Auditory-Visual Integration during the Perception of Spoken Arabic

Jehan Alsalmi

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

The University of Leeds

School of Medicine

January 2016

Declaration

The candidate confirms that the work submitted is her own and that appropriate credit has been given where reference has been made to the work of others.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Acknowledgments

It is my deep gratification to thank my primary supervisor and mentor, Dr. Nick Thyer and my entire supervision team, Dr. Barry Heselwood, Dr. Wahab Owolawi, and Dr. Sarah Isherwood for their tremendous efforts and support throughout the duration of my study. They directed me toward the completion of my research project, and my thesis would not have been completed without their extraordinary help and support.

I would also like to seize the opportunity to thank King Saud University College of Applied Medical Sciences for supporting me financially throughout my graduate studies.

It is my honour to dedicate this dissertation to my affectionate husband, Samer and to my children, Arjuwana, Faisal, and Batool. Their support, cooperation, understanding, and love have inspired me throughout my PhD program and made me more committed to this endeavour. They shared my difficult times, and today I am sharing with them my happy times. I would like to thank them for all that they have done for me, and I would like to assure them that I will always be there for them and working for their happiness. Finally, I would like to dedicate this dissertation to my parents for their encouragement and support.

Abstract

This thesis aimed to investigate the effect of visual speech cues on auditory-visual integration during speech perception in Arabic. Four experiments were conducted two of which were cross linguistic studies using Arabic and English listeners. To compare the influence of visual speech in Arabic and English listeners chapter 3 investigated the use of visual components of auditory-visual stimuli in native versus non-native speech using the McGurk effect. The experiment suggested that Arabic listeners' speech perception was influenced by visual components of speech to a lesser degree compared to English listeners. Furthermore, auditory and visual assimilation was observed for non-native speech cues. Additionally when the visual cue was an emphatic phoneme the Arabic listeners incorporated the emphatic visual cue in their McGurk response.

Chapter 4, investigated whether the lower McGurk effect response in Arabic listeners found in chapter 3 was due to a bottom-up mechanism of visual processing speed. Chapter 4, using auditory-visual temporal asynchronous conditions, concluded that the differences in McGurk response percentage was not due to bottom-up mechanism of visual processing speed. This led to the question of whether the difference in auditory-visual integration of speech could be due to more ambiguous visual cues in Arabic compared to English. To explore this question it was first necessary to identify visemes in Arabic. Chapter 5 identified 13 viseme categories in Arabic, some emphatic visemes were visually distinct from their non-emphatic counterparts and a greater number of phonemes within the guttural viseme category were found compared to English.

Chapter 6 evaluated the visual speech influence across the 13 viseme categories in Arabic measured by the McGurk effect. It was concluded that the predictive power of visual cues and the contrast between visual and auditory speech components will lead to an increase in the McGurk response percentage in Arabic.

Table of Contents

List of Figures

List of Tables

Abbreviations

A	Auditory
AVI	Auditory Visual Integration
ANOVA	Analysis of Variance
CI	Confidence Interval
CV	Consonant Vowel
fMRI	Functional Magnetic Resonance Imaging
MPEG	Movie Picture Experts Group
NL	Native Language
SPL	Sound Pressure Level
V	Visual
VOT	Voice Onset Time

Chapter 1

Auditory-Visual Integration

1.1 Introduction

The main rationale for conducting this research is that the literature on auditory-visual speech perception has shown differences in the use of visual speech cues across language (Hazan et al., 2006, Massaro et al., 1995, Sekiyama, 1997, Sekiyama and Burnham, 2008, Sekiyama and Tohkura, 1991, Sekiyama and Tohkura, 1993). Therefore, this research is aimed at investigating the influence of visual cues on speech perception in Arabic. This thesis attempts to make a step forward in the testing and practical exploration of auditory-visual integration during speech perception in Arabic. The findings of this research will make new contributions to the literature on auditory-visual speech perception.

The reason why the topic related to auditory-visual integration of speech is relatively new is because there was a historical bias toward an auditory only speech perception process due to the seemingly distinct perceptual systems. Until 1976, when the experiment of McGurk and Macdonald took place (see section 1.4.1), the predominant trend in the research on visual cues in speech perception was that vision had only a complementary role in speech perception when the auditory signal was degraded (Schwartz et al., 2004). The McGurk effect was a compelling example of the effect of visual cues on the perception of speech in optimal listening conditions.

Vision was then found to be more than a supplementary modality in the process of speech perception and even in optimal listening conditions visual speech cues

produced an advantage in speech perception, producing a faster and more accurate response (Buchwald et al., 2009). Further research has demonstrated that the brain has the ability to integrate speech information from both the auditory and visual modality into a unified percept which may not exactly match either auditory-only or visual-only percept (Baart and Vroomen, 2010). It has now been established with a considerable amount of behavioural and neurological research that even with optimal auditory input the visual modality is involved in the perception of face to face speech (Baart and Vroomen, 2010, Burnham and Dodd, 2004, Campbell, 2008).

Research has shown that there are differences cross-linguistically in the use of visual cues during speech perception (Hazan et al., 2006, Massaro et al., 1995, Sekiyama, 1997, Sekiyama and Burnham, 2008, Sekiyama and Tohkura, 1991, Sekiyama and Tohkura, 1993). However, the majority of languages investigated have been Indo-European languages which were found to be similar in their use of visual cues. To better understand the process of speech perception it is essential to evaluate the visual cue features that are incorporated during auditory-visual integration across different languages. Cross-linguistic investigation allows a comparison of different visual speech features to enable us to define which visual features are incorporated in the integration process.

