Audio-visual speech perception of plosive consonants by CG learners of English

Abstract

Second language (L2) speech perception can be a challenging process as listeners have to cope with imperfect auditory signals and imperfect L2 knowledge. However, the aim of L2 speech perception is to extract linguistic meaning and enable communication between interlocutors in the language of input. Normal-hearing listeners can perceive and understand the auditory message(s) conveyed effortlessly regardless of distortions and background noise as they can endure a dramatic decrease in the amount of spectral and temporal information present in the auditory signal. In their attempt to recognise speech, listeners can be substantially assisted by looking at the face of the speaker. Visual perception is important even in the case of intelligible speech sounds indicating that auditory and visual information should be combined. The present study examines how audio-visual integration affects Cypriot-Greek (CG) listeners' recognition performance of plosive consonants on word-level in L2 English. The participants were 14 first language (L1) CG users, who were non-native speakers of L2 English. They completed a perceptual minimal set task requiring the extraction of speech information from unimodal auditory stimuli, unimodal visual stimuli, bimodal audio-visual congruent, and incongruent stimuli. The findings indicated that overall performance was better in the bimodal congruent task. The results point to the multisensory speech-specific mode of perception, which plays an important role in alleviating the majority of the moderate to severe L2 comprehension difficulties. CG listeners' success seems to depend upon the ability to relate what they see to what they hear.

Keywords: audio-visual speech perception; plosive consonants; Cypriot-Greek; second language

1. Investigating the modes to hearing

1.1. Speech perception

An important question in the study of speech perception refers to the conversion of the continuously varying speech signal into a sequence of discrete linguistic units such as phonemes, phones, and/or allophones (Bien et al., 2009; Bien & Zwitserlood, 2013; Hickok & Poeppel, 2007; Kkese, 2016; Kkese & Petinou, 2017a,b; Kkese & Karpava, 2019; Obleser & Eisner, 2009). When listeners are exposed to speech, they tend to match each phoneme to a corresponding stretch of sounds in the utterance (Chomsky & Miller, 1963).. Nonetheless, factors such as coarticulation of adjacent phonemes and further contextual effects, influence speech perception while it is very difficult to identify the acoustic cues that match specific

phonemes irrespectively of the surrounding context (Stevens & Blumstein, 1981). Listeners, however, manage to perceive and understand the message(s) conveyed effortlessly by encoding the continuous acoustic cues, mapping these to phoneme categories, and accessing lexical entries. This suggests that phonemes could be audibly distinguished by several acoustic cues.

Segmental perception is particularly challenging for L2 learners, especially when the inventories of the L1 and L2 contain different phonemes or allophones (Dimitriou 2018; Iverson et al., 2003; Kkese, 2016; Lengeris, 2009). Just like L1 speakers, L2 learners also rely heavily on the acoustic signal during speech perception, although developing an acute awareness of acoustic distinctions in an L2 takes time and needs exposure to the L2 sounds (Flege & Liu, 2001; Flege, 2009). If their acoustic understanding of the L2 distinctions is not adequately developed, then their reliance on the acoustic signal may lead to inaccurate perception.

At this point, a brief description and comparison of the consonantal systems of British English and Standard Modern Greek (SMG)/Cypriot-Greek (CG), which constitute the language systems under investigation, seems mandatory to understand the differences between them. SSBE (Standard Southern British English) was chosen for comparison as a form of reference speech, since it is the variety most commonly used as the pronunciation model across the world (Deterding, 1997) while also being an extensively examined variety chosen in previous studies.

SSBE involves twenty-four consonants, with many having both voiced and voiceless pairs (Ladefoged & Ferrari Disner, 2012). Plosive or stop consonants are divided into the bilabial /p b/, alveolar /t d/, and velar /k g/. Fricative consonants are classified into the labiodental /f v/, dental /θ ð/, alveolar /s z/, palato-alveolar /ʃ ʒ/, and glottal /h/. Affricates include only two members, that is, the palato-alveolar /ʧ ʤ/. Nasal consonants are distinguished into the bilabial /m/, alveolar /n/, and velar /ŋ/. Approximants include the alveolar /ɹ/, the palatal /j/, and the labio-velar /w/. The last category consists of the alveolar lateral approximant /l/. On the other hand, SMG has a richer consonantal system of twenty-eight sounds, that is, /p b t d c ɟ k g ʦ ʣ f v θ ð s z ç ʝ x ɣ m n ɲ ŋ j r l ʎ/ while CG involves even more sounds and specifically fifty-one consonants (Kkese, 2020a). These consist of plosives, which could be further distinguished into the labial /p pʰː b/, alveolar /t tʰː d/, palatal /c cʰː ɟ/, and velar /k kʰː g/. Fricatives could be classified into the labial /f fː v vː/, (inter)dental /θ θː ð ðː/, alveolar /s sː z zː/, postalveolar /ʃ ʃː ʒ ʒː/, palatal /ç çː j jː/, and velar /x xː ɣ ɣː/. Affricates consist of the alveolar /ts/ and postalveolar /ʧ ʧː ʤ/. Nasals could be divided into the labial /m mː/, alveolar /n nː/, palatal /ɲ/, and velar /ŋ/. Lateral approximants include the alveolar /l lː/ and the palatal /ʎ/.

Lastly, CG consists of the alveolar tap /ɾ/ and the alveolar trill /r/; the alveolar tap seems to be the dominant allophone of /r/ in different prosodic contexts such as clusters, between vowels, and in singleton phrases and/or word-initially (Baltazani & Nicolaidis, 2013).

Since the perception of consonants depends on features such as voicing, manner and place of articulation, the differences between the phoneme inventories of the two languages are a source of difficulty for CG learners of English, especially in combination with the phonetic realisations of sounds in each language. For example, even though both languages have a distinction between voiced and voiceless plosives, SMG plosives are distinguished between voiceless unaspirated and fully voiced plosives (Botinis, Fourakis and Prinou, 2000; Arvaniti, 2001; 2007; Kainada, 2012), while CG has a third category as well, namely voiceless aspirated plosives. Furthermore, according to some descriptions of CG, voiced plosives do not exist in the dialect, or are realised as prenasalised voiced plosives, as in [mbaˈmbas] (Arvaniti, 2006; Kappa, 2002; Kkese, 2016; Terkourafi, 2001; Newton, 1972). The three realisations have phonemic status in CG, given that they can occur in minimal triplets (Kkese & Petinou, 2017a). English on the other hand, distinguishes voiceless aspirated and not fully voiced plosives. More specifically, even though English /b d g/ are phonologically described as voiced, they are realised as voiceless in initial position (Docherty, 1992). As a result of this difference in phonetic realisation between the two languages, CG and SMG listeners tend to identify the English voiced plosives as their voiceless counterparts /p t k/ (Kkese, 2016; Kkese & Petinou, 2017a; Lengeris & Nicolaidis, 2016).

Kkese (2016) examined the difficulties CG users of L2 English experience with the voiced plosive consonants /b d g/ in the target language. The two auditory tasks presented to the participants involved minimal pairs at the word and sentence level. The first task was a two-alternative forced task in which participants had to circle the word they could hear; the second was a words-in-sentences task in which they had to fill in two gaps in the same sentence involving a minimal pair. Performance was significantly better with reference to voiceless plosive consonants compared to their voiced counterparts at the word-initial position (i.e., pacing-basing, towering-dowering, crammer-grammar), word-medial position (i.e., calipers-calibers, sighting-siding, lacquered-laggard), and word-final position (i.e., tripe-tribe, squat-squad, broke-brogue). The findings of the study suggested that these difficulties may be the outcome of Voice Onset Time (VOT) cues since the VOTs are more difficult to produce in the lead voicing region (voiced plosive consonants).

