel Test

<u>L2 beginners: session 2</u>

- BKB-sentence test

<u>L2 advanced: session 1</u>

- Frequency discrimination test

- Vowel Test

English native speakers: session 1

- Vowel Test

The McGurk test, the BKB-sentence test and the vowel test were presented using CSLU Toolkit software, whereas the frequency discrimination test was presented on a platform developed at UCL. All instructions were presented in writing in English, to ENS, and in English and Spanish to L2 learners; the researcher was present in the room during the whole session.

a. Vowel test

This test was given to the three groups of participants. Although it might have been preferable to present all 11 vowels in a single test, in order to get an 'unconstrained' pattern of vowel confusions, it was felt that the inclusion of 11 response options would be too confusing for L2 speakers at beginner level. Therefore, all /bVt/ and /hVd/ words were grouped into three separate sets, each including 3 or 4 vowels. The selection of the vowels to be included in each set was based on research reporting the most frequent confusions of English vowels for native Spanish speakers learning English (Garcia-Lecumberri & Cenoz, 1997, Ortega-Llebaria *et al.* 2001, Iverson & Evans 2007). These studies showed that the most frequent English vowel confusions for L2 learners are: {/æ/, /ʌ/}, {/e/, /ɪ/, /iː/}, {/ɒ/, /ɔː/}. In the Vowel test, English vowels were clustered in three sets (*Set 1*[/æ/, /ʌ/, /ɑː/], S*et 2*[/ɪ/, /iː/, /e/, /ɜː/], and *Set 3* [/ɒ/, /ɔː/, /ʊ/, /uː/]). Because the tokens used in the test were "hVd" and "bVt" words which may not all have been known by the learners, more familiar words were used as "response buttons" in the practice phase and in the test itself in order to make the response procedure easier. They were simple words and were clustered in three sets (set1 [*cat, cup, card*], set2 [*sit, pet, feel, word*] *and* set3 [*pot, caught, full, food*]). 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 by asking them when they had had the first practice phase, No one seem 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. In the instructions, participants were told they were going to watch/listen to a male speaker saying English words containing an English vowel as "example-words" ("response button). After that, they would watch/hear different words using the English vowels said by other 4 people (2 male, 2 female) and they had to click on the word containing the vowel they had just heard/watched choosing from the "response buttons" (Fig. 2.1). There was no feedback on participants´ answers. This test took around 30 minutes.

White noise was added to the Vowel test given to ENS 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.

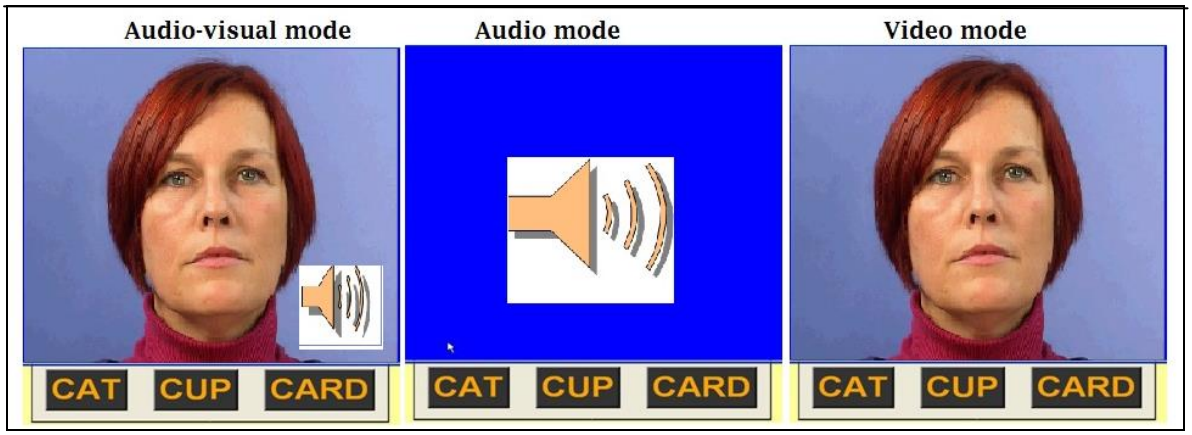


Figure 2.1 Three screens of the Vowel test (Set1). Screens from left to right show the Audio-visual (AV), the Audio (A) and Video (V) mode. Response buttons were displayed at the bottom but appeared after the token had been played.

b. Frequency discrimination test

This test was given to the L2 Beginner and L2 advanced group. In the frequency discrimination test, participants were presented with a forced-discrimination task (three-choices). They saw three frogs on the computer screen; each one jumped making a sound (nonspeech token). Participants were told that two of the frogs would make the same sound

and one would be different. They should click on the one that was different (odd-one-out task).

Participants could see immediately whether their answers were right or wrong as ticks or crosses appeared on the chosen frog as a way of feedback. One endpoint of the continuum (1250 Hz) was used as the 'standard' or reference token and the frequency of the other token was roving. The inter-stimulus interval time was 200 milliseconds. To find the just noticeable difference (jnd) a three-down/one-up adaptive procedure (Levitt, 1971) was used to assess the frequency difference at which tokens were discriminated from the standard 79% of the time. The test ended after seven reversals or after 50 trials. The "jnd" was obtained from the mean of the last four reversals. The test took around 3-5 minutes.

c. McGurk test

This test was given to the L2 Beginner group only. The test presented participants with randomised stimuli under three main conditions in the following order: audio-visual congruent (in clear/noise), audio-visual incongruent (in clear/noise), audio (in clear & noise) and video-only stimuli. After each presentation, participants were given a close-set choice of "BA, DA, GA" response buttons; these buttons only appeared once the video or audio sequence had finished. This test took around 12 minutes.

d. BKB-sentence test

This test was given to the L2 Beginner group only. Two lists of BKB-sentences were used: List A was presented in A mode and list B in AV mode to half of the participants, while the other half received list A in AV mode and list B in A mode. Participants had to click on a button to watch/hear a sentence and were asked to repeat as many words as they could, immediately after each sentence was heard or watched. The researcher kept a record on paper of all the words they repeated correctly. No feedback was provided on participants' responses. The number of correctly repeated keywords for each sentence was computed later (maximum of 50 per list). This test took around 15 minutes.

Table 2.2 Test battery summary. Test, time, stimulus mode and participant group. The McGurk test and BKB-sentence test were not given to L2 learners due to availability problems.

Test	Stimuli mode	Participants
Vowel test 30 minutes	Audio Audio-visual Video-only (in clear for L2 learners/in noise for ENS)	L2 beginners L2 advanced English native speakers (ENS)
Frequency test 3-5 minutes	Audio	L2 beginners L2 advanced
McGurk test 12-15 minutes	Audio (clear/noise) Audio-visual (clear/noise/congruent/incongruent) Video	L2 beginners
BKB-sentences 15 minutes	Audio Audio-visual	L2 beginners

2.3. Results

2.3.1 Vowel Test

To measure the perception of English vowels by the two groups of L2 learners (Beginners and Advanced) and English native speakers (ENS), a Vowel test that presented 11 English monophthongs in three different conditions in clear (A, AV and V) was used. Due to a technical problem, data for L2 beginners in A mode Set 1 (/æ/, /ʌ/, /ɑ/) was not available for this group. Thus, data for eight vowels in three conditions (A, AV and V) was used to compare overall results for the three groups of participants.

Overall correct identification means for eight vowels per group showed L2 beginners obtained the lowest scores (A: 58%, AV: 58%, V: 40%), followed by L2 Advanced learners (A: 65%, AV: 67%, V: 47%). The highest results were scored by the ENS (A: 82%, AV: 88%, V: 56%), even though a relatively low SNR was used in their test. These scores were higher than the ones obtained in the pilot test when the -10 dB SNR was decided for the ENS test. Means were calculated per participant and then per group.

A Mixed-effects model was chosen to analyse all the data presented in this thesis. The reason for choosing this type of analysis was that it has the advantage of considering the fixed effects which influence the mean in the data and the random effects which influence the variance. Thus, the analysis becomes more powerful when the variability usually introduced by individuals or the stimulus becomes part of the model (Crawley, 2007).

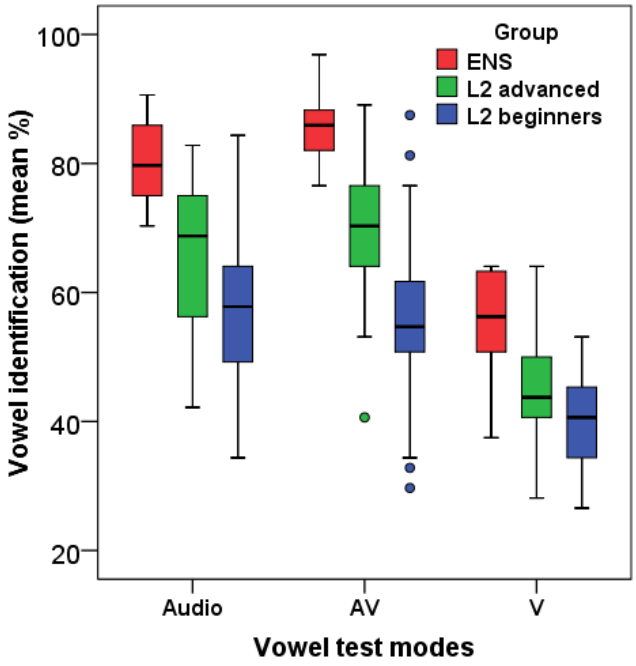


Figure 2.2 Boxplots of identification of English vowels in the Vowel Test for 8 vowels in three modes (A, AV, V). The groups tested were L2 beginners, L2 advanced learners and English native speakers (ENS). All comparisons per condition between groups were significant p<0.001.

Table 2.3 Results for the Vowel test data analysis for L2 beginners, L2 advanced learners and English native speakers (ENS). All effects and interactions were significant.

Effects	**F**	**Sig.**
Group	(2, 101)= 40.069	<.001
Stimulus	(7, 707)= 99.878	<.001
Mode	(2, 17.818)= 373.037	<.001
Mode*group	(4, 17.818)= 11.325	<.001
Stimulus*group	(14,707)= 5.814	<.001
Stimulus*mode	(14, 17.818)= 14.645	<.001
Stimulus*mode*group	(28, 17.818)= 4.644	<.001

A Logistic regression was used to analyse the data (individual response data) using the statistical software R (glmmPQL function). Group (L2beginner, L2advanced, ENS), mode (A, AV, V), vowels (8 vowels), vowel*mode, vowel*group and vowel*mode*group were used as fixed effects. Participants and stimulus were used as random factors. Effects and interaction values are reported in Table 2.3. The between-subject effect of listener group was found to be significant with higher overall scores for ENS (*M*: 75%) than the two non-native groups, even though the vowel identification task had been made more difficult by the addition of noise. The L2 advanced (M: 60%) group outperformed the L2 beginners (M: 51%). All between-group mean comparisons were significant at p<.001 (Fig. 2.2).

The effect of mode (A, AV, V) was significant, overall means for V were lower than A and AV but A and AV mode did not differ. However, this effect was modified by a mode per group interaction which showed that for the ENS AV scores were higher than A and V, with A higher than V mode. Whereas for the L2 groups, there was no significant difference between A and AV scores, both higher than V mode (Table 2.4).

Table 2.4 Overall means per mode (A, AV, V) in the Vowel test for each of the three groups tested. Level of significance between modes is shown in the Mode difference column (p value).

Group	Mode (Mean; SD)	Mode diff. (p)
L2beginners	A (58;11), AV (57;12), V (40;7)	A-AV: >.05 A-V : <.001 AV-V: <.001
L2advanced	A (65;11), AV (67;10), V (47;7)	A-AV: >.05 A-V : <.001 AV-V: <.001
ENS	A (82;6), AV (88;5), V (56;8)	A-AV: <.05 A-V : <.001 AV-V: <.001

The vowel effect (8 vowels) was significant, with scores ranging from 45% for vowel /ʊ/ to 77% for /e/. There were also a vowel*group, vowel*mode and a vowel*mode*group interaction. Post hoc tests were conducted to explore the three-way interaction which included the other two interactions. The results of this analysis, based on data for eight vowels, showed that for ENS there was a visual advantage for three vowels (/ɪ/, /ɜ:/, /u:/) with higher scores in AV than A and V (AV>A>V) and no significant difference between

A and AV for the other five vowels (/iː/, /e/, /ɒ/, /ɔː/, /ʊ/). In general, the L2 groups obtained similar scores in A and AV mode (A=AV>V), but significantly higher scores were found for the L2 advanced group with a visual advantage for two vowels (/ɜː/, /ʊ/) in AV mode (Table 2.5).

Table 2.5 Means for vowel identification per vowel, mode and group. The significant level for the difference between A, AV& V modes is presented in the Mode difference column (p value). Values for Set 1 vowels were included for general reference for ENS, and L2 advanced group. There was missing data for L2 beginners in A mode, so only 8 vowels were used in the comparison of the three groups.

Stimulus	Mode	L2beg.(M)	Mode diff. (p)	L2adv. (M)	Mode diff. (p)	ENS (M)	Mode diff. (p)
/iː/	A	43	A-AV : >.05	55	A-AV : >.05	79	A-AV : >.05
	AV	44	A-V : <.001	58	A-V : <.001	85	A-V : <.001
	V	28	AV-V : <.001	39	AV-V : <.001	44	AV-V : <.001
/ɪ/	A	63	A-AV : >.05	76	A-AV : >.05	88	A-AV : <.05
	AV	58	A-V : <.001	76	A-V : <.001	96	A-V : <.001
	V	38	AV-V : <.001	40	AV-V : <.001	46	AV-V : <.001
/e/	A	83	A-AV : >.05	87	A-AV : >.05	82	A-AV : >.05
	AV	83	A-V : <.001	90	A-V : <.001	83	A-V : >.05
	V	56	AV-V : <.001	63	AV-V : <.001	73	AV-V : <.05
/ɜː/	A	65	A-AV : >.05	76	A-AV : <.001	80	A-AV : <.05
	AV	66	A-V : >.05	85	A-V : >.05	89	A-V : >.05
	V	70	AV-V : >.05	74	AV-V : <.001	85	AV-V : >.05
/ɔː/	A	48	A-AV : >.05	65	A-AV : >.05	91	A-AV : >.05
	AV	44	A-V : <.001	63	A-V : <.001	96	A-V : <.001
	V	20	AV-V : <.001	27	AV-V : <.001	43	AV-V : <.001
/ɒ/	A	64	A-AV : >.05	73	A-AV : >.05	89	A-AV : >.05
	AV	67	A-V : <.001	66	A-V : <.001	89	A-V : <.001
	V	36	AV-V : <.001	45	AV-V : <.001	51	AV-V : <.001
/uː/	A	47	A-AV : >.05	48	A-AV : >.05	78	A-AV : <.001
	AV	45	A-V : <.001	51	A-V : <.001	93	A-V : <.001
	V	33	AV-V : <.001	30	AV-V : <.001	50	AV-V : <.001
/ʊ/	A	39	A-AV : >.05	50	A-AV : <.05	52	A-AV : >.05
	AV	42	A-V : >.05	66	A-V : <.05	52	A-V : >.05
	V	31	AV-V : <.05	36	AV-V : <.001	51	AV-V : >.05
/ɑː/	A			63	A-AV : >.05	72	A-AV : >.05
	AV			65	A-V : <.001	79	A-V : >.05
	V			50	AV-V : <.001	71	AV-V : >.05
/æ/	A			56	A-AV : >.05	85	A-AV : >.05
	AV			57	A-V : >.05	91	A-V : >.05
	V			55	AV-V : >.05	89	AV-V : >.05
/ʌ/	A			58	A-AV : >.05	65	A-AV : <.001
	AV			61	A-V : <.05	87	A-V : >.05
	V			48	AV-V : <.001	64	AV-V : <.001

All together, these results showed that there was visual information available in V mode for English vowel identification for the 11 vowels tested. Scores in V mode for ENS were all above chance, and showed to contribute to higher scores for vowel identification in AV mode (in noise) for most vowels; although it only became significantly higher for four vowels (Table 2.5). L2 learners were also able to identify vowels in V mode with lower scores than ENS, but their use of visual cues varied as a function of learning experience. One issue that stands out is that both L2 learners obtained very high scores in V for vowels /e/ and /ɜ:/ but their scores in AV were not higher than A. This suggests that even though there were some salient visual gestures that distinguished these vowels in their set, L2 learners were not able to integrate the visual information available to improve their vowel identification in AV mode and relied mainly on the audio information

There was also a tendency of higher scores for lax vowels in both non-native groups. This may suggest that L2 learners with Spanish L1 may be biased towards English vowels that show no tenseness or duration feature because those features are not used in their L1 vowel inventory.

2.3.2 Video mode: ENS versus L2 learners

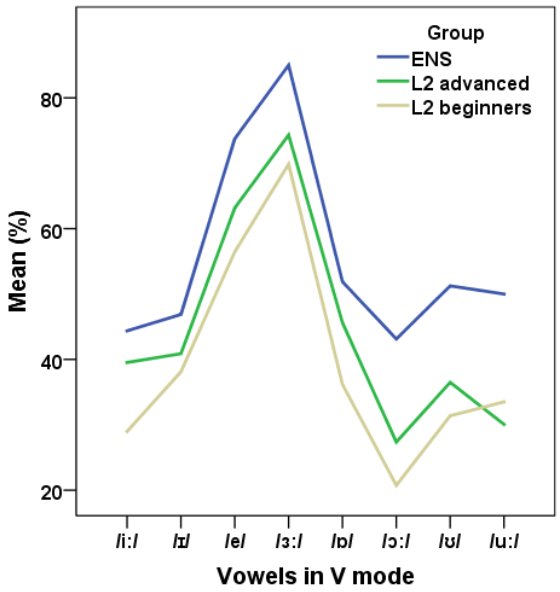


Figure 2.3 Line graph for identification of English vowels in the V mode by ENS, L2 beginners and L2 advanced groups. The data used was based on the eight-vowel data from the main analysis for the Vowel test.

The main analysis for the vowel test presented above (2.3.1) showed significant difference for all modes between groups, with lower scores for the L2 learners. With regards to the use of visual cues for English vowel perception in V mode (Fig. 2.3), ENS scores were above chance level for the eight vowels tested with overall means ranging from 43% for /ɔ:/ to 85% for /ɜ:/. L2 learners were able to attend to visual information in the "forced" experimental condition in the V mode; however, their identification capacity was not as good as the ENS' and varied as a function of their level of proficiency (Fig. 2.3). The L2 advanced group showed similar scores to ENS for three vowels (/i:/, /ɪ /, / ʊ/) in V mode, with scores above chance level for 7 vowels and near chance for vowel /ɔ:/. L2 beginners obtained similar scores to ENS only for one vowel (/ɪ/), and scores above chance level for six vowels, and two near chance (/i:/, /ɔ:/) . In spite of this slight advantage in scores for the advanced L2 group, no difference was found between learners in their lack of capacity to integrate visual cues to auditory information in AV mode. One implication of this lack of integration of visual information was that L2 learners could not improve their perception of English vowels in a more native-like fashion in AV mode (Fig. 2.3).

ENS were not only able to attend to visual cues in V mode but also managed to integrate this visual information available to identify English vowels in AV mode (noise added) better than in A mode (noise added). These findings support the idea of the availability of visual cues for English vowels to native speakers and the potential benefit of this visual information for non-native speakers if they could learn to interpret it.

2.3.3 Confusions in Video mode by native and non-native speakers in the Vowel test

To explore the confusions found when identifying English vowels in Video-alone (V) mode, confusion matrices are presented for the English native speakers (ENS), L2 beginners (L2beg.) and L2 advanced (L2adv.) learners. A general pattern of confusion between tense and lax English vowels was expected based on confusions reported in audio-alone condition from previous L2 speech perception research and the five-vowel Spanish

system. However, there are no previous studies that have included vowel identification in video-alone condition for L2 learners. The predictions about possible confusions may be based on visual speech-reading studies on English vowels with native speakers that report that phonemes may share the same viseme (Owens & Blazek, 1985). This may lead to expected confusions between tense-lax vowels for native speakers and L2 learners as well.

The figures in the cells represent percentage of response per stimulus. Stimulus presented on the horizontal axis and participants' response (PR, downwards). The information in V mode was available for the three groups, so vowels in Set 1 were also included.

Table 2.6 Confusions for vowels in Set 1, V mode by ENS and L2 learners. Participants' responses are presented in percentages. Results should be read downwards; chance level is 33.3%.

Video	**Response % (ENS)**			**Response% (L2Beg.)**			**Response% (L2Adv.)**		
Stim.	**/ɑ:/**	**/æ/**	**/ʌ/**	**/ɑ:/**	**/æ/**	**/ʌ/**	**/ɑ:/**	**/æ/**	**/ʌ/**
/ɑ:/	71.2	2.5	29.4	44.1	19.4	21.8	49.7	12.1	25.3
/æ/	2.5	89.4	6.2	18.6	47.1	32.2	12.2	55.1	26.7
/ʌ/	26.2	8.1	64.4	37.2	33.5	46	38.2	32.8	48

Set 1 (Table 2.6). Identification scores based on visual information only was above chance level (33.3%) for the ENS and the non-native groups. Based on visual gestures of the Spanish vowel /a/, which is central and open but not as open as English /ɑ:/, it was expected that L2 learners would show more confusions for /ɑ:/ and /ʌ/ and vowel /æ/ would be easier to tell apart because it shares less visual gestures with the Spanish /a/. The results for L2 learners showed bidirectional confusions (/ɑ:/←→/ʌ/, /æ/←→/ʌ/) with lower amount of confusion between /ɑ:/ and /æ/. This suggested that, indeed, these two vowels share less visual gestures for L2 learners. In general, the pattern of confusions was similar for both L2 groups and suggested that longer experience with the language may have contributed to better identification of visual cues for two vowels in this set (/ɑ:/, /æ/) .

Table 2.7 Confusions for vowels in Set 2, V mode by ENS and L2 learners. Participants' responses are presented in percentages. Results should be read downwards; chance level is 25%.

Video	Response (ENS)				Response (L2Beg.)				Response (L2Adv.)			
Stim	/ɪː/	/ɪ/	/e/	/ɜː/	/ɪː/	/ɪ/	/e/	/ɜː/	/ɪː/	/ɪ/	/e/	/ɜː/
/ɪː/	44.4	25.6	7.5	0.6	29	20.8	15.4	5.9	39.5	26	12.5	2.7
/ɪ/	45	47	13.8	5.6	42.5	38	16	10.4	40.2	41	16.6	8.1
/e/	10.6	23.7	73.8	8.8	25	35.3	56.4	14	18.6	28.6	63.2	14.9
/ɜː/	0	3.8	5	85	3.5	5.9	12.2	69.7	1.7	4.4	7.8	74.3

In *Set 2* (Table 2.7), identification scores were mostly above chance (25%) for the three groups. The pattern of confusion for /iː/←→/ɪ/ was similar for the three groups, the confusions were bidirectional but stronger for /iː/→/ɪ/. This pattern of confusion was only expected in non-native speakers but it may suggest that the two vowels may be represented by the same viseme for ENS. The L2 beginners also confused /ɪː/and /ɪ/ with /e/, this may be related to a shared visual gesture with the Spanish lip position for /e/. The most visually salient vowel in this set was /ɜː/ for the three groups. This is a non-existent vowel in Spanish and its lip-rounding may have contributed to make it more visually salient when presented in this group of vowels. However, it could have been different if presented together with vowels in set 3 which show different degrees of lip-rounding. The high scores for /ɜː/ may suggest that new visemes were easier to perceive for L2 learners, as an analogue to what the Speech Learning Model (SLM) suggests for L2 phonemes that are non-existent in the learners' L1.

Table 2.8 Confusions for vowels in Set 3, V mode by ENS and L2 learners. Participants' responses are presented in percentages. Results should be read downwards; chance level 25%.

Video	Response % (ENS)				Response % (L2Beg.)				Response% (L2Adv.)			
Stim.	/ɔː/	/ɒ/	/uː/	/ʊ/	/ɔː/	/ɒ/	/uː/	/ʊ/	/ɔː/	/ɒ/	/uː/	/ʊ/
/ɔː/	43.1	20.6	20.6	13.8	20.7	38.3	15.7	33	27.4	39	10.5	29.1
/ɒ/	14.4	52	9.4	6.2	16.8	36.2	19.7	10.6	15.2	45.6	17.5	12.1
/uː/	27	6.2	50	28.8	30.6	11.7	33.5	25	32.4	6.7	30.1	22.3
/ʊ/	15.5	21.2	20	51.2	31.9	13.8	31.1	31.4	25	8.7	41.9	36.5

Most of the English back vowels were presented in *Set 3* (Table 2.8). Due to these vowels' place of articulation, visual information was expected to be less informative for perceivers. The lowest scores in this set were for vowel /ɔ:/ by ENS and L2 learners. ENS showed scores above chance level (25%) for the four vowels. L2 learners' results were mostly just above chance. Confusion patterns by L2 learners showed that the identification of the tense-lax contrast was really hard, not being able to tell the difference between the contrastive pair.

The aim of this subset test in the Vowel test was to find how much visual information was there for English native speakers (ENS) in this "forced-choice condition" and to compare it with the L2 learners' capacity to attend to visual information for English vowels in V alone mode. The results showed that all scores for ENS were above chance with some vowels being more visually salient, at least within the set in which they were presented. These were mainly open/mid-open front/central vowels (/æ/, /e/, /ɜ:/) and open back (/ɑ:/) due to more visible articulation and liprounding or spreading of the lips as well. The two non-native groups showed similar patterns of confusions in V mode, in spite of the slightly better identification performance by the L2 advanced group. Most of the confusions for non-native participants may be related to their lack of experience with the tense-lax vowel contrasts which are non-existent in Spanish and may suggest that some tense-lax contrasts (e.g. /i:/–/ɪ/) share visual gestures which cause problems for the native and also the non-native perceivers. This has been reported that /i:/–/ɪ/ vowels fall within the same viseme categories for American English perceivers (Lesner & Kricos, 1981). There were some back vowels which were less visually salient for ENS; thus, it is not surprising that non-native speakers also struggle in identifying them in V mode. After establishing the degree of visual information available for English vowels in V mode for ENS and L2 learners, the use of this visual information to improve vowel identification in AV mode needs to be explored.

2.3.4 Confusions in vowel identification in A and AV mode in the Vowel test

In this section, an analysis of the patterns of vowel confusions in the A and AV conditions will be provided as an additional view to the main analysis conducted so far. Vowel confusions and the contribution of visual cues in AV to improve vowel identification in native and non-native speakers will be explored. Patterns of confusions for contrastive tense-lax pairs were expected for the L2 learners.

In *Set 1* (A and AV mode), ENS reduced their degree of confusion in AV (noise) for the three vowels, in spite of already having high scores in A mode in noise (Table 2.9). This suggested that ENS benefited from visual information in AV mode (noise) even when vowel perception was not so affected by the added noise. For L2 advanced learners (vowels in quiet), the patterns of confusions were slightly shifted from A to AV mode for two vowels (/ɑ/→/ʌ/, /æ/→/ʌ/) but the amount of confusion remained the same. Data for A mode for L2 beginners was not available due to a technical problem; the AV scores are presented only to illustrate that the confusions they showed in AV mode were similar to the pattern in the advanced group (Table 2.10a,b).

Table 2.9 Confusions for vowels in Set1 by ENS in A and AV mode (in noise). Participants' responses are presented in percentages; chance level 33.3%.

	A Response % (ENS)			AV Response % (ENS)		
Stim	**/ɑ:/**	**/æ/**	**/ʌ/**	**/ɑ:/**	**/æ/**	**/ʌ/**
/ɑ:/	71.9	6.9	8.8	79.4	1.8	2
/æ/	9.3	85.6	23.8	1.2	91.2	11
/ʌ/	18.8	7.5	65	19.4	7	87

Table 2.10a Confusions for vowels in Set1 by L2 beginners in A and AV mode (missing data in A). Participants' responses are presented in percentages; chance level 33.3%.

	A Response % (L2 Beg.)			AV Response % (L2 Beg.)		
Stim	**/ɑ:/**	**/æ/**	**/ʌ/**	**/ɑ:/**	**/æ/**	**/ʌ/**
/ɑ:/				62	17.7	12.2
/æ/				19	51.3	37.8
/ʌ/				19	31	50

Table 2.10b Confusions for vowels in Set1 by L2 advanced groups in A and AV mode. Participants' responses are presented in percentages; chance level 33.3%.

	A Response % (L2 Adv.)			AV Response % (L2 Adv.)		
Stim	**/ɑ:/**	**/æ/**	**/ʌ/**	**/ɑ:/**	**/æ/**	**/ʌ/**
/ɑ:/	62.8	18.6	9.8	65.5	12.5	9.5
/æ/	19.3	55.7	31.8	9.8	56.8	29.1
/ʌ/	17.9	25.7	58.4	25	30.7	61.5

For vowels in *Set 2* (A and AV mode), there was only one vowel (/e/) for ENS which did not reduce the amount of confusion in AV mode (Table 2.11). For the L2 learners, it was expected that the high scores for /e/ and /ɜ:/ in V mode would help learners reduce the confusions in AV but this did not happen (Table 2.12a,b). Also a strong unidirectional confusion for/i:/ perceived as /ɪ/ was not reduced in AV mode. This amount of confusion suggested that L2 learners cannot hear nor see a difference between the English /i:/–/ɪ/ contrast. Similar difficulty has been found in previous studies for Spanish learners; García-Lecumberri and Cenoz (1997) found that Spanish learners of English confused /i:/ with /ɪ/ around 50% of the time but the confusion was unidirectional (/ɪ/–>/i:/only 8%).

Table 2.11 Confusions for vowels in Set 2 by ENS in A and AV mode (in noise). Participants' responses are presented in percentages; chance level 25%.

	A Response % (ENS)				AV Response % (ENS)			
Stim	**/ɪ/**	**/i:/**	**/e/**	**/ɜ:/**	**/ɪ/**	**/i:/**	**/e/**	**/ɜ:/**
/ɪ/	88.8	13.8	13	10	96.2	13.8	14.4	7.4
/i:/	8.8	79.4	0	0.6	3.8	85	0.6	0
/e/	1.2	4.4	82	9.4	0	1.2	83.1	3.8
/ɜ:/	1.2	2.5	5	80	0	0	1.9	88.8

Table 2.12a Confusions for vowels in Set 2 by L2 beginners in A and AV mode. Participants' responses are presented in percentages; chance level 25%.

	A Response % (L2 Beg.)				AV Response % (L2 Beg.)			
Stim	**/ɪ/**	**/i:/**	**/e/**	**/ɜ:/**	**/ɪ/**	**/ i :/**	**/e/**	**/ɜ:/**
/ɪ/	63	50	2.4	2.1	58.2	48.7	3.7	1.3
/l:/	9.3	43.4	6.9	9	10.6	44.9	6.6	8.2
/e/	26.1	5.6	83.8	23.2	30.4	5.6	83.8	23.7
/ɜ:/	1.6	1	6.9	65.7	0.8	0.8	5.9	66.8

Table 2.12b Confusions for vowels in Set 2 by L2 advanced group in A and AV mode. Participants' responses are presented in percentages chance level 25%.

	A Response % L2 (Adv.)				AV Response % (L2 Adv.)			
Stim	**/ɪ/**	**/iː/**	**/e/**	**/ɜː/**	**/ɪ/**	**/ iː/**	**/e/**	**/ɜː/**
/ɪ/	76	40	4.1	2.4	76.7	39	2.7	1
/iː/	11.1	55.4	4	5.1	11.7	58	2.7	3.4
/e/	12.5	3	87.8	17	11.6	1.5	90.2	10.1
/ɜː/	0.4	1.6	4.1	75.5	0	1.5	4.4	85.5

In *Set 3* (Table 2.13), confusions for two tense vowels (/ɔː/, /uː/) were reduced in AV for ENS, in spite of their already high scores in V. Confusions for the L2 learners (Table 2.14a,b) remained across modes , with higher scores for the advanced group. There was a strong bidirectional confusion for /ʊ/–/uː/ and weaker for /ɔː/–/ɒ/ by both L2 groups. This clearly suggests that L2 learners have difficulty identifying the tense-lax distinction. The only vowel that decreased its confusions in AV was vowel /ʊ/ for the L2 advanced group only. Interestingly, this vowel was also problematic for ENS.

Table 2.13 Confusions for vowels in Set 3 by ENS in A and AV mode (in noise). Participants' responses are presented in percentages; chance level 25%.

	A Response % (ENS)				AV Response % (ENS)			
Stim	**/ɔː/**	**/ɒ/**	**/uː/**	**/ʊ/**	**/ɔː/**	**/ɒ/**	**/uː/**	**/ʊ/**
/ɔː/	92.5	4.4	0.6	10	97.5	8.8	0.6	10
/ɒ/	1.3	89.4	3.8	1.2	0.6	89.4	0	1.2
/uː/	0	0	78.1	36.2	0	0	93.1	36.2
/ʊ/	6.2	6.2	17.5	52.5	1.9	1.9	6.3	52.5

Table 2.14a Confusions for vowels in Set 3 by L2 Beginners in A and AV mode. Participants' responses are presented in percentages; chance level 25%.

	A Response % (L2 Beg.)				AV Response % (L2 Beg.)			
Stim	**/ɔː/**	**/ɒ/**	**/uː/**	**/ʊ/**	**/ɔː/**	**/ɒ/**	**/uː/**	**/ʊ/**
/ɔː/	48.1	29.3	8.8	9	43.7	25.5	7.2	11.7
/ɒ/	23.9	64.6	8	8.5	22.4	67	13.3	5.3
/uː	17.3	3.5	47.9	43.1	22.7	5.6	44.7	40.4
/ʊ/	10.6	2.7	35.4	39.4	11.2	1.9	34.8	42.6

Table 2.14b Confusions for vowels in Set 3 by L2advanced group in A and AV mode. Participants' responses are presented in percentages; chance level 25%.