In this thesis, a series of experiments were conducted to investigate auditory-visual integration in Arabic during speech perception and comparing it to English. The aim is to examine whether the use of visual cues during auditory-visual integration of speech in Arabic is different to that of English. Furthermore, the characteristics of Arabic that may lead to the different use of visual cues during auditory-visual integration compared to English are evaluated. Additionally, whether auditory-visual

integration of speech can be shaped by native language visual speech cues in Arabic was examined. The influence of native language visual cues during speech perception in Arabic is the focus of this thesis.

The aim of this chapter is to introduce the main terminology and processes involved in this research. Consequently, the chapter is structured as follows. First of all, auditory and visual cues are explained. Secondly, advantages of auditory-visual speech, including such aspects as confusion within auditory and visual speech cues, dichotic listening paradigm, speech in noise, coarse visual speech and complex speech are examined. Next, evidence of auditory-visual integration are outlined, such as the McGurk effect, the effect of auditory stimulus on visual perception, and auditory-visual neurophysiological studies. The chapter then proceeds with the influence of native language on the development of auditory and visual cues. In this regard, auditory speech development, visual speech development, auditory-visual development of speech and auditory-visual neural development are explained. Then, cross-linguistic studies of auditory-visual integration are analysed, followed by explanation of Arabic in the context of auditory-visual speech. Finally, a detailed summary is given at the end of this chapter.

1.2 Introduction to Auditory and Visual Cues

1.2.1 Auditory Cues

In order to understand the processes by which visual and auditory speech information are combined, it is first necessary to have some understanding of the nature of speech processing within each of the two modalities separately. The basic

aural unit of auditory speech is the phoneme. Phonemes are the smallest segment of sound for which, if that segment is replaced with another, the meaning of the word changes (International Phonetics Association, 1999). The study of phonetics and speech has a long history. In 1887 development was started on a phonetic alphabet, known as the International Phonetic Alphabet. The phonetic alphabet established by the Association rapidly developed, and demonstrates an agreement on a set of phonemes for use in describing speech in various languages (Table 1.1). Phonemes are split into two groups vowels and consonants. Vowels are phonemes produced

Table 1.1 IPA chart showing consonants grouped by the place of articulation (International Phonetic Association, 2005)

THE INTERNATIONAL PHONETIC ALPHABET (2005)

CONSONANTS (PULMONIC)

	Bilabial	Labio-dental	Dental	Alveolar	Post-alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Epi-glottal	Glottal
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ			
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʡ	ʔ
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	H ʢ	h ɦ
Approximant		ʋ		ɹ		ɻ	j	ɰ				
Trill	ʙ			r					ʀ		ʜ	
Tap, Flap		ⱱ		ɾ		ɽ					ʡ	
Lateral fricative				ɬ ɮ								
Lateral approximant				l		ɭ	ʎ	ʟ				
Lateral flap				ɺ								

Where symbols appear in pairs, the one to the right represents a modally voiced consonant, except for murmured *fi*.
Shaded areas denote articulations judged to be impossible. Light grey letters are unofficial extensions of the IPA.

without obstructing air flow out of the mouth. Consonants are phonemes produced by obstructing the flow of air out of the mouth (International Phonetics Association, 1999).

This thesis is not an investigation of auditory cues, which is why in this section only the main auditory cues relevant for the distinction between consonants are discussed. The main auditory cues that differentiate consonants are place of articulation,

manner of articulation, voicing, voice onset time (VOT), formants and formant transitions. The place of articulation is the point of contact where the articulation is being produced; i.e., places where the obstruction occurs in the mouth where articulators such as the tongue move relative to the roof of the mouth. Articulation can be produced by forming bilabials (on the lips), labiodentals (lower lip against the upper teeth), dentals (teeth and the tongue), alveolar (tongue tip and alveolar ridge), palatal-alveolar (tongue blade and alveolar ridge), palatals (body of tongue and hard palate), velars (back part of tongue and soft palate), uvular (back part of tongue and uvula), pharyngeal (root of the tongue against the uvula), glottal (obstructing airflow at the glottis) and emphatic (back of the tongue approaching the pharynx) (see Figure 1.1).

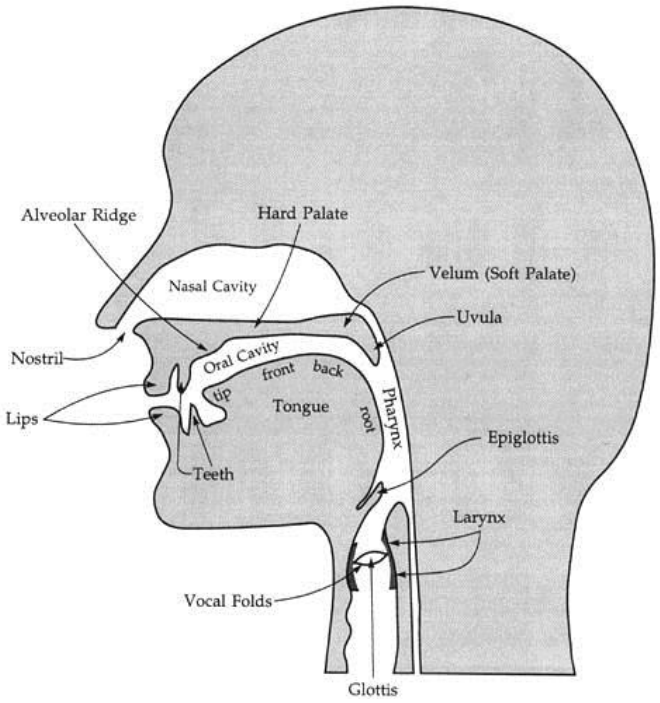

Figure 1.1 Section of the vocal tract, with places of articulation labelled.