Kkese and Petinou (2017a) investigated the perception of plosive consonants by CG speakers. Participants were asked to listen to recorded sentences containing the target words

and fill in the blanks with the word they believed they had heard. The results of this study suggest that participants' performance was considerably better in the perception of voiceless plosives. As concerns voiced plosives, most of the incorrect responses provided involved substitutions with voiceless consonants. The findings of this study supported predictions concerning the perception skills of these learners. The participants were more successful in identifying voiceless compared to voiced plosives. It was also observed that CG learners tended to substitute voiced plosives with either a voiceless or a prenasalised counterpart.

Lengeris and Nicolaidis (2016) examined the identification and production of the full set of English consonants by native SMG learners in one quiet and two noise conditions. The participants were found to have more difficulties with English plosives (mainly voiced plosives which were identified as their voiceless counterparts), affricates (especially /dʒ/) and fricatives (especially /ʒ/), across noise conditions. Most successfully produced consonants were /b/, /n/, /l/, /ɹ/, /j/ and /w/, while the most problematic English consonants were /p/, /t/, /k/, /ʧ/, /ʤ/, /θ/, /s/, /ʃ/ and /ŋ/ (mostly confused with /b/, /d/, /g/, /t/, /d/, /f/ and /s/, /z/ and /ʒ/, /s/, and /g/ respectively). A comparison between the results from the English consonant identification by the Greek speakers and the results from the identification scores of English listeners revealed that difficulties with English consonants are not always the same across modalities, since some consonants were easy to identify but difficult to produce, such as /p/, /k/ and /θ/, while others were difficult to identify but easy to produce, such as /d/.

This brief comparison between SSBE and SMG/CG also suggests that there is a considerable discrepancy between phonemes and graphemes in the two language systems, creating several difficulties to listeners (Kkese, 2020a,b,c). In SSBE, the forty-four phonemes (twelve-four consonants, twelve monophthongs, eight diphthongs) can be written with only twenty-six graphemes; adhering to a one-to-one correspondence between phonemes and graphemes may not be ideal. In SMG, the grapheme-phoneme correspondence could be applied to a larger extent since the thirty-three phonemes (twenty-eight consonants, five vowels) could be written with twenty-four graphemes. Even though CG employs the same graphemes as SMG, it employs more phonemes. Specifically, in CG there are approximately fifty-six phonemes (fifty-one consonants, five vowels), which are written with twenty-four graphemes. Nevertheless, vowels in SMG and CG may be more challenging as some may have different graphical representations (i.e., /i/ spelled as ι, η, υ, οι, ει, υι and /e/ spelled as ε, αι). This discrepancy between phonemes and graphemes creates difficulties to listeners that suggests a problem for their orthographic skills. Given that the present study is about speech perception

as a process based on auditory and visual information, examining these difficulties may provide useful insights into the L2 acquisition process.

1.2. Visual perception

Whereas the previous section briefly described the auditory aspect associated with the production of speech sounds, this section refers to the visual aspect and how information about speech sounds is obtained through a viseme. Visemes refer to visual representations that are produced when the speaker is talking; in these cases, the listeners are trying to map visemes to phonemes to help understand what a speaker is saying (Binnie et al., 1974). Visemes result from groups of phonemes, which have the same visual appearance (Bozkurt et al., 2007; Neti et al., 2000; Potamianos et al., 2003; Saenko, 2004). Such visual information may involve the speaker's teeth, lips, and tongue and it is available even when listeners encounter hearing loss and/or under noisy conditions, where some phonemes are lost or become difficult to perceive (MacLeod & Summerfield, 1987; Summerfield et al., 1989). Evidence is provided by Vatikiotis-Bateson et al. (1998) as well as Lansing and McConkie (2003), who suggest that individuals with hearing loss continue to look at the speaker's eyes when faced with conditions of highest levels of noise rather than merely looking at the speaker's mouth. Concerning English, Lucey et al. (2004) provide a table mapping the possible phonemes to 14 viseme classes (Table 1). Other accounts, however, support a different number of viseme classes and a total phoneme number (Bozkurt et al., 2007; Hazen et al., 2004; Jeffers & Barley, 1971; Lee & Yook, 2002; Neti et al., 2000).

Overall, vowels are believed to be easier to distinguish when compared to consonants since each vowel is produced with a separate oral cavity shape (Jackson, 1988; Markides, 1989). Co-articulation, however, may affect the shape of a viseme; adjacent phonemes and their shape of the viseme may result in the blending of speech sounds as in *pull /pʊl/* and put /pʊt/ due to the following consonant. With reference to consonants, an example may refer to the phonemes /l/ and /r/, which are acoustically close in English but visual difference may help distinguish which one is being pronounced by the speaker since these are generated using different visemes as in the low-frequency minimal pairs of *enamelling* [ɪˈnæməlɪŋ] and *enamouring* [ɪˈnæməɹɪŋ], *Mauritius* [məˈrɪʃəs] and *malicious* [məˈlɪʃəs]. Nonetheless, the mapping between viseme-to-phoneme is not usually straightforward but it may involve a many-to-one correspondence (Lucey et al., 2004). This is attributed to the fact that several phonemes may not be distinguished using only visual cues (Cappelletta & Harte, 2012). An example could be consonants differing in terms of voicing since these refer to very similar visemes even

though they are audibly different; however, it would be very challenging for listeners to distinguish visually the voiced /v/ and its voiceless counterpart /f/ as in *vault* [vɒlt] and *fault* [fɒlt] as well as *vinery* [ˈvaɪnəɹi] and *finery* [ˈfaɪnəɹi]. In such cases, when visual information is available before the auditory signal onset, phoneme perception is facilitated (Mitterer & Reinisch, 2016). A last point refers to homophenes, which are words that seem visually similar when spoken. According to Auer and Bernstein (1997), 40-60% of the words in English are similar when spoken while an example may involve the words *yes* [jes] and *grade* [gɹeɪd], which are audibly distinct but visually similar.

Visual information, as a result, seems to aid speech perception for three main reasons (Summerfield, 1987). These involve the enhancement of the speaker auditory source localisation, the inclusion of speech segmental information that supplements the auditory information, and the provision of complementary information concerning the place of articulation. With reference to the place of articulation, this occurs because the articulators are partially or fully visible and can help disambiguate consonants that are similar in terms of the rest of the parameters. Examples may involve the voiceless plosive /p/ (place of articulation: bilabial) and /k/ (place of articulation: velar), the voiced plosive /b/ (place of articulation: bilabial) and /d/ (place of articulation: alveolar), and the voiced nasal /m/ (place of articulation: bilabial) and /n/ (place of articulation: alveolar) (Massaro & Stork, 1998). Depending merely on acoustics will lead the listeners to several difficulties, while taking visibility of the articulators (*i.e.*, jaw and lower face muscle movement) into consideration will enhance speech perception (Smeele, 1996; Summerfield et al., 1989).