	A Response % (L2 Adv.)				AV Response % (L2 Adv.)			
Stim	**/ɔː/**	**/ɒ/**	**/uː/**	**/ʊ/**	**/ɔː/**	**/ɒ/**	**/uː/**	**/ʊ/**
/ɔː/	65.5	22.3	2.4	3.3	63.5	30	2.4	2.7
/ɒ/	17.9	73.3	3.6	8.8	19.3	67	1.6	2
/uː	13.6	2	49	37.2	15	2.7	51.4	29.8
/ʊ/	3	2.4	45	50.7	2.2	0.3	44.6	65.5

The aim of presenting the confusions made by ENS and L2 learners in the Vowel test in A and AV mode was a) to explore what they perceived when they chose a different response than the correct vowel in A mode and b) to see whether confusions were reduced in AV mode aided by visual information. The results showed that ENS' confusions were reduced in AV mode for most vowels whereas no benefit of visual information was observed for the L2 learners who showed similar patterns of confusions regardless of their difference in amount of language experience (i.e. level of proficiency).

In general terms, the confusion patterns found in this study are in line with those reported in A and AV mode in previous studies with L2 learners with Spanish as L1 (Flege, Bohn, & Jang, 1997; García-Lecumberri & Cenoz, 1997; Ortega-llebaria, Faulkner & Hazan, 2001). Better identification of vowels in V mode by the L2 advanced group partly agreed with the idea that more experience with the L2 contributes to better sensitivity to visual cues, advanced by Wang et al., (2008). However, the fact that similar patterns of confusions in A and AV mode were found for both the L2 groups suggested that L2 participants failed to integrate the visual information in AV to reduce their vowel confusions. Thus more experienced learners were no better at integrating visual cues for speech perception.

2.3.5 Individual differences and vowel perception in L2 learners

The L2 learners' results in the Vowel test showed a wide range of individual differences in the perception of English vowels (Fig.2.2). To explore this individual variability, three

additional tests were used to find possible sources for these differences in individual scores: a frequency discrimination test, a McGurk test for visual bias and a BKB-sentence test for discrimination of key words. The L2 Advanced group was only available for a one-hour session, so a decision was made to only give them the main test (Vowel test) and the frequency discrimination test. Finally, L2 participants' level of proficiency was used to explore any relation with their vowel identification results.

a. Frequency discrimination test

A frequency discrimination test (FDT) was used as an auditory measure of the participants' capacity to discriminate frequency differences in non-speech. Data are presented in terms of the number of frequency steps between the standard and jnd: the lower the value the better the frequency discrimination. The overall mean was very similar for the two non-native groups (Fig.2.4): L2 Beginners (M: 40, SD: 12.5) and L2 Advanced (M: 42, SD: 13). A mixed-model analysis was run in R (lme function) with frequency test scores and group as fixed factors and participants as random factor. There was no significant difference between groups $F(1,28)=0.5416$, $p>.05$.

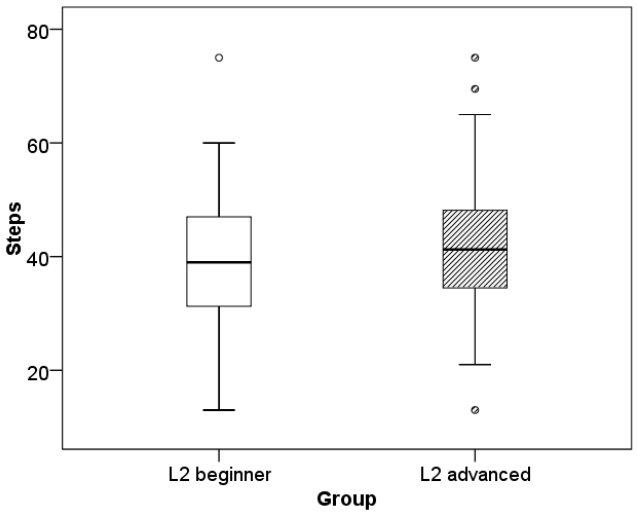


Figure 2.4 Boxplots with overall scores of the steps in the Frequency discrimination test for L2 beginner and L2 advanced group. One outlier from each group was removed from the analysis as their scores were 3 St. Dev. from the group mean. Steps were obtained from the average 7 reversals or after 50 trials.

Table 2.15 Results of the Pearson correlations run separately for L2 Beginners and L2 Advanced group on the Frequency discrimination test scores and the A, AV & V scores of the Vowel Test (8vowels).

	Vowel test L2 Beginners			**Vowel test L2 Advanced**		
Frequency discrimination test	**A_all**	**AV_all**	**V_all**	**A_all**	**AV_all**	**V_all**
Pearson correlation	-.252	-.265	.061	.080	-.117	.242
Sig. (2-tailed)	.099	.082	.693	.643	.496	.155
N	44	44	44	36	36	36
*Correlation significant at 0.05 level						

Separate Pearson product-moment correlation coefficients were computed per group to assess the relationship between overall percent correct scores for the A, AV and AV mode (Vowel test) and the frequency jnd. There was no significant correlation found between vowel identification scores in any mode and the participants' frequency discrimination abilities (Table 2.15). These results suggest the participants' identification of English vowels was not related to their capacity to discriminate frequency differences in a non-speech continuum.

These results differ from previous findings which have suggested that auditory frequency discrimination capacity was related to vowel perception both in the native language and in L2 contrasts (Lengeris & Hazan, 2010) and to a non-native consonant voicing contrast (Kim & Hazan, 2010).

b. McGurk test

A McGurk test was used to get a measure of the participants' visual bias when audio and video inputs are discrepant. This test was given to the L2 beginners. A measure of the 'Visual Effect' (VE) was calculated to show the degree of visual influence in consonant intelligibility as in Chen and Hazan (2009) and Sekiyama, Burnham et al. (2003). This is calculated as follows: a) the positive effect of visual cues (AV+) is estimated as the difference between scores for AVclear (AV) and Aclear (A). Then, b) the negative effect of visual cues (AV-) is obtained from the difference between A and AVincongruent (AVinc).

Finally, the VE is obtained from the sum of (AV+) and (AV-). The results showed that VE (Fig. 2.5) varies across speakers from -0.03 to 0.75 proportion correct, with a mean of 0.34 (SD: 0.24). In Chen & Hazan (2009) the VE in clear ranged from 0 to 55% and from 60 to 95% in noise for Chinese adult learners. In the same study, English adults showed VE in clear ranged from 5 to 35% and from 65 to 95% in noise.

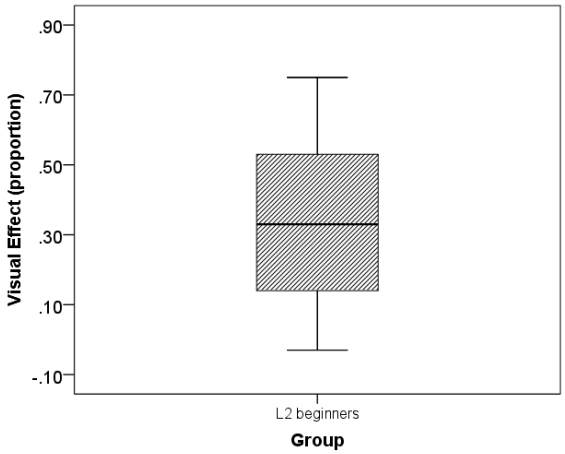


Figure 2.5 Boxplot for the proportion of visual effect (VE) obtained from the McGurk test given to the L2 beginners group.

The extent to which visual speech influences perception was also measured in the amount of McGurk effect (McGurk & MacDonald, 1976) found when an incongruent "visual_ga + audio_ba" token is presented: that is to say, the percentage of time participants chose "da" as a response label for this incongruent combination. Fig. 2.6 shows the boxplot for the McGurk effect which varied from 0 to 100% mean (M: 51.6 SD: 31.3). The visual bias in noise when presenting "visual_ba+audio_ga" was found to be higher than the visual bias in clear (though for a different token). The favoured response was mainly "ba", ranging from 33.3 to 100% (M: 81, SD: 22.2). A higher score relates to a greater visual bias when assigning a label to the perceived token.

These results suggest that there was a visual bias in participants when perceiving incongruent input in AV mode in clear and it was increased when perceiving incongruent stimulus in noise. There was also a lot of individual variability in the amount of visual bias

for each participant. Based on these findings, we may hypothesize that visual cues could be exploited in the perception of L2 novel contrasts.

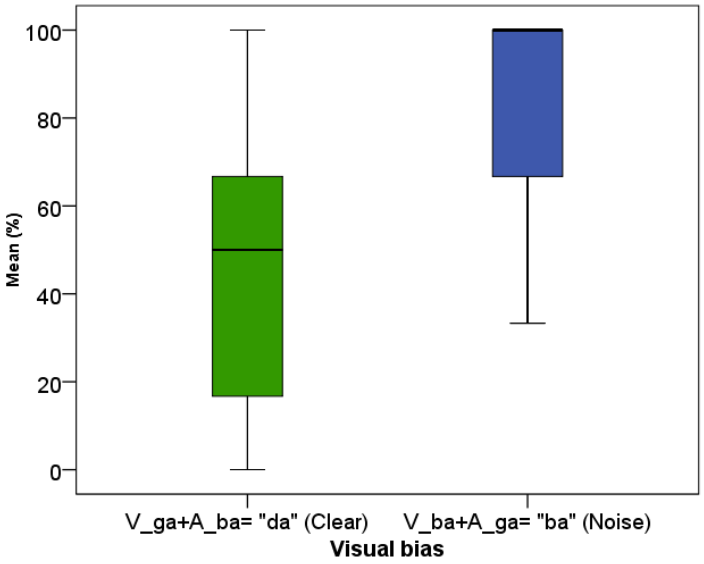


Figure 2.6 Boxplot for the visual bias in the McGurk test given to L2 learners. Visual bias in clear (the McGurk effect) and visual bias in noise (SNR -12 dB).

Table 2.16 Pearson correlations between Visual effect (%) and the AV-A relative difference (%) in the Vowel Test (8 vowels) for L2 Beginners group.

Visual Effect % (VE)	**Vowel test AV-A relative Difference (%)**
Pearson correlation Sig. (2-tailed) N	.112 .481 42

To explore whether individuals showed any effect of visual cues in the Vowel test, the difference between the scores in A and AV modes relative to A was calculated ((AV-A)/A) showing the relative change in identification when visual cues were available. A range of AV-A difference scores was found: positive from 2.6 to 55% (18 participants) and negative from 0 to -29% (29 participants). Pearson correlations were run to explore a possible relation between the Visual Effect (VE) percentage and the AV-A relative difference in the

Vowel Test. The results showed no significant correlation between the VE and the AV-A difference (Table 2.16). These results suggest that there was no direct relation between the L2 learners' visual bias found in the VE (McGurk test) and visual advantage in AV-A difference in the vowel test.

c. BKB-sentence test

A BKB-sentence test was used to measure whether participants' perception of key words in short sentence-material in English was improved when visual information was present. This test was given to the L2 beginner group. Two different lists of sentences were presented, one in A (BKB_A) and the other in AV mode (BKB_AV). Each BKB-sentence list contained 50 key words. Correct repetition of key-words were computed to obtain overall correct percentages. Means for the BKB_AV (78%, SD=14.3) and BKB_A (76%, SD=16.5) were very similar. However, the boxplots show a wider range of variability of scores in A mode (Fig. 2.78).

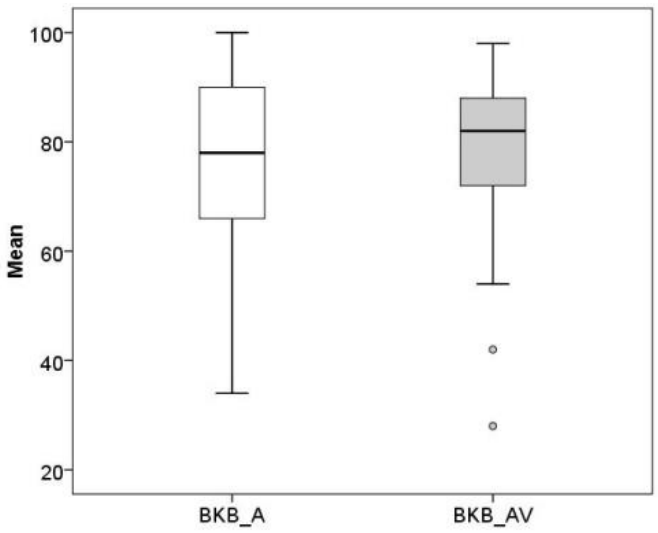


Fig. 2.7 Boxplots with overall scores in the BKB test in A and AV mode for the L2 beginners group. The mean correct scores were estimated by the number of correct key words per sentence participants were able to repeat.

To find the visual advantage from AV over A, the percentage of relative change in scores across these two conditions was calculated (AV-A/A)*100). Mean relative change between A and AV was 4.7%; 26 participants showed positive values ranging from 2.3% to 70.6% while 20 participants showed negative values ranging from 0% to -29%; that is, their scores were higher in the A than the AV condition. Only one participant obtained the same scores in A and AV mode. The scatter plot (Fig. 2.8) shows the strong correlation between A and AV scores which suggests that the perception of key words was strongly determined by their auditory modality.

To explore the relation between word identification in the BKB-sentence test and vowel identification (Vowel test), Pearson correlations were run for BKB-A and BKB-AV mode and the scores in the Vowel test in A, AV & V mode. No strong correlation was found between the variables, a significant but weak correlation was found between the BKB_A and the AV scores in the Vowel test (Table 2.17). This weak correlation could only account for a very small amount of the variance in the data. Thus the capacity to perceive key words at sentence level was not related with the capacity to identify vowels presented in words.

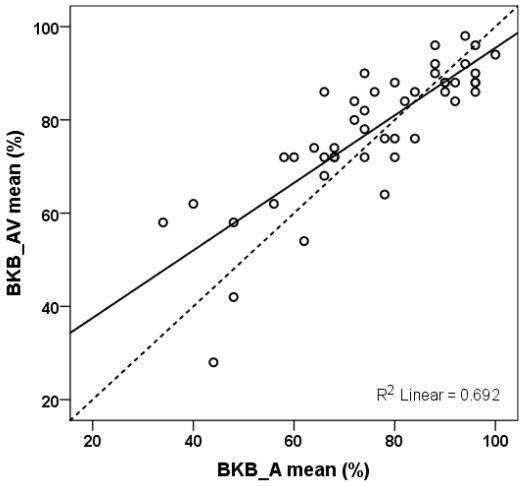


Figure 2.8 Scatter plot for the BKB-A mean and BKB-AV mean scores. The continuous line shows the correlation between BKB-A and BKB-AV scores. The dotted line is the line at total.

Table 2.17 Values for Pearson correlations between the BKB-sentence test (A, AV) and the overall scores in the Vowel test in three modes (A, AV, V) for the L2 Beginners group. Correlation between the AV advantage in the BKB test and in the vowel test was included.

	Vowel test (L2 Beginners.)						
BKB-A	**A_all**	**AV_all**	**V_all**	**BKB-AV**	**A_all**	**AV_all**	**V_all**
Pearson correlation	.228	.341*	.052		.142	.260	.062
Sig. (2-tailed)	.123	.019	.731		.342	.078	.677
N	47	47	47		47	47	47
*Correlation significant at 0.05 level							
BKB AV advantage vs. Vowel test AV advantage							
Pearson correlation				-0.031			
Sig. (2-tailed)				.837			
N				47			

The aim of this test was to explore the relation between sentence-level word identification and vowel identification at word-level. The results showed evidence of different degrees of visual advantage for word identification at sentence-level for more than half of the L2 learners but this capacity did not show to be related to the Identification of English vowels in the vowel test in any modality.

d. Level of proficiency and vowel identification scores

The L2 learners in this study varied in the length of time they had studied English at university level. The L2 Beginners had just started their English teaching degree (Major in English language) and they had had one month of lectures, involving 20 hours of English a week, when they took part in this study. The second group (L2 advanced) had more advanced proficiency learners. They were students in their fourth year at university, also studying an English Teaching degree with an average of 20 hours of English lessons per week. A measure of their level of proficiency was established by using the results from the Preliminary English test (PET) given to the beginners group as a diagnostic test in their third week at university. The final mark in the English subject from the recently completed term was used for the advanced group as a measure of English language proficiency.

The scores from their English test/course were used to run correlations with their overall identification scores in the Vowel test (8 vowel comparison) across modes (A, AV & V).

Table 2.18 Pearson correlation values for the level of proficiency (English test) and the Vowel identification overall scores (8 vowels) for the L2 Beginners and L2 Advanced group. (* correlation significant at .05 level)

	Vowel test	
English test	**L2 Beginners**	**L2 Advanced**
Pearson Correlation	.311*	-.076
Sig. (2-tailed)	.035	.657
N	46	37

Pearson correlations were run for English test scores and Vowel test overall means (8vowels) for the two learners groups. A significant but weak correlation was found for L2 beginners but not for the advanced group (Table 2.18). The scatter plots (Fig. 2.9) illustrate the individual variability in each group. Participants who had higher results in their English test, therefore more proficient in each group, did not always obtain the highest scores in the Vowel identification test. In the beginners group, some learners with very little knowledge of English were able to perform as well as those who were more proficient. Thus, level of proficiency alone cannot explain the individual variability in English vowel perception in this study. It is important to bear in mind that though the axis for English proficiency level goes from 1 to 100 for each group, this refers to a different reality per group. That is to say, a 50 in the beginners group does not compare with a 50 in the advanced group. Thus, the scatter plots should not be compared along the X-axis.

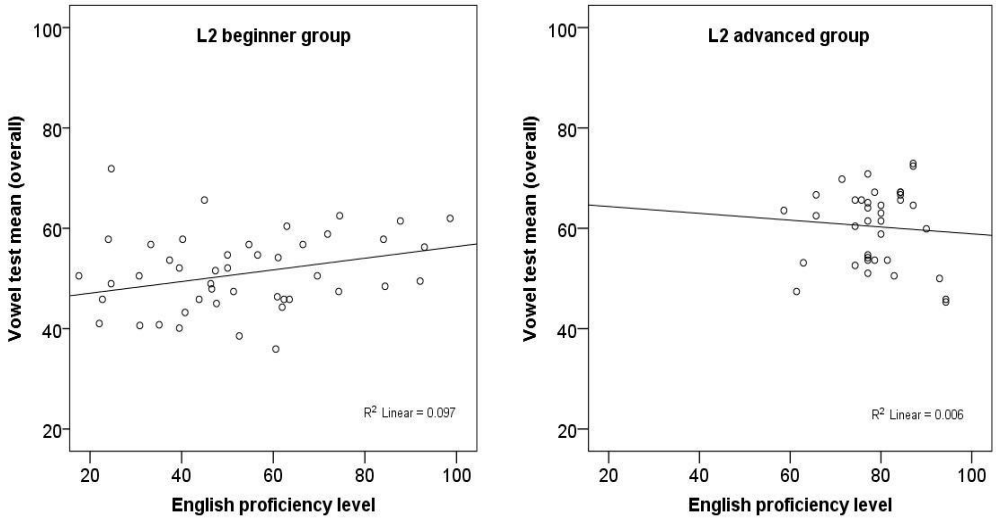


Fig. 2.9 Scatterplots with the scores from the English test versus the Vowel test scores for the L2 beginners and L2 advanced group. There was missing data for one participant from the beginners group.

2.4 Discussion

This study investigated whether L2 learners with Spanish as L1 were sensitive to visual cues for the perception of English vowels. To do this, establishing the availability of visual information for English vowel identification to English native speakers was needed. Once the availability for ENS was explored, results from the use of visual cues in vowel identification by the L2 groups were compared between the non-native groups and against ENS'. Additional tests were used to explore individual differences in L2 learners and their possible relation to English vowel identification in the three modalities under study (A, AV, V).

The results in the Vowel test given to ENS showed that there were different degrees of visual information for English vowels that improved their vowel perception in AV mode relative to A mode within the fixed sets in which they were presented. English native speakers were able to integrate the visual information available for English vowels in AV with better scores than in A mode when vowels were presented in background noise. Generally speaking, these results showed that the addition of visual information contributed to better vowel perception; this benefit was present when data were examined at the level of individual vowels, with significantly higher AV than A scores obtained for four English vowels out of 11. The rest of the vowels showed higher scores in AV even when their scores were already near ceiling effect in A mode and no further improvement was expected.

In general, L2 learners' vowel identification varied as a factor of their proficiency level, with better performance for the advanced group. Overall, they showed poorer vowel perception than ENS in all modes (A, AV, V). Even though learners' vowel identification in V mode was above chance for some vowels, there was no significant difference between A and AV mode, suggesting a strong reliance on the auditory modality. Potential use of visual cues was found as L2 learners were able to attend to visual information for vowel identification in the Video-alone (V) task, as identification scores were generally above chance level. However, they failed to integrate visual cues in the AV mode to improve vowel perception. It could be hypothesized that visual gestures for vowels in the type of

tokens used in this study ("hVd", "bVt" words) are more difficult to perceive than visual gestures for consonants in initial position of a syllable, for instance. Most of the studies conducted on the contribution of visual cues in L2 have tested consonants in initial position in one-syllable words as in the McGurk effect (Bovo et al.2009; Cebrian & Carlet, 2012; Fuster-Duran, 1996; Hazan et al., 2005; Massaro et al., 1993; Massaro, Cohen & Smeele, 1995). Therefore, the influence of the material type may have played a role in making visual information for vowels more difficult to access.

Previous studies in L2 speech perception have concluded that the benefit of visual cues is lower when the contrasts to be perceived are less visually salient (Hardison, 1999; Hazan et al., 2005; Sennema et al., 2003; Alm, Behne, Wang & Eg, 2009) or when the contrastive categories have shared gestures (Ortega-Llebaria et al., 2001). Moreover, the weighting of visual cues may vary among perceivers and its availability seems to depend on speakers' speech readability as well. For instance, there are speakers whose visual gestures are easier to speechread and perceivers who are better at interpreting visual cues for speech perception (Hazan, Kim & Chen, 2010; Lesner & Kricos, 1981; Owens & Blazek, 1985).

It has also been suggested that lack of experience with the use of visual cues in the L1 may make non-native perceivers less sensitive to the use of visual information for speech perception. In this scenario, non-native perceivers will need to learn how to weigh visual information in an L2 to establish visual representations of visemes in the same way as learners acquire L2 auditory categories (Wang et al. 2008; Hazan et al., 2005; Hazan, Sennema, Faulkner et al., 2006). An issue that needs consideration is whether in a small vowel inventory, as the Spanish one, visual cues are likely to be redundant. Since Spanish has five well acoustically defined vowels, it may be the case that though vowels are visually distinct, native speakers can do without the visual information. However, this argument would somehow conflict with the well-established theory of the bimodality of speech which claims that speech is auditory and visually perceived with these components being automatically integrated, whether an early or later integration approach is advocated (Braida, 1991; Massaro & Cohen, 1983; 1993; Rosenblum, 2005).

The lack of use of visual cues by L2 learners may lay on the nature of the visual contrast for vowels. Vowels are typically visually distinguished by their degree of rounding and spreading of the lips (Ladefoged, 2006). Nothing is said about the visual salience of the length difference between tense and lax vowels. However, the fact that ENS were, in general terms, able to use visual information to distinguish vowels may also suggest that a length and lip rounding/spreading may both contribute to the visual identification of vowels. Given that L2 learners are not sensitive to visual cues as a way to mark a difference in vowel quality in their L1, cues for liprounding/spreading or lengthening become irrelevant information in the speech perception process.

Research on L2 speech perception has found that the amount of use of visual information in speech perception may vary with linguistic experience in L2 learners (Wang et al., 2008). Suggesting that more experienced learners make more use of visual information for speech perception. In the current study, the two groups of learners differed significantly in their level of proficiency (i.e. amount of experience in learning English). Though the L2 advanced group showed significantly better identification of English vowels, they did not show a significant advantage when visual cues were available in AV mode to reduce the amount of confusion in A mode. Their results in A and AV suggested that they relied mainly on the audio input for vowel perception. Therefore, more experience with the language did not make a difference in the capacity to integrate visual cues to the auditory input in AV. Wang and colleagues investigated the use of visual cues for English fricatives by Mandarin Chinese living in Canada. The visual advantage they found in participants with longer residence in Canada (i.e. more experienced) may have been driven by the type of material used. Namely, three contrastive consonant pairs followed by /a, i, u/ vowels ("/fi, vi, ði, Θi, si, zi/"). The gestures for fricative consonants in initial position in the syllable are more salient than those for vowels. Another issue to consider is that Wang et al. (2008) tested participants living in the country where the language is spoken, unlike participants in our study whose contact with English is mainly in the classroom context. It may be argued that interacting face to face with native speakers in a bilingual environment benefits the learning of visual information for L2 speech perception.

In the current study, confusions in the identification of English vowels by L2 learners were expected to decrease with the aid of visual information; however, the pattern of confusions across L2 groups was very similar without showing any visual benefit in AV. That is, confusions remained across A and AV mode. Given that scores in A and AV did not differ, these confusions may be the result of an auditory assimilation strategy of an English tense-lax pair to a one-vowel category in Spanish, as observed in previous studies (Fox, Flege & Munro, 1995; Iverson & Evans, 2007). This makes sense for the less experienced learners who had a very low proficiency level at the time of testing. In the case of the more advanced learners, these findings illustrate that some of the confusion problems may persist after many years of linguistic experience. That said, it is uncertain to what extent the quality and quantity of their linguistic experience, namely English as a foreign language mostly through non-native speakers, exerts an influence on the learners' vowel identification capacity. It was not possible to control for which English accent variety the advanced learners received more input in; all of them reported having had instruction in British and American English at university. It could be the case that some advanced participants may have been more familiarised with Standard American English which has somewhat different vowel configuration. As for the beginner learners, they received mainly British English input at university but they had surely received American and British English at school through their teachers and audio material.

The results in the vowel identification test revealed a great amount of individual differences that was explored by testing participants' auditory and visual bias and capacity to perceive key words in sentences. Previous studies have suggested (Hazan & Kim, 2010; Lengeris & Hazan, 2010), a relation between learners' English vowel identification and frequency discrimination capacity. The results in the current study showed that though learners exhibited an overreliance on the auditory channel for vowel perception, their auditory capacity to perceive small spectral changes (Frequency discrimination test) was not related to their vowel identification performance. This difference with Lengeris & Hazan's study may have an origin in the type of material used. They used two two-vowel 40-step synthetic continua (/iː/−/ɪ/ and /æ/−/ʌ/ in quiet and in noise)while in our study, the number of vowels tested was larger (11 vowels),the number of speakers was also different (four

speakers in pre & post test) and tokens were produced with natural speech. This may suggest that having more variability in the current study may have caused more dispersion of the individual identifications scores and thus the correlation between the frequency discrimination test and the vowel identification results is lost.

Regarding participants' visual weighting, the results showed that L2 beginner learners' consonant identification was affected by information in the visual modality in in the McGurk test, in line with previous findings (Bobo, et al., 2009; Cebrian & Carlet, 2012; Fuster-Duran, 1996; Hazan et al., 2005; Massaro et al., 1993). However, this "visual effect" (VE) was not correlated with the participants' relative AV advantage in the Vowel test (difference between A and AV mode). One possible explanation may be the nature of the stimulus used in the McGurk test; namely, monosyllabic words like "ba", "da", "ga". However, identifying consonants' place and manner of articulation as in the McGurk test (bilabial, alveolar or velar plosives) seems to pose less difficulty for L2 learners than perceiving the visual cues to English vowels. Additional evidence that L2 learners could attend to visual information for vowel perception was found in the V mode (Vowel test). But their overreliance on audio input seems to point to a lack of integration of visual cues in AV mode. It may be hypothesized that this lack of integration may have a cognitive explanation related to exerting an overload of the working memory resources while having to attend to auditory and visual channels of information for speech perception in a non-native language. This issue will be explored further in the final discussion of this thesis.

In conclusion, evidence of the availability of visual cues for English vowels and the benefit of its integration with the audio input to improve ENS' perception (in noise) was found. This study also showed the lack of integration of visual cues to auditory input in the perception of English vowels by L2 learners, irrespective of their level of proficiency. Thus, if experience with the language did not make a difference between L2 groups with regards to the integration of visual information in AV, it remains to be seen whether L2 learners can be trained to attend to visual cues in a similar way as L2 vowel training studies have shown to improve L2 vowel perception (Iverson & Evans, 2007; 2009; Wang & Munro, 2004). This study also revealed a wide range of individual variability;

unfortunately, no relation between auditory acuity and visual bias with speech perception could be established.

In summary, this study is original in establishing the potential contribution of visual cues for English vowel perception in L2 learners of different proficiency levels, using a large set of vowels. It is also novel in suggesting the lack of integration of visual cues for which L2 learners showed some sensibility in V mode (Vowel test) as the cause for non-native speakers' inability to perceive a unitary audio-visual percept for English vowel identification.

Chapter 3

Impact of vowel training modality on English vowel perception

In the study presented in Chapter 2, it was found that L2 learners (Spanish_L1) were not able to integrate visual cues for vowel identification in audio-visual (AV) mode, though they were able to attend to visual information for English vowels in a video-alone (V) vowel identification task. These learners also showed different degrees of visual bias when tested on the perception of words presented with incongruent input (McGurk test). In the same study, different degrees of visual benefit in the perception of English vowels were found for a group of English native speakers (ENS). These finding suggested that there is visual information available for the perception of English vowels that L2 learners could use to aid their English vowel perception. The aim of the current study (study 2) was to find whether L2 learners can be trained to attend to visual information for English vowels to aid their vowel perception accuracy.

To be able to establish the benefit of visual information for English vowel training, three types of training modalities were devised: an auditory (A) training, an audio-visual (AV) training and a video-alone (V) training modality. 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, training would lead to differences across training groups in the perception of English vowels in A, AV and V mode post-training. For instance, vowel identification accuracy in A mode could be better for learners trained with audio mode, whereas identification in AV and V mode could be better for learners who received AV or V training.

3. 1 Aims

The aim of this study was to compare three vowel training modalities to assess which, if any, may be more beneficial in improving vowel perception in L2 learners. The research questions that guided this study were:

a. Is there a difference in the improvement in English vowel perception that is achieved across these training modalities?

b. Can L2 learners be trained to attend to visual cues for the perception of English vowels?

> i. Can L2 learners attend to visual cues for those English vowels that were found to be more visually salient to ENS (Chapter 2)?

c. To what extent does perceptual training lead to improvement in the perception English vowels at word-level?

d. To what extent does perceptual training lead to improvement in the perception of more 'naturalistic' speech in sentence material (TF sentence test)?

e. To what extent are individual differences in vowel identification accuracy related to visual bias and auditory frequency discrimination?

f. How does the participants' level of proficiency relate to their perception of English vowels and use of visual cues before and after training?

3.2 Methods

3.2.1 Participants

The main group of participants was an L2 beginners group who took the vowel training programme and were tested before and after training. A group of 47 L2 beginner learners were recruited and tested in Chile at Universidad de Concepcion (15 males, 32 females). Participants were first-year university students from a Teacher of English Training programme who had just completed their first semester at university (20 hours of English a

week) and their age ranged from 18 to 24 years old (M: 19.6, SD: 1). 34 of these participants had taken part in study 1 (Chapter 2) six months earlier. Thirteen new participants were recruited from a first-year Translation programme at the same university. This group matched the other 37 students in level of English and similar number of hours of English instruction.

Thirteen L2 learners with beginner level of English and characteristics which matched the main L2 beginners group (same university programme) were used as control for the True-or-False sentence test (described in section 3.2.2).

Thirty-seven L2 advanced learners (3 males and 34 females) took the True-or-false sentence test. This was the same advanced learners group that participated in a previous study reported in Chapter 2 (2.2.1). The results from the Vowel test presented in Chapter 2 were also used in this chapter for group comparison purpose.

Twenty English native speakers (6 males, 14 females) with Southern British English accent took the True-or-false sentence test. These were the same ENS who participated in a previous study reported in Chapter 2. The results from the Vowel test presented in Chapter 2 were also used here for group comparisons

Participants' level of English was determined by converting their final marks in their English class (first semester recently completed) into percentages. This final mark comprised written tests which assessed listening, reading, writing and vocabulary knowledge, and also marks from oral presentations and oral interviews throughout the semester. All participants were right-handed and self-reported not having any hearing impairment.

3.2.2 Test battery

a. Pre and post test battery

The stimuli used in this study were grouped into pre and post test battery and training material. Table 1, at the end of the procedure part presents a summary of all tests and

participant groups. Most of the pre and post test materials were used in the study reported in Chapter 2; namely the Vowel test (for description, see section 2.2.2a), the frequency discrimination test (see section 2.2.2b), the McGurk test (see section 2.2.2c) and the BKB-sentence test (see section 2.2.2d).

To test whether learners' perception of English vowels in isolated words was related to perception of vowels in words embedded in sentence-material, a 'True-or-false' sentence test was devised.

Generalization tests in training studies typically use word-level material (e.g. hVd words or real words) to measure if the improvement achieved in speech perception after training can transfer to new material (new tokens) and to talkers that were not encountered during training (Hazan et al., 2005; Iverson et al., 2005; Iverson & Evans, 2009; Lively et al., 1993; Logan et al., 1991; Nishi-Kewly-Port, 2007; Wang & Munro, 2004). One of L2 learners' typical challenges is to perceive language in naturalistic contexts that go beyond the mere identification of phonemes in isolated words. For this reason, it was felt that a more realistic speech task was needed to evaluate the accuracy of vowel perception within a sentence context.

A list of 66 sentences containing minimal-pair words for nine vowels (Appendix A) was recorded by three Standard Southern British English (SSBE) native speakers (1 male, 2 females; the same speakers used for the recording of the vowel test materials). Two versions of the same sentence were made using one of the minimal-pair words chosen so that one of the sentences would make sense but the other would not be acceptable from the point of view of its meaning. For instance, the sheep-ship minimal pair was presented in sentences like "the ***sheep*** is eating the grass" and "the ***ship*** is eating the grass" where 'sheep' is meaningful and 'ship' meaningless, and the sentences "The ***ship*** is in the sea" and "the ***sheep*** is in the sea". It was necessary for participants to accurately perceive the vowel in the keyword in order to make the right decision in terms of whether a sentence was semantically-anomalous or not. The test included 146 sentences, with at least two different speakers' version per sentence. The condition for a set of minimal pairs to be used was that they could fit into a sentence without changing the grammatical category of the

word i.e. if the “true” sentence contained a key word that was a noun, the “false” sentence should have its minimal pair used in the same category and in the same slot in the sentence. Only two out of the 11vowels in the study were not included due to the impossibility of finding a minimal pair that could fall in the same grammatical category. The vowels excluded were /u:/ and /ʊ/ (e.g. fool, full).