The manner of articulation describes how the articulators interact to produce the phoneme. The different categories of manner are nasal (the air passes through the nose), stop (the vocal tract is blocked so that all airflow ceases), fricative (partial occlusion hinders but does not block airflow in the vocal tract), affricate (begin as stops but are released as fricatives), and approximant (the articulators approaching each other but not narrowly enough to create turbulent airflow).

Phonemes are either voiced or voiceless, voicing occurs at the larynx which houses the vocal folds. Voiced phonemes are produced when the vocal folds are close together loosely so they can vibrate, for example the phoneme /**b**/. Most vowels and nasal stops are voiced. Voiceless phonemes are produced when the vocal folds are wide apart so that air passes freely and the vocal folds are not vibrating, for example the phoneme /**p**/.

Another auditory cue is VOT; it refers to the time interval between the release of an occlusion and the beginning of voicing. The existence of this interval is caused by the fact that the voicing and closure frameworks are distinct. The oral occlusion occurs at the region that is above the larynx, while voicing occurs at the larynx that houses the vocal folds. Since the occlusion and voicing frameworks are distinct; therefore, their operations may have a temporal mismatch measured in milliseconds (ms).

Formants for vowels and formant transitions for consonants are also another type of auditory cue. Formants are regions of frequency space on a spectrogram where phonemes carry a lot of energy. Formants are produced due to resonances in the vocal tract. The place of articulation and manner will change the dimensions of the resonance cavities in the vocal tract and therefore change the formant frequencies.

Therefore, different vowels will have different formant frequencies which can be used as an auditory cue to identify them. Usually, the first two formants (F1 and F2) are sufficient to distinguish between two vowels (see Figure 1.2). F1 varies as a consequence of vertical tongue movement, therefore the lower the tongue the higher the value of F1. While, F2 reflects the horizontal movement of the tongue and it is also influenced by the rounding of the lips.

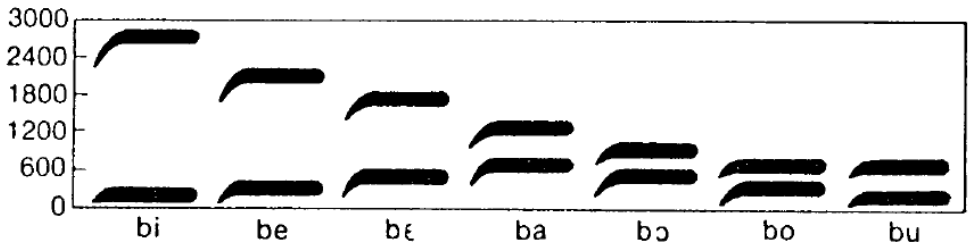

Figure 1.2 Formants for different vowels (Liberman, 1957).

Consonants also have formants but they are not as easily recognizable as compared to vowels. This is due to the constriction in the oral cavity when producing a consonant the resonance is reduced. Formant transitions are auditory cues that can better help to identify stop constants. The movement of the formant transition whether upward or downward for each formant in the spectrogram helps in discriminating which stop constant preceded or followed the vowel. The formant transition for F1 reflects manner of articulation and place of articulation is reflected by F2 and F3. In Figure 1.3, the formants after the initial transition shifts are the same indicating that all the speech samples have the same vowel phoneme. However, they are all preceded by different formant transitions which would assist the listener in differentiating auditorily between them.

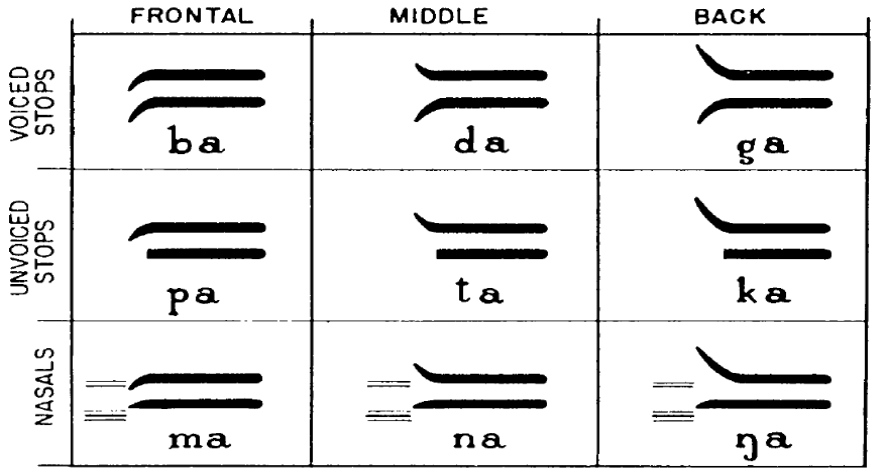

Figure 1.3 Formant transitions for stop consonants (Liberman, 1957).

1.2.2 Visual Cues

This part of the chapter addresses the aspects of visual speech which are the most relevant for this research. The main connection between visual cues and actual speech is that the mouth changes its shape depending on the basal structure of the word being spoken (Holden and Owens, 2000). The main function of visual cues is thought to be in complementation and cross-verification of the auditory information (Altieri et al., 2011, Peelle and Sommers, 2015). In this regard, it is considered to have three fundamental roles in speech perception. First, it helps to localise a speaker, which gives an opportunity to accommodate listening according to the location of the speaker (Carlyon et al., 2001). Secondly, it provides additional, environmental and contextual information about the place of articulation (Peelle and Sommers, 2015). Finally, it provides temporal information about the speech signal which increases the precision of predicting the acoustic signal (Peelle and Davis, 2012).