Table 1. Phoneme to viseme correspondences (based on Lucey et al., 2004)

Phoneme	Viseme	Phoneme	Viseme
P(p)	/p/	K(k)	/k/
B(b)		G(g)	
M(m)		N(n)	
EM(m̩)		L(l)	
F(f)	/f/	NX(ɾ̃)	
V(v)		HH(h)	
T(t)	/t/	Y(y)	
D(d)		EL(l̩)	
S(s)		EN(n̩)	
Z(z)		NG(ŋ̩)	
TH(θ)		IH(ɪ)	/iy/
DH(ð)		IY(i)	
DX(ɾ)		AH(ʌ)	/ah/
W(w)		AX(ə)	

WH(w)	/w/	AY(aɪ)	
R(r)		ER(ɝ)	/er/
CH(ʧ)	/ch/	AO(ɔ)	
JH(ʤ)		OY(ɔɪ)	
SH(ʃ)		IX(ɨ)	
ZH(ʒ)		OW(oʊ)	
EH(ɛ)	/ey/	UH(ʊ)	/uh/
EY(eɪ)		UW(u)	
AE(æ)		AA(ɑ)	/aa/
AW(aʊ)		SIL & SP	/sp/

2. Combining information associated with the L2 production of speech sounds

Individuals can produce speech due to the combined influence that elements of the vocal tract have on the air, which is being exhaled from the lungs. Sound waves are created when speakers vibrate the air being exhaled by using the vocal folds muscles. By changing the frequency of the vibrations, speakers can change the pitch of the sound. Moreover, articulators are employed to produce certain sounds such as vowels and consonants. The auditory characteristics of speech are, as a result, of paramount importance for forming an understanding of what a speaker is saying (Best, 1995; Flege, 1995). However, the visual characteristics of speech are also very important; it seems that there is a direct relationship between auditory and visual characteristics of speech (Cappelletta & Harte, 2012; Massaro et al., 1993; Potamianos et al., 2003). According to existing research, speech is processed in a bimodal nature where both the auditory and visual characteristics of speech are processed by the brain so that a more comprehensive understanding of the message being conveyed is achieved (Besle et al., 2004; Goldstein, 2013). The Ganong effect, where *?*esk is often realised as *desk* when an ambiguous /d/-/t/ blend replaces /d/ illustrates the influence of the word (lexical information) on the phoneme (Ganong, 1980). The McGurk effect (McGurk & MacDonald, 1976) further supports that auditory information integrates to the extent visual information influences what the hearers report listening (Rosenblum, 2019). In this effect, when there is conflicting auditory and visual input, listeners may report perceiving a sound that is absent in both types of input. For instance, when listeners hear /ba/ but at the same time they see the speaker articulating /ga/, they will most probably perceive this as /da/ combining the information from the two sources (McGurk & MacDonald, 1976). In Bertelson et al. (2003), listeners when seeing a speaker pronouncing /aba/ paired with an auditory /aba/-/ada/ blend, reported listening /aba/ when the ambiguous blend was not paired with the auditory context. Also, visual information is important for speech intelligibility in noise (Sumby & Pollack, 1954). Furthermore, visual information is very

important for the hearing-impaired since mouth movement has an important role in sign language and simultaneous communication (Marschark et al., 1998).

Concerning L2 perception, auditory and visual information could be a useful source of guidance (Burnham, 1998; Green, 1998; McGurk & MacDonald, 1976; Rosenblum, 2005; Summerfield, 1987), but it could also help with perceptual adjustment, which is the use of contextual cues to reconfigure internal representations of phoneme categories in the L2 leading to the adjustment and understanding of the conveyed linguistic message more easily. In this context, the use of auditory, visual, and audio-visual cues when perceiving sound contrasts that have different status in the L1 (first language) and L2 seem to be extremely important. There is disagreement, however, as to the use of visual cues by L2 listeners. Even though audiovisual perception in L2 learners has not been extensively investigated, studies of the McGurk effect suggest that non-native listeners use visual cues (Hardison, 1999). The McGurk effect was found to be weaker when the learners of English were Japanese and Chinese than for American English subjects (Sekiyama, 1997). However, Hardison (1998) examining audiovisual syllables with listeners from four different L1 backgrounds, found that the influence of the visual cue was 'dependent upon its information value, the intelligibility of the auditory cue and the assessment of similarity between the two cues' and was further affected by linguistic experience. Further, Hardison (1999) investigating the perception of /r/-/l/ by Japanese and Korean listeners of L2 English suggests that when visual cues are present relative to the auditory cue, the target phonemes are considerably improved in terms of intelligibility. On the other hand, Ortega-Llebaria et al. (2001) by studying Spanish learners of L2 English concluded that participants improved in consonants but not in vowels; visual cues may have different weights when cueing phonemic and allophonic distinctions.

Given the conflicting information as to the use of visual cues by L2 listeners, the present study aims to investigate the effectiveness of auditory and audio-visual cues for L2 perception for contrasts in which the visual cues vary in terms of informativeness for the L2 listeners. The study seems to be the first to investigate the audio-visual perception of the distinction between English /p b t d k g/ in CG L2 learners of English. Previous research comparing L1 CG and L2 English refers to auditory perception studies including plosive consonant perception on a word (Kkese, 2016; Kkese & Petinou, 2017a,b) or utterance level (Kkese, 2016), as well as consonant and vowel perception on a word level (Kkese & Karpava, 2019, 2021), without any examination of the effect of visual or combined audio-visual cues. The present research seeks to fill these gaps by investigating the following research questions:

1. What is the effect of visual cues in enhancing the perception of certain types of phonetic information such as voicing and place of articulation, particularly for plosive consonants?
2. To what extent combined audio-visual cues induce effects larger than those elicited by either cue on its own?

Based on the findings of previous studies, it is expected that listeners will mainly rely on the auditory cues while visual cues will facilitate the perception of plosives, particularly when the audio matches visual information. Furthermore, based on the differences between the L1 and L2 consonantal systems and the findings of previous studies, it is expected that these learners will have more difficulties with voiced compared to voiceless plosive identification. However, previous studies have not examined the role that visual cues may play in the identification of plosives.

3. Methodology

3.1. Demographic profile of the participants

One group of adult participants took part in this study. Before taking part in the study, the participants completed a questionnaire (see Appendix) aiming to gather information about their linguistic and sociolinguistic background. The experimental group included 14 non-native speakers of English consisting of seven males and seven females between the ages of 18 and 26 (M = 21.4, SD = 2.85), who were native speakers of CG. A larger sample was neither practical nor necessary given the nature of the study and the amount of data collected from each participant. All participants had normal hearing and vision; the population included second-, third-, and fourth-year undergraduate university students attending a variety of BA programmes (i.e., Accounting and Finance, Business Administration, Web Design and Development). They all had similar exposure to English as the participants reported living in Cyprus all their lives and being in constant contact with people who use CG while they graduated from public schools. Furthermore, as university students, all participants fulfilled the minimum English language proficiency requirements for undergraduate programs in English-speaking universities in Cyprus, which is the B1-B2 intermediate level of the Common European Framework. As a result, their L2 English proficiency was from low intermediate to advanced (based on their IELTS scores). Non-probability convenience sampling was used since participants were selected based on their possession of targeted characteristics and availability. Participation was on a completely voluntary basis and participants gave their written consent for their participation to the research. Further, they had the right to withdraw at any time while they were ensured about the confidentiality of their personal details. By

following this procedure, the ethical criteria of the research were assured. Table 2 offers some additional information about the participants.