To test the knowledge of the minimal pairs used in the TF-sentence test, a Vocabulary test which consisted of a list with all the words used in this test was given on paper to L2 beginners and the L2 control group, after completing the TF test in the post test session. Participants were asked to give an equivalent of the words in Spanish or English to show they knew the meaning of the words in the list. Scores were transformed into percentages (1-100).

b.Vowel trainer

The word list was taken from the Vowel Trainer developed by Iverson & Evans (2009). The stimuli consisted of real words containing 14 British English vowel sounds grouped into 4 clusters ([/e/, /ɑ:/,/æ/, /ʌ/]; [/i:/, /ɪ/, /aɪ/, /eɪ/]; [/ɒ/, /əʊ/, /ɔ:/]; [/u:/, /aʊ/, /ɜ:/]) based on findings of confusions made by German and Spanish speakers found in previous research (Iverson & Evans 2007). The first three clusters contain vowels that are mutually confusable and the remaining vowels form the last cluster. The total number of words used was 140 (10 sets of minimal pairs per word/vowel). Video recordings were made by five SSBE native speakers (2 males, 3 females). The same video filming procedure was used as for the vowel test material described in Chapter 2 (part 2.2.2a). 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 version of each token, either the video or audio were stripped out from the filmed versions of each token using the software Virtualdub.

3.2.3 Procedure

Participants were assigned to one of the three different vowel training groups: audio training (AT), audio-visual training (AVT) or video-only training (VT). Results from the

pre test audio (A) mode in the Vowel test were used to assign participants to one of the three training groups with the aim of balancing groups in terms of their pre test means and standard deviation (Fig. 3.1). 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, the AVT group had 14 and VT group had 16 participants who completed all the pre and post tests and training sessions. It would have been desirable to have a control group for this training study. Unfortunately, there are only 50 students per level in the teacher training programme at Universidad de Concepción. All the learners available with the same level of English (beginner) were needed to form the three training groups, so there were no remaining possible participants for a control group.

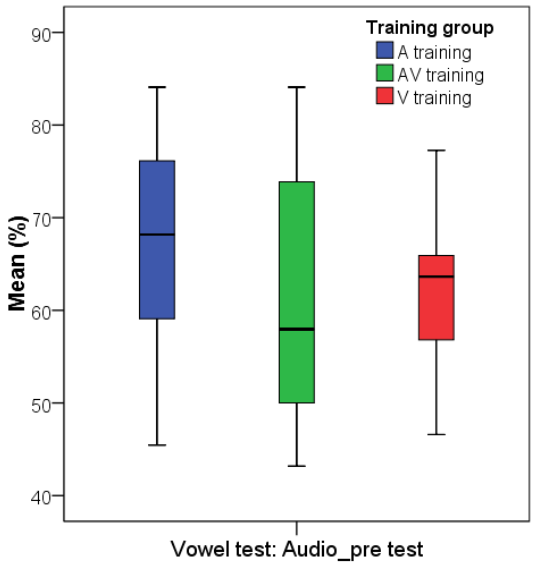


Figure 3.1 Boxplots for vowel identification accuracy (mean %) in the Audio mode of the Vowel Test at Pre test. These results were used to form the three vowel training groups: A training (M: 67, SD: 12), AV training (M: 61.3, SD: 13) and V training (M: 63, SD: 8).

The Vowel test, frequency discrimination test, McGurk test and BKB-sentence test were presented on desktop computers in a computer Laboratory in the Foreign Languages Department at Universidad de Concepcion to the L2 beginners. 15" monitors were placed 40 centimeters away from participants. Headphones were used for all tests and participants

could adjust the volume at a comfortable level. The BKB-sentence test was run in private individual sessions as the researchers needed to write down the key words repeated from each sentence. The rest of the tests were run in small groups in the computer lab. The researcher was always present while the participants were doing their tests.

The True-or-False test (TF test) was presented in audio-visual mode (Fig.3.2) on a desktop computer using headphones. It was first piloted with 5 ENS (1 male, 4 females) to test for instructions and difficulties which may arise in the procedure. Results showed the test was easy for ENS (mean 96%). It was then given to L2 beginners (47) and L2 advanced (37) learners and ENS (20). Participants were told they would watch videos of 3 native speakers of English saying a sentence that could be "likely" (True) or "unlikely" (False) and they would have to press a key on the keyboard to indicate if the sentence was likely (TRUE) or not (FALSE). They were told to answer as quickly as possible. Different colour paper labels (T, F) were stuck onto two keys on the keyboard (Z, M). Half of the participants had the "True" key on the right of the keyboard (key M) and the others on the left-hand side (key Z) to control for bias due to righthandedness. A practice phase presented 20 sentences which were not used in the test itself, to allow familiarisation with the material and test procedure. To check whether participants were following instructions, they were asked what their criterion to respond true or false was and they were also allowed to ask questions. In general, they were able to spot that minimal pairs which were being used, though this information was in the instructions. No feedback was provided on the correctness of their answers. A total of 144 sentences were presented (Appendix B), including 16 repetitions per vowel (9 vowels; 8 True sentences, 8 False sentences) were used. The test took around 25-30 minutes to complete.

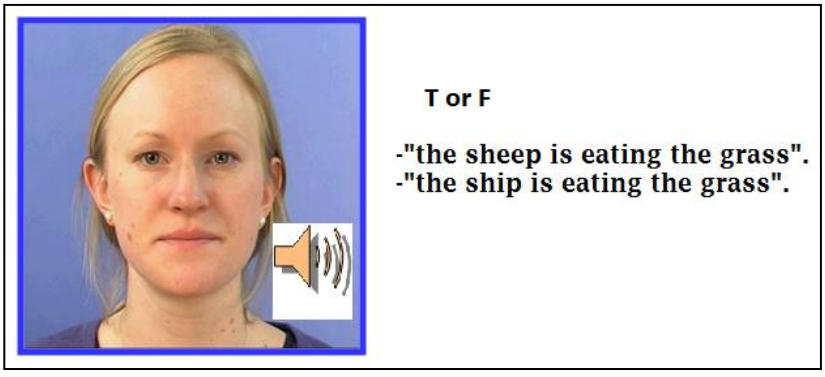


Figure 3.2 A snapshot and two sentences used in the True-or-False test.
Participants were presented the sentences in audio-visual mode with no access to their written form.

Vowel Trainer

The vowel trainer programme was identical to the one used in Iverson & Evans (2009). The same software and token list were used with the only change that the tokens were filmed to be able to implement the AV and the V training programmes. This training programme consisted of five sessions of high-variability phonetic training (HVPT, Logan et al., 1991) given to participants on a desktop computer and headphones. At each session, 225 training tokens were presented, and a different talker was used per session, alternating from female to male. Training sessions each lasted between 45 to 60 minutes. At each session, the same 225 tokens were presented in audio, audio-visual or video-only (without sound) mode depending on the training group participants were allocated to (AT, AVT, VT). The training software uses an adaptive procedure. In the first phase of the session, 70 fixed tokens are presented (five repetitions of 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 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 computer room availability, so it took 3 to 4 weeks to complete the 5 sessions. Participants heard or watched a speaker saying a word and had to click on the correct answer choosing from the alternatives (3 or 4 minimal-pairs) that appeared on the screen some seconds after the stimulus was presented. All alternatives were accompanied by a "help-word" on the side (Fig. 3.3); 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.

Figure 3.3 Screenshots of the three different types of Vowel training programmes used. Each alternative (words in white font) was accompanied by a "help word" (blue fonts on the right) as a way to help participants identify the key vowel sound just played.

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.

Table 3.1 Summary of tests, presentation mode of the stimuli and participants per test. (*) indicates the group was tested in this study described in Chapter 3. (**) shows the group was tested in Chapter 2 and their results are being used in Chapter 3 for comparison with L2 beginners.

Materials	**Presentation Mode**	**Participants**
Vowel test – (30 minutes)	A, AV, V	L2 beginners *, L2 advanced**, ENS**
Frequency Discrimination test 3-5 min	A	L2 beginners*
McGurk test (12-15 min)	A, AV, V (Clear, noise, congruent/incongruent	L2 beginners*
BKB-sentence test (15 min)	A, AV	L2 beginners*
True-or-false sentence test 25-30 minutes	AV	L2 beginners*, L2 control*, L2 advanced*, ENS*
Vowel trainer 5 sessions of 45-60 minutes	A training, AV training V training	L2 beginners*

3.3 Results

3.3.1 Vowel test

The Vowel test was used as the main test in the pre and post test battery that assessed whether L2 learners vowel perception had been affected by the training they received. To explore significant differences in perception per mode and training modalities (AT, AVT, VT), the 11 vowels were tested in A, AV and V modes. As described in Chapter 2, the Vowel test comprised stimuli for 11 English vowels presented in clear in three sets: Set 1{/æ/, /ʌ/, /ɑ:/}, Set 2 {/e/, /ɜ:/, /ɪ/, /i:/} and Set 3 {/ɒ/, /ɔ:/, /ʊ/,/u:/}.

A general analysis for the Vowel Test results, pre and post training, will be presented first. Due to the presentation of the stimuli in sets, data were first analysed separately per set to explore whether the same effects and interactions were present in all the sets. Confusion patterns will also be included to explore if training contributed to reducing confusions. All the data were analysed with a logistic mixed model with time, group, mode and vowel as fixed effects and participant and stimulus as random effects.

The overall results (Table 3.2) showed no significant effect of Training group (Fig. 3.4); there was no overall difference in performance across groups (Appendix A). There was a significant effect of Time: vowel identification (Fig. 3.5) in the post test (*M*: 64, *SD*: 24.2) was better than in the pre test (*M*: 58, *SD*: 25.5), showing overall improvement in English vowel identification after training (Fig. 3.5). The effect of mode was also significant with overall similar results in A and AV mode, both scores higher than V (Table 3.3). There was also a time*mode interaction; though all modes improved significantly after training, the V mode showed smaller amount of improvement (Fig. 3.5). Percentage of improvement relative to pre test was estimated for A (M: 14.7%, SD:14), AV (M: 14.2%, SD:15) and V(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.

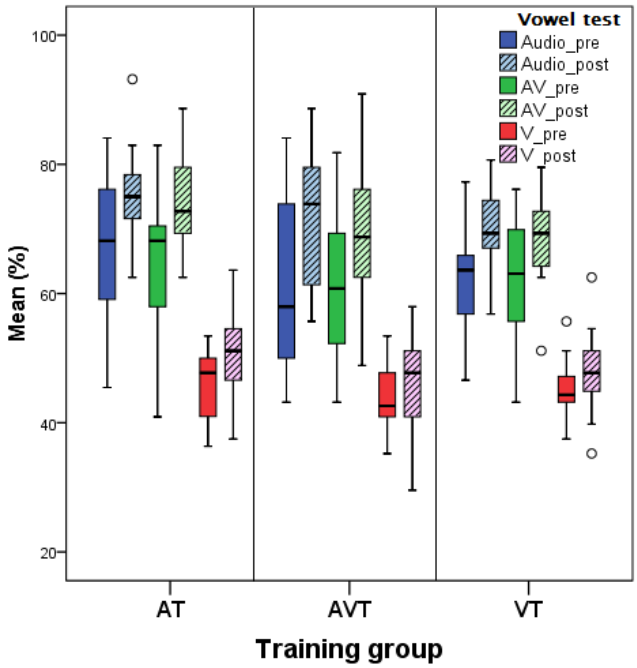


Figure 3.4 Boxplots for vowel identification in the Vowel Test (Pre & Post) per mode (A, AV, V) and Training Group: Audio Training (AT), Audio-visual training (AVT) and Video training (VT) groups. There was no significant difference between training groups.

Table 3.2 Vowel Test (pre & 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

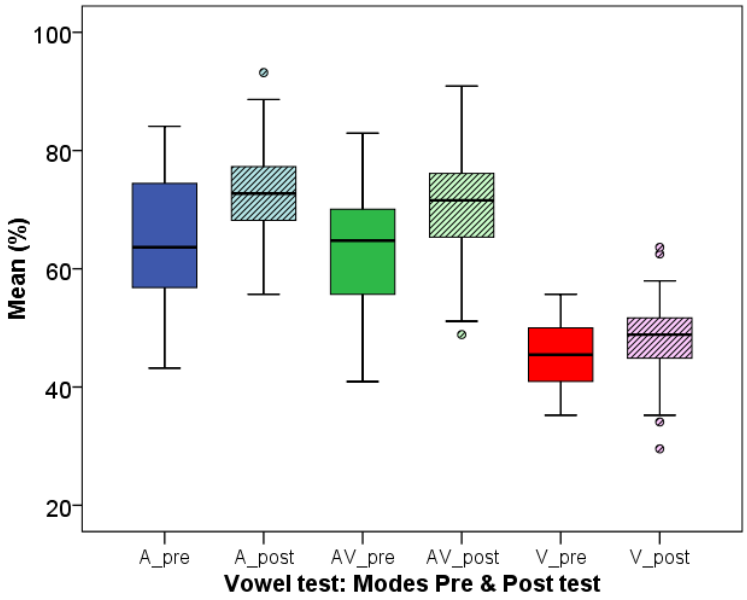


Figure 3.5 Vowel identification means in the Vowel test in A, AV and V modes (pre and post test), averaged over all three training groups. All the mean differences per mode (pre, post) were significant.

There was a significant effect of vowel, overall means varied from 45% for vowel /ʊ/ to 75% for /ɜ:/. There was also a vowel*mode interaction which will be described per set as it was later found to be present in all the three individual set analysis.

Table 3.3 Vowel test (pre & 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 & 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 percentages of change for the Vowel test were also estimated [((post test-pre test)/pre test)*100] for each of the conditions included (A, AV, V). The amount of improvement in the perception of English vowels 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

3.4). It seems odd that the VT group improved less in V mode than the other groups given that their training consisted of video-alone stimuli.

Table 3.4 Percent change (improvement) relative to pre test for the Vowel test per mode (A, AV, V).

Trainig Groups	Vowel test: A % change	Vowel test: AV % change	Vowel test: V % change
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)

This analysis showed that 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 tokens during training. Concerning their use of auditory and visual cues, participants showed no overall greater benefit of visual cues in AV mode compared to A, as their scores at pre and post test revealed no significant difference between A and AV mode (Table 3.3: Pre & Post (M, SD). These findings suggested that L2 learners are not sensitive to visual cues that are available for the identification of English vowels and that they relied mainly on the auditory input for the perception of English vowels. One finding that came to a surprise was that the group that was trained with visual information only (VT) improved in their vowel identification including in the A mode as much as the other participants who did have access to auditory and visual input. However, they did not show greater sensitivity to visual cues in AV mode than the rest of the participants.

3.3.2 Vowel Test: Set 1

Table 3.5 Vowel test, Set 1. Mean (%) and standard deviation (SD) per vowel in A, AV &V mode for Pre and Post test, and overall mean (OM) and standard deviation (SD) are presented.

Set 1	A pre	SD.	A post	SD	AV pre	SD	AV post	SD	V pre	SD	V post	SD	OM *SD*
/ɑ:/	61.2	*17.7*	69.7	*18.9*	65	*18.2*	74.7	*15.1*	50.2	*14.1*	55	*17*	62.6 *18.7*
/æ/	58	*18.3*	62.2	*20.9*	58.7	*18.7*	62	*21.5*	62.7	*18.1*	62	*16.8*	61 *19.1*
/ʌ/	56.3	*26.9*	66.2	*20.3*	62.5	*21.8*	69.7	*18.6*	44.9	*18.2*	51.3	*17.1*	58.5 *22.2*

Set 1 included vowels /æ/, /ʌ/, /ɑ:/ in A, AV and V modes (Table 3.5). All result for the fixed effects are presented in Table 3.6. The effect of group was not significant. Time effect (pre, post) showed that there was an overall significant improvement in vowel identification for this set after training (Pre (M: 60.7, SD: 20.2), Post (M: 67.4, SD: 19.7)). There was also a significant effect of mode; this was caused by significant differences between all three conditions with higher scores for AV, then A and V mode (Table 3.7). There was no significant time*mode interaction; this means that improvement in all modes was similar from pre to post test.

Table 3.6 Vowel test, Set 1 (pre & post). Fixed effects and interactions for vowel identification in A, AV and V mode for the three training groups.

Fixed effects	
Group	F(2,44)= 1.657, p>.05
Time	F(1,6582)= 23.073, p<.001
Mode	F(2,6582)= 31.231, p<.001
Vowel	F(2,88)= 1.871, p>.05
Time*group	F(2,6582)= .729, p>.05
Time*mode	F(2,6582)= 1.257, p>.05
Time*vowel	F(2,6582)= 2.500, p>.05
Group*mode	F(4,6582)= .416, p>.05
Group*vowel	F(4,88)= .887, p>.05
Vowel*mode	F(4,6582)= 11.606, p<.001
Time*group*mode	F(4,6582)= .111, p>.05
Time*group*vowel	F(4,6582)= .746, p>.05
Time*mode*vowel	F(4,6582)= .159, p>.05
Group*mode*vowel	F(8,6582)= .438, p>.05
Time*group*mode*vowel	F(8,6582)= .501, p>.05

Table 3.7 Vowel test , Set1 (pre & post). Overall means (M) and standard deviation (SD) for A, AV & V mode and results for the mode effect.

Mode, Overall M (%) & SD	**Mode effect**
A (62.3, 23.7)	A-AV F(1,4362)= 5.308, p<.05
AV (65.4, 23.7	A-V F(1,4362)= 30.148, p<.001
V (54.4, 21.5)	AV-V F(1,4362)= 59.264, p<.001

Table 3.8 Vowel test , Set1 (pre & post). Vowel*mode interaction: results for the comparisons between modes per vowel in Set1 (Overall means per mode were used).

Vowel	Vowel*Mode interaction
/ɑ:/	A-AV F(1,1448)= 3.482, p>.05 A-V F(1,1448)= 25.271, p<.001 AV-V F(1,1448)= 46.542, p<.001
/æ/	A-AV F(1,1448)= 0.011, p>.05 A-V F(1,1448)= 0.833, p>.05 AV-V F(1,1448)= 0.672, p>.05
/ʌ/	A-AV F(1,1448)= 4.140, p<.05 A-V F(1,1448)= 27.946, p<.001 AV-V F(1,1448)= 50.651, p<.001

The effect of vowel was not significant; overall scores were very similar for the three vowels (Table 3.5) but this was modified by a significant vowel*mode interaction which came from different patterns for vowel scores per condition (Table 3.8). Vowel /ɑ:/ had similar scores in A and AV, both higher than V. Vowel /æ/ had similar scores in all three modes (A= AV= V), while vowel /ʌ/ had higher scores in AV than A and V, A higher than V (Table 3.5).

3.3.3 Vowel test Set 2

Table 3.9 Vowel test, Set 2 (Pre & Post). Mean (%) and Standard deviation (SD) per vowel in A, AV &V mode for Pre and Post test, and overall mean (OM) and SD are presented.

Set 2	A pre	SD	A post	SD	AV pre	SD	AV post	SD	V pre	SD	V post	S D	OM SD
/i:/	49.7	*14.8*	59.3	*15.3*	50.5	*14.5*	61.7	*13.8*	35.3	*16.3*	37.7	*14.3*	*49 17.7*
/ɪ/	80.3	*17.6*	82.7	*14.8*	75	*18.2*	80	*16.3*	42.8	*16*	44.4	*17.8*	*67.5 23.9*
/e/	93.3	*12.4*	94.4	*11*	91	*15.7*	94.1	*11.9*	63.3	*17.9*	64.1	*16.4*	*83.4 20*
/ɜ:/	75.5	*17.8*	84.5	*16.7*	77	*17.9*	85.3	*14.6*	68.6	*15.2*	70	*14.6*	*76.8 17.2*

Set 2 presented vowels /i:/, /ɪ/, /e/, /ɜ:/ in A, AV and V mode. All result for the effects and interactions are presented in Table 3.10. There was no significant effect of group, this means all participants improved in similar amounts regardless of their training modality. The time effect was significant, vowel identification improved after training Pre (M: 72.1, SD: 24.1), Post (M: 77.3, SD: 23.1).

The effect of vowel was significant; overall scores ranged from 49% (/iː/) to 83% (/e/). There was also a vowel*mode interaction which came from different patterns of scores per vowel and mode (Table 3.9). For instance, vowels /iː/ and /ɜː/ had similar overall scores in A and AV mode, both higher than V.

Table 3.10 Vowel test, Set 2 (pre & post). Fixed effects for vowel identification in A, AV and V mode for the three training groups.

Fixed effects	
Group	F(2,8768)= 1.630, p>.05
Time	F(1,8768)= 13.849, p<.001
Mode	F(2,8768)= 211.536, p<.001
Vowel	F(3,138)= 24.240, p<.001
Time*group	F(2,8768)= .661, p>.05
Time*mode	F(2,8768)= 5.459, p<.05
Time*vowel	F(3,8768)= 1.573, p>.05
Group*mode	F(4,8768)= .544, p>.05
Group*vowel	F(6,8768)= 2.030, p>.05
Vowel*mode	F(6,8768)= 20.624, p<.001
Time*group*mode	F(4,8768)= .637, p>.05
Time*group*vowel	F(6,8768)= 1.014, p>.05
Time*mode*vowel	F(6,8768)= .339, p>.05
Group*mode*vowel	F(12,8768)= 1.583, p>.05
Time*group*mode*vowel	F(12,8768)= .540, p>.05

There was an effect of mode with similar scores for A and AV, higher than V (Table 3.11). This effect was modified by a mode*time interaction which revealed that there was significant overall improvement for vowels in this set in A and AV but not in V mode.

Table 3.11 Vowel test, Set 2 (pre & post). Overall means (M) and standard deviation (SD) for A, AV & V mode and results for the mode effect are included.

Mode, Overall Mean (%) & SD	**Mode effect**
A (77.5, 21)	A-AV F(1,5784)= 0.010, p>.05
AV (76.8, 20.6)	A-V F(1,5784)= 247.386, p<.001
V (53.3, 21.1)	AV-V F(1,5784)= 265.161, p<.001

Table 3.12 Vowel test, Set 2 (pre & post). Overall mean scores (M) and standard deviation (SD) per mode and mode*time effect results are presented.

Mode: Pre & Post Mean(%), SD	**Mode*Time**
A : Pre (74.7, 22.2) Post (80.2, 19.4)	F(1,2808)= 10.269, p<.05
AV: Pre (73.4, 22.1) Post (80.4, 18.4)	F(1,2808)= 17.021, p<.001
V : Pre (52.5, 21.3) Post (54.3, 20.7)	F(1,2808)= 0.484, p>.05

The vowel effect was significant, scores ranged from 49% (/iː/) to 83% (/e/) (Table 3.10). There was also a vowel*mode interaction (Table 3.13), which was due to similar scores in A and AV, higher than V for vowels /iː/, /e/ and /ɜː/ (A=AV>V) and higher A than AV and V for /ɪ/ (A>AV>V) (Table 3.13).

Table 3.13 Vowel test, Set 2 (pre & post). Vowel*mode interaction: results for the comparisons between modes per vowel in Set 2 (Overall means per mode were used).

Vowel	Vowel*Mode interaction
/iː/	A-AV F(1, 1448)= 0.396, p>.05 A-V F(1, 1448)= 47.869, p<.001 AV-V F(1, 1448)= 56.861, p<.001
/ɪ/	A-AV F(1, 1448)= 3.922, p<.05 A-V F(1, 1448)= 241.560, p<.001 AV-V F(1, 1448)= 172.277, p<.001
/e/	A-AV F(1, 1448)= 1.328, p>.05 A-V F(1, 1448)= 162.052, p<.001 AV-V F(1, 1448)= 165.741, p<.001
/ɜː/	A-AV F(1, 1448)= 0.393, p>.05 A-V F(1, 1448)= 22.255, p<.001 AV-V F(1, 1448)= 28.512, p<.001

3.3.4 Vowel test Set 3

Table 3.14 Vowel test, Set 3 (Pre & Post). Mean (%) and standard deviation (SD) per vowel in A, AV &V mode and overall mean (OM) and SD are presented.

Set 3	A pre	SD	A post	SD.	AV pre	SD.	AV post	SD	V pre	SD	V post	St. Dev.	OM SD
/ɔː/	60.4	*27.8*	71.5	*23.1*	54.2	*26.5*	65.1	*21.5*	21.8	*17.3*	25	*17.6*	49.6 *29.4*
/ɒ/	69.1	*21.4*	80	*13*	70	*18.8*	78.1	*15.7*	43.8	*16.1*	48.6	*14.1*	65 *21.6*
/uː/	47	*17.1*	52.1	*15.5*	48.4	*20.1*	53.1	*17.6*	33.5	*14.7*	37.5	*14.5*	45.3 *18.2*
/ʊ/	53.1	*34.3*	71.8	*25.3*	42	*33.8*	57	*28*	31.3	*21.1*	34	*20.4*	48.2 *31*

Set 3 presented vowels /ɔː/, /ɒ/, /uː/, /ʊ/. Table 3.15 shows the results for the fixed effects and interactions for vowels in this set. There was no significant effect of Training group, participants improved in similar amounts after training regardless of the training

modality they took. The time effect was significant, vowel identification scores were higher in the post test (Pre: 48.7, SD: 27; Post: 56.4, SD: 25.7). There was also a mode effect, scores in A were higher than AV and V; V mode showed the lowest scores (Table 3.16).There was also a mode*time interaction which showed that there was significant improvement after training in A and AV mode only (Table 3.17).

Table 3.15 Vowel test, Set 3 (pre & post). Fixed effects for vowel identification in A, AV and V mode for the three training groups.

Fixed effects	
Group	F(2,7640)= 0.862, p>.05
Time	F(1,7640)= 35.603, p<.001
Mode	F(2,7640)= 236.024, p<.001
Vowel	F(3,138)= 29.920, p<.001
Time*group	F(2,7640)= .295, p>.05
Time*mode	F(2,7640)= 3.182, p<.05
Time*vowel	F(3,7640)= 1.771, p>.05
Group*mode	F(4,7640)= 1.300, p>.05
Group*vowel	F(6,7640)= 1.587, p>.05
Vowel*mode	F(6,7640)= 14.764, p<.001
Time*group*mode	F(4,7640)= .646, p>.05
Time*group*vowel	F(6,7640)= .511, p>.05
Time*mode*vowel	F(6,7640)= .631, p>.05
Group*mode*vowel	F(12,7640)= 1.660, p>.05
Time*group*mode*vowel	F(12,7640)= .625, p>.05

Table 3.16 Vowel test, Set 3 (pre & post). Overall means (M) and standard deviation (SD) for A, AV & V mode and results for the mode effect are included.

Mode, Overall M (%) & SD	**Mode effect**
A (63.2, 25.3)	A-AV F(1,5032)= 6.933, p<.05
AV (59.8, 25.7)	A-V F(1,5032)= 381.704, p<.001
V (34.7, 18.8)	AV-V F(1,5032)= 298.952, p<.001

Table 3.17 Vowel test, Set 3 (pre & post). Overall mean scores (M) and standard deviation (SD) per mode and mode*time effect results are presented.

Mode: Pre & Post Mean(%), SD	**Mode*Time**
A : Pre (58, 27) Post (68, 22.2)	F(1,2424)= 28.731, p<.001
AV: Pre (55.3, 27.3) Post (64.2, 23.1)	F(1,2424)= 21.214, p<.001
V : Pre (32.8, 18.8) Post (36.6, 18.7)	F(1,2424)= 3.856, p>.05

The effect of vowel was significant; scores ranged from 45.3% (/uː/) to 65% (/ʊ/); but this effect was modified by a vowel*mode (Table 3.18). This interaction was due to two vowels (/ʊ/, /uː/) having similar scores in A and AV, higher than V (A=AV>V) and two vowels

(/ɔː/, /ʊ/) showing higher scores in A than AV and V, with the lowest scores in V (A>AV>V).

Table 3.18 Vowel test, Set 3 (pre & post). Vowel*mode interaction. Results for the comparisons between modes per vowel in Set 3 (Overall means per mode were used).

Vowel	Vowel*Mode interaction
/ɔː/	A-AV F(1, 1448)= 7.344, p<.05 A-V F(1, 1448)= 259.711, p<.001 AV-V F(1, 1448)= 200.813, p<.001
/ɒ/	A-AV F(1,1448)= 0.746, p>.05 A-V F(1, 1448)= 122.801, p<.001 AV-V F(1, 1448)=116.890, p<.001
/uː/	A-AV F(1, 1448)= 0.220, p>.05 A-V F(1, 1448)= 30.036, p<.001 AV-V F(1, 1448)= 35.566, p<.001
/ʊ/	A-AV F(1, 696)= 15.300, p<.001 A-V F(1, 696)= 64.480, p<.001 AV-V F(1, 696)= 21.903, p<.001

It is important to notice that the /ʊ/ vowel was not included in the vowel trainer used in this study (Iverson & Evans, 2009); yet, participants were able to improve their identification scores from pre test (M: 42.6, SD: 31.6) to post test (M: 54.3, SD: 29.1) without having access to this phoneme in the same way as the rest of the vowels (i.e. through training). This improvement for vowel /ʊ/ suggests participants may have used strategy of elimination based on improved identification performance of the rest of the vowels in the set. Thus the least identifiable was labelled with /ʊ/, as it was the only option left.

3.3.5 Confusions in vowel identification in A and AV

In chapter 2, vowel perception in AV mode (noise) showed a tendency of higher scores in most vowels for English native speakers (ENS), though only statistically significant for four vowels (/ɪ/, /ɜː/, /uː/, /ʌ/). The results for the L2 groups revealed similar patterns of confusions between the L2 groups and no visual benefit in AV mode to disambiguate vowel confusion. In this section, separate confusion matrices for each set with pre and post test scores were used to explore if confusions were reduced as a result of perceptual improvement after training. For the purpose of exploring whether there was improvement

in the integration of visual cues in AV mode by L2 beginner learners, only A and AV mode scores (Vowel test) were used for the confusion matrices. The stimulus is presented across and participants' responses (Resp.) downwards. Raw counts were transformed into percentages.

3.3.5.1 Confusions in Set 1 in A and AV mode, Pre and Post test

Table 3.19 Set 1. Confusion matrices for A and AV mode in Pre and Post test. Scores represent percentages.

Set1 Audio Pre Test				**Set1 Audio Post Test**			
Stimulus	/ɑ:/	/æ/	/ʌ/	Stimulus	/ɑ:/	/æ/	/ʌ/
Resp.	%	%	%	Resp.	%	%	%
/ɑ:/	61.2	13.6	13.3	/ɑ:/	69.7	9.6	10
/æ/	19.7	58	30.3	/æ/	9.3	62.2	23.8
/ʌ/	19.1	28.5	56.4	/ʌ/	21	28.2	66.2
Set1 AV Pre Test				**Set1 AV Post Test**			
Stimulus	/ɑ:/	/æ/	/ʌ/	Stimulus	/ɑ:/	/æ/	/ʌ/
Resp.	%	%	%	Resp.	%	%	%
/ɑ:/	65	13.2	13.3	/ɑ:/	74.7	10.4	12.2
/æ/	14	58.8	24.2	/æ/	4.3	62	18.1
/ʌ/	21	28	62.5	/ʌ/	21	27.7	69.7

The expected confusion for this set was bidirectional confusions for /æ/–/ʌ/ found for L2 beginners and advanced learners in the study presented in Chapter 2. In the same study, the three vowels showed some visual advantage and reduced confusions in AV for ENS with significant visual advantage for /ʌ/. The matrices in Table 3.19 show that vowel /ɑ:/ and /ʌ/ had slightly higher scores in AV than in A mode, both in the pre and post test. This may suggest that the visual information for these vowels is the source of the reduction in confusions with the other vowels in the set, though the pattern of confusions for /ɑ:/–>/ʌ/ and /æ/<–>/ʌ/ remained after training.

3.3.5.2 Confusions Set 2 in A and AV mode, Pre and Post test

Table 3.20 Set 2. Confusion matrices for A and AV mode in Pre and Post test. Scores correspond to percentages.

Set2 Audio Pre Test					**Set2 Audio Post Test**				
Stimulus	/iː/	/ɪ/	/e/	/ɜː/	Stimulus	/iː/	/ɪ/	/e/	/ɜː/
Resp.	%	%	%	%	Resp.	%	%	%	%
/iː/	49.7	6.6	1.8	1.9	/iː/	59.3	9	0	0.3
/ɪ/	47.6	80.3	2.9	1.3	/ɪ/	40	82.7	2.4	0.8
/e/	1.9	12.5	93.4	21.3	/e/	.03	8	94.4	14.4
/ɜː/	0.8	0.5	1.9	75.5	/ɜː/	0.4	0.3	3.2	84.6
Set2 AV Pre Test					**Set2 AV Post Test**				
Stimulus	/iː/	/ɪ/	/e/	/ɜː/	Stimulus	/iː/	/ɪ/	/e/	/ɜː/
Resp.	%	%	%	%	Resp.	%	%	%	%
/iː/	50.5	9.2	1.6	2.9	/iː/	61.7	12.2	0.5	0.8
/ɪ/	45.7	75	2.7	1.6	/ɪ/	37	80.1	3.4	0.5
/e/	2.1	15	91	18.4	/e/	1.3	7.2	94.1	13.3
/ɜː/	1.6	0.8	4.8	77.1	/ɜː/	0	0.5	2	85.4

Based on the results found for beginners in study 1 (Chapter 2), the confusions expected for L2 learners were /ɪ/–>/e/, /iː/–>/ɪ/and /ɜː/–>/e/. For ENS, visual advantage was found for /iː/, /e/, /ɜː/ in AV mode (Chapter 2). In the current study, the /ɪ/–>/iː/ confusion appears as the only strong confusion this set. No visual benefit was observed for any vowel in the set, though scores for most of the vowels are quite high in A mode after training which may indicate a ceiling effect in learning (Table 3.20).

3.3.5.3 Confusions Set 3 in A and AV mode, Pre and Post test

Table 3.21 Set 3. Confusion matrices for A and AV mode in Pre and Post test. Scores correspond to percentages.