One of the examples of how visual cues can provide additional information is that they help to identify the place of articulation of phonemes such as /**b**/ versus /**d**/ (Munhall et al., 2004). While auditory recognition might be unclear, visual speech cues assist in distinguishing some phonemes such as /**b**/ versus /**w**/ articulation (Jiang and Bernstein, 2011). When phonemes are spoken, they correspond to a specific change in mouth shape including movement of the lips, tongue and appearance of teeth. Thus, motion of speech articulators like lips, tongue and jaws create visual cues. According to the similarity of visual movements that produce phonemes, they are grouped together. These groups are known as visemes (Fisher, 1968). Consequently, different groups of visemes are visually distinguishable, while separate phonemes within a group are not (Jackson, 1988). One of the sources of information distortion is the fact that there can be more than one phoneme in a viseme group. For example, in English /**p, b, m**/ are all bilabial consonants; although they are acoustically and phonetically different visually they look the same and the same is the case for /**d,t**/ or /**f,v**/ (Bozkurt et al., 2007).

Viseme groups are often established through the ability of observers to recognise consonant phonemes in sequences of consonant vowels (CV). Furthermore, these clusters are distinguished in confusion matrices and then they are labelled as visemes (Chen, 1998, Goldschen et al., 1994, Owens and Blazek, 1985). Table 1.2 shows one of the most recent categorization of all the English consonants corresponded to 10 viseme categories (Bozkurt et al., 2007). Since every language consists of different phonemes and thus have different phonetics; visemes have to be identified for each language separately. Visemes have been investigated in many languages

Table 1.2 Viseme categories for consonants (Bozkurt et al., 2007).

Viseme Category	Consonants
1	/ p, b, m/
2	/f, v/
3	/w/
4	/ θ, ð /
5	/t, d, n, l /
6	/s, z/
7	/ʃ, dʒ/
8	/r/
9	/j/
10	/k, g/

such as German (Aschenberner and Weiss, 2005), French (Werda et al., 2007), Swedish (Engström, 2003), and Italian (Magno Caldognetto et al., 1997). However viseme classification of all Arabic consonants by speechreading has not been performed.

1.3 Advantages of Auditory-Visual Speech

In this section the advantages of auditory -visual speech will be discussed. The benefits of using both modalities has been examined under different experimental conditions such as confusion within auditory and visual speech cues, dichotic listening paradigm, speech in noise, coarse visual speech and complex speech. There is one consistent factor found for all of the different experimental conditions which is that there is a clear advantage of auditory-visual speech compared to auditory only. These experiments highlight the importance of visual speech in the process of speech perception.

1.3.1 Confusion within Auditory and Visual Speech Cues

The following two experiments analysed CV syllables to determine which phonemes are most confused in auditory only (Miller and Nicely, 1955) and visual only condition (Walden et al., 1977). Both experiments used hierarchical cluster analysis that expresses similarities between the consonants by a measure based on correlation. In the first experiment the auditory identification of all the English consonants was measured at different signal to noise ratios under the condition of white noise (Miller and Nicely, 1955). In Figure 1.4 the horizontal lines demonstrate the range of signal to noise ratios from -18 dB to +18 dB, which is calculated in terms of the peak level of the vowel. The results demonstrated that below -18 dB, no syllable was identified. Then, in the interval of -15 to -12 dB the first branching can be seen between consonants indicating a distinction between voiceless, voiced and nasal consonants.

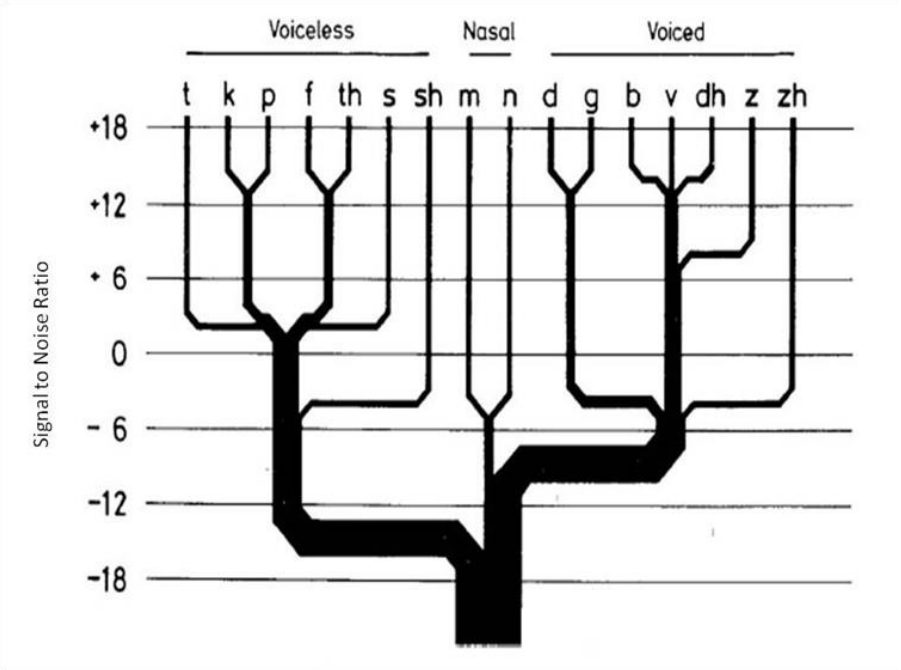

Figure 1.4 Auditory confusions between consonants presented as CV syllables in white noise (Kryter, 1970, from Miller & Nicely, 1955).

For example at the signal to noise ratio of -15 dB /**d**/ and /**t**/ were easy to distinguish since /**d**/ is voiced and /**t**/ is voiceless. However, to distinguish between the voiced consonants /**d**/ and /**g**/ a signal to noise ratio of +15 dB was needed. The increase in signal to noise ratio is linear to the improvement of distinguishability between consonant groups. At the signal to noise ratio of +15 dB all the consonants were distinguished from one another. It can be concluded that in the auditory modality the most salient cue is voicing. This can be observed by the initial separation of consonants into voiced and voiceless groups which means that they are the easiest to differentiate in terms of the auditory modality.