Table 2. Participant information

Participant	Gender	Age
001	F	20
002	F	25
003	M	20
004	F	19
005	M	25
006	F	20
007	M	25
008	M	18
009	F	20
010	M	26
011	F	19
012	F	18
013	M	22
014	M	23

3.2. Stimuli

For the present study, four conditions were created; these consisted of the auditory- and visual-only stimuli constituting the unimodal part of the perceptual task and the congruent and incongruent audio-visual stimuli constituting the bimodal part of the study. The four conditions were created using digital audio and video recordings of a female adult who was a simultaneous bilingual speaker of CG and English. For the video recordings, the speaker was recorded from her shoulders up and was instructed to speak naturally in an emotional passive tone without moving her head. She was recorded at a 44.1 kHz sample while speaking into a microphone that fed directly into the sound card (IDT High-Definition Audio CODEC) of a laptop computer. The four conditions consisted of the same disyllabic pseudowords, with the addition of the video of the speaker pronouncing the words in the visual mode, and the integration of the audio and video for the audio-visual mode.

Twelve minimal sets were recorded for each condition; three minimal sets were included for each category, namely for the voiceless/voiced bilabial /p b/, the alveolar /t d/, and the velar /k g/ making up nine minimal sets in total while distractors focusing on the voicing contrast were also intermixed and made up three of the minimal set words. The consonants /p b t d k g/ were embedded within nonsense words of **C**VCV structure, where the target consonant was found word-initially; vowel was one of the following: /a e o/. Each minimal set

consisted of six pseudowords (in each set, 6x12=72 pseudowords in total for each task) and was parallel in distribution and contrastive differing in only one sound that could be found word-initially (i.e., **p**aga, **b**aga, **t**aga, **d**aga, **k**aga, **g**aga). Concerning distractors, two distractors were used for every ten presentations that were presented in random intervals. These included fricative consonants such as the labiodental [f] and [v], dental [θ] and [ð], and alveolar [s] and [z]. Auditory-only, video-only, and audio-visual congruent conditions were saved from the same formatted file. The audio-visual incongruent condition was made by dubbing the audio of one word onto the audio-visual congruent stimulus of its minimal set (*i.e.*, audio of *tero* dubbed onto video of *pero*). PsychoPy was used to create the experimental presentation of the edited stimuli and collect response and reaction time data. A PowerPoint presentation of 12 minimal sets for each condition was created that was presented using a Dell computer.

The phonemic contrasts emphasising on plosive consonants in L2 were selected due to their relative difficulty for CG speakers (Kkese, 2016, 2020c; Kkese & Petinou, 2017a,b). Plosive consonants are present in L1 and L2, but their phonetic identifications vary among the two languages. This type of task was chosen to eliminate any semantic information from the input (context-free) as it may have occurred if a conversation was presented instead. Although English is commonly used in Cyprus (Kkese & Lokhtina, 2017), examining auditory plosive consonant perception in L2 English and how it is influenced by visual cues is a corollary question.

3.3. Procedure

The perceptual task was carried out in a sound-attenuated room. Once participants completed the consent process, a background questionnaire was distributed collecting demographic information on the participants' linguistic and sociolinguistic background. Participants then completed the perceptual tasks in which four conditions were created for plosive perception: auditory-only, video-only, audio-visual congruent, and audio-visual incongruent. They were presented with the recordings of 72 plosive words in the different conditions through the program PsychoPy via a Dell computer. Participants had to select the word they could hear from six options (a target and five foils). A small pilot study involving three CG users of L2 English before the perceptual tasks ensured the naturalness and prototypicality of pronunciation. The findings of this small case preliminary study were not included in the results.

For the present study, participants were given a general introduction and instructions while they could ask any clarification questions at any point during the perceptual task. They

were seated in front of a computer in a quiet testing room with audio presented over earphones set to a comfortable volume (with a listening volume at 75dB). For the unimodal part of the task, participants listened to the stimuli (auditory-only condition) or watched the speaker's lips on the screen (visual-only condition) and selected what they perceived by clicking on the appropriate label, choosing their response from a set of six options. Participants could have a short two-minutes break between the two short unimodal tasks. After the unimodal part, there was a short pause halfway (5mins); after that, participants completed the bimodal part of the perceptual task. Participants for the two conditions had to report what they could perceive in the audio-visual condition. However, for the congruent bimodal stimuli, auditory and visual information matched while for the incongruent bimodal stimuli, the auditory and visual information did not match. Participants could also have a two-minutes break between the bimodal part if they wanted. They could control the transition between the slides; they had to indicate whether the pseudowords began by /p b t d k g/ by pressing the corresponding button on a response pad. Overall, testing lasted about 15 minutes while no feedback was provided in the four conditions, and the target stimuli were not repeated.

4. Results

4.1 Performance across conditions

Table 3 shows the total number of correct and incorrect responses as well as the mean correct score of participants and standard deviation in each of the four conditions (N = 14). Figure 1 offers a visual representation of the number of correct responses per test.

Table 3. Correct and incorrect responses to stimuli in each condition

Condition	Total correct	Total incorrect	Std. Deviation
Unimodal Audio	75 (59.52%)	51 (40.48%)	1.336
Unimodal Video	46 (36.51%)	80 (63.49%)	2.016
Bimodal Congruent	118 (93.65%)	8 (6.35%)	.756
Bimodal Incongruent	74 (58.73%)	52 (41.27%)	.914

[Fig.1]

A one-way repeated measures analysis of variance (ANOVA) was conducted in order to assess participants' performance across the four conditions (Unimodal Audio, Unimodal

Video, Bimodal Congruent and Bimodal Incongruent) and evaluate the null hypothesis that there is no change in participants' performance in each condition. The independent variable CONDITION was set as the within-subject factor with four levels. For comparing the main effects for CONDITION, Confidence Interval Adjustment was selected and a Bonferroni correction was performed. The results of the ANOVA indicated a significant effect of CONDITION ($p < .001$, Wilks' Lambda = .026, $F = 137.811$, $\eta^2 = .974$), providing significant evidence to reject the null hypothesis.

Pairwise comparisons showed significant differences between Bimodal Congruent and each of the other conditions with $p < .001$ in all comparisons. A significant difference was also found between Unimodal Audio and Unimodal Video ($p = .020$), but the differences between Unimodal Audio and Bimodal Incongruent, and between Unimodal Video and Bimodal Incongruent did not reach significance ($p = 1$ and $p = .064$, respectively).

4.2 Identification of voiced vs. voiceless plosives

Table 4 shows the total number of correct and incorrect responses to stimuli with word-initial voiced or voiceless plosives in each condition. Figure 2 offers a visual representation of the mean correct responses of participants to voiced and voiceless plosives in each of the four conditions.

Table 4. Total correct and incorrect responses in voiced and voiceless plosives

Condition	**Voiced**		**Voiceless**	
	Correct	Incorrect	Correct	Incorrect
Unimodal Audio	44 (34.92%)	12 (9.52%)	31 (24.6%)	39 (30.95%)
Unimodal Video	22 (17.46%)	48 (38.1%)	24 (19.05%)	32 (25.4%)
Bimodal Congruent	68 (53.97%)	2 (1.59%)	50 (39.68%)	6 (4.76%)
Bimodal Incongruent	50 (39.68%)	6 (4.76%)	24 (19.05%)	46 (36.51%)

[Fig.2]

A two-way repeated measures ANOVA was conducted to evaluate the influence of CONDITION, VOICING and the CONDITION*VOICING interaction on participants' performance. CONDITION with four levels (Unimodal Audio, Unimodal Video, Bimodal Congruent and Bimodal Incongruent) and VOICING with two levels (Voiced and Voiceless) were set as the within-subject factors. For comparing the main effects for CONDITION,

Confidence Interval Adjustment was selected and a Bonferroni correction was performed. A significant main effect of CONDITION, VOICING and CONDITION*VOICING was observed (CONDITION: $p < .001$, $F = 38.101$; VOICING: $p < .001$, $F = 62.52$; CONDITION*VOICING: $p < .001$, $F = 6.674$).