Set3 Audio Pre Test					**Set3 Audio Post Test**				
Stimulus	/ɔː/	/ɒ/	/uː/	/ʊ/	Stimulus	/ɔː/	/ɒ/	/uː/	/ʊ/
PResp.	%	%	%	%	PResp.	%	%	%	%
/ɔː/	60.4	25.5	7.4	8.5	/ɔː/	71.5	18.1	0.8	4.3
/ɒ/	22.3	69.1	2.7	2.1	/ɒ/	17.8	80.1	0.5	0.5
/uː/	10.6	3.5	47.1	36.2	/uː/	6.9	0.8	52.1	23.4
/ʊ/	6.6	1.9	42.8	53.2	/ʊ/	3.7	1.1	46.5	71.8
Set3 AV Pre Test					**Set3 AV Post Test**				

Set3 AV Pre Test					**Set3 AV Post Test**				
Stimulus	/ɔː/	/ɒ/	/uː/	/ʊ/	Stimulus	/ɔː/	/ɒ/	/uː/	/ʊ/
PResp.	%	%	%	%	PResp.	%	%	%	%
/ɔː/	54.3	25.8	4.8	12.8	/ɔː/	65.2	21.3	2.4	5.3
/ɒ/	22.6	70	3.4	1.6	/ɒ/	18.8	78.2	1.3	2.1
/uː/	14.1	2.7	48.4	43.6	/uː/	12.8	0.5	53.2	35.6
/ʊ/	9	1.5	43.4	42	/ʊ/	3.2	0	43.1	57

Bidirectional confusions for /ɔː/–/ɒ/ and /uː/–/ʊ/ were expected as they were found in Study 1. Besides, visual information could have been used for /ɒ/ and /ʊ/, as it was found to be available for ENS in the confusion matrices in Study 1. The confusions found in the current study for L2 beginners revealed bidirectional patterns for /ɔː/–/ɒ/ and /uː/–/ʊ/, with stronger scores for confusions in the latter pair of vowel contrast. No benefit of visual information in AV mode for any vowel was found (Table 3.21).

3.3.6 Overall comparison: L2 beginners and L2 Advanced groups

To compare the L2 beginners' perceptual improvement after the training with the L2 advanced group (tested in Chapter 2), a logistic regression using R (glmmPQL function) was run on the L2 beginners' post test scores and the L2 advanced scores. Group (2) vowel (11) mode (A, AV, V) and group*mode, group*vowel, vowel*mode and group*mode*vowel were introduced in the model as fixed effects. Participants and stimulus were the random factors in the model. The results (Table 3.22, Fig. 3.6) showed a significant group effect; the L2 beginners' overall post test scores (M: 64, SD: 7) were higher than the L2 advanced group (M: 60, SD: 9.3). The mode effect (Table 3.22) was significant; no difference was found between A and AV mode, with lower scores for V. The group*mode interaction, revealed that both L2 groups had similar scores in A and AV mode, higher than V mode (Table 3.24).

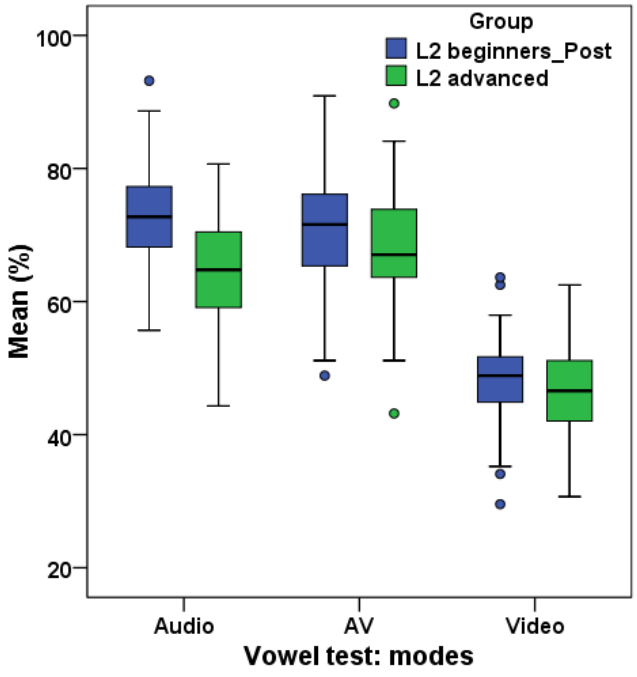


Figure 3.6 Boxplots for mean vowel identification accuracy (mean %) for the L2 beginners (Post test) and the L2 advanced (scores reported in Chapter 2).

Table 3.22 Main effects and interactions for the Vowel test data analysis using L2 beginners (Post test data) and L2 advanced and ENS data (one-time test).

Effects	
Group	F(1,82)= 4.259, p<.05
Vowel	F(10,21020)= 98.308, p<.001
Mode	F(2,21020)= 349.289, p<.001
group*mode	F(2,21020)= 5.743, p<.05
group*Vowel	F(10,21020)= .940, p>.05
Vowel*mode	F(20,21020)= 15.533, p<.001
group*mode*vowel	F(20,21020)= 1.637, p<.05

Table 3.23 Mode results for L2 beginners (Post test). Means (M) and Standard deviations (SD) per mode and F values for the comparisons are presented.

Mode (M,SD)		
A (69, 9.3)	A-AV	F(1,13166)= .491, p>.05
AV (70, 9)	A-V	F(1,13166)= 513.642, p<.001
V (48, 6.1)	AV-V	F(1,13166)= 535.356, p<.001

Table 3.24 Scores per mode for the L2 beginners and L2 advanced groups. Comparisons for between-mode difference for the group*mode interaction.

L2 beginners (Post test)		**L2 advanced**	
Mode (M,SD)	Mode comparison	**Mode (M,SD)**	Mode comparison
A (72.2, 9.1) AV (71.6, 8.7) V (48.8, 5.4)	A-AV F(1, 7828)= .251, p>.05 A-V F(1, 7828)= 328.180, p<.001 AV-V F(1,7828)= 311.684, p<.001	A (65.2, 9.5) AV (67.4, 9.3) V (46.8, 6.8)	A-AV F(1,6158)= 1.911, p>.05 A-V F(1,6158)= 172.658, p<.001 AV-V F(1,6158)= 208.938, p<.001

The vowel effect, group*vowel, vowel*mode and group*mode*vowel were significant. The three-way interaction included all the single and two-way interactions; Post-hoc analysis showed that, in general, vowels had similar scores in A and AV mode with lower scores in V mode. However, some vowels broke this pattern. For instance, in the L2 beginners group some vowels showed higher scores in A than AV mode (/ɪ/, /ɔː/, /ʊ/), one vowel showed higher AV than A (/ʌ/) (see Tables 3.8, 3.13, 3.18). The L2 advanced group showed higher scores in AV for two vowels (/ɜː/, /ʊ/) (Chapter 2, Table 2.5).

In summary, the L2 beginner learners improved their vowel identification accuracy with the training sessions and became significantly better than the L2 advanced group. The improvement achieved by the L2 beginners in a period of six weeks is striking given that they differ in about three years of intensive learning. This suggests that regarding perception of English vowels, five training sessions (L2 beginners) in a month could equate four years of learning (L2 advanced). These results showed a strong impact of perceptual training though no improvement in the integration of visual cues in AV mode for vowel perception was observed.

3.3.7 Training improvement versus classroom learning

To compare the contribution of training and classroom learning to English vowel perception, the data for 34 participants tested in the study presented in Chapter 2 and in Chapter 3 were used. These 34 L2 beginner learners were actually given the “Vowel test” three times. The first time will be referred to as pre test1 (data in Chapter 2). The second time will be referred to as pre test2, this corresponds to the pre test before training

described in this chapter (Chapter 3). And the third time corresponds to the post test (after training) in the current study (post test, Chapter 3). The length of time between pre test1 and pre test2 was six months, and between pre test2 and post test was four to six weeks. The level of English of these participants was described as beginners, but at pre test1 they were just starting their course at university. After pre test1, they had six months of quite intensive classroom learning (20 hours English instruction per week) without vowel training. After these six months, they managed to make progress in their level of English and were probably at the highest level of the beginner category. It was at this point when they were given the vowel training programme and, thus, we obtained the pre test 2 (before training) and post test (after training). It must be acknowledge that a learning effect for those participants who took the test battery for a second and third time was possible; so the overall scores in the Vowel test for the new (13) and the old participants (34) was compared.

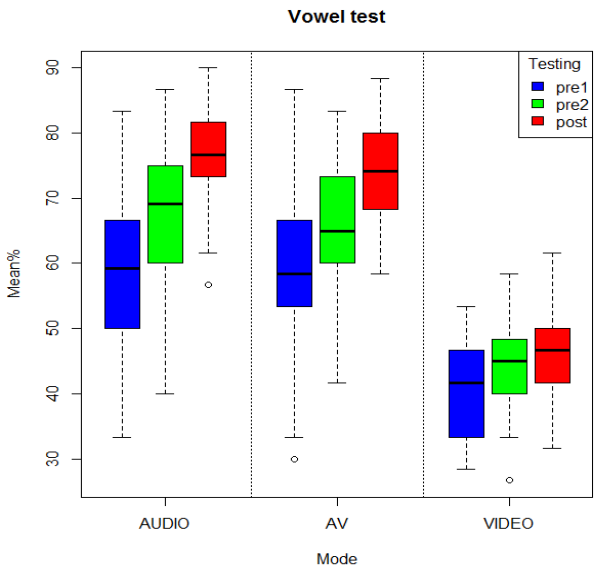


Figure 3.7 Boxplot for vowel identification accuracy (mean%) in the Vowel test. Results show change over time from Pre test 1 (pre1), Pre test 2 (pre2) and Post test (post) in Audio, audio-visual (AV) and Video mode. (Data for 8 vowels and 34 participants).

The overall results in the Vowel test per participant and group (old, new) were submitted to a mixed model analysis in R (lme function). Group (new, old) and time (pre, post) were the

fixed factors and participant was the random effect. The results showed no significant effect of group; overall scores were similar ($F(1, 45)= 2.352$, $p>.05$). Time was significant ($F(1, 45)=90.607$, $p<.001$); this was expected as there was improvement after training. The group per time interaction was not significant ($F(1, 45)= .591$, $p>.05$). These results suggested that there was no learning effect for the 34 participants who had taken the vowel test in study 1; their results did not differ significantly from the new participants' performance.

Due to technical problems at pre test 1, data were lost for the A mode for some vowels so the comparison presented here will be based on data for 8 vowels in three modes (A, AV, V) from the Vowel test. The vowels included in this analysis will be: /e/, /ɜː/, /ɪ/, /iː/, /ɒ/, /ɔː/, /ʊ/,/uː/.

To compare vowel identification from six-month classroom instruction and the vowel training programme, data from the Vowel test at three times were used to run a logistic mixed model in R (glmmPQL function). The fixed effects were time (pre1, pre2, post), mode (A, AV, V) and vowel (8), time*mode, time*vowel, vowel*mode, time*mode*vowel. The results (Table 3. 25, Fig.3.7) showed there was a significant effect of time; overall mean for pre test2 (M: 59) was higher that pre test1 (M: 53), post test mean (M: 65) was higher than pre test2. There was also a mode effect; this was caused by no significant difference between A and AV mode, both higher than V mode. The vowel effect was also significant, mean scores ranged from 44 (/ʊ/) to 80.5 (/e/).

The time*mode interaction was significant; the post hoc analysis showed that A and AV mode had significant differences between pre test1 and pre test2, and between pre test2 and post test. The video mode showed significant difference only between pre test1 and pre test2 (Fig. 3.7). There was also a mode*vowel interaction, this was caused by vowels showing different patterns per mode. However, this interaction has been described in Chapter 2 and Chapter 3 (above), so it won't be detailed here. No significant results were found for the time*vowel and time*mode*vowel interactions; this indicates that there was similar improvement for all vowels across the three testing times.

Table 3.25 Results for the logistic regression analysis for the Vowel test for 34 participant (L2 beginners), tested at three times: Pre test1, Pre test2 and post test.

Fixed effects	
Time	F(2,18254)= 60.650, p<.001
Mode	F(2,18254)= 332.879, p<.001
Vowel	F(7,18254)= 172.051, p<.001
Time*mode	F(4,18254)= 11.739, p<.001
Time*vowel	F(14,18254)= 2.246, p>.05
Mode*vowel	F(14,18254)= 13.613, p<.001
Time*mode*vowel	F(2,18254)= 1.041, p>.05

The overall percent improvement between the three test points was calculated per mode and overall. Improvement from pre test1 to pre test2 is estimated relative to pre test1 (([pre test2-prestest1]/pre test1). Improvement from pre test2 to post test is calculated relative to pre test2 (([post test-pre test2]/pre test2)). The same procedure was used to estimate the percent improvement per mode (A, AV, V). The results in Table 3.26 show that there were similar amounts of improvement between the testing times. This suggested that the amount of vowel identification improvement achieved with the Vowel training sessions (in four to six weeks) equalled the improvement obtained after six months of classroom learning without perceptual training. This finding highlights the effectiveness of high-variability perceptual training, though it is not clear whether the improvement comes from learners' capacity to create new phonetic categories or simply improving the identification and labelling of certain vowel contrasts. This issue will be revisited later in the discussion part (final chapter).

Table 3.26 Improvement in the Vowel test from pre test1 to pre test2 and from pre test2 to post test. Data for 8 vowels and 34 participants.

Vowel test **Mode means** **(Pre test1 - Pre test2 – Post test)**	**Mean %** **improvement** **Pre test1 – Pre test2**	**Mean %** **improvement** **Pre test2 – Post test**
A (58.5 - 67.7 – 76.3)	15.7	12.7
AV (59 – 65.3 – 73.6)	10.6	12.7
V (40.5 – 43.7 – 46.1)	7.9	5.4
Total improvement	**11.4**	**10.2**

3.3.8 Individual differences and vowel perception in L2 learners

In order to find possible sources for individual differences in the participants' vowel identification capacity, different measures were obtained before and after training. This test had been given to 34 of the participants six months before this training study; nonetheless, all tests were given to these participants again. A frequency discrimination test (FDT) was used to test participants' ability to discriminate auditory frequency differences. A McGurk effect test was used to obtain a measure of visual bias in speech perception. In addition to auditory and visual bias measures, a BKB-sentence test was used to measure word in sentence perception in two modalities (A and AV). Participants' English proficiency level was established by transforming their final mark in their English class (end of term 1) into percentage.

a. Auditory frequency discrimination test

The overall means for frequency "jnd" (steps relative to the standard) obtained from the pre (M: 35.5, SD: 12.3) and post test (M: 36.5, SD: 18) were used to run separate correlations with the pre and post scores in vowel identification in the vowel test (A, AV and V mode). Results showed no correlation between the measures neither before nor after training (Table 3.27). This would confirm the findings in Chapter 2; as no relation between the individuals' capacity to discriminate auditory frequency differences and vowel identification capacity was found. It also provides more evidence for the difference with a previous study (Lengeris & Hazan, 2010) which did find a correlation for these two measures. As suggested earlier in Chapter 2, the lack of correlation in the current study may have to do with more variability in the stimuli and a larger number of contrasts tested.

Table 3.27 Correlations for the frequency discrimination test (FDT) and the Vowel test (A, AV, V). Two cases were excluded from the FDT-post test due to instructions misunderstanding.

	Vowel test: A mode	**Vowel test: AV mode**	**Vowel test: V mode**
FDT (Pre test)			
Pearson correlation	.176	.232	.013
Sig. (2-tailed)	>.05	>.05	>.05
N	47	47	47
FDT (Post test)			
Pearson correlation	.094	.150	.016
Sig. (2-tailed)	>.05	>.05	>.05
N	45	45	45

b. Visual bias

The participants' visual bias was measured with a McGurk test as explained in Chapter 2. In the current study, the visual effect (VE) measures ranged from 0 (5 cases) to 0.92 before training (pre test) and from 0 (5 cases) to 0.94% after training (post test) with 2 participants with negative scores (i.e. more auditory biased). The VE overall mean at pre test (M: .31, SD: .23) and post test (M: .30, SD: .24) was very similar but the individual results showed a lot of variability in the amount of VE found (Fig. 3.8), as seen in the Standard Deviation.

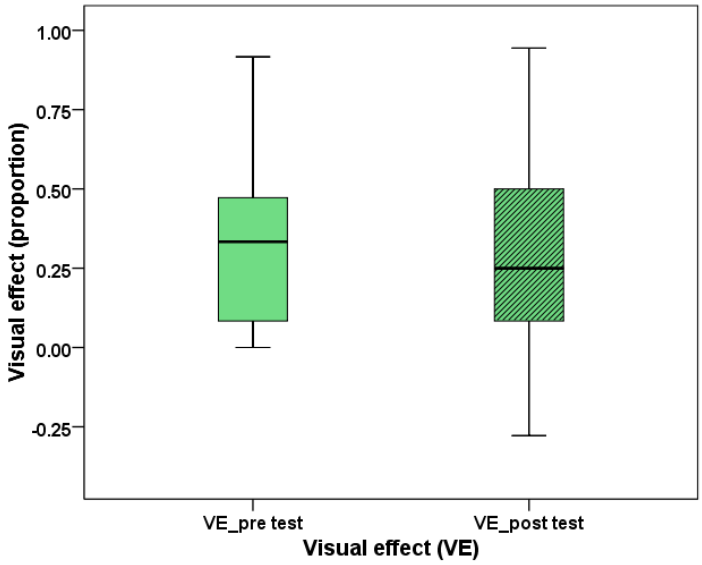


Figure 3.8 Boxplots for the Visual Effect (VE) at pre and post test (before and after training) for the L2 beginner learners (47 participants).

Correlations were run for the VE and the scores in the Vowel test. The results suggested that participants' visual bias for speech perception was not related with the vowel identification measures before or after training in any modality (Vowel test: A, AV, V) (Table 3.28). These results also suggest that due to the difference in the degree of salience of the visual cues for English consonants and vowels, the use of the McGurk test as a tool to measure visual bias for vowels is less effective. Therefore, an alternative way to measure sensitivity to visual cues, specifically for vowel perception may be needed.

Table 3.28 Correlations between the Visual Effect (VE) measures and the vowel identification accuracy in the Vowel test (A, AV, V) before and after training (pre, post test).

	Vowel test: A mode	**Vowel test: AV mode**	**Vowel test: V mode**
VE (Pre test)			
Pearson correlation	.183	.129	.160
Sig. (2-tailed)	>.05	>.05	>.05
N	47	47	47
VE (Post test)			
Pearson correlation	.112>.05	.106	.231
Sig. (2-tailed)	47	>.05	>.05
N		47	47

Another indicator of visual bias is the amount of McGurk effect (McGurk & MacDonald, 1976); that is when participants show a fused response for AV incongruent stimulus in noise (stimulus: A-ba +V-ga, response: "da"). The boxplots in Fig. 3.9a show a slight change from pre to post test in the amount of McGurk effect which was explored with a mixed model analysis in R (lme function). The fixed effects were training group, McGurk effect and time, and the possible interactions. The results revealed no significant effect of group ($F(3,44)=.219$, $p>.05$); the amount of visual bias did not differ significantly between training groups. Time effect was not significant ($F(1,44)=.105$, $p>.05$); the post test increment in McGurk effect was not significant. The group*time interaction was not significant ($F(2,44)= 2.602$, $p>.05$), no significant difference was found between groups over time. However, the AVT and the VT groups showed a tendency of higher overall mean McGurk effect at post test whereas the AT group reduced the amount of McGurk effect.

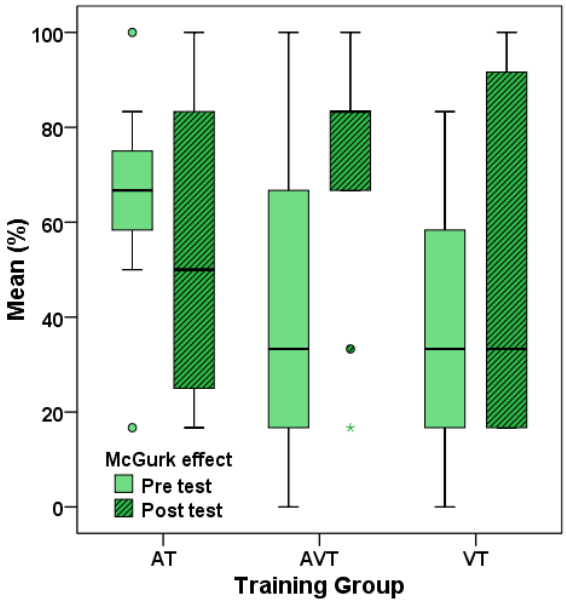


Figure 3.9a Visual bias in clear. Boxplots for the McGurk effect at pre and post test for the L2 beginner learners per training group. The AVT and VT groups showed a tendency of higher overall McGurk effect.

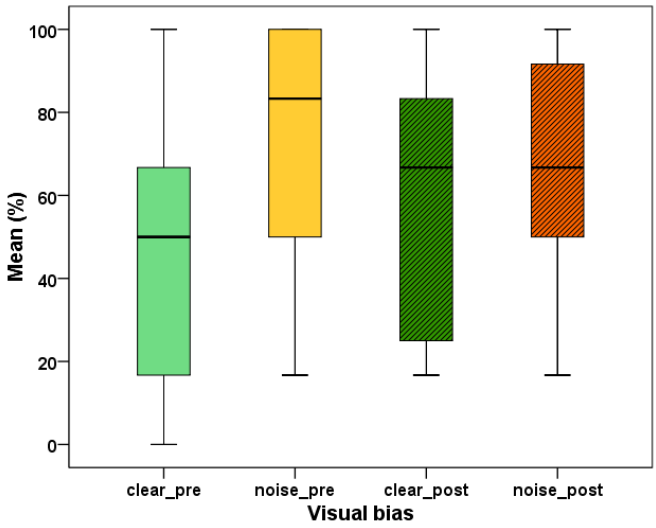


Figure 3.9b Visual bias in clear (McGurk effect) and in noise for the L2 beginners group**.** Higher visual bias was found in noise condition.

The McGurk effect is a measure of visual bias in clear condition. The visual bias was also measured in noise for stimulus "A_ga+V_ba" perceived as "ba". A mixed model analysis

was run on the scores for "ba" responses (in R, lme function) with group (AT, AVT, VT), time (pre,post), mode (clear, noise) as fixed effects and participants as random effect. The only significant effect was condition ($F(1,131)= 54.984$, $p<.001$); this was due to higher overall scores for visual bias in noise (M: 69.7, SD: 21.1) than in clear (M: 53.2, SD: 24.2).

c. BKB-sentence test

Learners' capacity to perceive key words in short sentences was measured before and after training with the BKB-sentence test presented in audio (BKB_A) and audio-visual (BKB_AV) mode. A mixed model was used for the data analysis in R (lme function). Group (AT, AVT, VT), time (pre, post) and mode (A, AV) and all possible interactions were the fixed effects (Table 3.29). Participants were the random effect. The results showed an effect of time; with higher scores in the post test (pre M: 85, SD: 12.7; post M: 91, SD: 9.1). No group effect was found; training groups improved in similar amounts. The mode effect was significant; the overall mean for AV (M: 88.6, SD: 9.8) was slightly higher than A (M: 87, SD: 12). There were no significant interactions.

Table 3.29 Results for the mixed model analysis. Fixed effects results for the BKB-sentence test. Data from pre and post test. Group refers to training groups (AT, AVT, VT).

Effect	
Group	$F(2,44)= 1.421$, $p>05$
Time	$F(1,1.32)= 61.375$, $p<.001$
Mode	$F(1,132)= 3.979$, $p<.05$
Group*time	$F(2,132)= .182$, $p>.05$
Group*mode	$F(2,132)= .633$, $p>.05$
Time*mode	$F(1,132)= .703$, $p>.05$
Group*time*mode	$F(2,132)= .140$, $p>.05$

Table 3.30 Correlations between BKB-sentence test and Vowel test (A, AV, V) in the pre and post test measures. **Correlation is significant at 0.01 level (2-tailed).

	Vowel test: A mode	**Vowel test: AV mode**	**Vowel test: V mode**
BKB_pre (Pre test)			
Pearson correlation	.581**	.758**	.216
Sig. (2-tailed)	<.001	<.001	>.05
N	47	47	47
BKB_post (Post test)			
Pearson correlation	.584**	.577**	.426**
Sig. (2-tailed)	<.001	<.001	<.05
N	47	47	47

Overall scores for the pre and post BKB-sentence test were used to run correlations between the vowel identification accuracy in the pre and post measures using the Vowel test (A, AV, V). The results showed that scores in the BKB_pre correlated with the scores in A and AV mode in the Vowel test. The BKB_post scores correlated with results in A, AV and V mode in the Vowel test (Table 3.30).

These results showed that the L2 beginner learners were able to attend to key words in simple short sentences and that this capacity was related to vowel identification accuracy for isolated words in the Vowel test before and after training. In the previous study (Chapter 2), no relation was found between the BKB-test scores and the performance in the Vowel test. It could be speculated that the correlation found now may be mediated by higher level of proficiency, as BKB-scores were also higher now (before and after training) than in the previous study (Chapter 2) where the same test was used but participants had less knowledge of English.

d. Learners' level of proficiency

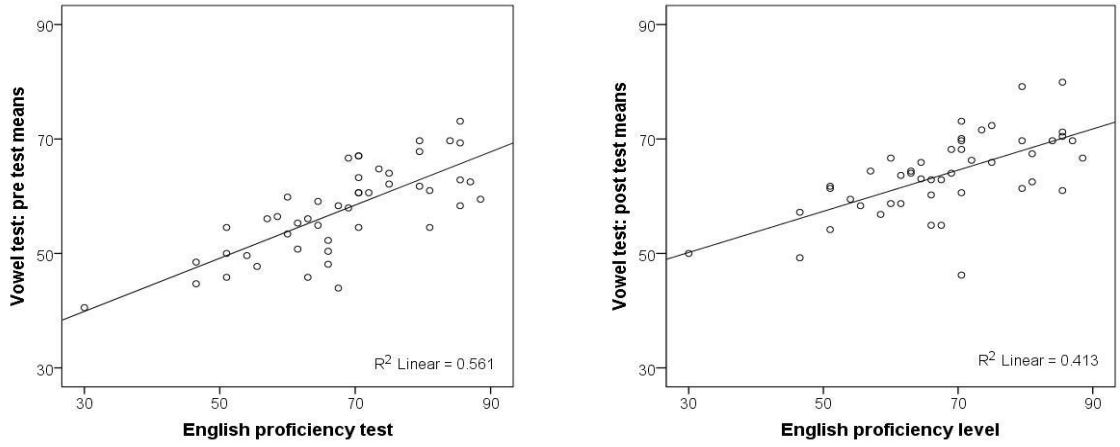


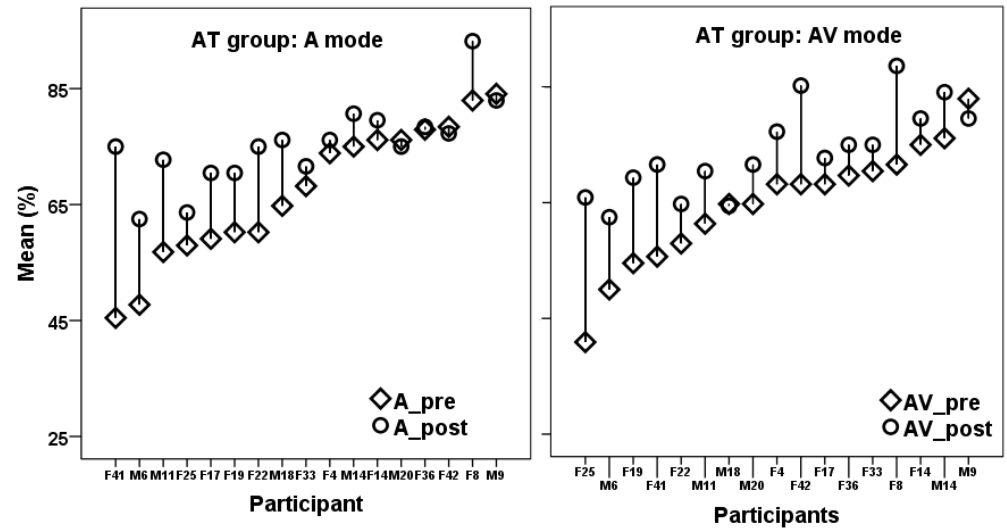
Figure 3.10 Scatter-plots showing correlation between English proficiency level (%) and vowel identification accuracy in the Vowel test (A, AV, V together) before and after training. Both axis show percent correct overall means.

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 the vowel identification accuracy (Vowel test) before and after training (pre, post tests). The results showed a strong correlation (Fig.3.10) 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 the performance in vowel identification depends on the amount of knowledge and experience with the language (English) rather than any other of the measures used in the current study.

In a previous study (Chapter 2), this index of level of proficiency did not show any correlation with vowel perception (Vowel test) for the L2 beginners or L2 advanced group; both groups differed in three years of learning English at university level. However, the beginner group tested here (Chapter3) had improved their level of English compared to their first time tested six months earlier (data in Chapter 2). Also, the beginner group in the current study outperformed the advanced group in their vowel identification after training. These results reveal the strong impact of perceptual training which boosted learning in the beginner participants to the level which they would potentially achieve after three years of learning.

e. Who benefits the most from training?

When exploring individual variability in vowel identification from pre to post test , a general pattern emerged which showed that learners with lower scores achieved greater change in their vowel perception after training (Fig. 3.11, 3.12, 3.13). This finding would suggest that having more room for improvement plays a role in the amount of benefit the learners can obtain from a perceptual training programme. Some of the L2 beginner learners may have experienced some kind of "learning plateau"; particularly those who scored between 70% and 80 % in the pre test did not show much improvement after training. Another issue of interest is that the VT group was expected to make more improvement in the V mode since they trained on video-alone in their sessions. However, they did not particularly improve more in the V mode than the other participants in the AT and AVT groups.



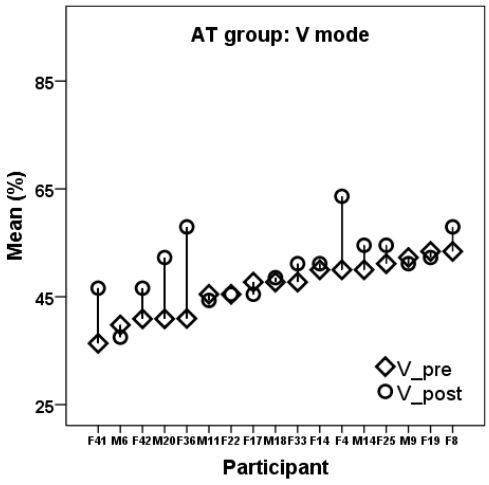


Figure 3.11 Line graphs for the Vowel test, mean identification scores in the Pre and post test in A, AV and V mode for the Audio Training (AT) group.

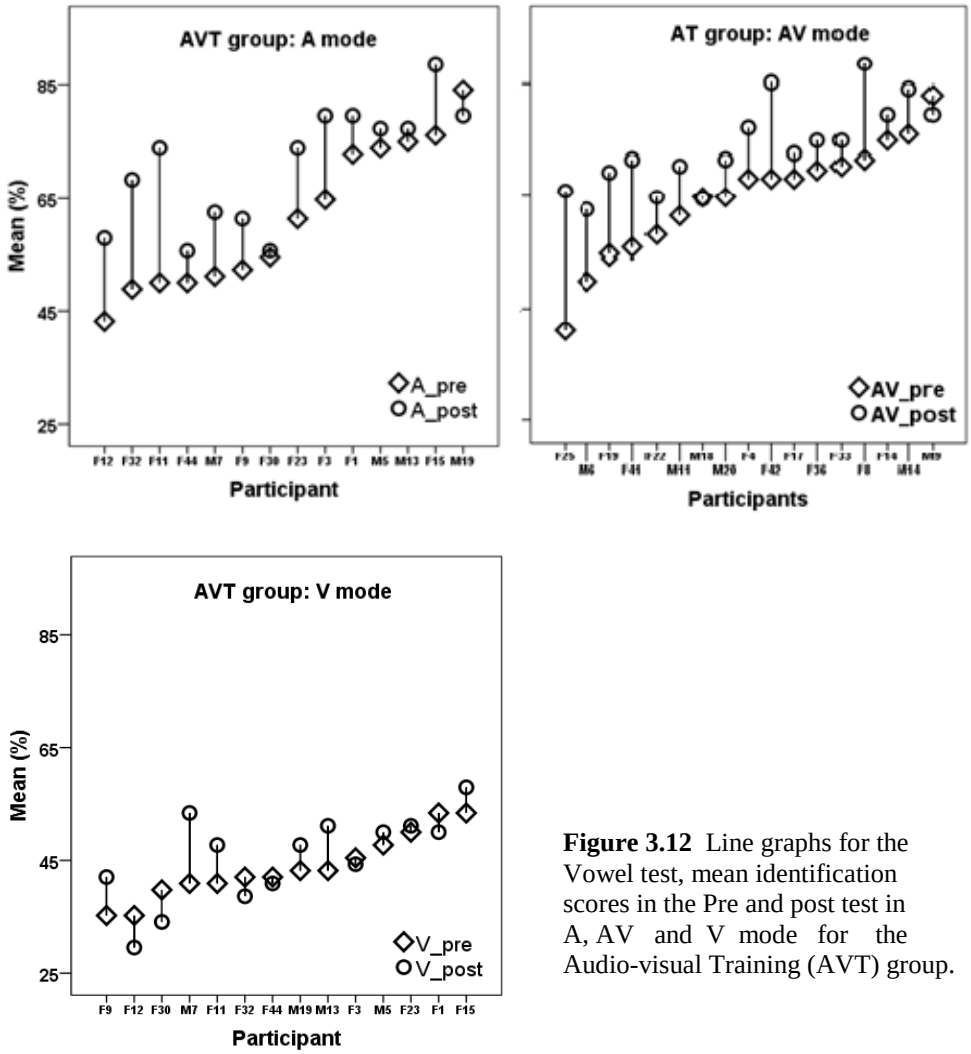


Figure 3.12 Line graphs for the Vowel test, mean identification scores in the Pre and post test in A, AV and V mode for the Audio-visual Training (AVT) group.