On the other hand, the second experiment demonstrates visual confusion for the CV syllables (Walden et al., 1977). In Figure 1.5 the horizontal lines correspond to visual correlation between syllables from level 15 at 0% correlation to level 1 at 100% correlation. For example, eight groups of visemes were distinguishable on the 11^{th} level this corresponded to 75% correlation. The lowest visual confusion is among consonants that are created as a result of different external mouth shapes. For instance, /**b**/ and /**g**/ are visually distinct at level 15 which means there is 0% visual correlation between these phonemes. There is no visual confusion between these two phonemes because /**b**/ is produced by the closure of lips and /**g**/ is produced by an open mouth, and so this makes it easy to differentiate between them visually.

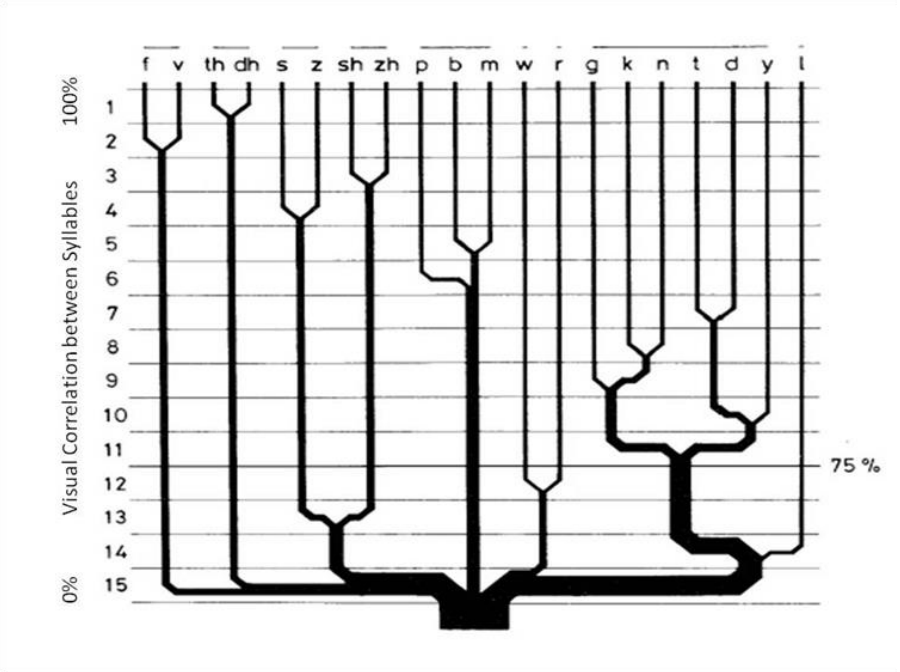

Figure 1.5 Visual confusion among consonants presented as CV syllables (Walden et al., 1977).

The main outcome of this observation is that confusion among English consonants largely differs between auditory and visual perception. For instance, /**b**/ and /**v**/ phonemes were auditorily highly confused since they are both voiced phonemes. However, they were easily distinguished visually since /**b**/ is a bilabial phoneme and /**v**/ is a labiodental phoneme. In contrast, /**p**/ and /**b**/ phonemes had the opposite characterisation. They were difficult to identify visually because they are both bilabial phonemes. However, they were easily discriminated in the auditory modality since /**b**/ is voiced and /**p**/ is voiceless.

Figure 1.4 and Figure 1.5 demonstrated the primary advantage of auditory-visual integration, meaning mutual complementation of the information gained. In this regard, auditory information is complemented by visual information. To a certain extent, the process of cross-verification of information is taking place (Khalil, 2013).

What is not clearly identified in the auditory modality can be clarified through the visual modality. The main outcome is that auditory-visual perception of speech can result in a more accurate identification of speech rather than auditory only modality (Chen and Rao, 1998, Potamianos et al., 2004, Hazan et al., 2005).

1.3.2 Speech in Noise

The advantage of visual speech cues during speech perception can be clearly observed in noisy surroundings or with more than one person speaking simultaneously. This occurs in everyday situations such as being in traffic, at a restaurant, or attending a meeting. The importance of the visual element in auditory-visual speech perception is demonstrated in many studies by the fact that the presence of visual face movement information significantly improves speech perception, which reduces the signal-to-noise ratio required for participants to identify speech against a background noise mask (Bernstein et al., 2004b, Tye-Murray et al., 2007). Signal to noise ratio is a measure used to compare the level of a signal to the level of background noise and it is measured in decibels (dB).

Sumby and Pollack (1954) were the first to establish that visual speech cues improve the perception of speech presented in noise. In their experiment, speech was presented in background noise at one of seven signal to noise ratios through a headphone and the participants were asked to report what they heard. Half of the presentations had the speaker's face visible to the participant and the other half the speaker was facing away from the participant. The participants consistently performed better when they received visual speech information in addition to the auditory information for all the signal to noise ratios. The greatest improvement

found by the addition of visual speech information was for low signal to noise ratios where the noise was louder than the speech (Sumby and Pollack, 1954).

MacLeod and Summerfield (1987) also compared speech perception thresholds in auditory only and auditory-visual conditions. They found that visual speech cues improved speech perception thresholds by an average of 11 dB. Figure 1.6 shows results obtained by Chen (2001) illustrating the effect of auditory noise. It can be seen that as the signal to noise ratio decreases, the auditory recognition accuracy decreases. However when the recognition is auditory-visual there is a statistically significant improvement in recognition ratio compared to auditory only (Chen, 2001). These findings are highly relevant to everyday speech perception since in normal listening situations speech is usually accompanied by background noise.