Pairwise comparisons showed significant differences between Voiced and Voiceless plosive scores overall, with $p < .001$. Significant differences in the scores of participants between voiced and voiceless plosives were observed in all conditions except for Unimodal Video (Unimodal Audio: $p = .009$; Bimodal Congruent and Bimodal Incongruent: $p < .001$). Concerning Voiced plosives, significant differences were observed between Unimodal Video and all other conditions (with Audio: $p=.025$; with Bimodal Congruent: $p < .001$; with Bimodal Incongruent: $p = .004$), between Bimodal Congruent and all other conditions ($p < .001$ in all cases), but not between Unimodal Audio and Bimodal Incongruent ($p = .322$). In Voiceless plosives, significant differences were only observed between Bimodal Congruent and all other conditions (with Audio: $p = .010$; with Video: $p = .002$; with Bimodal Incongruent: $p < .001$).

4.3 Identification of plosives based on place of articulation

Figure 3 offers a visual representation of the mean correct responses of participants to target bilabial, alveolar and velar plosives in each of the four conditions.

[Fig.3]

A two-way repeated measures ANOVA with CONDITION and PLACE as the within-subject factors with four and three levels respectively (Unimodal Audio, Unimodal Video, Bimodal Congruent and Bimodal Incongruent for CONDITION and Bilabial, Alveolar, Velar for PLACE) was conducted to evaluate their influence on participants' scores. For comparing the main effects for CONDITION, Confidence Interval Adjustment was selected and a Bonferroni correction was performed. A significant main effect of CONDITION ($p < .001$, $F = 38.101$), PLACE ($p < .001$, $F = 41.777$) and the CONDITION*PLACE interaction ($p <.001$, $F = 6.381$) was observed. Pairwise comparisons showed significant differences between Bilabial and Alveolar plosives and between Alveolar and Velar plosives overall ($p < .001$ in both cases).

In the Unimodal Audio condition, differences reached significance between Bilabial and Alveolar plosives, and between Alveolar and Velar plosives ($p < .001$ in both cases), but not between Bilabial and Velar plosives. The same pattern of identification of plosives was observed in the Bimodal Incongruent condition: Bilabial-Alveolar and Alveolar-Velar plosive identification differed significantly ($p = .001$ and $p < .001$, respectively), but Velar-Bilabial

plosive identification did not. In the Unimodal Video and Bimodal Congruent conditions, no significant differences were observed based on the place of articulation of the plosive.

Pairwise comparisons within Bilabial plosives showed significant differences between Unimodal Audio and Unimodal Video ($p = .049$), between Unimodal Video and Bimodal Congruent ($p = .003$), and between Bimodal Congruent and Bimodal Incongruent ($p = .016$). Within Alveolar plosives, only the differences between Bimodal Congruent and each of the other conditions reached significance ($p < .001$ in all cases). In Velar plosives, significant differences were found between Unimodal Video and each of the other conditions (with Unimodal Audio: $p = .010$; with Bimodal Congruent: $p < .001$; with Bimodal Incongruent: $p = .007$), and between Unimodal Audio and Bimodal Congruent ($p = .034$).

5. Discussion

The purpose of this study was to evaluate the role of visual cues in speech perception by L1 CG users of L2 English in order to better understand the role of visual information in L2 speech acquisition. To achieve this, two questions were formulated:

1. What is the effect of visual cues in enhancing the perception of certain types of phonetic information such as voicing and place of articulation, particularly for plosive consonants?
2. To what extent combined audio-visual cues induce effects larger than those elicited by either cue on its own?

Overall, participants underwent three forms of phoneme boundary adjustments using auditory, audiovisual, and visual stimuli. Results for consonant perception tasks were analysed to answer these questions since consonants are characterised by a high-frequency structure and vocal-tract constriction (Ladefoged & Disner, 2012). This study focused on the perception of plosive consonants given that CG learners of English face difficulties with the specific phonemes. Because of the native phonology, the acoustic difference between two plosive consonants such as /p/ and /b/ may not be as easily recognised as that between two consonants of different manner and voicing such as /p/ and /v/. An acute awareness of acoustic distinctions takes time and needs exposure to the L2 sounds (Flege & Liu, 2001; Flege, 2009), and if L2 learners' acoustic understanding of the L2 distinctions is not adequately developed, then relying on the acoustic signal for this distinction may lead to inaccurate perception.

With reference to the first research question, the vast majority of studies into L2 speech perception has emphasised the auditory domain (Best, 1995; Flege, 1995). The results of the

study, though, indicate that the inclusion of visual speech perception could affect auditory L2 speech perception (Burnham, 1998; Green, 1998; McGurk & MacDonald, 1976; Rosenblum, 2005; Summerfield, 1987). Incorporating visual speech could provide the necessary multiple redundant cues to make L2 speech perception more robust. The study indicated that visual cues could enhance the perception of certain types of phonetic information such as voicing and place of articulation of plosive consonants. Regarding voicing, participants performed better in identifying voiced compared to voiceless plosives across all conditions (unimodal audio: voiced 34.92% *vs.* voiceless 24.6%), bimodal congruent: voiced 53.97% *vs.* voiceless 39.68%), bimodal incongruent: voiced 39.68% *vs.* voiceless 19.05%), except in the video-only condition, where there was no significant difference in the identification of voiced and voiceless plosives (unimodal video: voiced 17.46% *vs.* voiceless 19.05%). Differences in the identification of voiced and voiceless plosives did not reach significance in the unimodal video condition, suggesting that participants rely mostly on auditory information when distinguishing between them, while in the absence of such information, they seem unable to differentiate between the two categories. Even though better performance was expected for voiceless consonants since this category also exists in the L1 CG, participants obtained high accuracy in voiced plosives. Consequently, voiced plosives were not substituted with L1 sounds that most closely resemble them in place and manner of articulation and which participants could perceive based on their knowledge of the L1 phonetic inventory (Carlisle, 1994; Weinreich, 1953; Eckman, 1977).

When it comes to place of articulation, bilabial and velar plosives were more accurately identified compared to alveolar plosives (Mean: bilabial 8.14, velar 9.00 but alveolar 5.21) especially in the Unimodal Audio and Bimodal Incongruent conditions. Better performance in velar plosives can be explained since according to speech perception, velar plosives are produced with longer VOT values compared to bilabial and alveolar plosives (Ng et al., 2011; Liu et al., 2007; Lisker & Abramson, 1964); however, bilabial plosives are associated with the shortest VOT values (Klatt, 1975; Lisker & Abramson, 1964). The findings obtained for both voicing and place of articulation could imply that the perception of certain types of phonetic information could also be enhanced by visual cues; on the other hand, in the Unimodal Video and Bimodal Congruent conditions, no significant differences were observed. Given that sustaining voicing during a plosive is usually difficult, "[l]owering the velum during the stop closure allows air to be vented through the nose, slowing the build up of oral pressure, and thus facilitating voicing. In addition, voicing during an oral stop is radiated only through the neck and face, resulting in a low intensity acoustic signal, whereas lowering the velum allows sound to be radiated from the nose, resulting in greater intensity". (Flemming, 2005: 165). Taken

together, the effect of visual cues seems to be quite important when it comes to speech perception of plosive consonants.