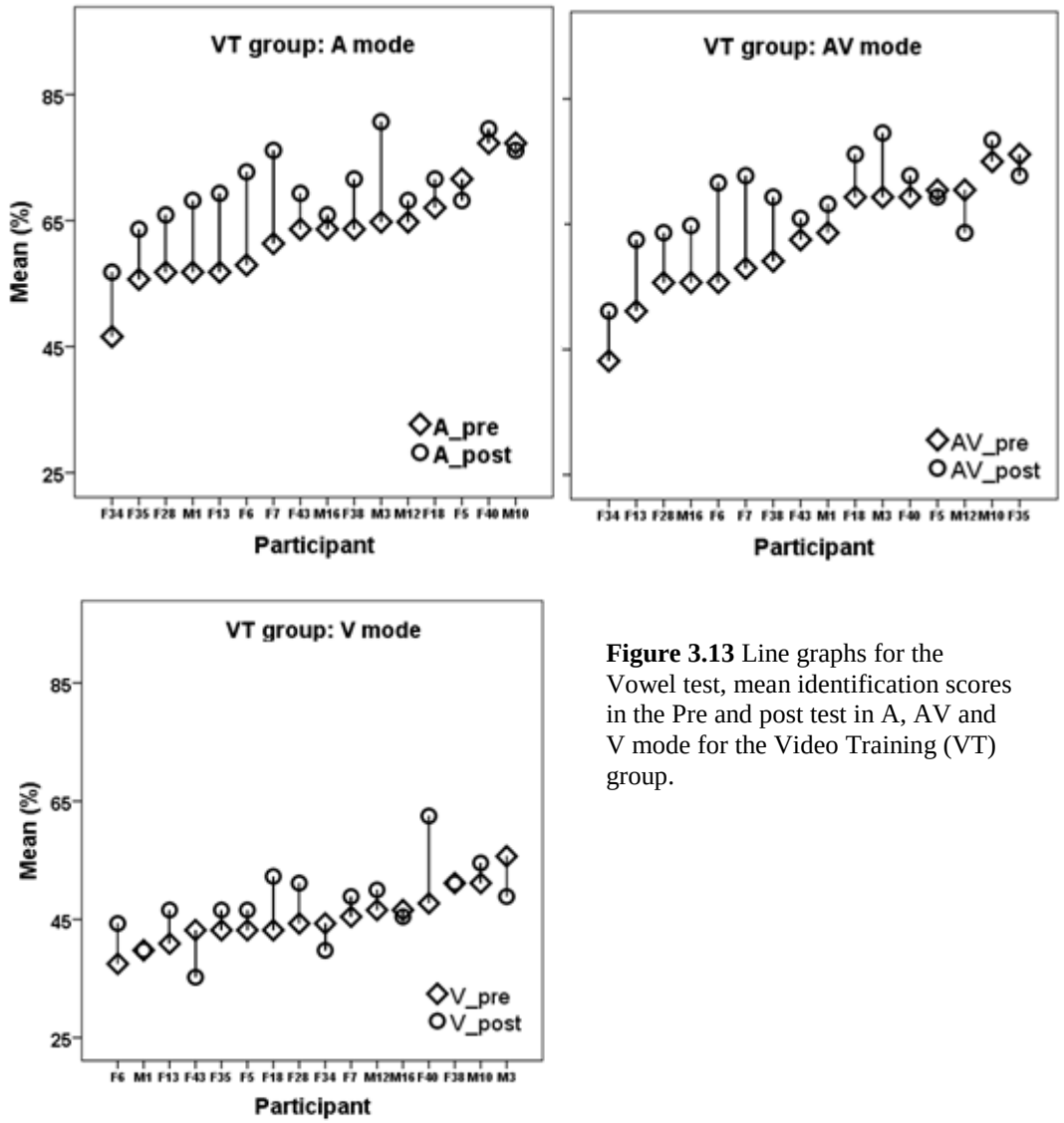


Figure 3.13 Line graphs for the Vowel test, mean identification scores in the Pre and post test in A, AV and V mode for the Video Training (VT) group.

3.3.9 True-or-False sentence test

To test the generalisation of the vowel learning to vowel processing in a context where accurate perception of the vowel was needed to make a decision about the status of a sentence as meaningful or semantically anomalous, a sentence test was used. This test will be referred to as the "True-or False" (TF) test; it presented sentences in AV mode using minimal pair words. A logistic regression analysis was used (glmmPQL function). The fixed effects were time (pre, post), training group (AT, AVT, VT) vowel (9) and speakers

(3); participants (47) and stimulus were the random effects. The results showed that the time effect was significant; the scores in the post test (M: 57.4, SD: 16.4) were significantly higher than the pre test (M: 54, SD: 14.3). There was no effect of training group; participants' overall scores were similar regardless of their vowel training modality (Fig. 3.14). The vowel effect was significant; overall mean scores ranged from 51% SD: 13 (/e/) to 59% SD 15.2 (/ɑ:/). Though the correct identification of a vowel determined the correct meaning of a sentence, sentences may have potentially varied in difficulty for reasons other than the vowel itself (e.g. lexical meaning of other words in the sentence).

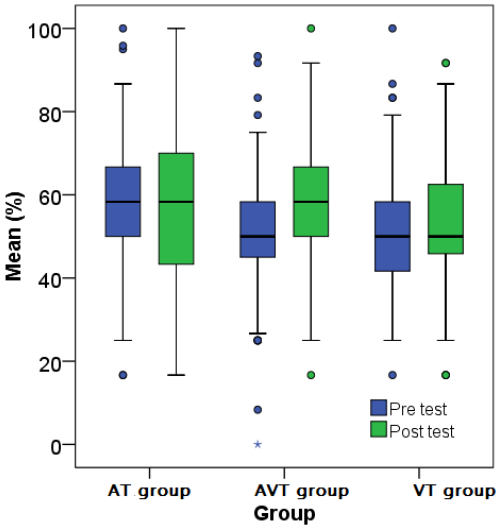


Figure 3.14 Boxplots for the True-or-false sentence test per training group before and after training.

No effect of speaker was found, this means that participants sentence accuracy did not differ as a function of speaker who produced the sentence presented in the videos. The only significant interaction was group*time; this was caused by the AVT group showing more improvement from pre to post than the other groups (Fig 3.14); overall mean scores for this group remained near chance level (Pre test: 52.3. SD: 14; Post test: 58.7, SD: 16). All the fixed effect values are presented in Table 3.31.

Table 3.31 Fixed effects for the True-or-False sentence test, L2 beginners group.

Effect	
Group	F(2,44)= 1.406, p>.05
Time	F(1,11295)= 20.693, p <.001
Vowel	F(1,11295)=12.221, p<.001
Speaker	F(2,11295)= 0.307, p>.05
Group*time	F(2,11295)= 3.168, p<.05

To explore the relation between the TF sentence test and the L2 learners' knowledge of the minimal-pair words used in the sentences, a vocabulary test which contained a list of the minimal-pair words was given to the L2 beginner and the Control group. On average, overall mean scores were very high for both groups; L2 beginners (M: 72, SD: 14) and Control group (M: 69, SD: 17). Correlations were run between the scores in the TF sentence test and the Vocabulary test. Some significant but moderate correlations were found for the L2 beginners and significant strong correlations for the Control group (Table 3.32). The scores in the TF sentence test and the English test scores were also used to run correlations; the L2 beginners group showed some weak but significant correlations. Taken altogether, knowing the words used in the TF sentence test did not contribute much to account for the perception of the minimal pair-words used in the sentences of this test.

Table 3.32 Pearson-moment correlation results for the TF sentence test, Vocabulary test and English test.

TF sentence test (pre & post)	**Vocabulary test**	**English test**
L2 beginner group		
Pre test	r=342*, N=47, p<.05	r=373**, N=47, p<.05
Post test	r=.405**, N=47, p<.05	r=390**, N=47, p<.05
Control		
Pre test	r=.589*, N=13, p<.05	r=.450, N=13, p>.05
Post test	r=.678*, N=13, p<.05	r=.310, N=13, p>.05

The True-or-false (TF) sentence test was also given to an L2 control group (13 participants); they were learners with similar language experience and proficiency level (beginner) and they did not receive training. They were tested twice, with six weeks in between testing sessions, in parallel time with the vowel training study. The data of these

two groups was compared using a mixed model analysis in R (lme function) with group (2), time (pre, post) and group*time as fixed effects and participants as random factor. The results showed that none of the effects was significant (Table 3.33). The group effect was not significant, groups had similar overall scores (L2 beginner: 55.7, control: 55.5). The absence of significant time effect showed there was no improvement after the six weeks. The group*time effect was not significant (Figure 3.15); this means there was no difference in scores between groups per time. These results suggest that training did not make an impact on participants' capacity to perceive vowel differences to detect the anomalous sentences in the TF sentence test, as they performed similar to a group of learners who did not receive the vowel perceptual training.

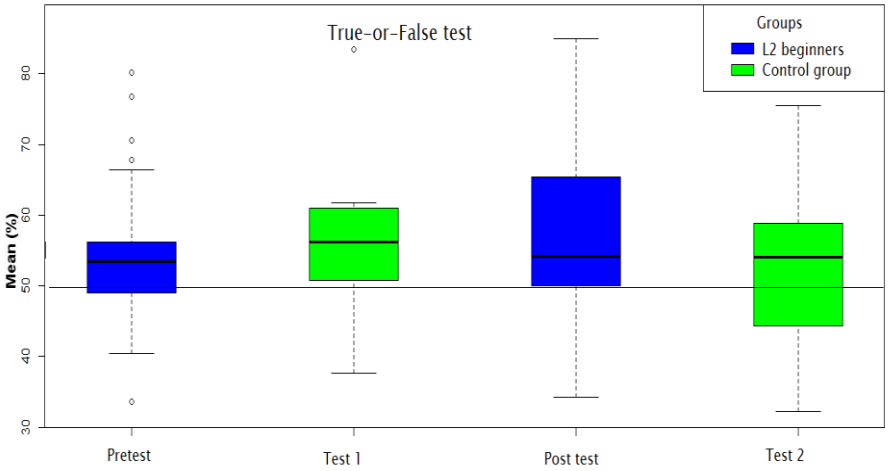


Figure 3.15 Boxplots for the True-or False test for the L2 beginners (pre, post test) and Control group (test1, test2). The time between test was 6 weeks. The line indicates the chance level at 50%.

Table 3.33 Overall means in the True-or-False sentence test per group and time. Fixed effects values for the analysis comparing L2 beginners and the control group.

Means per group (M, SD)	**Fixed effects**
,L2 beginner Pre (54, 9.1), Post (57.4, 11.4)	Group (F(1,59)=.007, p>.05)
Control Pre (56.3, 10.8), Post (54.7,14.5)	Time (F(1,57)= 1.745, p>.05)
	Group*time (F(1,57)= 1.408, p>.05)

The TF sentence test was also given to the L2 advanced group (same as in Chapter 2) and a group of English native speakers (ENS, same as in Chapter 2). A comparison between L2 beginners (only post test scores), the Control group, the L2 advanced group and ENS was explored with a mixed model analysis run in R (lme function). The overall scores per group were used for this analysis; group (4) was the fixed effect and participants the random factor. The results showed there was a significant effect of group ($F(3,114)= 96.919$, $p<.001$); the ENS (M: 95.5) had significantly higher scores than the non-native group -as expected. The L2 advanced group (M: 63.6) obtained higher scores than the L2 beginners post test (M: 55.6) and the Control group (M: 56.2). Post hoc analysis showed that the scores for the L2 advanced group were significantly higher than the Control group $p<.05$, and also higher than the L2 beginners $p<.001$ (Fig. 3.16).

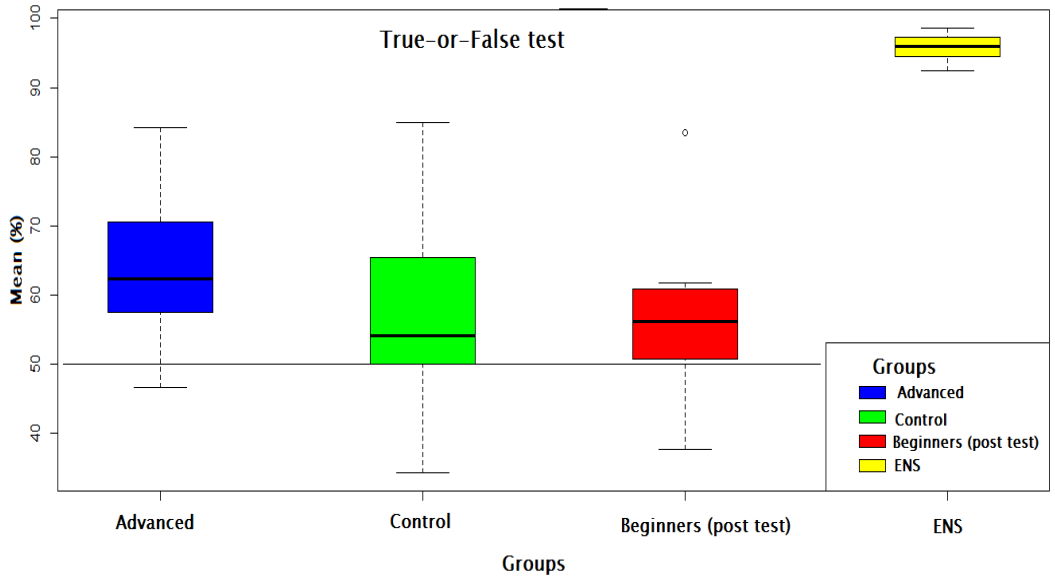


Figure 3.16 Boxplots for the TF sentence test for the L2 beginners (post test), L2 control (test 1 & 2) and for the L2 advanced group (test 1). The ENS overall mean was added with a star as a baseline (test 1).

L2 beginners' overall mean scores in the Vowel test and the TF sentence test (before and after training) were used to run correlations. Significant correlations were found at the pre

(r: .504**, N: 47, p<.001) and post test (r: .330*, N: 47, p<.05) but both correlations could only account for a small amount of the variance in the data.

The results in this test suggest that the perception of English vowel contrasts in words embedded in sentence material was really difficult for L2 learners, irrespective of the level of proficiency (e.g. beginners vs. advanced). The group that took the vowel training sessions showed some improvement in the TF sentence test after training, but their results remained near chance level. This suggested that the improvement in the perception of English vowels at word level did not generalise to perceiving vowel contrasts in a more naturalistic context in sentence length material.

3.3.10 Vowel Training

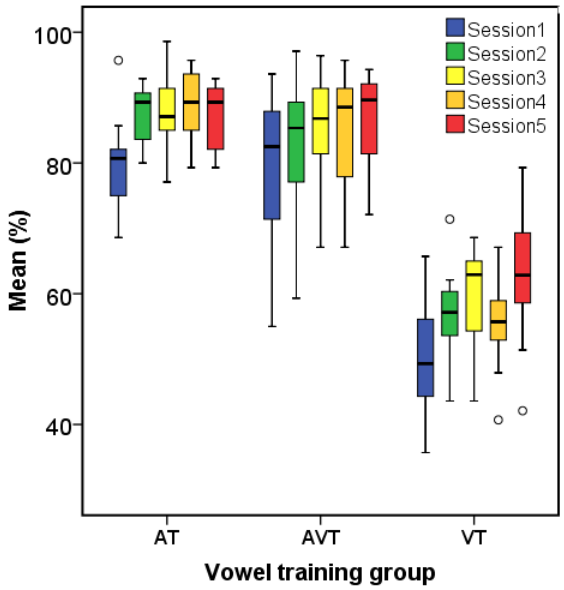


Figure 3.17 Boxplots for vowel identification accuracy (mean %) per training group (AT, AVT, VT) in the Vowel Trainer sessions (Sessions 1 to 5). Outliers are marked with a small circle.

To explore what happened in the three different training modalities (AT, AVT, VT), the training data will be presented here. Due to the adaptability nature of the training software, mean scores per vowel per session were computed from the first 70 (phase 1) and last 70

responses (phase 3), 10 total repetitions per vowel out of the total 225 trials per session. These 140 trials were the fixed items in the adaptive training procedure. A logistic regression analysis was run on the data with training group (3), session (5), vowel (14), group*session, group*vowel and group*session*vowel as fixed effects and participants and stimulus as random effects. The results (Table 3.34) showed a significant effect of Training group (Fig. 3.17): AT and AVT groups did not differ in their overall means, but the VT group had significantly lower scores (Table 3.35). There was a session effect, mean scores were generally different across sessions and ranged from 49.6% to 89% increasing consistently, but this effect was modified by a group*session interaction. The interaction was caused by a significant between-group difference in mean scores per session, with similar higher scores for AT and AVT groups across sessions and significantly lower means by the VT group (Table 3.34).

Table 3.34 Fixed effects and interactions in the results from the Vowel Trainer by the three training groups (AT, AVT & VT).

Fixed effects	
Group	F(2,44)= 53.099, p<.001
Session	F(4,26312)= 60.796, p<.001
Vowel	F(13,6334)= 75.086, p<.001
Group*session	F(8,26312)= 6.111, p<.001
Group*vowel	F(26,6334)= 14.636, p<.001
Vowel*session	F(52,26312)= 8.017, p<.001
Group*session*vowel	F(104,26312)=3.481, p<.001

The vowel effect was also significant, mean scores varied from 63% for vowel /ʌ/ to 98% for /au/. There were also significant interactions for group*vowel, vowel*session and group*session*vowel. The three-way interaction modifies all the two-way interactions and it was caused by some vowels having the same mean per session per groups. For example, the AT group's scores for vowel /u:/ had the same overall mean (98%) in sessions 2, 3 and 4; the AVT group had the same scores for vowel /i:/ (88%) in sessions 2 and 3. The VT group had the same scores for vowel /æ/ (63%) in sessions 4 and 5.

Table 3.35 Vowel identification in the Vowel Training sessions: overall means (M) and Standard Deviation (SD) per session and training group; session*group interaction and group effect.

Session	A Training Group (M)	SD	AV Training Group (M)	SD	V Training Group (M)	SD	Session*group interaction
1	80%	7.4	79%	12.8	49.6%	10.8	F(2,44)= 40.249, p<.001
2	87.6%	10.2	83.2%	12.7	55.3%	10.4	F(2,44)= 38.393, p<.001
3	88.4%	12.5	85.4%	8.4	60%	8.4	F(2,44)= 50.300, p<.001
4	89%	7.1	85.6%	8.6	55.3%	7.7	F(2,44)= 61.290, p<.001
5	87%	5.6	86.2%	7.6	63%	10	F(2,44)= 20.140, p<.001
Mean	**86.4%**	1.78	**84%**	1.96	**56.7%**	1.8	
Training Group Effect	A training vs. AV training group, F(1,29)= 0.415, p>.05 A training vs. V training group, F(1,29)= 160.401, p<.001 AV training vs. V training group, F(1,29)= 57.169, p<.001						

Though no significant mean difference was found between the AT and AVT groups, the boxplots in Figure 3.16 show the amount of variability among groups for each session and the learning tendency from the first session to the fifth. Generally speaking, for the AT and AVT group not much improvement was found after the second session as scores remained very similar from session 2 to session 5. The V training group showed an increase from session 1 to 3, scores go down in session 4 and rise again in session 5. This decline in scores in session 4 might have been caused by a talker's effect (each session had a different talker). There was a mild pursing of the lips in that particular talker in session 4 which may have interfered with participants' visual vowel perception for this group but did not affect the other training groups.

3.3.10.1 Individual differences

Figures 3.18, 3.19 and 3.20 illustrate the learning curve for each participant per session and separate per training groups. A range of individual differences can be observed in the scores per session as illustrated in the lines (means) going up or down from session to session. Improvement between sessions 1 to 5 was observed for most of the participants but the individual variability suggested that there were differences in vowel perception based on the talker and also on the perceivers.

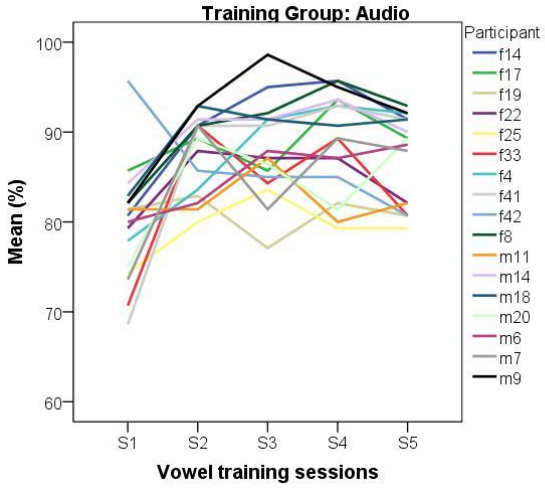


Figure 3.18 Individual performance in the Vowel Training sessions for the AT group. Mean correct identification scores per session and participant.

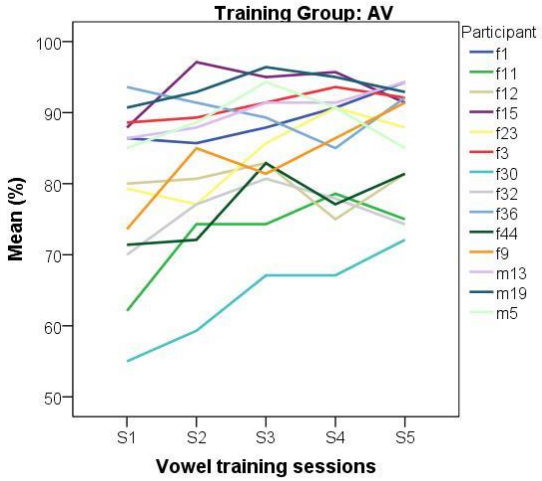


Figure 3.19 Individual performance in the Vowel Training sessions for the AVT group. Mean correct identification scores per session and participant.

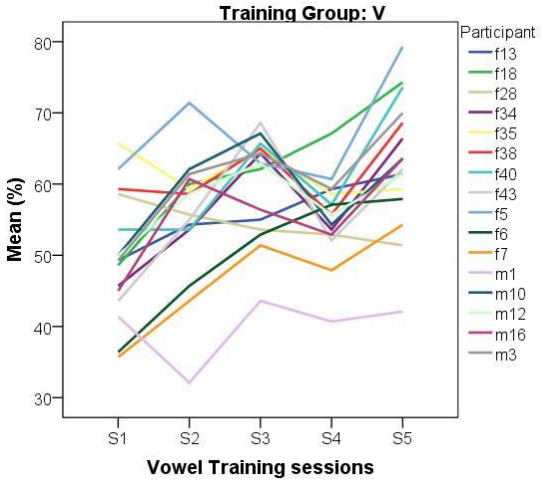


Figure 3.20 Individual performance in the Vowel Training sessions for the VT group. Mean correct identification scores per session and participant.

3.3.10.2 Improvement

The per cent change per Training group was calculated considering the first 70 and last 70 responses (10 total repetitions per vowel). Those were the fixed items in the adaptive training sessions. Mean scores for Session 1 and Session 5 were used to establish the amount of improvement per participant [((Session5 - Session1)/Session1)*100)].The overall improvement found ranged from -15% to 59%, with only 6 participants showing no improvement and all the remaining 41 scoring some amount of improvement. The AT

group obtained an average of 8.7% of change, the AVT group scored 11.2% and the VT group scored 29.4% of change (Table 3.36). This large improvement in the VT group suggested that once participants were able to learn about the training procedure and were familiarised with the tokens, they were able to make quick improvement. It may be the case that because they had low scores in their first session, there was more room for improvement as well.

Table 3.36 Percent change (improvement) relative to session 1 for the Vowel Training and improvement relative to the pre test for the Vowel test per mode (A, AV, V).

Trainig Groups	Vowel Training % change	Vowel test: A % change	Vowel test: AV % change	Vowel test: V % change
A_Traing group	8.7% (10.6)	14.2% (16.7)	16% (15)	9.2% (14)
AV_Training group	11.2% (9.5)	17.5% (15)	15.8% (18.8)	4.4% (13.7)
V_Training group	29.4% (21.6)	12.6% (9.8)	10.8% (10.8)	5.8% (12.8)

3.3.10.3 Correlations for amount of improvement

In order to find whether there was any relation between the amount of change in the Vowel training sessions and the improvement in the perception of English vowels in the Vowel test, correlations were run separate per training group. No correlation was found for any of the three training groups for the improvement in the Vowel training sessions and their improvement in their vowel identification accuracy in the Vowel test used as pre and post test (Table 3.37).

Table 3.37 Correlations for amount of improvement in the Vowel training sessions and the Vowel test (vowel identification accuracy, pre & post training measure).

Training Group	Audio, Vowel Test	AV, Vowel Test	Video, Vowel Test
A-Training Group (N=17)	.420, p>.05	-.061, p>.05	.241, p>.05
AV-Training Group (N=14)	.050, p>.05	-.416, p>.05	.207, p>.05
V-Training Group (N=16)	.070, p>.05	.439, p>.05	-.055, p>.05

In conclusion, the three training groups improved their vowel identification accuracy during training from session 1 to session 5, regardless of the training modality they had during their Vowel training programme. The AT and AVT groups did not differ significantly in the amount of improvement from their first to the last session, whereas the V training group showed more improvement but with lower overall results. Individual differences were also

found but no relation between the participants' vowel training improvement and the initial and final vowel identification ability (Vowel test: pre & post test) was found. This is the first time that this vowel training programme has been used to compare these three modalities, as it was originally designed for auditory training only.

3.4 Discussion

The aim of this study was a) to find out whether there was a training modality that could be more beneficial for L2 learners to improve their vowel identification accuracy. Another aspect that was examined in this study was b) the possible sources of individual differences in vowel identification. Additionally, c) the extent to which perceptual training using isolated words may generalise to vowel identification at sentence-level material was also explored.

Three high-variability (HV) vowel training modalities were given to different participants in this study: an auditory training (AT), audio-visual training (AVT) and video-alone training (VT) mode. Each group took five sessions of one vowel training modality. The results of the training sessions showed that the A and AV training modalities had a similar impact on the L2 learners. That is to say, they achieved the same overall results by the end of the fifth session and showed a very similar pattern of learning which revealed quick improvement in the first three sessions and no greater improvement after that. Both training groups improved by about 10% from their initial scores. This amount of learning was in line with previous HV vowel training studies that have reported between 10% and 25% of improvement (Iverson & Evans, 2007; Iverson & Evans, 2009; Lengeris & Hazan, 2010; Nishi & Kewley-Port, 2007; Wang & Munro, 2004).

The V training group showed a greater amount of learning after training in spite of not having access to any auditory input, but their learning curve looked similar to those of the other two training groups. The 29% of relative improvement came to a surprise as there was no previous data on this type of training with L2 learners. This high percentage of

improvement in the V mode during perceptual training showed that the visual gestures for vowels contain a lot of information that contributed to improve their vowel identification in the same training modality and may also be the result of having more room for learning, as their performance in the first training session was the lowest of the three groups. These results may also suggest that learners' were able to perceive visually duration differences for the tense-lax distinction and became more efficient at matching visual cues to the written forms of the stimuli, as mentioned in the discussion section in Chapter 2. However, they failed to integrate this information in the AV mode for better vowel identification compared to A mode.

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 was the first time that three different types of HV vowel training modalities (Auditory, Audio-visual & Video-alone) were used. 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 (Vowel test). The results in vowel identification accuracy in AV mode after training did not show any advantage for the two groups that had access to visual information during their vowel 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.

To assess the improvement in vowel identification accuracy due to training, a Vowel test was used as pre and post test. This test measured participants' capacity to identify 11 English vowels in A, AV and V presentation mode. It was expected that some difference would be found due to training modalities if the material used had allowed trainees to focus their attention on specific cues used in their training mode and had promoted stronger

learning in either audio, audio-visual or video-alone contexts. The results in this test revealed that training modality did not make a difference in the amount of learning after training, as all groups improved their vowel identification accuracy with similar results. The expected attention focus to different cues promoted by the A, AV and V training modalities did not result in better vowel identification in the trained contexts. That is to say, learners trained with auditory, audio-visual or visual-only cues did not show an advantage in vowel identification for their specific training modality.

Together with testing overall vowel identification accuracy after training, the Vowel test allowed us to measure vowel perception not only in an auditory context but also in AV and V mode. L2 speech perception studies comparing A and AV modalities have found some advantage when the stimulus was presented in AV mode, suggesting L2 learners can improve their perception when they had access to visual information for visually salient contrasts (Hardison, 2003, 2005; Hazan et al., 2006). Contrary to what was expected, participants in the current study did not show any visual benefit in AV mode before or after training; similar results were found in A and AV modes, suggesting that participants were only using the auditory cues for their vowel identification, even by those participants who were trained in AV or V mode. These two groups in particular did not show better performance in AV or V mode than the A training group. This raises the question of what is being trained with these three types of perceptual training programmes for L2 learners.

It could be assumed that A training would improve auditory perception because learners were focusing their attention on auditory cues for speech. However, auditory training was not the only kind of training that promoted better speech perception in auditory mode. Learners trained with visual cues only (VT group) were also able to improve their auditory perception of vowels at the same level of the other groups (AT, AVT) who did have access to auditory cues. This would suggest that the training with just the gestures for the vowel phonemes was as strong an input as the audio itself and contributes as much as the traditional auditory perceptual training to improve English vowel perception in L2 learners. This finding may be related to an "Analysis-by-synthesis model" for speech perception (Stevens & Halle, 1967; Stevens, 1972) and to more recent findings on AV speech

perception in neuroscience research (Skipper, van Wassenhove, Nusbaum & Steven, 2007; Van Wassenhove, Grant & Poeppel, 2005) which claims that visual speech information enables the prediction of the auditory input.

In the Analysis-by-synthesis model (Stevens & Halle, 1967; Stevens, 1972), the speech signal (auditory input) first goes through a peripheral auditory analysis to proceed to the master control unit where a hypothesised representation is generated. This hypothesis is knowledge-based, interprets the speaker's "intended speech gestures". Motor commands are activated based on the hypothesised representation but remain inhibited during perception. The commands produce a hypothetical auditory pattern which is passed to a Comparator module where the pattern is matched to the original input and kept in a temporary store. Regarding this model, van Wassenhove et al., (2005) have suggested that it could be extended to audio-visual (AV) perception by specifying that the sensory input is made up of sound and observed facial gestures and so should be the hypotheses generated in the control unit. The motor commands would predict an auditory and somatosensory pattern instead of an auditory pattern alone (Skipper et al., 2007). These models use a loop as a way of feedback to check if the matching of the hypotheses and internal representation has been successful, or else the process needs adjustment.

Though the models mentioned above do not account for the video-alone perception of speech, they could be used to explain the improvement in auditory perception by the Video training (VT) group. It could be hypothesised that when the auditory signal is absent from the input, the process follows the same steps as in the Analysis-by-synthesis (AbS) and van Wassenhove et al. (2005) model, only based on the visual input. The visual gestures, in the form of a hypothesis, are sent to the control unit where it is mapped onto motor commands and a richer representation is generated (visual gestures and auditory representation). A new hypothesis based on experience with speech is generated (i.e. to which sound these gestured can best match). Then, the decision in the form of a hypothesis or prediction is compared with the original input to check for the best match. The process may continue if the prediction does not match the input. Altogether, learners may have been using the process of matching the visual input to gesture commands and auditory representations of

English vowels during training. The feedback received during training may also have contributed to make the matching process more efficient.

The fact that different degrees of visual cues were found to be available for English Native speakers in a previous study (Chapter 2) suggested that L2 learners could improve their English vowel perception if they could make use of these visual information. However, the learners' lack of integration of visual cues for English vowel perception found in the current 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 related to their experience in using visual cues in their L1 (Hazan et al., 2006; Wang at al., 2008). Due to a smaller L1 vowel system with less ambiguity in its five vowels (/i/, /e/, /a/, /o/, u/), the existing visual cues may become secondary or even redundant. When perceiving vowels in an L2 they may focus on the channel that provides more information in the L1 for vowel perception, the audio channel. Information as lip-rounding or tenseness for L1 vowels has less weight as it would make no difference in vowel category or in meaning in Spanish. For example, a word like "silencio" (silence) could be pronounced with a length change or lip rounding for a specific Spanish vowel and sound like /sile:nsio/ but making /e/ longer does not change the meaning of the word. It would only be interpreted as emphasising the idea of silence. Thus, the /e/ sound remains the same in its essential interpretation.

It has been suggested that L2 learners may learn to attend to visual cues as they gain experience with the language (Wang et al., 2008). The results in the Vowel test for the L2 advanced group (i.e. more experienced) did not show any evidence of greater integration of visual information for vowel perception than the beginner group. As discussed in Chapter 2, the advantage that Wang and colleagues found in the more advanced learners (Mandarin-Chinese speakers living in Canada) may be related to the visual salience of the material tested (/f/–/v/, /s/–/z/, /θ/–/ð/) in three vowel contexts (/i/, /a/, /u/). A small number of visually salient contrasts, together with less variability (one speaker) may have contributed to the visual benefit found. Another important issue to bear in mind is that in

the Wang et al. study, participants were currently living in Canada at the time of testing; therefore, they had access to interact with native speakers of English. The Chilean participants would rarely interact with a native speaker on regular basis; thus they have considerable less access to natural visible speech. This difference may suggest that having access to face to face interactions with native speakers may have contributed to the AV advantage found in the more advanced learners in the study by Wang et al. (2008). However, it could be the case that when more contrasts and larger variability were added, less AV benefit would be found as in the current vowel study.

Individual differences are commonly reported in perceptual studies and the sources for these differences seem varied. Some studies have linked individual variability to the type of training paradigm, larger or smaller amount of stimuli (Perrachione, Lee, Ha & Wong, 2011), amount of cognitive resources engaged for a perceptual task (Diaz, Mitterer, Broersma & Sebastián-Gallés, 2012), learners' phonological short-term memory (MacKay, Meador & Flege, 2001) and auditory processing capacity (Lengeris & Hazan, 2010). To account for individual variability in the current study, visual and auditory acuity measures and a frequency discrimination test were used. Also, a sentence-repetition test (BKB-sentence test) and English level of proficiency were used to run correlations with the learners' vowel identification accuracy before and after training. The auditory and acuity measures did not seem to be related to the learners' vowel identification capacity before or after training. The only measure that showed some correlation with English vowel perception before and after training was the BKB-sentence test which may have to do with the participants' level of proficiency. The sentences used very simple vocabulary which may not have posed high demand on learners for repetition (test aim). The English level of proficiency did show more significant relation with English vowel perception before and after training, though this relation was weaker after training but still significant. The benefit of the vowel training sessions was found to benefit weaker learners more than more proficient learners in general. This may be related to having more room for learning but it could also have been aided by motivational factors.