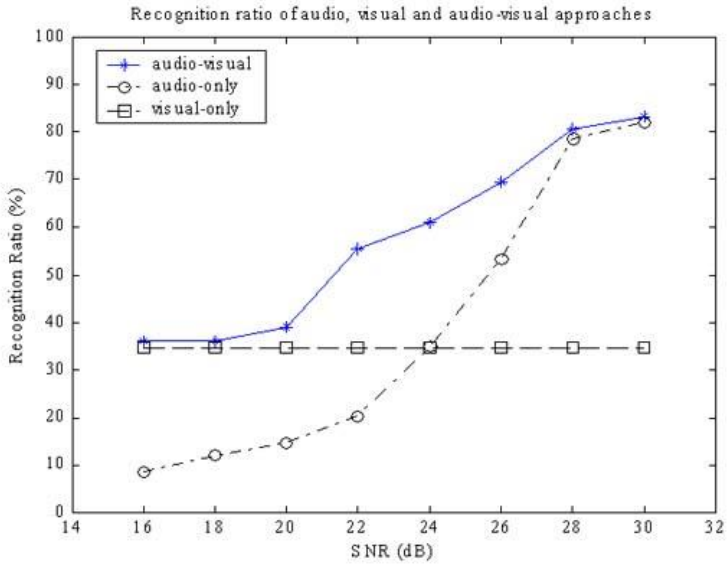

Figure 1.6 Recognition ratio versus signal to noise ratio (SNR) of the auditory signal (Chen, 2001).

1.3.3 Coarse Visual Speech

Another aspect of how visual information can improve speech perception in contrast to auditory stimulus alone was presented in the research of Rosenblum et al., (1996). They studied the influence of coarse visual input on auditory-visual integration. In their experiment, instead of using natural face, they created a point-light display to correspond for the visual input. Their findings suggested that a coarse visual stimulus was more productive in information perception and argued in favour of auditory-visual integration rather than an auditory stimulus only (Rosenblum et al., 1996). As it is demonstrated in Figure 1.7, point light display consisted of reflective dots that were situated on the places of articulators meaning lips, teeth, mouth, and chin.

Figure 1.7 Schematic reflection of point-light display used by Rosenblum & Saldana (1996), aimed at de-contextualisation of speech perception.

In order to create systematic video stimuli of motion, special lighting was used. The main rationale for using this display instead of a natural face is to explore participants' perception of speech without the context of facial identity. Thus, the secondary aim of the experiment was also to see the importance of the facial context

for speech perception. The fact that participants could perceive speech in its auditory-visual integration without actual special details of the face suggests that in order to perceive speech, viewing the face is not necessary (Rosenblum et al., 1996).

Rosenblum's research (1996) demonstrated that comprehension increased linearly in accordance with the increase in the number of reflective points detailed on the face. Although comprehension was achieved by highlighting 14 points on the lips and mouth, the increase of highlighted points improved understanding. The final part of the experiment demonstrated that fully highlighted display resulted in the best threshold. The conclusion of this research is that coarse visual stimuli can improve speech perception; however, the best comprehension is achieved through auditory-visual integration based on the observation of a natural face. The implications of these findings are that since even coarse visual stimuli can improve speech perception then it is not only the visual movement of the mouth that matters but also a mental representation of these movements. Visual mental representation is the realisation of key details of predicting a potential word visually through past experience (Barnard et al., 2002).

1.3.4 Complex Speech

Arnold & Hill (2001) used an alternative method to evaluate the effect of visual cues on speech perception. In this study the quality of the auditory signal was not degraded but there was an increase in the cognitive load by presenting speech that was semantically and syntactically complex. The participants' comprehension was measured both in auditory only and auditory-visual condition. The comprehension

performance was scored by a judge blind to the condition of presentation (auditory or auditory-visual). Speech perception was significantly better when speech was presented auditory-visual rather than auditory-only. The authors concluded that perceiving intact auditory input can also be aided by visual cues (Arnold and Hill, 2001).

One might argue that this advantage is solely due to there being two separate sources of information available by which to identify speech. However, Reisberg et al., (1987) found that for the same stimuli, auditory-alone presentation produced 6% word identification and visual-alone presentation 1%, while auditory-visual presentation produced performance of 45%. If the advantage of visual cues was simply a complementary one, we would then expect to see a combined improvement of 7%. However, the results showed a combined improvement of 45%, which is much larger than the sum of the individual modalities speech identification scores. This result suggests that speech perception by the auditory and visual modalities is integrated rather than independently sampled (Reisberg et al., 1987). The above studies strongly suggest that visual speech cues play an essential role in the process of speech perception. More recent research on auditory-visual speech perception has mainly focused on using the McGurk effect to evaluate the influence of visual speech.

1.4 Evidence of Auditory-Visual Integration

1.4.1 The McGurk Effect

For speech perception research studies it has been challenging to produce a behavioural test to evaluate the process of auditory-visual integration during speech perception. In the previous section behavioural studies were reviewed which tested participants in auditory alone and then in auditory and visual condition to measure the effect of visual cues on speech perception. These studies have shown that our ability to understand speech is better when we can hear and see the speaker under many different conditions such as noise and complex speech (Grant et al., 1998, Sommers et al., 2005, Arnold and Hill, 2001, Chen, 2001). There have also been equations created to try to quantify the amount of improvement in auditory-visual condition when compared to auditory alone (Sumby and Pollack, 1954, Rabinowitz et al., 1992, Grant and Seitz, 1998). Although the previous section has clearly shown that there is a great benefit from the addition of visual cues during speech perception these tests cannot assist us in understanding a framework underpinning auditory-visual integration of speech.