Finally, concerning manner of articulation, the emphasis of the study was on plosive consonants, which are generally difficult sounds for L1 CG users of L2 English (Kkese, 2016). Even though other categories of consonants were not examined, the acoustic difference between plosives may not be as easily recognised as that between sounds of different manner and/or voicing, such as plosives and fricatives. Comparing plosives and fricatives could lead to different findings in terms of the place/manner of articulation since contrasts such as the /b/-/v/ are marked by visual cues which are highly contrastive for English listeners; the voicing contrast, on the other hand, may not be visually marked. Developing an awareness of the acoustic distinction between sounds of different manner and/or voicing takes time along with experience to accurate models of L2 speech. Even though several studies have suggested that L1 and L2 listeners rely on the acoustic signal as the primary source of information during speech perception (Best, 1995; Flege, 1995), depending merely on the acoustic signal may be difficult for the L2 listeners if they do not have a sufficiently developed sense of acoustic distinction between voiceless and voiced plosives. Different models within L2 speech perception have described how the phonetic and phonological organisation of the L1 affects the perception and production of L2 sounds, including the Perceptual Assimilation Model (PAM; Best, 1995) and its L2 extension (PAM-L2; Best & Tyler, 2007) and the Speech Learning Model (SLM; Flege, 1995) and its recent revision (SLM-r; Flege & Bohn, 2021). Learning an L2, though, could imply establishing an L2 specific representation involving L2 auditory and visual cues; auditory-visual training could, therefore, provide an optimal solution. L2 confusions associated with L1 allophonic relations could be a main target for auditory-visual training.

When it comes to the second research question, it seems that the combined audio-visual cues induce effects larger than those elicited by either cue on its own as overall performance was found to be better in the bimodal congruent task (correct responses 93.65% *vs.* incorrect responses 6.35%), when the auditory information matched that of the visual stimulus. Specifically, participants performed significantly worse in all other conditions, especially in the unimodal video condition (correct responses 36.51% *vs.* incorrect responses 63.49%), and their performance was similar in the unimodal audio (correct responses 59.52% *vs.* incorrect responses 40.48%) and bimodal incongruent conditions (correct responses 58.73% *vs.* incorrect responses 41.27%). The lack of a significant difference between the Unimodal Audio or the Unimodal Video compared to the Bimodal Incongruent condition suggests that the positive

effect that combined audio-visual cues entail in the identification of L2 plosives may be eliminated when these cues are not aligned. Several studies have examined the influence of the likelihood of an integrated multisensory percept by studying the timing of single auditory and visual events (Fujisaki & Nishida, 2005; Zampini et al., 2005) or simple periodic modulations of stimulus features (Recanzone, 2003; Spence & Squire, 2003; Fujisaki & Nishida, 2005). However, the temporal dynamics of one event or a repeating sequence are quite different compared to natural sounds, such as speech. This was addressed by Denison et al. (2013) by creating randomly timed sequences of discrete auditory-visual events. The study indicated that coherence discrimination was better for the unpredictable sequences than for predictable ones.

During speech perception, both L1 and L2 speakers seem to depend on the auditory signal as the primary source of information; however, a less developed sense of acoustic dimension could make this dependence on this signal alone more challenging for L2 speakers. Utilising visual information in speech perception may be the answer to these difficulties as the group in the current study was found to perform better in congruent stimulus. Integrating auditory and visual signals benefits L2 speech perception, while the fact that participants' performance was similar in the audio only and the bimodal incongruent conditions suggests that learners are strongly affected by the McGurk effect, since the added visual information does not improve learners' performance unless it is aligned with auditory cues. The findings, therefore, point to the multisensory speech-specific mode of perception, which plays an important role in alleviating the majority of the moderate to severe L2 comprehension difficulties. CG listeners' success seems to depend upon the ability to relate what they see to what they hear suggesting that L2 speech perception could benefit from explicit instruction of visual and auditory distinctions especially in the early stages of acquisition. While it is evident that learners can greatly benefit by incorporating visual alongside auditory cues, as expected, it is surprising that visual cues are rarely employed in the L2 classroom. Consequently, this study further supports that L2 teachers should focus on providing input that combines audio-visual information, rather than simply relying on auditory input as is usually the case with in-class listening tasks. Explicit and systematic instruction to phonological awareness could further benefit L2 learners as phonological awareness could have a positive impact on developing literacy and specifically the skills of spelling, writing, and reading.

6. Conclusion

This is the first study to investigate the role of visual cues as well as the role of the McGurk effect in the perception of segments by this group of learners. The findings of the current study

suggest that visual information plays an important role in L2 speech perception. Even though the participants were second to fourth year undergraduate students attending an English-speaking university and were, therefore, probably more 'language aware' and linguistically experienced than listeners at the early stages of L2 acquisition, the main question, which was whether visual cues would have a greater effect in disambiguating the plosives' contrast could still be examined.

Together, the findings of the current study along with the existing L2 audio-visual literature suggest a shared mechanism in L2 speech perception for the auditory and visual information. However, at this point, a distinction needs to be made between the L2 listeners who have acquired the ability to distinguish between English phonemic categories and those who are still in the process of acquisition. A successful learner seems to be sensitive both to the acoustic and visual cues marking the distinction; a learner at the early stages of acquisition is more likely to confuse the L2 consonants in both the auditory and visual modalities while s/he may not benefit from seeing the speaker. In the current study, even though participants appear to be 'language aware' and linguistically experienced, the fact that they have never lived in an English-speaking country may have affected their overall performance. Various studies have demonstrated the effect of length of residence on L2 audio-visual perception (Wang et al., 2008) as well as auditory speech learning (Flege, Yeni-Komshian, & Liu, 1999; McAllister, Flege, & Piske, 2002; Riney & Flege, 1998). In the effort to pinpoint the effect of audio-visual processing research, techniques such as eye-tracking (Lansing & McConkie, 1999) could provide quantitative measures of where exactly the perceivers' eyes focus. The general observations obtained from the current study, thus, could provide a good basis for further, more controlled investigation of the importance of visual cues in L2 speech perception.

References:

Arvaniti, A. (2001). Comparing the phonetics of single and geminate consonants in Cypriot and Standard Greek. *Proceedings of the fourth international conference on Greek linguistics.* In Y. Aggouraki, A. Arvaniti, J. I. M. Davy, D. Goutsos, M. Karyolaimou, A. Panayiotou, A. Papapavlou, P. Pavlou, & A. Roussou (Eds.), University Studio Press, Thessaloniki, pp. 37-44.

Arvaniti, A. (2006). *Linguistic practices in Cyprus and the emergence of Cypriot Standard Greek.* Department of Linguistics, UCSD. San Diego Linguistic Papers, 2. Paper 2. San Diego, CA: University of California.

Arvaniti, A. (2007). Greek phonetics: The state of the art. *Journal of Greek Linguistics, 8*(1),

97-208.

Auer, E.T.Jr., & Bernstein, L.E. (1997). Speechreading and the structure of the lexicon: Computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness. *Journal of the Acoustical Society of America, 102*(6), 3704–3710.

Baltazani, M., & Nicolaidis, K. (2013). Production of the Greek rhotic in initial and intervocalic position: an acoustic and electropalatographic study. *Selected papers of the 10th International Conference of Greek Linguistics*. In Z. Gavriilidou, A. Efthymiou, E. Thomadaki, & P. Kambakis-Vougiouklis (Eds.), University of Thrace, Komotini, pp. 141-152.

Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Science, 14*(6), 592–597.