Most of the L2 training studies report successful learning with retained changes even months after training (Bradlow, 1999; Iverson & Evans, 2009; Lively et al., 1994; Nishi & Kewley-Port, 2007) and suggest that the learning promoted by L2 perceptual training leads to the establishment of new phonemic categories. This way, the new L2 phonemes are incorporated into the learners' perceptual space. However, there is a different approach that suggests that through training learners do not establish new categories but only become better at using their existing L2 phonemic system (Iverson & Evans, 2009). In this view, learners use strategies to perceive the L2 sounds based on how similar or different the L2 sounds are from the nearest L1 phoneme. In the current study, learners were able to improve their perception of English vowels after training at word level and this learning generalised to new tokens and new speakers. However, when tested on vowel contrasts in minimal-pair words embedded in sentence material, no transfer of improvement to judge when a sentence was using the wrong minimal-pair word was found. This finding poses the question of whether learners were able to establish new categories for English vowels or not, as their strategy for the perception of vowel contrasts failed in a more naturalistic context. It is also possible that having to attend to a short sentence in search for the "wrong word" made learners' working memory overload and this made them fail the task. But again, this brings the question of what kind of training would be needed to improve vowel perception in a more naturalistic context which resembles the reality learners' face in having to perceive vowels in longer chunks of language than just a word.

Chapter 4

Production of English vowels by L2 learners

Research on L2 speech has shown that the link between perception and production is not fully understood yet. In general terms, L2 learners seem to rely on duration more than on spectral information for the perception of English vowel contrasts (Cebrian, 2006; Flege, Bohn & Jang, 1997; Kondaurova & Francis, 2008), though attention to spectral and temporal cues has also been found (Iverson & Evans, 2009). With regards to English vowel production, it has been found that learners also tend to over-rely on duration to produce the tense-lax distinction (Cebrian, 2007; Chen, 2006; Mora & Fullana, 2007; Rallo-Fabra & Romero, 2012), though some exceptions have been reported (Bent, Bradlow & Smith, 2008). It has also been reported that L2 learners usually fail to implement the spectral characteristics of English vowels in their production (Flege, Bohn & Jang, 1997; Mora & Fullana, 2007). However, it has been shown that perception and production of vowel contrasts can be improved with perceptual training.

In general, perceptual training studies have reported improvement in both perception and production after training, in the absence of production training for consonants and vowels (Bradlow, Akahane-Yamada, Pisoni & Tohkura, 1999; Bradlow, Pisoni, Akahane-Yamada &Tohkura, 1997; Hazan et al., 2005; Iverson et al., 2012; Lambacher et al., 2005; Lengeris & Hazan, 2010; Thomson, 2011). Vowel production improvement after training has been found to vary as a factor of experience with the language (Flege, Bohn & Jang, 1997), learner's attention capacity (Kondauruva & Francis, 2010), learners' capacity to

discriminate differences in auditory frequencies (Lengeris & Hazan, 2010) and learners' capacity to change cue weighting (Lambacher, 2005), among others.

The improvement in production after perceptual training has often been suggested as the consequence of changes in the perceptual space, possibly caused by the modification or creation of new perceptual categories. Problems in production would reflect poor perception of the contrasts (Flege, 1995). It may also be the case that perceptual changes are implemented faster than the "motor commands" that guide production; thus, perception would lead production (Bradlow et al., 1997). However, production training with improvement in the production but not in the perception of the contrasts trained revealed that this direct link between perception and production is not so straight forward (Hattori & Iverson, 2009; 2010). Another aspect that adds complexity to the perception-production link is the amount of individual difference found in the perception and production of L2 contrasts which seems to obscure the understanding of the relation between perception and production.

To explore the effect of L2 vowel perceptual training on the production of English vowels, three types of data were collected. First, the English vowels produced by L2 learners before and after training were judged by English native speakers (ENS) in a goodness-rating test. It was hypothesised that if high-variability perceptual training had had a positive effect on learners' production, the productions recorded after the training period would be rated higher by ENS. Next, if learners had learnt to focus on the right spectral features for English vowels, this would be reflected in spectral changes in the production of English vowels after training. Spectral difference is the primary cue for English vowel perception for ENS, though tense vowels tend to be relatively longer than lax vowels in English (Ladefoged & Maddieson, 1996). Finally, the duration of vowels at pre and post test was compared to find whether these learners followed a duration strategy to differentiate the tense-lax vowels. An overreliance on duration has been suggested as the most accessible strategy for L2 learners to perceptually distinguish tense-lax vowels regardless of the status of duration in their L1 (Bohn, 1995; Cebrian, 2006; Flege et al, 1997; García-Lecumberri & Cenoz, 1997; Wang & Munro, 1999). If this strategy was transferred from perception to

production, it may be expected that duration would be used as a primary cue to make the difference between English tense-lax vowels after training.

4.1 Aims

The research questions that guided this study were:

a) What acoustic-phonetic measures differentiate English tense-lax vowels produced by L2 learners?

b) To what extent does L2 vowel perception improvement transfer to vowel production?

- If improvement is found in the production of English vowels after training, what dimension of vowel production is changed?

c) Was there any difference in the impact of vowel training modality on the production of English vowels after training?

d) Is there a direct relation between the amount of perceptual and production improvement after training?

4.2 Method

4.2.1 Participants

47 L2 learners with beginner level of English who had completed an English vowel training programme. These were the same participants as described in Chapter 3 (3.1.1).

Eleven English native speakers (ENS) with Southern British English accent were recruited as raters to judge the English vowels recorded by the L2 learners group. The ENS were

postgraduate students at University College London who had taken either English phonetics or speech science modules. The aim was to recruit raters who were familiar with the classification of English vowels. The raters' age ranged from 21 to 40 years (mean: 26.7 years) and there were nine females and two males. They were all right handed and self-reported normal hearing. They were paid a small contribution for their participation.

4.2.2 English vowel recordings

L2 learners recorded 33 sentences before and after training (pre & post test). 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 word in the box is *book*"). 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]. This selection of words was made based on a second set of recordings of a picture description task which was collected with the idea of comparing isolated words and less controlled production of the same words which appeared in the pictures. Due to time restrictions, the latter recordings were not analysed. As a consequence, vowels obtained from the recordings had different consonant environments which need to be taken into consideration to interpret the results with caution.

The recordings were made individually in a quiet room in the Spanish Phonetics Laboratory at Universidad de Concepcion. 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 were used for the recordings which were made at a sampling rate of 44.1 kHz and recorded in mono.

The vowel in each keyword was tagged using the Praat software version 5.2.17 (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. Some words had been mispronounced by a few participants due to strong orthographic interference. When that was the case, the token was not included in the sample but was replaced by a repetition of the target vowel from one of the other words. Examples of excluded tokens: "cut" if pronounced as /kʊt/ or "word" as /wɒrd/. This decision was made to avoid distortions in the acoustic measures and avoid bias when submitted to goodness rating judgement by English native speakers.

The extracted vowels were used as stimuli for the goodness rating task and to obtain spectral (Formant 1, Formant 2) and duration measures. The decision of using only the extracted vowel instead of the keyword as stimuli for the goodness rating test was made to avoid listeners (raters) being influenced by the mispronunciation of the preceding or following consonants. The L2 learners were bound to produce foreign accented consonants which could have made the judgement of the vowel-alone more difficult to English native speakers.

4.2.3 Goodness rating test

To obtain a measure of the accuracy of the English vowels produced by the L2 learners, vowels produced by the learners before and after (pre and post test) taking the five vowel-training sessions were given to ENS in a goodness rating test. Four tokens per vowel (vowels extracted from keywords) from each participant were selected: two tokens from the pre test and two from the post test recordings. Tokens were selected so as to have the same two tokens (same vowel extracted from the same keywords) from the three available per vowel from the pre and post tests for all participants. The tokens chosen were [cap, flag], [cut, luck], [park, part], [peach, beach], [tin, bin], [shoes, food], [book, foot], [surf, girl], [net, shell], [sock, dog] and [shorts, shore].

A total of 2068 tokens (2 x 11 vowels x 47 participants x 2 times) were obtained which were randomised to create three different tests. The three tests were given to the raters (ENS) in three separate sessions of approximately 35 minutes each. All raters heard all 2068 tokens. The vowel tokens were fully randomised across the three sessions. The goodness rating test was given individually to ENS using a laptop and headphones, using an MFC Praat experiment format. Listeners were asked to rate 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 listeners had the option to hear the sound again if needed (Fig. 4.1). A break option was introduced every 10 minutes. The presentation was blocked per vowel.

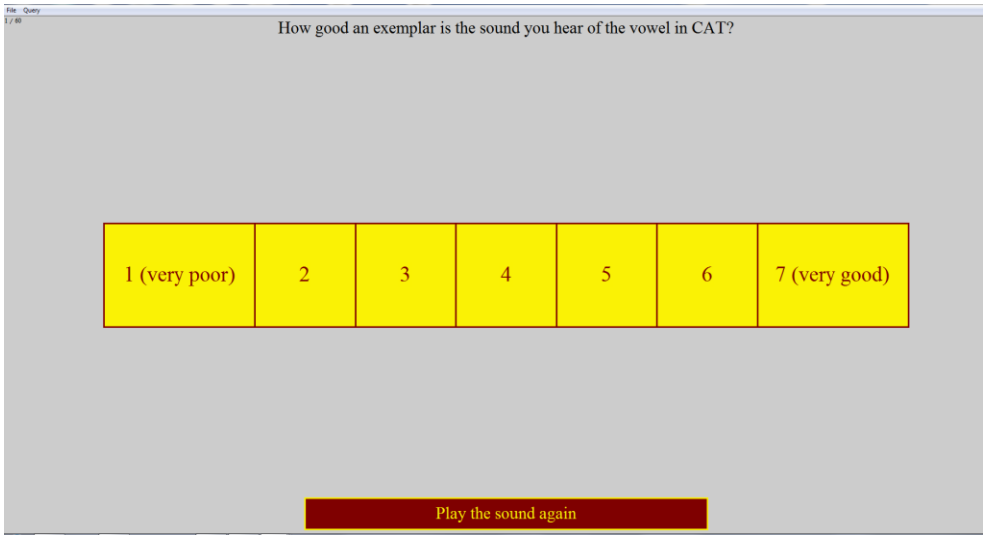


Figure 4.1. Shows a screen of the goodness rating test, exactly as raters saw it. The prompt sentence appeared at the top, the rating scale in the middle and the "*play the sound again*" option at the bottom of the screen.

4.2.4 Vowel measures

Three tokens per vowel from the pre test and three tokens from the post test were used for the formant measures (F1, F2) and duration measures. Token were vowels extracted from keywords as described in 4.2.2.

a. Vowel formant measures

The vowels extracted from the keywords in the read sentences were processed to change boundaries at the nearest zero-crossing, check sample frequency (44.1 Hz) and number of channels (mono), before formant frequencies were measured. Formant values for F1 and F2 were measured at the mid-point of the token specifying minimum and maximum reference values for male and female recordings. All this was done with different scripts in Praat (version 5.2.17).

Raw values for vowel formants (F1, F2) were checked to spot if values fell outside the normal range. If so, files were individually checked and manually corrected. After that measures for F1 and F2 were normalised, separately for male and female data, using "The Vowel Normalisation and Plotting Suite" website (Thomas & Kendall, 2007). The vowel normalization method used was the Lobanov method. This method was chosen as it has been found to be the best to factor out physiological differences (Adank, Smits & Van Hout, 2004) within-participant groups.

b. Vowel duration measures

Vowel duration was measured from the start to the end point of the vowel (extracted from the key-words) using a script in the Praat software. Values were obtained in milliseconds. The values were obtained in the process described in "a)" for formant measures.

4.3 Results

4.3.1 Vowel goodness rating test

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) and vowel (/iː/, /ɪ/, /e/, /ɜː/, /æ/, /ʌ/, /ɑː/, /ɔː/, /ʊ/, /uː/, /ʊ/) and training group*vowel, training group*time and training group*vowel*time interactions; participant was treated as a random effect. Table 4.1 shows all the F values.

The analysis revealed no significant effect of training group: overall means per group did not differed. There was an effect of time (Fig. 4.2), 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 4.2).

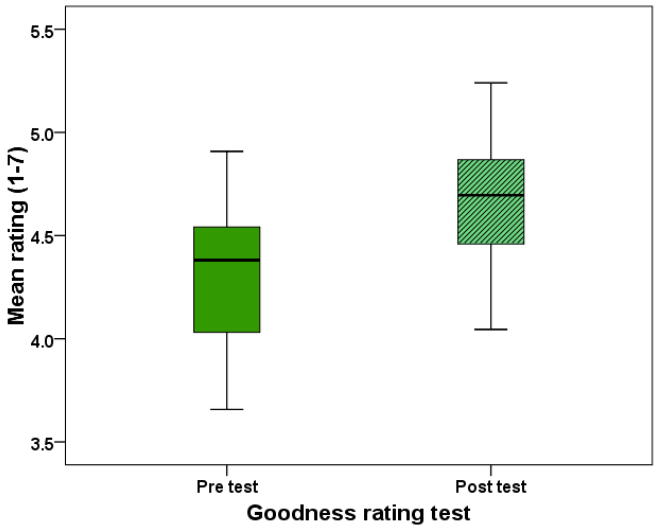


Figure 4.2. Boxplots for the overall means in the goodness rating test before (pre test) and after training (post test). ENS rated L2 learners' vowel production with a scale from 1 (poor) to 7 (very good).

Table 4.1 Results for the vowel goodness rating test. L2 learners' vowel production from the pre and post test recordings from 11 English monophthongs were assessed by ENS. 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 4.2 Mean rating (M) and standard deviation (SD) for pre and post test vowel production per vowel (11) and level of significance for the vowel*time interaction (time effect).

Vowel (Pre & Post M, SD)	Time effect
/ɑ:/ (M: 4.7, SD: .69; M: 5.4, SD: .67)	F(1,46)= 44.903, p<.001
/æ/ (M: 3.9 , SD: .94; M: 4.5 , SD: .82)	F(1,46,)= 13.780, p<.001
/ʌ/ (M: 4.1, SD: .69; M: 4.2, SD: .72)	F(1,46,)= 2.445, p>.05
/e/ (M: 3.7, SD: .68; M: 4.1 , SD: .87)	F(1,46,)= 9.824, p<.05
/ɜ:/ (M: 3.8, SD: .57; M: 4.2, SD: .60)	F(1,46,)= 18.299, p<.001
/ɪ/ (M: 4.7, SD: .69; M: 4.7, SD: .82)	F(1,46,)= .001, p>.05
/i:/ (M: 4.6, SD: .73; M: 4.9, SD: .76)	F(1,46,)= 5.285, p<.05
/ɔ:/ (M: 4.3, SD: .94; M: 5.0, SD: .91)	F(1,46,)= 23.243, p<.001
/ɒ/ (M: 4.2, SD: .66; M: 4.6, SD: .73)	F(1,46,)= 14.740, p<.001
/ʊ/ (M: 4.5, SD: .57; M: 4.6, SD: .62)	F(1,46,)= 1.051, p>.05
/u:/ (M: 4.5, SD: .66; M: 4.8, SD: .52)	F(1,46,)= 6.596, p<.05

To explore whether the 11 raters had been consistent in their ratings for the L2 learners' vowel production, 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 consistency in the raters with a Cronbach's Alpha α=.844. The maximum value for the Cronbach's Alpha is 1, and values from .70 and above are considered good indicators of consistency. These results confirm that there was very good amount of consistency and absolute agreement on the scores within and between raters.

Together, the results of the goodness ratings test showed that L2 learners had improved the quality of their English vowel production after training in a way that was perceptible by native listeners. To explore the source of this improvement, acoustic and duration measures were conducted and analysed.

4.3.2 Vowel spectral measures and Euclidean distance for vowel contrasts

To investigate whether L2 learners' production of English vowels experienced any significant spectral change after training, the Euclidean distance (ED) between seven contrastive pairs (/i:–ɪ/, /ɜ:–e/, /æ–ʌ/, /ɑ:–æ/, /ɑ:–ʌ/, /ɔ:–ɒ/, /u:–ʊ/) was estimated

at pre and post test. The pairs were chosen based on the frequent perceptual confusions for tense-lax vowels and the /æ-ʌ/ pair by learners of English with Spanish L1. The mean F1 and F2 values per vowel and participant were used to calculate the ED between contrastive pairs at pre and post test. The ED is equivalent to the square root of the difference of the F1 and F2 values of two vowels [SQRT(V1_F1-V2_F1)^2 + (V1_F2-V2_F2)^2]. Figures 4.3, 4.4 and 4.5 show boxplots of the ED for seven contrastive pairs. The /æ-ʌ/ contrast was included to the tense-lax vowel pairs due to its "highly-confusable" status for Spanish-L1 learners of English.

In Figure 4.3, the ED for the vowel pairs /æ-ʌ/, /ɑ:-ʌ/and /ɑ:-æ/ at pre and post test is shown. L2 beginner learners were expected to have problems in producing the distinction between these vowel pairs as these English vowels are likely to all be assimilated to the Spanish /a/. However, L2 learners showed different degrees of spectral differences between vowels at pre and post test, larger when contrasting vowels with /ɑ:/ and smaller between /æ-ʌ/. Thus, learners had some idea of the spectral difference between these three English vowels even before training.

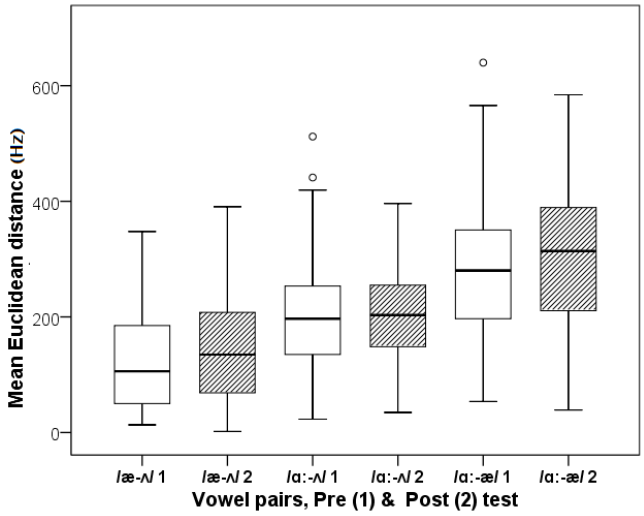


Figure 4.3 Euclidean distance (Hz) for vowel pairs /æ-ʌ/, /ɑ:-ʌ/ and /ɑ:-æ/ produced at pre and post test.

The ED between vowel pairs /iː–ɪ/ and /ɜː–e/ at pre and post test are shown on Fig. 4.4. Because Spanish does not have a tense lax distinction for its vowels, it was predicted that beginner learners would make no spectral distinction between these pairs at pre test. However, Fig. 3 shows that learners made some distinction for the two pairs with a wide range of individual differences and less distinction for the /iː–ɪ/ contrast.

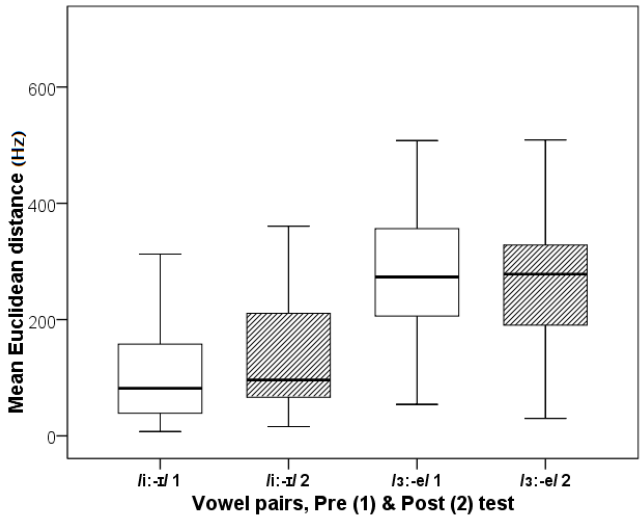


Figure 4.4 Euclidean distance for vowel pairs /iː–ɪ/ and /ɜː–e/ produced at pre and post test.

Figure 4.5 shows boxplots for the ED for the pairs /ɔː–ɒ/ and /uː–ʊ/ at pre and post test. As observed in the previous sets of vowels, some distinction was found between the pairs before and after training. It seems that the distinction between /ɔː–ɒ/ was larger than /uː–ʊ/. Individual difference was also found for these pairs.

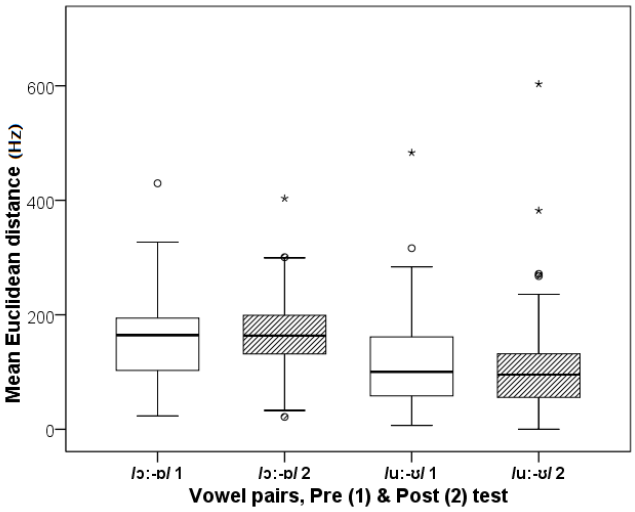


Figure 4.5 Euclidean distance for vowel pairs /ɔ:–ɒ/ and /u:–ʊ/ produced at pre and post test.

Once the spectral distance between vowel pairs was established, these Euclidian distance (ED) values were submitted to a statistical analysis to find if any spectral change had occurred following the vowel training programme. A mixed model analysis was run in R (lme function); the fixed effects were training group (AT, AVT, VT), time (pre, post), vowel pair (7), training group*time, training group*vowel pair, training group*vowel pair*time. Participants were the random factor. The results (Table 4.3) showed there was no effect of training group; results were generally similar across groups. The time effect was not significant; the overall ED between the pairs did not become significantly different after training. The effect of vowel pair was significant; the overall ED between the pairs ranged from overall ED mean (EDM) of 120 Hz for /i:–ɪ/ and /u:–ʊ/ pairs to EDM: 296 Hz for the /ɑ:–æ/ pair.

Table 4.3 Main effects and interactions for the Euclidean distance between seven contrastive pairs of English vowels.

Effect	
Training Group (TGroup)	F(2,44)= 2.522, p>.05
Time	F(1,568)= 1.773, p>.05
Vowel pair	F(6,568)= 63.440, p>.05
TGroup*time	F(2,568)= 1.110, p>.05
Vowel pair*TGroup	F(12,568)= 2.042, p<.05
Vowel pair*time	F(6,568)= 1.056, p>.05
Vowel pair*TGroup*time	F(12,568)= 1.397, p>.05

There was a vowel pair*group interaction which was explored with the same linear mixed-model (lme function) analysis. There was only one significant difference between groups for the contrast /iː–ɪ/; the Video training (VT) group had significantly lower scores than the Audio training (AT) group while the A and Audio-visual training (AVT) groups did not differ for this vowel pair.

Overall, there was no significant spectral change (ED) in the way L2 learners produced the difference between a tense-lax pair after training. Yet, L2 learners were able to make some spectral difference between tense-lax vowels before and after training for the seven vowel pairs, as can be seen in Fig. 4.3, Fig. 4.4 and Fig. 4.5.

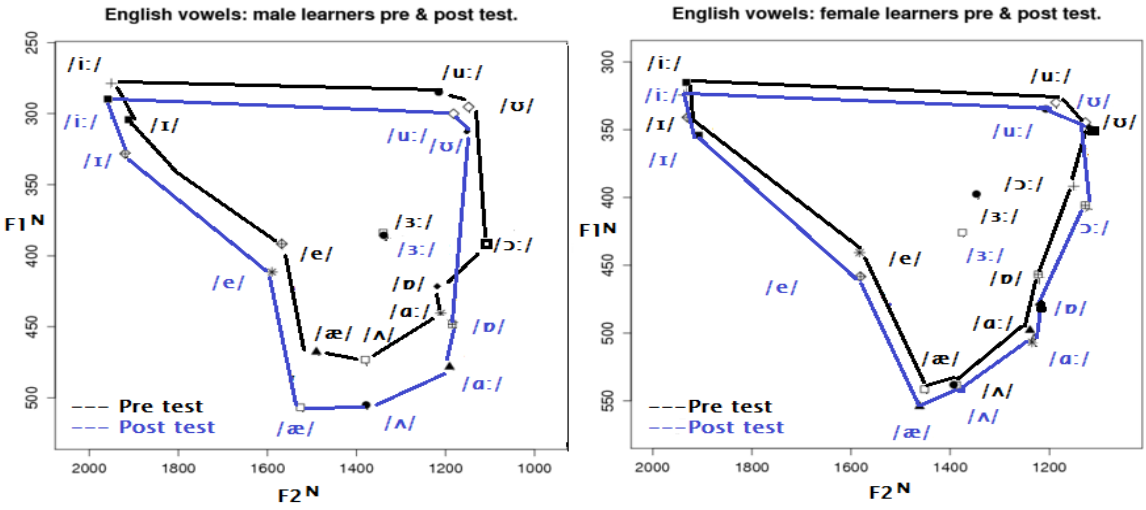


Figure 4.6 Vowel plots for English vowels produced by male and female L2 learners at the pre and post test. Mean values for English vowels (F1 & F2, normalised) at the pre and post test.

Figure 4.6 shows the normalised formant values for the 11 English vowels measured at pre and post-training for this data analysis for the male and female participants. Visually, these plots confirm the lack of spectral change that had been shown in the statistical analysis. In general, some spectral differences between contrastive pairs can be seen but the distance does not change as an effect of the perceptual training sessions.

To explore the apparent change for male speakers observed in Fig.4.6, a separate mixed analysis was run on the Eucledian distance (ED) values between pairs from pre to post (R function: lme). Time, contrasts and time*contrast were the fixed factors and participants the random effect. The time effect was not significant ($F(1,182)=3.072$, $p>.05$), the time*contrast was not significant ($F(6,182)= .960$, $p<.05$. Only the contrast effect was significant ($F(6,182)=19.389$, $p<.001$. These results indicate that the spectral differences for the contrasts did not change significantly after training.

In order to investigate whether the spectral differences between vowel pairs produced by L2 learners were similar to those found in English native speakers, data for English vowels produced by English native speakers (ENS) with Standard Southern British English accent were collected from two sources. Data from Hawkins and Midgley (2005) were used for the 5 male speakers' comparison (20-25 years old) while data from Moreiras (2006) were used for the 5 female speakers plot (Fig. 4.7). The criteria to choose these papers, in comparison with the current study, was based on their similarity in number of speakers (five), word context (hVd) and their reflection of more recent vowel change for young population (20-25 years old).

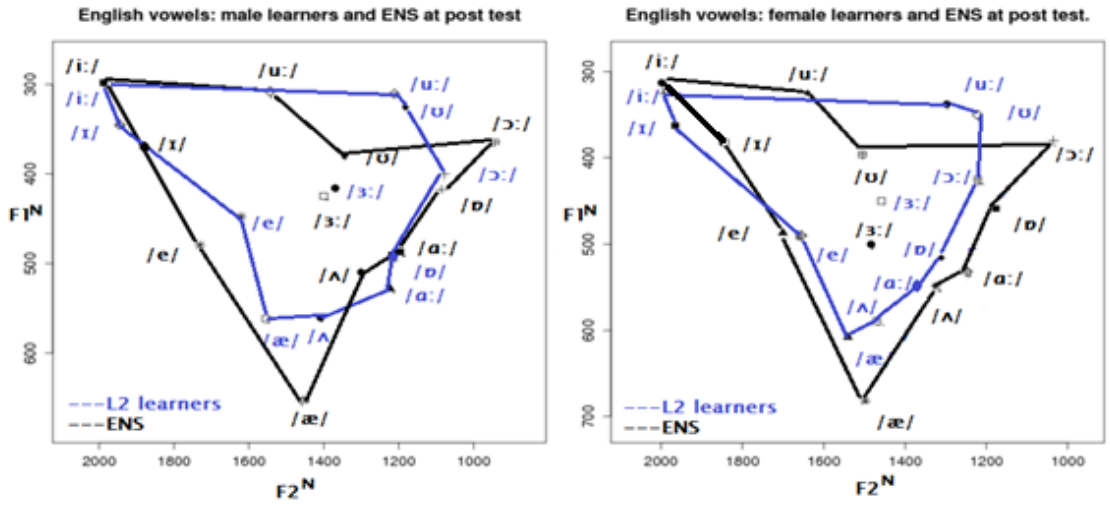


Figure 4.7 Vowel plots for English vowels produced by 5 male and 5 female ENS and L2 learners. Mean values for English vowels (F1 & F2, normalised).

The vowel plots in Fig. 4.7 show that L2 learners distribute the 11 English vowels into a smaller vowel space than the ENS. There was a general tendency for L2 learners to make less spectral difference for the tense-lax vowel pairs compared to the ENS, especially for the high front and high back vowels (/iː-ɪ/, /uː-ʊ/). However, the data used for the plots is not totally comparable due to the unequal vowel context for the L2 learners' data.

Overall, no change was found in the way spectral difference was expressed for English vowel contrasts from pre to post training. In the next section, the duration of English vowels produced by the L2 learners will be explored as another source of possible change after training.

4.3.3 Vowel duration

To find whether there was any change in the length of the English vowels produced by L2 speakers or in the difference in duration between tense and lax vowels, measures of vowel duration were obtained from the vowel tokens extracted from the recorded keywords (three tokens per vowels at pre test and three tokens at post test). These measures were submitted to a linear mixed-effect model using the R software (lme function). Training group (AT, AVT & VT), time (pre, post test), vowel (11 monophthongs), training group*time, training group*vowel, training group*vowel*time interactions were the fixed effects and participant the random factor. All results are shown in Table 4.4.

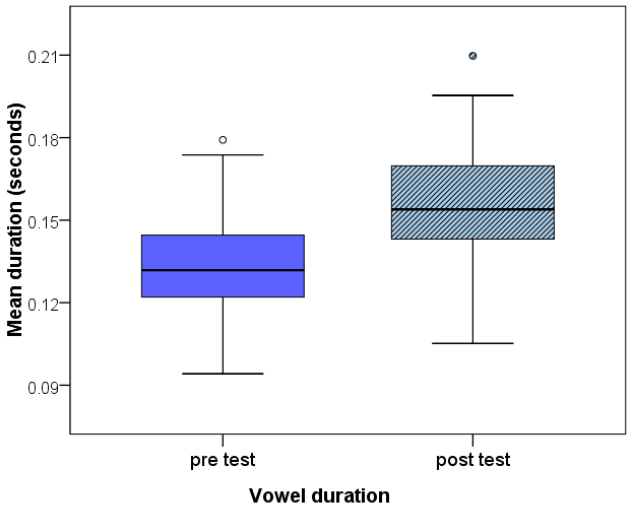


Figure 4.8 Overall mean duration for 11 English vowels produced by L2 learners at pre and post test. The overall difference was significant p<.001.

No effect of training group was found: vowel duration was similar across participants regardless of the type of vowel training they had taken. There was a significant effect of time (Fig. 4.8); the vowels produced post-training were longer than those produced pre-training (pre M: 0.135, SD: 0.019; post M: 0.155, SD: 0.020). The vowel effect was also significant but it was modified by a vowel*time interaction. The same main analysis was used to explore this two-way interaction; the results revealed that most vowels were produced with overall longer duration after training, except for /ʊ/ and /ʌ/, which showed no significant change in duration from pre to post test (Table 4.5).

Table 4.4 Fixed effects for the duration of vowels produced by L2 learners (beginners) before and after training.

Effect	
Group	F(2,44)= 0.190, p>.05
Time	F(1,922)= 140.858, p<.001
Vowel	F(10,922)= 51.704, p<.001
Time*group	F(2,922)= 2.975, p>.05
Vowel*group	F(20,922)= 1.321, p>.05
Vowel*time	F(10,922)= 5.919, p<.001
Vowel*group*time	F(20,922)= 0.541, p<.05

Table 4.5 Results for the vowel*time interaction for vowel duration (11 monophthongs) at pre and post test production. Mean (M) duration and standard deviation (SD).

Vowel	Pre test M (SD)	Post test M (SD)	Time effect
/ɑː/	0.124 (0.033)	0.137 (0.027)	F(1,46)= 21.547, p<.001
/æ/	0.138 (0.027)	0.159 (0.041)	F(1,46)= 42.910, p<.001
/ʌ/	0.121 (0.020)	0.124 (0.022)	F(1,46)= 2.445, p>.05
/e/	0.149 (0.035)	0.191 (0.051)	F(1,46)= 9.654, p<.05
/ɜː/	0.123 (0.024)	0.162 (0.023)	F(1,46)= 40.944, p<.001
/ɪ/	0.109 (0.021)	0.117 (0.022)	F(1,46)= 13.648, p<.001
/iː/	0.140 (0.031)	0.158 (0.033)	F(1,46)= 7.377, p<.05
/ɔː/	0.149 (0.031)	0.188 (0.041)	F(1,46)= 49.592, p<.001
/ɒ/	0.130 (0.024)	0.141 (0.028)	F(1,46)= 10.508, p<.05
/ʊ/	0.133 (0.023)	0.141 (0.026)	F(1,46)= 6.596, p>.05
/uː/	0.167 (0.044)	0.189 (0.043)	F(1,46)= 16.919, p<.001

Given that most vowels were produced with longer duration after training, a comparison of the difference between tense-lax vowel pairs at pre and post test was conducted to see if there was any change after training. To estimate the percentage of relative change in duration between the lax (VD1) and tense (VD2) vowel within a pair, the following formula was used: (VD2-VD1)/VD1)*100.

A linear mixed-effect model using the R software (lme function) was used to analyse this relative change data for seven vowel pairs (/iː–ɪ/, /ɜː–e/, /æ–ʌ/, /ɑː–æ/, /ɑː–ʌ/, /ɔː–ɒ/, /uː–ʊ/). Training group (3), time (pre, post), vowel pair (7), training group*time, training group*vowel pair and training group*vowel pair*time were used as fixed effects and participant was the random factor. The results (Table 4.6) showed that there was no significant effect of training group; tense-lax duration difference did not differ significantly as a factor of vowel training mode. There was a significant effect of time: tense-lax duration difference increased after training. That is, L2 learners produced tense vowels with longer duration than the lax counter-part after training. The vowel pair effect was also significant, the overall mean duration difference between pairs ranged from 10% to 40%.