However one test that has been used to evaluate a framework of auditory-visual speech is the McGurk effect (McGurk and MacDonald, 1976). The perception of clear unambiguous speech has been shown by the McGurk effect to depend on both the auditory and visual modality. McGurk and MacDonald (1976) demonstrated this by dubbing incongruent auditory and visual stimuli which differed in place of articulation. For example auditory /**ga**/ velar consonant is superimposed over the video of /**ba**/ bilabial consonant. Surprisingly, the participant perceives a new

response that differs from both the auditory and visual stimuli. When the participant looks away from the video screen, the auditory stimulus is heard correctly. The integration of the visual and auditory modalities is called the McGurk effect.

The effect occurs when there is a mismatch between the visual and auditory speech stimuli. The syllable /**ga**/ is produced by air being pushed up through the glottis stopping at the velum. It is made at the back of the mouth therefore it is difficult for an observer to see. The syllable /**ba**/ is produced similarly but the place of articulation is the lips. The outcome of the two conflicting places of articulation is the perception of a new syllable for example /**da**/, which is made between the lips and the velum at the alveolar ridge (see Figure 1.8). When this occurs it is considered a fusion of the auditory and visual stimuli because the place of articulation for the /**d**/ alveolar consonant lies between velar /**g**/ consonant and bilabial /**b**/ consonant.

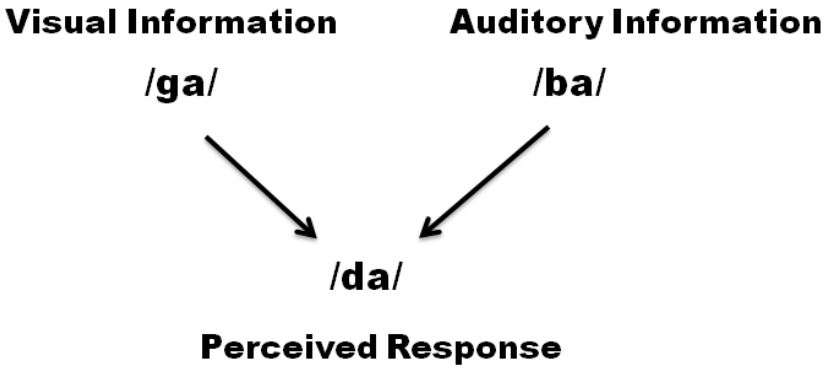

Figure 1.8 The McGurk Effect

McGurk and McDonald also reported a second type of response which is called a combination response. This occurs when both the auditory and visual stimuli are perceived. For example when the visual stimulus is /**ba**/ and the auditory stimulus

is /**ga**/, the observer perceived the sound to be /**bga**/, a combination of both stimuli. The reason for this variation in integration is thought to be due to the fact that visual bilabial information is clearly seen because of its highly visible place of formation at the lips, in comparison to a velar placement at the back of the oral cavity. However, only fusion is considered to be a McGurk effect, because the response is different from the visual and auditory stimulus. This provides evidence that speech perception is multimodal and not just auditory. The McGurk effect demonstrates that visual speech cues even in optimal listening conditions cannot be ignored during the speech perception process.

The explanation given for the McGurk effect is during the process of speech perception when there is conflicting information coming from the visual modality and the auditory modality the listener perceives an alternative sound. This alternative sound is a compromise between the incompatible cues perceived from the visual modality and the auditory modality (McGurk and MacDonald, 1976). The McGurk stimuli can be perceived according to the auditory stimulus, visual stimulus or a combination of both. Speech perception seems to be a flexible process which depends on the relative weighting between the auditory and visual stimuli (Jiang and Bernstein, 2011).

The McGurk effect has been perceived in many different conditions. Even when participants are informed of the mismatch between the auditory and visual cues the McGurk effect is still perceived (McGurk and MacDonald, 1976). Likewise, when participants are instructed to only report what they hear the McGurk effect still occurs (Summerfield and McGrath, 1984). Furthermore, when the visual stimulus was reduced to only three frames the McGurk effect was still present. Therefore,

even with degraded visual input the effect occurs (Whalen et al., 1996). Easton and Basala (1982) proposed that only nonsense syllables would elicit the McGurk effect, but the effect was also elicited when real words were used as stimuli (Dekle et al., 1992).

Functional Magnetic Resonance Imaging (fMRI) has been used to study the McGurk effect. The fMRI permits identification of brain areas that show task related cerebro-vascular responses (Ogawa et al., 1990). fMRI is used to evaluate changes in the levels of haemoglobin in the different brain areas activated by speech and other cognitive events. vanWassenhove et al. (2007) fMRI study provided thought provoking insights into the McGurk effect. In this experiment, a visual /**ka**/ was dubbed onto an auditory /**pa**/. The resulting brain activation of the fused perception of /**ta**/ correlated more closely with the activation of the perception of a true /**ta**/ than with either the visual /**ka**/ or the auditory /**pa**/ stimuli (van Wassenhove et al., 2007). Moreover the McGurk effect is not a measure of speech reading ability as there is no correlation between the two (Munhall et al., 2004).

The McGurk effect is a striking revelation of the powerful role visual speech cues can play during auditory-visual speech perception. Even a degraded visual input and a conscious awareness of the mismatch between the visual and auditory speech does not diminish the strength of the effect. The McGurk effect has clearly demonstrated that visual speech cues have an integrative role during speech perception and not simply a complementary one. The McGurk effect is a powerful method to investigate the complementary nature of the separate auditory and visual information sources and this can be subsequently applied to understand the integration stage of auditory-visual speech perception. Furthermore, the strength of the McGurk effect can be

taken to reflect the strength of auditory-visual integration (Altieri, 2014, Jiang and Bernstein, 2011). In this thesis three experiments using the McGurk effect were performed to investigate auditory-visual speech perception in Arabic.