Besle, J., Fort, A., Delpuech, C., & Giard, M-H. (2004). Bimodal speech: early suppressive visual effects in human auditory cortex. *European Journal of Neuroscience, 20*(8), 2225–2234.

Best, C. T. (1995). A direct realist view of cross-language speech, perception. In W. Strange (Ed.), *Speech Perception and Linguistic Experience* (pp. 171–204). Timonium, MD: York.

Bien, H., Lagemann, L., Dobel, C., & Zwitserlood, P. (2009). Implicit and explicit categorization of speech sounds—dissociating behavioural and neurophysiological data. *European Journal of Neuroscience, 30*, 339–346.

Bien, H., & Zwitserlood, P. (2013). Processing nasals with and without consecutive context phonemes: evidence from explicit categorization and the N100. *Frontiers in Psychology, 4,* 21.

Binnie, C.A., Montgomery, A.A., & Jackson, P.L. (1974). Auditory and visual contributions to the perception of consonants. *Journal of Speech, Language, and Hearing Research 17*(4), 619–630.

Botinis, A., Fourakis, M., & Prinou, I. (2000). Acoustic structure of the Greek stop consonants. *Glossologia, 11-12*, 167-199.

Bozkurt, E., Erzin, E., Erdem, C.E., & Ozkan, M. (2007). Comparison of phoneme and viseme based acoustic units for speech driven realistic lip animation. In *3DTV Conference, 2007*, 1–4.

Burnham, D. (1998). Language specificity in the development of auditory-visual speech

perception. In R. Campbell, & B. Dodd (Eds.), *Hearing by Eye II: Advances in the psychology of speechreading and auditory-visual speech* (pp. 27-60). London: Erlbaum.

Cappelletta, L., & Harte, N. (2012). Phoneme-to-viseme Mapping for Visual Speech Recognition. *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods*, 322-329.

Carlisle, R.S. (1994). Markedness and environment as internal constraints on the variability of interlanguage phonology. In M. Yavas (Ed.), *First and Second Language Phonology*. San Diego: CA: Singular Publishing Group Inc, pp. 223–249.

Chomsky, N., & Miller, G.A. (1963). Introduction to the formal analysis of natural languages. In R.D. Luce, R.R. Bush, & E. Galanter (Eds.), *Handbook of Mathematical Psychology*, Vol. 2, Wiley, Amsterdam, pp. 269-321.

Denison, R.N., Driver, J., & Ruff, C.C. (2013). Temporal structure and complexity affect audio-visual correspondence detection. *Frontiers in Psychology, 3*, 619.

Deterding, D.H. (1997). The Formants of Monophthong Vowels in Standard Southern British English Pronunciation. *Journal of the International Phonetic Association, 27*, 47 - 55.

Dimitriou, D. (2018). L2 acquisition and production of the English rhotic by L1 Greek-Cypriot speakers: The effect of L1 articulatory routines and phonetic context. *Philologia*, *16*(1), 45-64.

Docherty, G.J. (1992). *The timing of voicing in British English obstruents.* Berlin: Foris Publications.

Eckman, F. (1977). Markedness and the contrastive analysis hypothesis. *Language Learning, 27,* 315–330.

Flege, J. E. (1995). Second language speech learning: Theory, findings and problems. In W. Strange (Ed.), *Speech Perception and Linguistic Experience* (pp. 233– 277). Timonium, MD: York

Flege, J. E. (2009). Give input a chance!. In T. Piske, & M. Young-Scholten (Eds.), *Input Matters in SLA.* Bristol: Multilingual Matters, pp. 175-190.

Flege, J. E., & Liu, S. (2001). The effect of experience on adults' acquisition of a second language. *Studies in Second Language Acquisition*, *23*(4), 527-552.

Flege, J.E., Yeni-Komshian, G., & Liu, S. (1999). Age constraints on second language learning. *Journal of Memory and Language, 41*, 78–104.

Flemming, E. (2005). Speech perception and phonological contrast. In D.B. Pisoni, & R.E.

Remez (Eds.), *The handbook of speech perception*. Malden, MA: Blackwell, pp. 156-81.

Fujisaki, W., & Nishida, S. (2005). Temporal frequency characteristics of synchrony–asynchrony discrimination of audio-visual signals. *Experimental Brain Research, 166,* 455–464.

Ganong, W.F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance, 6*(1), 110–125.

Goldstein, E. (2013). *Sensation and Perception*. Cengage Learning, Independence, KY.

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 (Eds.), *Hearing by Eye II: Advances in the psychology of speechreading and auditory-visual speech* (pp. 3-25). London: Erlbaum.

Hardison, D. (1998). Acquisition of Second-Language Speech: Effects of Visual Cues, Context and Talker Variability. *Doctoral Dissertation*, Indiana University, Bloomington.

Hardison, D. (1999). Bimodal speech perception by native and nonnative speakers of English: Factors influencing the McGurk effect. *Language Learning, 49,* 213-283 Suppl.1.

Hazen, T.J., Saenko, K., La, C-H., & Glass, J.R. (2004). A segment-based audio-visual speech recognizer: data collection, development, and initial experiments. *Proceedings of the 6th international conference on Multimodal interfaces*, 235–242, State College, PA, USA. ACM.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8,* 393–402.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y. I., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition, 87*(1), B47-B57.

Jeffers, J. & Barley, M. (1971). *Speechreading (Lipreading).* Springfield, IL: Charles C. Thomas Pub Ltd.

Jackson, P.L. (1988). The theoretical minimal unit for visual speech perception: Visemes and coarticulation. *The Volta Review, 90*(5), 99-115.

Kainada, E. (2012). The acoustics of post-nasal stop voicing in Standard Modern Greek. *Selected papers of the 10th International Conference of Greek Linguistics*. In Z. Gavriilidou, A. Efthymiou, E. Thomadaki, & P. Kambakis-Vougiouklis (Eds.), University of Thrace,

Komotini, pp. 320-329. Kappa, I. (2002). On the Acquisition of Syllable Structure in Greek. *Journal of Greek Linguistics, 3*, 1–52.

Kkese, E. (2016). *Identifying plosives in L2 English: the case of L1 Cypriot Greek speakers*. Bern: Peter Lang.

Kkese, E. (2020a). *L2 Writing Assessment: the neglected skill of spelling*. Cambridge Scholars Publishing.

Kkese, E. (2020b). Phonological Awareness and Literacy in L2: Sensitivity to Phonological Awareness and Phoneme-Grapheme Correspondences in L2 English. In G. Neokleous, A. Krulatz, & R. Farrelly (Eds.), *Handbook of Research on Cultivating Literacy in Diverse and Multilingual Classrooms*. IGI Global Press, pp. 62-81.

Kkese, E. (2020c). Categorisation of plosive consonants in L2 English: evidence from bilingual Cypriot-Greek users. In L. Sciriha (Ed.), *Comparative Studies in Bilingualism and Bilingual Education*, pp. 179-199. Cambridge Scholars Publishing.

Kkese, E., & Karpava, S. (2021). Challenges in the perception of L2 English phonemes by native speakers of Cypriot Greek. *Journal of Monolingual and Bilingual Speech, 3*(1), 1-39.

Kkese, E., & Karpava, K. (2019). Applying the Native Language Magnet Theory to an L2 setting: Insights into the Cypriot Greek adult perception of L2 English. In E. Babatsouli (Ed.), *Proceedings of the International Symposium on Monolingual and Bilingual Speech 2019* (pp. 67–74).