Table 4.6 Results from the linear mix-model analysis on duration differences between tense-lax vowel pairs at pre and post training.

Effects	
Training group (Tgroup)	F(2,44)= 1.139, p>.05
Time	F(1,568)= 49.514, p<.001
Vowel pair	F(6,568)= 12.503, p<.001
TGroup*time	F(2,568)= 0.694, p>.05
Vowel pair*Tgroup	F(12,568)= 1.093, p>.05
Vowel pair*time	F(6, 568)= 1.883, p=.081
Vowel pair*time*Tgroup	F(12, 568)= 0.086, p>.05

There was a marginally significant effect of pair*time interaction which was also explored with a linear-mixed model analysis using R. The results showed that two pairs did not change their durational difference at a significant level after training. These pairs were the /ɑ:-æ/ (pre M: 11.6, SD: 29; post M: 20, SD: 25) and /i:-ɪ/ (pre M: 31, SD: 31: post M: 37.5, SD: 30) which showed some small change from pre to post in duration (Table 4.7).

Table 4.7 Mean (M) and standard deviation (SD) for pre and post duration differences between vowel pairs. The time effect from pre to post is also shown. The difference is expressed in percentage for the tense vowel relative to the lax pair (except for /æ-ʌ/). Most pairs increased their duration difference.

Vowel pair	**Pre test M (SD) Relative duration difference %**	**Post test M (SD) Relative duration difference %**	**Time effect**
/ɑ:-æ/	11.5 (29)	20 (25)	F(1,46)= 3.118, p>.05
/ɑ:-ʌ/	25 (37)	56 (48)	F(1,46)= 37.362, p<.001
/æ-ʌ/	15 (33)	31.5 (32)	F(1,46)= 18.149, p<.001
/ ɜ:-e/	1.4 (25)	19.6 (31)	F(1,45)= 14.543, p<.001
/i:-ɪ/	31 (31)	37.5 (30)	F(1,46)= 1.490, p>.05
/ɔ:-ɒ/	16 (24)	36.8 (35)	F(1,46)= 18.539, p<.001
/u:-ʊ/	28 (43)	39.6 (53)	F(1,46)= 5.339, p<.05

To summarise, L2 learners increased their use of duration as a strategy to produce the distinction between tense-lax English vowels after training. This was observed for most of the tense-lax pair comparisons and also for the /æ–ʌ/pair.

4.4 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 from the Vowel test (pre and post test, Chapter 3) and the goodness ratings collected for vowel production in this study (also pre- and post-test, 4.2.1) were used to run a Pearson moment-correlation analysis. The results (Table 4.8) 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. Only a weak correlation was found between vowel perception and production at post test. To further explore this weak relation, correlations were run separately per training group; a significant correlation was found only for the Video training group at the post test ($r=.506$, $N=16$, $p<.05$). This group had no access to audio during their vowel training; yet, they improved as much as the other groups in vowel perception and production.

Table 4.8 Pearson correlation values for pre and post test scores for the Vowel identification (Vowel test) test and vowel production (goodness-rating test). (*) significant at .05 and (**) significant at .001 level.

	Perception		**Production**
Test	**Vowel ID Pre test**	**Vowel ID Post test**	**Vowel Production Pre**
Vowel ID Pre test	-	-	-
Vowel ID Post test	.820** .001 47	-	-
Vowel Production Pre test	-	-	-
Vowel Production Post test	-	.339* .020 47	.579** .001 47

Initial English vowel production (pre test) was significantly correlated with the post test production ratings (Table 4.8). That is to say, the ranking of learners in terms of the quality of their vowel production remained similar after training (Fig. 4.9). The correlation between pre and post vowel identification scores (Vowel test) did not seem to be related to the learners' capacity to produce the 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.

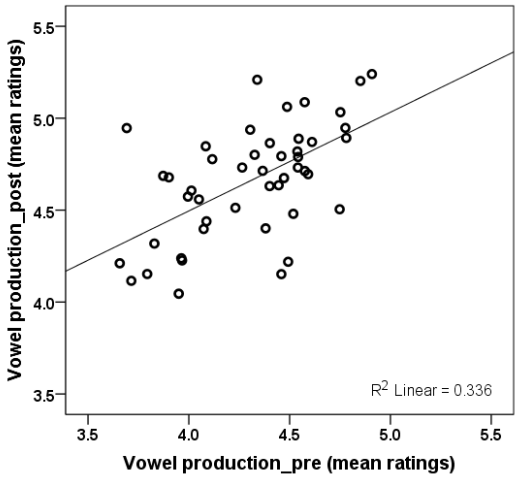


Figure 4.9 Scatter plot showing mean ratings for the L2 learners' vowel production at pre and post test.

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 (1-100%), per Training group. The results (Table 4.9) showed that there was a weak correlation between vowel production and proficiency level before and also after training. Therefore, the participants' level of proficiency could only account for a small amount of variability in the production data.

Table 4.9 Pearson correlations for the goodness-ratings (vowel test, pre & post) and English proficiency level. (*) significant at .05 (2-tailed) and (**) significant at .001 level (2-tailed).

Vowel production Goodness-ratings		**English proficiency Level**
Pre test	**r:**	.339*
	p:	.020
	N:	47
Post test	**r:**	.379**
	p:	.009
	N:	47

4.5 Vowel duration and goodness ratings

Given that duration was the only significant change in learners' English vowel production after training, the relation between duration and goodness-ratings was explored with separate Pearson product-moment correlation for each vowel. The aim of this analysis was to find whether ENS rated the tense vowels higher if they were produced with longer duration than the lax pairs.

The results for the tense vowels (Table 4.10) showed significant correlations between duration and goodness-ratings for three vowels at pre test and for three vowels at post test. A negative correlation was found for three lax vowels at pre test and for two in the post test comparison. These results suggest that ENS rated tense vowels with longer duration as a positive feature and lax vowels with long duration negatively.

Table 4.10 Correlation between goodness-ratings and duration at pre and post test. ** Correlations is significant at .01 level (2-tailed), *significant at the .05 level (2-tailed).

Goodness-ratings vs. Duration					
Tense vowels	**Pre test**	**Post test**	**Lax vowels**	**Pre test**	**Post test**
/iː/ pre	.371*	-	/ɪ/ pre	-.496**	-.399**
/ɜː/ pre	.387**	.312*	/e/ pre	-	-
/ɑː/ pre	-	.441**	/æ/ pre	-	-
/ɔː/ pre	.567**	.466**	/ʌ/ pre	-.352*	-
/uː/ pre	-	-	/ɒ/ pre	-	-
			/ʊ/ pre	-.555**	-.592**

4.6 Discussion

The aim of this study was to explore the impact of perceptual vowel training on the L2 learners' production of English vowels. The results showed an improvement in vowel production quality as judged by ENS using a goodness-rating test. Spectral and duration measures were also conducted on the vowel production recordings made pre- and post-training.

Improvement in the L2 learners' production of English vowels after perceptual training was confirmed by ENS who gave higher ratings to vowels in the post test recordings. These results are in line with previous L2 training studies which have reported improvement in production after perceptual training for English consonants (Akahane-Yamada, 1996; Bradlow et al., 1997; Hazan et al., 2005) and vowels (Iverson et al., 2012; Lambacher et al., 2005; Lengeris & Hazan, 2010), without giving participants any explicit production training on the contrasts tested. 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. Although, there is no agreement on how exactly this relation works (Iverson & Evans, 2009; Iverson et al., 2012).

Learners in the current study showed some degree of spectral difference in their production of English tense-lax vowel contrasts before and after training. However, no substantial spectral change occurred as a result of the vowel training sessions; the spectral distance for the tense-lax pairs remained almost unchanged. L2 learners have been reported to be less sensitive to spectral information than English native speakers for vowel perception (Cebrian, 2006; Escudero, 2000; Lengeris, 2009; Wang & Munro, 2004). However, Iverson and Evans (2007) found that L2 learners were able to attend to both spectral and durational information for English vowel identification.

Another dimension explored in this study was the duration of vowels in production. L2 learners used duration to produce the tense-lax vowel distinction before and after training; with larger duration contrasts made in the post test recordings. These findings are consistent

with previous studies reporting that L2 learners over-rely on duration as primary cue for the tense-lax distinction, irrespective of the status of duration in their L1 (Bohn, 1995; Cebrian, 2006; Chen, 2006; Escudero, 2001; Ingram & Park, 1997; Kewley-Port et al., 1996; Lambacher et al., 2005; Mora & Fullana, 2007). It has been suggested that L2 learners use duration as a primary cue for L2 vowel perception, unlike ENS who use spectral information (Hillenbrand & Clark, 2000), when the acoustic information is not enough to perceive the vowel contrasts (Bohn, 1995). Furthermore, this preference for the use of a duration strategy for the tense-lax distinction seems to transfer to production, as reported in Chen (2006) and Ingram & Park (1997).

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 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. 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.

This lack of a training modality effect has been reported on studies that examined the improvement in L2 speech perception after training. Comparing A and AV perceptual training for the English /r/-/l/ contrasts given to Japanese learners, Hazan et al. (2005) found no difference in the amount of perceptual improvement between groups after training. However, the AV trained group showed more improvement in production. On the same English contrast, Iverson et al. (2005) trained Japanese learners using high variability phonetic training (HVPT) in three variations; using F3 (third formant frequency) contrast maximized, perceptual fading (with F3 reduced) and second cue variability (with F2 variation). The results showed similar amounts of perceptual improvement across training groups. The novelty of the current study is to compare the impact of three different perceptual training modalities (Audio, Audio-visual and Video-only) on L2 learners' production of English vowels, as most of the research on English vowels has mainly

studied the effect of auditory training on perception (Iverson et al., 2012; Lambacher et al., 2005; Lengeris & Hazan, 2010) and fewer HVPT studies have compared AV training modalities (identification and articulatory) without auditory training as baseline (Aliaga-Garcia, 2010).

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; Rallo-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 are these two measures comparable. Nevertheless, this is a common practice in L2 perception and production studies.

Some of the reasons to account for this lack of correlation suggested by Bradlow et al. (1997) were that learners may show different rates at which they experience "*the motor commands*" change to improve pronunciation. The authors suggested that the observed 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. On a different view, Iverson & Evans (2009) found that training made learners more efficient at applying existing categories to improve

vowel identification but their perceptual representations did not change after training. Furthermore, Iverson et al. (2012) suggested that perception and production may require different processes which may not share a direct link, as changes in one domain may not necessarily result in changes in the other (Hattori & Iverson, 2009). To our knowledge, there seems to be a lack of attention to this lack of correlation between L2 perception and production mechanisms in the L2 speech models that exist so far. 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.

In summary, findings in this study showed L2 learners improved the accuracy of their production of English vowels after training. Duration was the dimension that was favoured to express the tense-lax vowel distinction, though learners also showed different degrees of spectral differences for these contrasts. However, the only aspect that changed after training was the tense-lax duration contrast with tense vowels becoming longer than their lax counterpart.

Chapter 5

Discussion

The perception of English vowels by L2 learners has mainly been studied in relation to the use of auditory cues (Flege, 1997; Flege et al., 1997; Fox et al., 1995; Garcia-Lecumberri & Cenoz, 1997; Ingram & Park, 1997; Iverson et al., 2012; Iverson & Evans, 2007; 2009; Nishi & Kewley-Port, 2007; Lambacher et al., 2005; Lengeris & Hazan, 2010; Strange et al., 1998; Thomson, 2012; Wang & Munro, 2004). Given that speech perception is predominantly bimodal in a naturalistic environment (Rosenblum, 2005), the addition of visual speech information was thought to aid L2 learners in improving speech perception. The aim of the current thesis was to examine the sensitivity to visual cues for English vowel perception of L2 learners with Spanish as L1. Two studies were conducted, the first one to test sensitivity to visual cues for English vowels by ENS and L2 learners. The second study compared different types of training modalities and their impact on English vowel perception and use of visual cues. The impact of training on vowel production was also examined in the second study.

In the first study (Chapter 2), English native speakers (ENS) were tested on their vowel perception in noise to find out whether perception would be improved by the addition of visual cues. This was done by presenting word-tokens in A and AV mode in noise and exploring whether visual cues facilitated vowel identification in AV mode. The results for ENS showed that visual cues contributed to better vowel identification in AV mode compared to A mode. Besides, vowel identification in V mode (no sound) was also informative. Poorer vowel identification was found for L2 learners; although it varied

depending on learners' level of proficiency with better performance for the L2 advanced group. However, no fundamental difference in vowel identification was found between A and AV mode in L2 learners, irrespective of their big difference in level of proficiency. Thus, L2 learners showed no benefit of visual cues in AV mode, in spite of being able to identify some vowels in V mode (no sound).

In a second study (Chapter 3), three groups of L2 learners were given high-variability-perceptual training (HVPT) in one of three modalities: audio training (AT), audio-visual (AVT) and video-alone training (VT) modality. Participants were tested on vowel identification and vowel production before and after training. Vowel identification performance improved after training without a training modality effect. As discussed in Chapter 3, these results suggested that learners can improve their vowel perception with auditory input as well as with the video-alone input, as they seemed equally effective. Besides, results showed that AVT or VT training did not contribute to increase the use of visual cues for vowel perception. Learners continued to rely on auditory input for vowel identification which suggested that the problem may lie on the learners' ability to interpret and integrate visual information for vowel perception. Vowel production also improved after training as rated by ENS. However, results revealed that improvement in perception was not related to improvement in production. As discussed in Chapter three, these results are in line with previous research (Bradlow et al., 1999) and may support suggestions that these two processes may not have a direct relation (Hattori & Iverson, 2010).

The effect of L2 proficiency level on English vowel identification was clear in study one (Chapter 2) in which the more advanced group showed better performance than the L2 beginners. However, L2 beginners in the training study (Chapter 3) achieved better performance than the L2 advanced group after the former had completed their five training sessions. This finding revealed an interesting aspect of the powerful impact of perceptual training which boosted the L2 beginners' perceptual ability to a higher level than a group which had three or four years more of intensive English learning experience (i.e. the L2 advanced group). Having said that, it remains unclear for how long this learning gained through training would remain as retention measures were not used in this study. Although,

evidence of retention of learning after L2 speech training sessions has been reported (Iverson and Evans, 2009), so it could be hypothesised that learning may remain after months of training completion for the L2 beginner learners. Use of visual cues was another aspect that was expected to be mediated by level of proficiency as suggested in a previous study by Wang et al. (2008). The results showed no significant difference in the amount of visual information used by L2 learners when comparing beginners versus advanced, This finding, as discussed in Chapter 3, may have to do with the material tested here (vowels instead of consonants as in Wang and colleagues) and the L2 learners' environment with almost no access to direct interaction with native speakers.

In both studies presented in this Thesis (Chapter 2 and 3), measures of auditory frequency discrimination, visual bias, perception of key-words in sentences and language proficiency were used to see if a relation with individual variability in vowel identification could be shown. In general, the individual performance on vowel perception did not show to be correlated with any of the measures, except for the level of proficiency for L2 beginners in study 2 (Chapter 2). As discussed before, the auditory frequency discrimination test used here had shown correlation to vowel perception in a study by Lengeris and Hazan (2010). The lack of relation found in study one and two, presented in Chapter 2 and 3, may be due to a larger set of vowels tested in this thesis experiments. The measure of visual bias obtained from the McGurk test had also shown a relation with L2 speech perception before in Chen and Hazan (2009) but the material tested in their study was English consonants which may have been easier to visually identify than English vowels.

With the aim of measuring generalisation of vowel identification to new tokens and new talkers in sentence-length material, an audio-visual "True-or-false sentence test" was used in study 2 (Chapter 3) as pre and post test. The test was given to ENS, and L2 learners with beginner and advanced level of proficiency. As expected, ENS had no problem in identifying the correct and incorrect sentences. Overall, the L2 learners were not able to show any improvement in this task after training, their results remained at chance level. This finding suggested that vowel identification at sentence level is much more complex than perception of isolated words. Moreover, cognitive aspects like attentional overload

seem to play an important role in L2 speech perception in general and particularly at sentence level.

Summarising, one of the main findings in this thesis was that the L2 learners showed a lack of integration of visual information with the auditory component in audio-visual (AV) mode. Learners showed some capacity to identify English vowels using speech reading in video-alone (V) mode. However, unlike ENS, they were unable to use this information to improve their identification rates in the AV relative to A mode. An unexpected result was the lack of training modality effect: those learners who were given training in AV or V modality did not show any increased use of visual information post-training for vowel perception. Moreover, the group that was trained without audio (the VT group) improved as much as the other two groups (AT and AVT groups), including improvements in A mode.

Two main issues which are raised by the results of the current studies are the similarity in impact of the three training modalities on speech perception and the extent to which learning skills are brought to play in L2 speech training programmes. In the light of these findings, another major consideration that is brought to our attention is the lack of representation of visual information in L2 speech perception models. These issues will be discussed in more detail below.

5.1 Representation of visual information in L2 speech perception models

In the light of the current findings, a major issue which arises for discussion is the lack of representation of visual information in L2 speech perception models, in spite of the growing body of research on the role of visual cues in L2 speech perception (Chen & Hazan, 2007; 2009; Hardison, 1999; 2003; 2005; 2009; Hazan, et al., 2006; Hazan, Sennema, Iba & Faulkner, 2005; Hazan, Kim & Chen, 2010; Massaro et al., 1995; Navarra & Soto-Faraco, 2007; Navarra et al., 2010; Ortega-Llebaria et al., 2001; Sekiyama, 1997; Sekiyama, 2003; Wang et al., 2008; 2009). Neither of the most influential L2 speech perception models such as PAM (Best, 1995, Best & Tyler, 2007) or SLM (Flege, 1995)

has explicitly considered the visual information component in their accounts. Although PAM is framed within the Direct-Realist approach which considers the articulatory gestures of speech as its primitives, the model does not account for the lack of integration of visual information for L2 speech perception. If we take the SLM which was proposed to account for L2 perception and production by more experienced learners (as in the current study), it predicts different degrees of success in L2 speech perception based on the phonetic distance between an L1 and L2 phonemic categories (Flege, 1995). This model posits that learners relate allophones of the L2 to the nearest allophones of the L1, though new category creation is sometimes possible. The further an L2 contrast is from an L1 phoneme, the more easily a category is perceived as different and established as a new category. A parallel account in terms of how L2 visual cues may be perceived in relation to the L1 has been suggested in Hazan et al. (2006). The authors presented three types of possible scenarios for visual cues (VC) in L2 speech perception/acquisition: a) relatively similar visual cues for a viseme exist in the L1 and L2; b) the visual cues for a viseme exist only in the L2 but not in the L1, and c) the visual cues for a viseme exist in the L1 and L2 but are used to mark different phonetic distinctions.

In the first scenario (a), where the L2 viseme contrast has a similar counterpart in the L1, it is expected that, visually, this contrast will be assimilated to the L1 viseme category, and no new category will be formed. For instance, the English /iː/–/ɪ/ contrast would be assimilated to Chilean-Spanish /i/. Even though they are spectrally and visually different in English, their spectral and visual realizations conform to the range of naturally acceptable allophonic variability for the Chilean-Spanish (Ch-Spanish) /i/. In the second scenario (b), if the viseme does not exist in the L1, this would facilitate its perception or acquisition as a new viseme. For example, the English dental fricative /θ/ may be more easily perceived by Ch-Spanish learners, as no viseme or phoneme resembles the English viseme. Finally, in the third case (c) if the viseme exists in both the L2 and L1 but it is used to mark a different contrast in the L1, there would be no need to establish a new viseme category. However, a new association of the viseme with the corresponding phoneme will

be needed. Hazan et al. (2006) found that the existence of a labio-dental fricative viseme /f/ aided Spanish learners of English to perceive the contrast between English bilabials /b/–/p/ and the dental fricative /v/, even though the latter is not a phoneme category in Ch-Spanish.

A crucial aspect to bear in mind in L2 speech perception models is the need for more highlighting of the L2 visual cues together with the acoustic cues of a phoneme to create richer phoneme-viseme category representations. Research on visual cues in L2 speech perception has suggested that the lack of use of visual information for speech perception in learners may be caused by impoverished experience with visual speech information in their L1. Thus, sensitivity to visual cues for speech perception may not have been developed and needs to be acquired when learning a new language (Hazan et al., 2006; Wang et al., 2008). The Chilean learners may not have been familiarised with the use of visual information for vowel identification in their L1. It is difficult to imagine that a native speaker of Spanish may need extra information, other than auditory, when perceiving any of the Spanish five vowels in an optimal situation, i.e. without adverse conditions such as background noise. Each vowel constitutes a well-defined phoneme, with no competitor near its spectral centre-point which otherwise might have required additional cues for its identification. Even in adverse conditions, confusions across these five spectrally distinct vowels are unlikely. This is not the case for languages with a more crowded vowel inventory. For instance, visual information helps to distinguish between French vowels /i/ and /y/ (Benoît, Mohamadi & Kandel, 1994; Robert-Ribes et al., 1998). Also in Swedish, visual cues contribute to the perception of roundedness in vowels (Traünmuller & Öhrström, 2007).

The aim of presenting the proposal above, taken from Hazan et al., (2006), is to highlight the need for an L2 speech perception/acquisition model that accounts for the bimodality of speech as suggested by some researchers in the field (Hardison, 1999; Hazan et al., 2006; Wang et al., 2008). Taking into account the contribution of visual cues, as well as the acoustic cues and phonetic features, may allow models to make more specific predictions about assimilation patterns and difficulties in acquiring an L2 novel contrast.

5.2 Why no difference across the three training modalities?

Studies which compare the impact of different training approaches to improve L2 speech perception have shown quite similar degrees of improvement in trainees. For instance, García-Lecumberri and Cook (2008) used training in both quiet and in noise for 24 consonant contrasts and two different training groups (Spanish learners of English) with both groups showing similar gains after training. Hazan et al., (2005) gave Japanese L2 learners Auditory and Audio-visual training for English /r/–/l/ and found no difference in the amount of improvement per training modality between training groups. Similarly, Iverson et al., (2005) gave Japanese learners three different training programmes to learn /r/–/l/. The three training modalities differed in the amount of acoustic manipulation to F3 cues (enhanced, reduced) and the third modality introduced variability in F2. Their results showed improvement in all three training groups with no training modality effect. Another study which compared auditory perceptual training and articulatory training for English vowel reduction was conducted by Gómez-Lacabex, García-Lecumberri & Cook (2009). Spanish learners were found to improve in perception of the trained contrasts, regardless of the training programme they had received. Altogether, these findings suggest that the learning effect of training is quite robust but rather unaffected by the specific approach taken in training novel phoneme contrasts. There is a considerable number of studies on the effect of visual cues on L2 English consonant perception (Sennema, Hazan & Faulkner, 2003; Navarra & Soto-Faraco, 2007; Wang et al., 2008; 2009) but fewer studies which contrast auditory with audio-visual modality for L2 speech training (Aliaga-García, 2010; Hardison, 2003; 2005; Hazan et al, 2005). In general, most of the studies comparing audio (A) and audiovisual (AV) perception show an advantage for AV over A modality when the contrasts are visually salient (Hardison, 1999; Hazan et al., 2006). To our knowledge, the current study is original in the way it compares A, AV and video-alone (V) training modalities for English vowel perception with a high-variability perceptual training (HVPT) approach.

It was expected that participants who were given different types of vowel training programmes (AT, AVT and VT modality) would show some difference in performance in the identification of English vowels. For instance, better performance in auditory perception for learners trained in A modality, advantage in AV mode for AVT learners and poorer auditory perception for those who were given the VT sessions. The results of the post test revealed that all learners had improved their vowel identification capacity in similar amounts in A, AV and V mode. It may be hypothesised that training helped learners improved their auditory representations of the phonemic categories, irrespective of their training modality.

Learners in the AV training modality group did not show a visual advantage for speech in AV mode. A possible explanation is that during training and in the tests, they only attended to the audio component of the stimuli. Research in the area of working memory and language processing has suggested that AV perception seems to require more cognitive resources than A alone (Fraser, Gagné, Alepins & Dubois, 2010). Fraser and colleagues used a dual-task (tactile recognition) to compare the amount of effort needed to understand speech in AV and A mode in noise. They found that at the same levels of noise, AV information contributed to better speech perception. However, when the level of noise in AV modality was increased to make speech identification equally comparable, the speech perception task in AV mode required more effort. The effort was measured by the decrease in the dual-task performance. These findings suggested that under increased perceptual and cognitive demand, speech perception in AV mode is more effortful. In the same line of research, Alsius, Navarra & Campbell (2005) addressed the impact of high attention demands on audio-visual speech integration. They used the McGurk effect design in a dual-task for speech perception. They found that the binding of auditory and visual speech information is affected under higher attentional demands. In the current study, it could be argued that participants were unable to integrate the visual information in AV mode due to increased processing demands.

In the current study, no dual or extra task was explicitly given to participants when doing their pre or post tests. However, it could be hypothesised that learners were accessing other

resources at the lexical and semantic level together with their L1 and L2 phonetic-phonological knowledge for the speech perception task. Bernstein and Auer (2003) and Luce and Pisoni (1998) have found activation of different domains while undertaking word perception tasks in their research on auditory word recognition and lexical neighbourhood activation. Thus, as a consequence of the activation of different domains for the L2 speech perception task, the cognitive resources of the learners/perceivers may have been depleted. Perceivers may have been forced to rely on one channel of information (audio) and bimodal perception was impeded. This would account for the results for participants in the current study who received AV training and only showed improvement in A mode. The arguments presented so far would account for the similarity in performance among participants and their lack of integration of A and V information in AV mode for speech perception after training.

The learners in the VT sessions had to focus their attention on the visual gestures of vowels presented in short words (CVC-words). It could be speculated that they perceived these articulatory gestures and matched them with some kind of previously stored representation of phonemic categories for English vowels or for L1 vowels. However, to be able to do the matching based on the articulatory gestures only that representation needs to be necessarily bimodal; that is to say, made of spectral and of articulatory gestures. An account which may contribute to understand why perceiving only the articulatory gestures of speech aided L2 learners to improve their perception of English vowels may be the "Analysis-by-synthesis model" (Stevens & Halle, 1967; Stevens, 1972; see Chapter 3 for description). As discussed in Chapter 3, van Wassenhove et al. (2005) suggested that the Analysis-by-Synthesis model (AbS) would be the best to explain audio-visual perception of speech, provided the perceptual unit in the model is conceived as a "bimodal processor" as advanced in Wassenhove et al. (2005). The latter authors adapted the AbS model by incorporating a bimodal feature to the perceptual unit where the perceived gestures are used to build a hypothesis for the speech information perceived. They highlight the fact that visual speech allows the prediction of the auditory input. In our study, the VT group may have made predictions of possible "speech categories" based on the visual information presented, and then used the "loop" (feedback) in the AbS model which processes the

synthesized hypothesis and uses the sensory information (phoneme-viseme category) in search for a match. If the matched information is discrepant, the process is restarted until a perfect match of a bimodal category is found.

It is a fact that the learners in this study also had access to listening to English in the classroom context prior and while taking their vowel training sessions. It may also be the case that during training, they could have learnt articulatory gestures of the vowels and practised their matching with the auditory categories they already had developed in their vowel space –from classroom experience. Given that during training learners received feedback, they could immediately test a hypothesis of a possible match for the gestures they saw with the phoneme category they thought would match. Thus, training probably made learning to match visual cues to existing auditory categories more efficient. During the matching process, learners could have mentally rehearsed the phoneme while perceiving the gestures of the English vowels and this made up for the absence of auditory input during training.

Another possible explanation for the results of the VT group is that training did not contribute to better perception but rather, it was the intensive learning in the classroom what caused the improvement. There was no control group tested before or after training for the vowel test. However, in the comparison for vowel identification performance, beginner learners (who took the training sessions) outperformed the advanced learners at post test, with the VT group (M: 62.5) obtaining similar overall means to the L2 advanced group (M: 60). Though the difference did not reach a significant level $p>.05$, the fact that VT group achieved similar levels of vowel identification as a very advanced proficiency group confirms the effectiveness of training. The advanced learners had studied English for three or four more years than the beginners, in a similar intensive programme at university (major in English language) and were already quite proficient. Thus, it could be concluded that it was actually the effect of the vowel training sessions with video-only material which cause the improvement. This is in line with findings of improvement in identification performance by using video-alone training for English /l/–/r/ contrast in Japanese learners (Hazan & Sennema, 2007). The implications of the findings in the current study

point towards the possibility of Audio and Video training (AT, VT) being equally effective in improving English vowel identification. This holds, provided the learners have some previous experience with the language.

Individual variability could be another factor contributing to the lack of difference in the effect of training modes. It is not surprising that a lot of individual differences were found in the current study. For instance, the better perceivers at pre-test were not necessarily the only ones who achieved the highest scores at post test. These individual differences may be attributed to a number of different factors, including phonological short term memory (PSTM), attention capacity, associated-pair capacity and auditory frequency discrimination capacity. Some L2 studies have found that learners' PSTM, measured with a non-word repetition task, is related to speech perception of non-native contrasts: a higher PSTM capacity was related to better speech identification performance (Aliaga-García, 2010; Cerviño-Povedano & Mora, 2010; O'Brien, Segalowitz, Freed & Collentine, 2007). Another study on individual variability and its relationship with cognitive processes (Kim & Hazan, 2010) found that attentional switching and the ability to associate two unrelated items were most related to the capacity to learn a new phonetic contrast (Korean lenis and aspirated stops). Another possible source of difference in performance has been reported in Lengeris & Hazan (2010). Greek learners of English were tested on their auditory frequency discrimination capacity and their English vowel perception. A direct relation was found between higher capacity to discriminate small differences in auditory frequencies and a more accurate performance in the discrimination of two vowel pairs (/iː/–/ɪ/, /æ/–/ʌ/) in auditory mode.

In the present study, the complex individual variability in the results of the different tests may have cancelled out the possible effect of any training modality. Of the measures used to account for individual differences in vowel identification (auditory frequency discrimination test, visual bias and English proficiency) only the language proficiency level showed a strong relation to vowel identification capacity before and after training. However, it could be the case that there are intertwined relations between these factors

which bring more complexity to the already existing individual variability and thus establishing an exact relationship becomes more difficult. In future studies, different types of cognitive measures could be included to explore the sources of individual variability in L2 speech perception. Although, factors like motivation or even boredom could also affect the impact of training but are harder to measure.

5.3 What is actually being learnt as a result of training?

L2 training studies have demonstrated that L2 learners achieve identification improvement for novel L2 contrasts with generalization to new tokens and talkers, and in some studies this improvement has been shown to transfer to production as well (Bradlow et al., 1999; Chen, 2006; Hazan et al., 2005; Iverson et al., 2005; Iverson & Evans, 2009; Iverson et al., 2012; Lambacher et al., 2005; Lengeris & Hazan, 2010; Nishi & Kewley-Port, 2007; Thomson, 2012; Wang & Munro, 2004, among others). In most of the early training studies, researchers hypothesized that perceptual improvement after training was due to the change in the learners' perceptual space resulting in the creation of new phonemic categories. Failure to create new categories was explained by perceptual interference from the L1 repertoire (Iverson et al., 2003) or, most commonly, assimilation of an L2 contrast to an L1 category as in the SLM (Flege, 1995).

More recent training studies have shed light on a different perspective concerning training. Learners may not actually create new categories but instead simply become more accurate in the way they use their existing categories; using L1 or L2 categories they already knew before training. As discussed in Chapter 3, Iverson and Evans (2009) observed that learners improved their vowel perception without changing their best exemplars for English vowels after training. Thus, training did not modify the learners' mental representations for the English vowels they had prior to training. These results led the authors to suggest that during training learners focus their attention on the cues that allow them to better identify an L2 phoneme, and in doing so, improve the process of attaching a label to that phoneme

without creating new categories. In other words, perceptual improvement does not necessarily reflect new phonemic category establishment.

The findings in the current study did not provide evidence of new category formation. Though best exemplars for English vowels were not tested, the lack of transference of perceptual improvement to the perception of vowels at sentence level in the TF test suggests a lack of new categories being created. Once the task of labelling the phonemes presented in words by using pre-existing categories is changed to key-words in sentences, the labelling strategy was not enough to succeed in the identification of English vowels.

Many studies show that learners can retain their perceptual improvement when tested between three to six months later (Bradlow et al., 1999; Iverson & Evans, 2009; Lively et al., 1994; Wang & Munro, 2004). In this retention tests, they are usually given the same material as in the post test to make results comparable. But, would this improvement and retention transfer to more naturalistic perception contexts? For instance, would this learning allow participants to improve their perception when the contrasts are presented in a sentence or in a natural conversation? To our knowledge, there are no previous studies that have tested whether perceptual improvement obtained through training can transfer from syllable or word-level perception to sentence-level. In the present study, the aim of introducing the "True-or-False sentence test" was to measure perceptual improvement from word to sentence material. The overall results suggested there was not much transfer from the improvement of vowel perception in word-material to sentence-level material, though a small group of learners (14) showed some improvement. Overall, the design of the test revealed that some adjustments are needed to make the test more manageable in terms of how to help the learners focus their attention and allow better allocation of cognitive resources. However, these findings may encourage future researchers to include more naturalistic and cognitively challenging L2 speech perception tests to measure the generalization of perceptual improvement after training.

5.4 Limitations and future research

One of the limitations of this study is that there was no control group for the Vowel training study. Therefore, no comparison could be established with the improvement of other learners in perception as a consequence of natural classroom learning. This was due to logistic reasons; there were no remaining learners with similar level of English available to be tested. All first-year students in the Teacher training programme (UdeC) were included in the training study. To make up for this, a group of 37 advanced students (4^{th} and 5^{th} year students) were tested later and used in the comparison of results.

Measuring retention of learning after training was out of the scope of this study. As discussed earlier in this chapter, most studies which have measured retention show positive results (Bradlow et al., 1999; Iverson & Evans, 2009; Lively et al., 1994; Wang & Munro, 2004). The decision of not to measure retention was due to the intensive nature of the programme these participants were following. Participants had an average of 15 to 18 hours of English lessons a week, so any testing 6 months later would carry the risk of confounding the results if students had not continued having the same amount of input because they had either dropped out or failed a module taught in English resulting in having less English input for some months.

It would be desirable for future research on English vowels to consider different techniques to compare perception and training in A, AV and V modes, leading to better understanding of the phenomenon. Regarding training, it would be interesting to develop a training methodology that allowed the highlighting and integration of visual gestures for L2 vowel contrasts. This could possibly be done with a graded training procedure; combining first audio training, then a Video-alone phase and finally the AV phase. In addition, researchers could perhaps find a way to match the training to the A or V bias of the learner. As for designing tests to measure the transfer of perceptual improvement to more naturalistic contexts, it is a challenge researchers will need to tackle. It is clear that in everyday interactions learners need to improve their perception in a way that allows them to perform successfully beyond the word-level identification.

5.5 Summary

One of the main contributions of this thesis was the finding of a lack of effect of training modality which was only possible to establish by comparing the three training modalities and measure of vowel identification in three modes (A, AV, V). In particular, the fact that video-alone training (VT) can foster similar amounts of perceptual improvement of English vowels compared with auditory (AT) and audio-visual training (AVT) in learners who had some experience with the language. Concerning visual cues for English vowel perception, this study is novel in finding that L2 learners can identify some vowels visually but fail to integrate visual cues for their perception in AV modality and that more experienced learners did not show better capacity than less experienced learners to use visual cues for English vowel perception. Finally, another innovation that needs to be highlighted in this study is the use of a sentence-test to measure whether perceptual improvement transfers to a more naturalistic speech context. This new test was also a more cognitively challenging way of testing speech perception and a closer to a functional use of language.

References

Adank, P., Smits, R., & Van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *The Journal of the Acoustical Society of America*, *116*(5), 3099.

Akahane-Yamada, R., Tohkura, Y., Bradlow, A., & Pisoni, D. (1996). Does training in speech perception modify speech production? *4th International Conference on Spoken Language Processing*. pp. 606–609. Philadelphia, PA, USA.

Aliaga-García, C. (2010). Measuring perceptual cue weighting after training: a Comparison of auditory vs. articulatory training methods. *New Sounds 2010.* Poznán, Poland.

Aliaga-Garcia, C., Mora, J. C. & Cerviño-Povedano, E. (2010). Phonological short-term memory and L2 speech learning in adulthood. *New Sounds 2010.* Poznán, Poland.

Alm, M., Behne, D. M., Wang, Y., & Eg, R. (2009). Audio-visual identification of place of articulation and voicing in white and babble noise. *The Journal of the Acoustical Society of America*, *126*(1), 377–87.

Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Current biology : CB*, *15*(9), 839–43.

Aoyama, K; Flege, J. E; Guion, S. G.; Akahane-Yamada, R. & Yamada, T. (2004). Perceived phonetic dissimilarity and L2 speech learning: the case of Japanese /r/ and English /l/ and /r/. *Journal of Phonetics*, *32*(2), 233–250.

Bench, J., Kowal, A. & Bamford, J. (1979). The BKB (Bench-Kowal-Bamford) sentence list for partially-hearing children. *British Journal of Audiology*, 13, 108-112.

Benoît, C., Mohamadi, T., & Kandel, S. (1994). Effects of phonetic context on audio-visual intelligibility of French. *Journal of speech and hearing research*, *37*(5), 1195–203.

Bent, T., Bradlow, A. R., & Smith, B. L. (2008). Production and perception of temporal patterns in native and non-native speech. *Phonetica*, *65*(3), 131–47.

Bernstein, L.E, Demorest M.E & Tucker P.E. (2000). Speech perception without hearing. *Perception & Psychophysics. 62*, 233–252.

Bernstein, L.E & Auer, E.T. Jr. (2003). Speech perception ans spoken word recognition. In Marschark, M. & Spencer, P. E. (Eds.) *The Oxford Handbook of Deaf studies, Language and Education* (399-411).Oxford: OUP.

Best, C. (1993). Learning to perceive the sound pattern of English. *Haskins Laboratories Status Report on Speech Research. SR-114*, 31–80.

Best, C., Hallé, P., Bohn, O., & Faber, A. (2003). Cross-language perception of nonnative vowels: Phonological and phonetic effects of listeners' native languages. 15Th *International Congress of Phonetic Sciences*, 2889–2892.

Best, C. & M. Tyler. (2007). Non-native and second language speech perception. In Bohn, O-S. & M. J. Munro (Eds.). *Language Experience in Second Language Speech Learning. In honor of James Emil Flege (pp. 15-34).* Amsterdam & Philadelphia: John Benjamins.

Binnie, C. A., Montgomery, A. A., & Jackson, P. L. (1974). Auditory and visual contributions to the perception of consonants. *Journal of Speech and Hearing Research*, *17*, 619–630.

Bovo, R., Ciorba, a, Prosser, S., & Martini, a. (2009). The McGurk phenomenon in Italian listeners. *Acta otorhinolaryngologica Italica : organo ufficiale della Società italiana di otorinolaringologia e chirurgia cervico-facciale*, *29*(4), 203–208.

Boersma, Paul & Weenink, David (2012). Praat: doing phonetics by computer [Computer program]. Version 5.1.24. Retrieved from http://www.praat.org/

Boersma, P., Escudero, P., & Hayes, R. (2003). Learning abstract phonological from auditory phonetic categories: An integrated model for the acquisition of language-specific sound categories. *Proceedings of the 15th International Congress of Phonetic Sciences*, 1013–1016.

Bohn, O.-S. (1995). Cross-language speech perception in Adults: first language transfer doesn't tell it all. In W. Strange (Ed.), *Speech perception and linguistic experience: theoretical and methodological issues in cross-language speech research* (pp.275–300). Timonium, MD: York Press.

Bohn, O.-S., & Best, C. T. (2007). Testing PAM and SLM: Perception of American English Approximants by Native German Listeners. In K. Dziubalska-Kołaczyk, M. Wrembel, & M. Kul (Eds.), *New Sounds*, 43–48.

Bradlow, a R. (1995). A comparative acoustic study of English and Spanish vowels. *The Journal of the Acoustical Society of America*, *97*(3), 1916–24.

Bradlow, A. R., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech communication*, *20*(3), 255–272.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (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), 2299–310.

Bradlow, A., Akahane-Yamada, R., Pisoni, D., & Tohkura, Y. (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, A. R. & Pisoni, D. B. (1999). Recognition of spoken words by native and non-native listeners: talker-, listener-, and item-related factors. *The Journal of the Acoustical Society of America*, *106*(4 Pt 1), 2074–85.

Braida, L. D. (1991). Crossmodal integration in the identification of consonant segments. *The Quarterly Journal of Experimental Psychology Section A : Human Experimental Psychology*, *43:3*(July 2012), 647–677.

Burnham, D., Lau, S., Tam, H., & Schoknecht, C. (2001). Visual discrimination of Cantonese tone by tonal but non-Cantonese speakers, and by non-tonal language speakers. In D. Massaro, J. Light, & K. Geraci (Eds.), *Proceedings of the International Conference on Auditory–Visual Speech Processing*, 155–160.

Campbell, R., Dodd, B. & Burnham, D. (Eds.). (1998). *Hearing by Eye II: Advances in the psychology of speechreading and auditory-visual Speech*. Hove, UK: Psychology Press.

Cebrian, J. (2006). Experience and the use of non-native duration in L2 vowel categorization. *Journal of Phonetics*, *34*(3), 372–387.

Cebrian, J. (2007). Old sounds in new contrasts: L2 production of the English tense-lax vowel distinction. *Proceedings of the 16th International Congress of Phonetics Sciences*, Saarbrücken, 1637–1640.

Cebrian, J., & Carlet, A. (2012). Audiovisual perception of native and non-native sounds by native and non-native speakers. In Martín Alegre, S. (Coord. & Ed.), Moyer, M., Pladevall, E. & Tubau, S. (Eds.). *At a Time of Crisis: English and American Studies in Spain. Proceedings from 35th AEDEAN Conference UAB*. Departament de Filologia Anglesa i de Germanística, Universitat Autònoma de Barcelona (pp. 300-307). Retrieved from http://www.aedean.org

Cerviño-Povedano, E. and Mora, J.C. (2010). Investigating Catalan-Spanish bilingual EFL learners' over-reliance on duration: vowel cue weighting and phonological short-term memory. *New Sounds. 2010 Proceeding*, Poznán, Poland.

Chen, Y. (2006). Production of tense-lax contrast by Mandarin speakers of English. *Folia phoniatrica et logopaedica : official organ of the International Association of Logopedics and Phoniatrics (IALP)*, *58*(4), 240–9.

Chen, Y., & Hazan, V. (2007). Language effects on the degree of visual influence in audiovisual speech perception. *Proceedings of the 16th International Congress of Phonetics Sciences*. Saarbrücken, Germany, 2177–2180.

Chen, T. H., & Massaro, D. W. (2008). Seeing pitch: visual information for lexical tones of Mandarin-Chinese. *The Journal of the Acoustical Society of America*, *123*(4), 2356–66.

Chen, Y., & Hazan, V. (2009). Developmental factors and the non-native speaker effect in auditory-visual speech perception. *The Journal of the Acoustical Society of America*, *126*(2), 858–65.

Cohen, M. M., Walker, R. L., & Massaro, D. W. (1996). Perception of synthetic visual speech. *NATO ASI Series of Computer and Systems Sciences*, *150*, 153-168.

Cooke, M. and Ellis, D. (2001). The auditory organization of speech and other sources in listeners and computational models. *Speech Communication*, *35*(3-4), 141–177.

Crawley, M. (2007). *The R Book.*West Sussex: John Wiley & Sons Ltd.

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*, 680-689.

Escudero, P. (2000). Developmental patterns in the adult L2 acquisition of new contrasts: The acoustic cue weighting in the perception of Scottish tense/lax vowels in Spanish. Unpublished M. Sc. thesis, University of Edinburgh, UK.

Escudero, P. (2001). The role of the input in the development of L1 and L2 sound contrasts: language-specific cue weighting for vowels. *Proceedings of 25th Annual Boston University Conference on Language Development*, 250–261.

Escudero, P., & Boersma, P. (2004). Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition, 26*, 551–585.

Fisher, C. G., (1968). Confusions among visually perceived consonants. *J. Speech and Hearing Research*, *11*(4), 796-804.

Flege, J. (1995). Second-language Speech Learning: Theory, Findings, and Problems. In W. Strange (Ed.) *Speech Perception and Linguistic Experience: Issues in Cross-language research, 229-273.*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.

Flege, J. E., MacKay, I. R., & Meador, D. (1999). Native Italian speakers' perception and production of English vowels. *The Journal of the Acoustical Society of America*, *106*(5), 2973–87.

Flege, J. (2002). Interactions between the native and second-language phonetic systems. In P. Burmeister, T. Piske & A. Rohde (Eds.) *An Integrated View of Language development: papers in honor of Henning Wode*. Trier: Wissenschaftlicher Verlag, 217-244.

Flege, J. E., & MacKay, I. R. a. (2004). Perceiving Vowels in a Second Language. *Studies in Second Language Acquisition*, *26*(1), 1–34.

Fox, R. a, Flege, J. E., & Munro, M. J. (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 America*, *97*(4), 2540–2551.

Fraser, S., Gagné, J.-P., Alepins, M., & Dubois, P. (2010). Evaluating the effort expended to understand speech in noise using a dual-task paradigm: the effects of providing visual speech cues. *Journal of speech, language, and hearing research : JSLHR*, *53*(1), 18–33.

Fuster-Duran, A., 1996. Perception of conflicting audio-visual speech: an examination across Spanish and German. In: Stork, D.G., Hennecke, M.E. (Eds.), *Speechreading by Humans and Machines: Models Systems and Applications* (pp. 135–143). Springer-Verlag, New York.

García-Lecumberri, M.L. & Cenoz, J. (1997). L2 perception of English vowels: testing the validity of Kuhl's prototypes. *Revista alicantina de Estudios Ingleses*, *10*, 55–68.

Garcia-Lecumberri, M. L. G., Cooke, M., Cutler, A. (2010). Non-native speech perception in adverse conditions: A review. *Speech Communication*, *52* (11-12), 864–886.

Garcia Lecumberri, M. L., & Cooke, M. (2008). The effect of training in noise on foreign language consonant acquisition. *The Journal of the Acoustical Society of America*, *123*(5), 3881.

Gibson, J. J., & Gibson, E. J. (1955). Perceptual learning : Differentiation or enrichment? *Psychological Review*, *62*(1), 32–41.

Gómez Lacabex, E., García Lecumberri, M.L. & Cooke, M. (2009). Training and generalization effects of English vowel reduction for Spanish listeners. In Watkins, M.A., Rauber, A, & Baptista, B. (Eds.), Recent Research in Second Language

Phonetics/ Phonology: Perception and Production. Newcastle upon Tyne: CSP, 32-42.

Green, K. P. (1998). The use of auditory and visual information during phonetic processing: Implications for theories of speech perception. In R. Campbell, B. Dodd & D. Burnham (Eds.). *Hearing by Eye II: Advances in the psychology of speechreading and auditory-visual speech* (pp. 3–25). Hove, UK: Psychology Press.

Hardison, D. (1996). Bimodal speech perception by native and nonnative speakers of English: Factors influencing the McGurk effect. *Language Learning*, *46:1*, 3–73.

Hardison, D. M. (1999). Bimodal speech perception by native and nonnative speakers of English: Factors influencing the McGurk effect. *Language learning*, *49* (Supplement s1.), 213–283.

Hardison, D. M. (2003). Acquisition of second-language speech: Effects of visual cues, context, and talker variability. *Applied Psycholinguistics*, *24*(04), 495–522.

Hardison, D. M. (2005). Second-language spoken word identification: Effects of perceptual training, visual cues, and phonetic environment. *Applied Psycholinguistics*, *26*(04), 579–596.

Hardison, D. M. (2006). Effects of familiarity with faces and voices on second-language speech processing: components of memory traces. In *INTERSPEECH-2006*, *ISCA*, 2462-2465.

Hardison, D. M. (2009). Visual and auditory input in second-language speech processing. *Language Teaching*, *43*(01), 84.

Hattori, K., & Iverson, P. (2009). English /r/-/l/ category assimilation by Japanese adults: individual differences and the link to identification accuracy. *The Journal of the Acoustical Society of America*, *125*(1), 469–79.

Hattori, K., & Iverson, P. (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*. Retrieved from http://www.gavo.t.u-tokyo.ac.jp/L2WS2010/papers/L2WS2010_P2-04.pdf

Hawkins, S., & Midgley, J. (2005). Formant frequencies of RP monophthongs in four age groups of speakers. *Journal of the International Phonetic Association*, *35*(02), 183.

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.

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.

Hazan, V., & Sennema, A. (2007). The effect of visual training on perception of non-native phonetic contrasts. In Proceedings of the XVIth International Congress of Phonetic Sciences, pp.1585–1588.

Hazan, V., Kim, J., & Chen, Y. (2010). Audiovisual perception in adverse conditions: Language, speaker and listener effects. *Speech Communication*, *52*(11-12), 996–1009.

Herd, W., Jongman, A. & Sereno, J. (2013). Perceptual and production training of intervocalic /d, ɾ, r/ in American English learners of Spanish. J. Acoust. Soc. Am. 133(6), 4247- 4255.

Hillenbrand, J. M., Clark, M. J., & Houde, R. A. (2000). Some effects of duration on vowel recognition, *108*(6), 3013–3022.

Højen, A., & Flege, J. E. (2006). Early learners' discrimination of second-language vowels. *The Journal of the Acoustical Society of America*, *119*(5), 3072.

Ingram, J. C., & Park, S.-G. (1997). Cross-language vowel perception and production by Japanese and Korean learners of English. *Journal of Phonetics*, *25*(3), 343–370.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Kettermmann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, *87*, 47–57.

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.

Iverson, P. & Evans, B. G. (2007). Learning English vowels with different first-language vowel system: Perception of formant targets, formant movement, and duration. *The Journal of the Acoustical Society of America*, *122*(5), 2842–54.

Iverson, P., & Evans, B. G. (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, P., Pinet, M., & Evans, B. G. (2012). Auditory training for experienced and inexperienced second-language learners: Native French speakers learning English vowels. *Applied Psycholinguistics*, *33*(01), 145–160.

Jenkins, J.J. (1979). Four points to remember: A tetrahedral model of memory experiments. In Cermak, L.S. & Craik, F.I. (Eds) *Level of processing in human memory* (pp. 429-446). Oxford, England: Lawrence Erlbaum.

Kaplan, H., Bally, S. J. & Garrteson, C. (2010). *Speechreading: A way to improve understanding (2nd ed.)*. Washington, DC: Gallaudet University Press.

Kewley-Port, D; Akahane-Yamada, Reiko; Aikawa, K. (1996). Intelligibility and acoustic correlates of Japanese accented English vowels. *4th International Conference on Spoken Language Processing*. Pp.450-453. Philadelphia, PA, USA.

Kim, Y., & Hazan, V. (2010). Individual variability in the perceptual learning of L2 speech sounds and its cognitive correlates. *New Sounds. 2010 Proceeding.* Poznán, Poland.

Knight, R.-A. (2011). Assessing the temporal reliability of rhythm metrics. *Journal of the International Phonetic Association*, *41*(03), 271–281.

Kondaurova, M. V, & Francis, A. L. (2008). The relationship between native allophonic experience with vowel duration and perception of the English tense/lax vowel contrast by Spanish and Russian listeners. *The Journal of the Acoustical Society of America*, *124*(6), 3959.

Kondaurova, M. V, & Francis, A. L. (2010). The role of selective attention in the acquisition of English tense and lax vowels by native Spanish listeners : Comparison of three training methods. *Journal of Phonetics*, *38*(4), 569–587.

Kricos, P.B. & Lesner, A. (1982). Differences in visual intelligibility across talkers. *The Volta Review*, *84*, 219 – 225

Kuhl, P. K.**,** Williams, K. A., Lacerda, F., Stevens, K. N. & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, 255, 606-608.

Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(22), 11850–7.

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature reviews. Neuroscience*, *5*(11), 831–43.

Ladefoged, P., & Maddieson, I. (1996). *The Sounds of the World's Language*. Blackwell, Malden, MA.

Ladefoged, P. (2006). *A course in phonetics* (5th ed.). Boston, MA: Thomson Wadsworth.

Lambacher, S. G., Martens, W. L., Kakehi, K., Marasinghe, C. a., & Molholt, G. (2005). The effects of identification training on the identification and production of American English vowels by native speakers of Japanese. *Applied Psycholinguistics*, *26*(02), 227–247.

Lengeris, A. (2009). Perceptual assimilation and L2 learning: Evidence from perception of Southern British English Vowels by native speakers of Greek and Japanese. *Phonetica*, *66*(3), 169–87.

Lengeris, A., & Hazan, V. (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), 3757–68.

Lenneberg, E. H. (1967). *Biological foundations of language*. Oxford, England: Wiley.

Lesner, S., & Kricos, P. (1981). Visual vowel and diphthong perception across speakers. *Journal of the Academy of Rehabilitative Audiology*, *14*, 252–258.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical Society of America*, *49*, 467–477.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America*, *94*(3 Pt 1), 1242–55.

Lively, S. E., Pisoni, D. B., Yamada, R. a, Tohkura, Y., & Yamada, T. (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), 2076–87.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (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.

Luce, P., & Pisoni, D. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and hearing*, *19* (1), 1–36.

Massaro, D. W., & Cohen, M. M. (1983). Evaluation and integration of visual and auditory information in speech perception. *Journal of experimental psychology. Human perception and performance*, *9*(5), 753–71.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Erlbaum, Hillsdale, NJ.

Massaro, D. W., Cohen, M. M., Gesi, A., Heredia, R., & Tsuzaki, M. (1993). Bimodal speech perception: An examination across languages. *Journal of Phonetics, 21*, 445–478.

Massaro, D. W., Cohen, M. M., & Smeele, P. M. (1995). Cross-linguistic comparisons in the integration of visual and auditory speech. *Memory & cognition*, *23*(1), 113–31.

Massaro, D. W., Bosseler, A., & Light, J. (2003). Development and evaluation of a computer-animated tutor for language and vocabulary learning. *5th International Congress of Phonetic Sciences*, Barcelona.

MacKay, I. R., Meador, D., & Flege, J. E. (2001). The identification of English consonants by native speakers of Italian. *Phonetica*, *58*(1-2), 103–25.

McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.

Mora, J. C., & Fullana, N. (2007). Production and perception of English /iː/-/ɪ/ and /æ/-/ʌ/ in a formal setting : Investigating the effects of experience and starting age. *Proceedings of the 16th International Congress of Phonetics Sciences*. Saarbrücken, 1613–1616.

Moreiras, C. (2006). *An acoustic study of vowel change in female adult speakers of RP*. Unpublished undergraduate dissertation, University College London.

Morrison, G. S. (2002). Perception of English /i/ and /ɪ/ by Japanese and Spanish listeners: Longitudinal results. In G. S. Morrison & L. Zsoldos (Eds.), *Proceedings of the North West Linguistics Conference* (pp. 29–48). Burnaby, BC, Canada: Simon Fraser University Linguistics Graduate Student Association.

Morrison, G. S. (2008). L1-Spanish speakers' acquisition of the English /i/–/ɪ/ contrast: Duration-based perception is not the initial developmental stage. *Language & Speech*, 51, pp. 285–315.

Morrison, G. S. (2009). L1-Spanish Speakers' Acquisition of the English /i/–/ɪ/ contrast II: Perception of Vowel Inherent Spectral Change1. *Language and Speech*, *52*(4), 437–462.

Munro, M. J., Flege, J. E., & Mackay, I. R. a. (1996). The effects of age of second language learning on the production of English vowels. *Applied Psycholinguistics*, *17*(03), 313.

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.

Navarra, J., Alsius, A., Velasco, I., Soto-Faraco, S., & Spence, C. (2010). Perception of audiovisual speech synchrony for native and non-native language. *Brain Research*, *1323*, 84–93.

Nishi, K., & Kewley-Port, D. (2007). Training Japanese listeners to perceive American English vowels: influence of training sets. *Journal of Speech, Language, and Hearing Research : JSLHR*, *50*(6), 1496–509.

Nygaard, L. (2005). Perceptual integration of linguistic and non-linguistic properties of speech. In D.B. Pisoni & Remez, R.E. (Eds.) *The Handbook of Speech Perception,* (pp.390-414). Oxford: Blackwell Publishing Ltd.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & psychophysics*, *60*(3), 355–76.

O'Brien, I., Segalowitz, N., Freed, B., & Collentine, J. (2007). Phonological Memory Predicts Second Language Oral Fluency Gains in Adults. *Studies in Second Language Acquisition*, *29*(04), 557–581.

Ortega-llebaria, M., Faulkner, A., & Hazan, V. (2001). Auditory-visual L2 speech perception : effects of visual cues and acoustic-phonetic context for Spanish learners of English. *Speech , Hearing and Language : work in progress Vol.13*, 40–51.

Owens, E., & Blazek, B. (1985). Visemes observed by hearing-impaired and normal-hearing adult viewers. *Journal of speech and hearing research*, *28*(3), 381–93.

Perrachione, T. K., Lee, J., Ha, L. Y. Y., & Wong, P. C. M. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *The Journal of the Acoustical Society of America*, *130*(1), 461–72.

Piske, T., Flege, J.E., Mackay, R. A. & Meador, D. (2002). The production of English vowels by fluent early and late Italian-English bilinguals. *Phonetica*, *59*(1), 49–71.

Piske, T., Flege, J.E, MacKay, R. A & Meador (2010) Investigating native and non-native vowels produced. *New Sounds 2010*.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org.

Rallo Fabra, L., & Romero, J. (2012). Native Catalan learners' perception and production of English vowels. *Journal of Phonetics*, *40*(3), 491–508.

Reisberg, D., Mclean, J & Goldfield, A. (1987). Easy to hear but hard to understand. In Dodd, B. & Campbell, R. (Eds.) *Hearing by the eye: The psychology of lip-reading*, (pp. 97-114). Hillsdale, NJ, England: Lawrence Erlbaum Associates.

Remez, R. E., Fellowes, J. M., & Nagel, D. S. (2007). On the perception of similarity among talkers. *The Journal of the Acoustical Society of America*, *122*(6), 3688–96.

Robert-Ribes, J., Schwartz, J. L., Lallouache, T., & Escudier, P. (1998). Complementarity and synergy in bimodal speech: auditory, visual, and audio-visual identification of French oral vowels in noise. *The Journal of the Acoustical Society of America*, *103*(6), 3677–89.

Rosenblum, L. D. & Gordon, M. S. (2001). The generality of specificity: Some lessons from audiovisual speech. *Behavioral & Brain Sciences*, 24, 239–40.

Rosenblum, L. D. (2005) Primacy of multimodal speech perception. In D.B. Pisoni & R.E. Remez (Eds.) *The Handbook of Speech Perception*, (pp.51-78). Oxford: Blackwell Publishing Ltd.

Rosenblum, L. D. (2008). Speech Perception as a Multimodal Phenomenon. *Current Directions in Psychological Science*, *17*(6), 405–409.

Sadowsky, S. M. (2012). *Naturaleza fonética y estratificación sociolingüística de los alófonos vocálicos del castellano de Concepción (Chile)*. Unpublished PhD thesis, Universidad de Concepción, Chile.

Schwartz, J.-L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition*, *93*(2), B69–78.

Sekiyama, K., & Tohkura, Y. (1991). McGurk effect in non-English listeners: few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *The Journal of the Acoustical Society of America*, *90*(4 Pt 1), 1797–805.

Sekiyama, K., & Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics*, 21, 427–444.

Sekiyama, K. (1997). Cultural and linguistic factors in audiovisual speech processing: the McGurk effect in Chinese subjects. *Perception & psychophysics*, *59*(1), 73–80.

Sekiyama, K., Burnham, D., Tam, H., & Erdener, D. (2003). Auditory-visual speech perception development in Japanese and English speakers. *International Conference on Audio-Visual Speech Processing*.

Sekiyama, K., & Burnham, D. (2008). Impact of language on development of auditory and visual speech perception. *Developmental Science, 2,* 306– 320.

Sennema, A., Hazan, V., & Faulkner, A. (2003). The role of visual cues in L2 consonant perception. *Proc. 15th ICPhS,* 135–138.

Shinohara, Y., & Iverson., P. (2012). Effects of Identification and Discrimination Training Techniques on English /r/ - /l/: Perception and Production for Japanese Speakers. *Bilingual and Multilingual Interaction Conference*, Bangor, UK.

Skipper, J. I., Van Wassenhove, V., Nusbaum, H. C., and Small, S. L. (2007). Hearing lips and seeing voices:howcortical areas supporting speech production mediate audiovisual speech perception.Cereb.Cortex 17, 2387–2399.

Stevens, K. N., & Halle, M. (1967). Remarks on analysis by synthesis and distinctive features. In W. Wathen-Dunn (Ed.) *Models for the Perception of Speech and Visual Form* (pp. 88-102). MIT, Cambridge: MA.

Stevens, K. N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In Denes, P. B. Jr. and David, E. E.(Eds.). *Human Communication: A Unified View* (pp. 51–66). McGraw-Hill, New York.

Strange, W., & Dittmann, S. (1984). Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception & Psychophysics*, *36*(2), 131–45.

Strange, W. (1992). Learning non-native phoneme contrasts: interactions among subject, stimulus and tasks variables. In Tohkura, Y., Vatikiotis-Bateson, E. & Sagisaka Y. (Eds.) *Speech Perception, Production and Linguistic Structure* (pp 198-217). Oxford: IOS.

Strange, W., Trent, S. A., Nishi, K., Jenkins, J. J., & Strange, W., Akahane-yamada, Kubo, R., Trent, S., Nishi K & Jenkins, J. (1998). Perceptual assimilation of American English vowels by Japanese listeners. *Journal of Phonetics*, *26*(4), 311–344.

Sueyoshi, A., & Hardison, D. M. (2005). The Role of Gestures and Facial Cues in Second Language Listening Comprehension. *Language Learning*, *55*(4), 661–699.

Summerfield, Q. (1979). Use of visual information for phonetic perception. *Phonetica, 36,* 314-331.

Summerfield, A. Q. 1987 Some preliminaries to a theory of audiovisual speech processing. In B. Dodd & R. Campbell (Eds.) *Hearing by eye* (pp. 58–82). Hove, UK: Erlbaum Associates.

Summerfield, Q. (1992). Lipreading and audio-visual speech perception. *Philosophical transactions of the Royal Society of London. Series B, Biological Sciences*, *335*(1273), 71–8.

Thomas, E. R. & Tyler, K. (2007). NORM: The Vowel Normalisation and Plotting Suite. Retrieved from http://ncslaap.lib.ncsu.edu/tools/norm/

Thomson, R. I. (2011). Computer assisted pronunciation training: Targeting second language vowel perception improves pronunciation. *CALICO Journal*, *28*, 744–765.

Thomson, R. I. (2012). Improving L2 Listeners' Perception of English Vowels: A Computer-Mediated Approach. *Language Learning*, *62*(4), 1231–1258.

Traunmüller, H., & Öhrström, N. (2007). Audiovisual perception of openness and lip rounding in front vowels. *Journal of Phonetics*, 1–22.

Van Wassenhove, V., Grant, K.W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the Natural Academy of Sciences of USA. 102(4)*, 1181–1186.

Walden, B. E., Erdman, S. A., Montgomery, A. A., Schwartz, D. M., and Prosek, R. A., (1981). Some effects of training on speech recognition by hearing-impaired adults, *J. Speech and Hearing Research*, *24*, 207-216

Wang, X., & Munro, M. J. (1999). The perception of English tense-lax vowel pairs by native Mandarin speakers: The effect of training on attention to temporal and spectral cues. In J. J. Ohala et al. (Eds*.), Proceedings of the 14th International Congress of Phonetic Sciences*: San Francisco (pp.125–128). Berkeley: University of California Berkeley.

Wang, X., & Munro, M. J. (2004). Computer-based training for learning English vowel contrasts. *System*, *32*(4), 539–552.

Wang, Y., Behne, D. M., & Jiang, H. (2008). Linguistic experience and audio-visual perception of non-native fricatives. *The Journal of the Acoustical Society of America*, *124*(3), 1716–26.

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.

Wik, P. (2011). The Virtual Language Tutor: models and applications for language learning using embodied conversational agents. Unpublished doctoral thesis, KTH School of Computer Science and Communication, Stockholm.

Wik, P. & Hjalmarsson, A. (2009). Embodied conversational agents in Computer Assisted Language Learning. *Speech Communication, 51(10)*, 1024-1037.

Yamada, R. a, & Tohkura, Y. (1992). The effects of experimental variables on the perception of American English /r/ and /l/ by Japanese listeners. *Perception & psychophysics*, *52*(4), 376–92.

Appendix A

Overall mean at pre and post test per mode (A, AV, V) per training group (AT, AVT, VT).

Audio training group (AT). Pre and post test overall means

	N	Minimum	Maximum	Mean	Std. Deviation
Audio_pre test	17	45.45	84.09	**67.3**	**11.8**
AV_pre test	17	40.91	82.95	**64.7**	**10.4**
Video_pre test	17	36.36	53.41	**46.6**	**5.2**
Audio_post test	17	62.50	93.18	**75.3**	**7.1**
AV_post test	17	62.50	88.64	**73.9**	**7.6**
Video_post test	17	37.50	63.64	**50.6**	**6.2**

Audio-visual training group (AVT). Pre and post test overall means

	N	Minimum	Maximum	Mean	Std. Deviation
Audio_pre test	14	43.18	84.09	**61.2**	**12.9**
AV_pre test	14	43.18	81.82	**61.5**	**12.1**
Video_pre test	14	35.23	53.41	**43.7**	**5.7**
Audio_post test	14	55.68	88.64	**70.7**	**10.5**
AV_post test	14	48.86	90.91	**69.8**	**10.7**
Video_post test	14	29.55	57.95	**45.6**	**7.8**

Video-alone training group (VT). Pre and post test overall means

	N	Minimum	Maximum	Mean	Std. Deviation
Audio_pre test	16	46.59	77.27	**62.8**	**8.0**
AV_pre test	16	43.18	76.14	**62.7**	**9.2**
Video_pre test	16	37.50	55.68	**45.2**	**4.5**
Audio_post test	16	56.82	80.68	**70.2**	**6.0**
AV_post test	16	51.14	79.55	**68.8**	**7.0**
Video_post test	16	35.23	62.50	**47.7**	**6.4**

Appendix B

List of sentences used in the True-or-False sentence test presented in audio-visual mode. Each sentence was presented 3 times, with 1 male and 2 different female versions (total 146 sentences)

Sentences in the practice phase (20 presentations). Some were repeated (*) to have a male and female version.

1. The teabag is in the cap (*)
2. The teabag is in the cup (*)
3. They'll get money from the bank
4. They'll get money from the bunk
5. Cod is my favourite fish
6. Cord is my favourite fish
7. I can only sleep in my own bed
8. I can only sleep in my own bird
9. My main meal is lunch
10. My main mill is lunch
11. I hear birds singing now (*)
12. I hear beds singing now (*)
13. They live in a big house (*)
14. They leave in a big house (*)

Test

1. The ship is in the sea
2. The sheep is in the sea
3. The sheep is eating the grass
4. The ship is eating the grass
5. The hill is covered with trees
6. The heel is covered with trees
7. I have a blister on my heel
8. I have a blister on my hill
9. That was the best party of my life
10. That was the burst party of my life
11. The kid is crying because of the burst balloon
12. The kid is crying because of the best balloon
13. I have a large debt with the bank
14. I have a large dirt with the bank
15. You've got some dog dirt on your shoes
16. You've got some dog debt on your shoes
17. She didn't wear her new hat today
18. She didn't wear her new heart today
19. My heart is beating hard
20. My hat is beating hard
21. Lady Gaga is first in the charts today
22. Lady Gaga is first in the chats today

23. I've enjoyed our chats on the phone
24. I've enjoyed our charts on the phone
25. There are lots of trees in the park
26. There are lots of trees in the pack
27. I need to pack for my journey now
28. I need to park for my journey now
29. I need another hair cut
30. I need another hair cart
31. I'm putting the groceries in the cart
32. I'm putting the groceries in the cat
33. The cat was meowing all night
34. The cut was meowing all night
35. The cut on my leg is really deep
36. The cat on my leg is really deep
37. I want a cup of my favourite coffee
38. I want a cap of my favourite coffee
39. I need a cap to protect me from the sun
40. I need a cup to protect me from the sun
41. We heard some shots in the street
42. We heard some shorts in the street
43. The kids are wearing shorts
44. The kids are wearing shots
45. The boat is near the port
46. The boat is near the pot
47. You need a big pot to cook
48. You need a big port to cook