Kkese, E., & Petinou, K. (2017a). Perception Abilities of L1 Cypriot Greek Listeners – Types of Errors involving Plosive Consonants in L2 English. *Journal of Psycholinguistic Research, 46*(1), 1-25.

Kkese, E., & Petinou, K. (2017b). Factors affecting the perception of plosives in second language English by Cypriot-Greek listeners. In E. Babatsouli (ed.), *Proceedings of the International Symposium on Monolingual and Bilingual Speech 2017,* pp. 162-167. ISBN: 978-618-82351-1-3. URL: http://ismbs.eu/publications-2017.

Kkese, E. & Lokhtina, I. (2017). Insights into the Cypriot-Greek Attitudes toward Multilingualism and Multiculturalism in Cyprus. *Journal of Mediterranean Studies, 26*(2), 227-246.

Klatt, D.H. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics, 3,* 129–140.

Ladefoged, P., & Ferrari Disner, S. (2012). *Vowels and Consonants*. John Wiley & Sons.

Lansing, C.R., & McConkie, G.W. (2003). Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. *Perception & Psychophysics*, *65*(4), 536–552.

Lansing, C.R., & McConkie, G.W. (1999). Attention to facial regions in segmental and prosodic visual speech perception tasks. *Journal of Speech, Language, and Hearing Research, 42,* 526–539.

Lee, S., & Yook, D. (2002). Audio-to-Visual Conversion Using Hidden Markov Models. *Proceedings of the 7th Pacific Rim International Conference on Artificial Intelligence: Trends in Artificial Intelligence*, 563–570. Springer-Verlag.

Lengeris, A. (2009). Individual Differences in Second-Language Vowel Learning. *Doctoral Dissertation*, University College London, London.

Lengeris, A., & Nicolaidis, K. (2016). The identification and production of English consonants by Greek speakers. *Selected Papers of the 21st International Symposium on Theoretical and Applied Linguistics, 21,* 224-238.

Lisker, L., & Abramson, A.S. (1964). A cross-language study of voicing in initial stops: acoustical measurements. *Word, 20,* 384– 422.

Liu, H., Ng, M., Wan, M., Wang, S., & Zhang, Y. (2007). Effects of place of articulation and aspiration on voice onset time in Mandarin esophageal speech. *Folia Phoniatrica et Logopaedica, 59,* 147–154.

Lucey, P., Martin, T., & Sridharan, S. (2004). Confusability of phonemes grouped according to their viseme classes in noisy environments. *Proceedings of the 10th Australian International Conference on Speech Science & Technology*, 265–270.

MacLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology, 21*(2), 131–141.

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

Markides, A. (1989). Lipreading - Theory and Practice. *Journal of the British Association of Teachers of the Deaf, 13*(2), 29–47.

Marschark, M., Lepoutre, D., & Bement, L. (1998). Mouth movement and signed communication. In R. Campbell, B. Dodd & D. Burnham (Eds.), *Hearing by eye II*. Hove, UK: Psychology Press, pp. 245–266.

Massaro, D.W., Cohen, M.M., & Gesi, A.T. (1993). Long-term training, transfer, and retention in learning to lipread. *Perception & Psychophysics*, *53*(5), 549–562.

Massaro, D.W., & Stork, D.G. (1998). Speech recognition and sensory integration. *American Scientist, 86*(3), 236–244.

McAllister, R., Flege, J., & Piske, T. (2002). The influence of L1 on the acquisition of Swedish quantity by native speakers of Spanish, English, and Estonian. *Journal of Phonetics, 30*, 229–258.

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

Mitterer, H., & Reinisch, E. (2016). Visual speech influences speech perception immediately but not automatically. *Perception & Psychophysics, 79*(2), 660–678.

Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., Sison, J., Mashari, A., & Zhou, J. (2000). *Audio-visual speech recognition*. Technical report, Center for Language and Speech Processing, The Johns Hopkins University, Baltimore.

Newton, B. (1972). *The Generative Interpretation of a Dialect. A Study of Modern Greek Phonology*. Cambridge: Cambridge University Press.

Ng, M., Chen, Y., Wong, S., & Xue, S. (2011). Interarticulator timing control during inspiratory phonation. *Journal of Voice, 25*(3), 319–325.

Obleser, J., & Eisner, F. (2009). Pre-lexical abstraction of speech in the auditory cortex. *Trends in Cognitive Sciences, 13*(1), 14–19.

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: UCL Work in Progress, 13*, 39-51.

Potamianos, G., Neti, C., Gravier, G., Garg, A., & Senior, A.W. (2003). Recent advances in the automatic recognition of audio-visual speech. *Proceeding of the IEEE, 91*(9), 1306–1326.

Recanzone, G.H. (2003). Auditory influences on visual temporal rate perception. *Journal of Neurophysiology, 89*, 1078–1093.

Riney, T. J., & Flege, J. E. (1998). Changes over time in global foreign accent and liquid identifiability and accuracy. *Studies in Second Language Acquisition, 20*, 213–244.

Rosenblum, L.D. (2005). The primacy of multimodal speech perception. In D. Pisoni, & R. Remez (Eds.), *Handbook of Speech Perception* (pp. 51-78). Malden, MA: Blackwell.

Rosenblum, L.D. (2019). Audiovisual Speech Perception and the McGurk Effect. *Oxford Research Encyclopedia of Linguistics*. Oxford, UK: Oxford University Press.

Saenko, K. (2004). Articulatory Features for Robust Visual Speech Recognition. *Master thesis*, Massachusetts Institute of Technology.

Smeele, P.M.T. (1996). Psychology of human speechreading. In D.G. Stork & M.E. Hennecke (Eds.), *Speechreading by Humans and Machines*. Berlin, Germany: Springer, pp. 3–15.

Spence, C., & Squire, S. (2003). Multisensory integration: maintaining the perception of

synchrony. *Current Biology, 13*, R519–R521.

Stevens, K.N., & Blumstein, S.E. (1981). The search for invariant acoustic correlates of phonetic features. In P.D. Eimas, & J.L. Miller (Eds.), *Perspectives on the study of speech*. Hillsdale, N.J: Erlbaum.

Sumby, W.H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of Acoustical Society America, 26*(2), 212–215.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd, & R. Campbell (Eds.), *Hearing by Eye: The Psychology of Lip-Reading* (pp. 3-51). London, UK: LEA.

Summerfield, Q., MacLeod, A., McGrath, M., & Brooke, M. (1989). Lips, teeth, and the benefits of lipreading. *Handbook of research on face processing*, 223–233.

Terkourafi, M. (2001). *Politeness in Cypriot Greek: A Frame-Based Approach*. (Ph.D. dissertation). University of Cambridge, Cambridge, England.

Vatikiotis-Bateson, E., Eigsti, I-M., Yano, S., & Munhall, K.G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics, 60*(6), 926–940.

Wang, Y., Behne, D., & Jiang, H. (2008). Linguistic experience and audio–visual perception of nonnative fricatives. *Journal of the Acoustical Society of America, 124*, 1716–1726.

Weinreich, U. (1953). *Languages in Contrast: Findings and Problems*. The Hague: Mouton.

Zampini, M., Guest, S., Shore, D.I, & Spence, C. (2005). Audio-visual simultaneity judgments. *Perception & Psychophysics*, *67*, 531–544.

List of Figures

Appendix

Questionnaire:

Gender: _____

Age: _____

What are you studying: _____

Year of studies: _____

Mother tongue: _____

Years of living in Cyprus: _____

Contact with people who use Cypriot-Greek: yes _____ no _____