1.4.2 Auditory-Visual Neurophysiological Studies

Neural correlates of auditory-visual speech perception have also indicated that the integration of visual cues enhances speech perception. Studies using fMRI show that visual (silent) speech has been shown to activate the auditory cortex, which is involved in processing auditory speech (Besle et al., 2008, Kauramäki et al., 2010). Additionally, the auditory cortex was not activated by non-linguistic lip movements (Calvert et al., 1997). These findings suggest that cortical regions traditionally believed to be auditory processing areas for language are also accessed by visual speech; this is possibly due to neural networks involved in auditory-visual integration. Thus because visual and auditory speech seems to be processed in the same cortical areas, this would assist in the process of speech integration (Campbell, 2008, Okada and Hickok, 2009).

Sams et al. (1991) performed an electro-encephalography study to compare auditory only and auditory-visual speech processing. They found modification of the characteristic response of the auditory cortex by the inclusion of visual speech stimuli. The change in response by the addition of visual cues produced a change in the waveform pattern to appear in the primary auditory cortex (Sams et al., 1991). Additionally, cortical auditory evoked potentials have been used to evaluate the effect visual cues have on auditory-visual speech perception. Cortical auditory

evoked potentials are composed of a series of negative and positive peaks. The neural activity generated from different locations in the brain produces different peaks in the waveform. These average waveforms reflect electroencephalogram activity in response to specific stimuli. The main components for the auditory stimuli are N1 and P2 which occur between 60 and 200 ms after stimulus onset. A reduction in N1/P2 latency has been observed for auditory-visual speech compared to auditory only speech in native English listeners (van Wassenhove et al., 2005). This indicates that visual speech cues cause an increase in the speed of cortical speech processing of auditory information. These neurophysiologic results provide evidence that visual speech cues modify the functioning of the auditory cortex and processing of speech and suggest a strong association between auditory and visual speech.

Additionally many sub-cortical and cortical areas demonstrate multimodality to visual and auditory speech (Wallace and Stein, 2001, Campbell, 2008, Erickson et al., 2014). These studies propose that information from both visual and auditory modalities is integrated together at a sub-cortical and cortical level and that there are neurons and brain regions that respond maximally to auditory-visual stimuli. The existence of multisensory convergence sites suggests a regular maintenance of crosstalk between sensory specific streams, in what would create a multisensory mode of information processing. However, multisensory integration seems to occur after a certain amount of information has been extracted in the sensory specific streams that is late integration of the auditory-visual integration (Altieri et al., 2011).

1.5 The Role of Native Language on the Development of Auditory and Visual Cues

While behavioural and neurophysiological studies clearly show a strong association between auditory and visual speech input, this association is also dependent on the auditory and visual mental representations of the native language. For this statement to be true, one must find evidence of auditory and visual perceptual mental representations based on the internal mental representations corresponding to native language. There is evidence from developmental studies both from behavioural and neurological studies that suggests that infants' speech perception becomes fine-tuned to auditory and visual cues within the native language and reduction in sensitivity occurs for auditory and visual cues that are not present in the native language. This section will review studies on auditory and visual speech development.

1.5.1 Auditory Speech Development

Studies on speech perception in infants have led to the agreement that infants can discriminate between different sounds within the repertoire of the world's languages. This excellent ability of infants to discriminate between speech sounds provides them with the capacity for learning the different sound categories within their native language (Jusczyk et al., 1993). Initially infants can discriminate between any speech sounds whether native or non-native. Yet, gradually infants' auditory discrimination abilities become more tuned into speech sounds within their native language. Starting at six months of age, this universal perceptual phoneme ability in infants

begins to decrease because of increased exposure to their native language (Best, 1994).

For instance, Werker and Tees (2002) compared the ability of Salish (Native Indian) adults, Hindi adults, English adults, and English infants in discriminating between non native place distinction contrasts within the Salish language. The infants' ability to discriminate between the contrasts was evaluated by using a common preferential looking procedure called the head turn procedure. The procedure involved the infant sitting on a parent's lap while facing an assistant, who uses silent toys to attract the infant's attention. The infant is trained to respond to a change in the speech sound category by turning their heads away from the assistant and toward a loud speaker. Only correct head turn responses were reinforced with the presentation of a moving toy (e.g. monkey tapping on a drum). They found that the 6 to 8 month old English infants could discriminate between the Salish and Hindi contrasts. However, the 10 to 12 month old infants as well as the adults were not able to distinguish between the Salish and Hindi contrasts (Werker and Tess, 2002).

Furthermore, speech perception studies on infants have investigated prosodic features such as intonation, stress and tones. They found that infants in the first year of life demonstrate sensitivity to native prosodic properties (Gervain and Mehler, 2010). Kuhl et al. (2006) investigated the reduction in discrimination of non-native contrasts as well as the improvement in discrimination of native contrasts for the first year of life. They compared the ability of United States and Japanese infants in their ability to discriminate between the American English contrast /**r-l**/. They showed that by the age of 10 to 12 months the Japanese infants were not able to discriminate the non-native contrast. On the other hand the 10 to 12 month old American infants'

ability to discriminate between the native contrasts improved significantly (Kuhl et al., 2006). These results show that within the first year of life infants are beginning to demonstrate sensitivity to the organization and structure of the sound patterns within the native language. The developmental loss of speech perception abilities of non-native phonemes in the first year of life is not a loss but a reorganization of the speech perception framework to be finely tuned to the infant's native language. This fine tuning by infants might be the starting point for the construction of perceptual auditory speech mental representations based on the native language.

1.5.2 Visual Speech Development

A significant amount of evidence suggests that the development of visual speech perception begins during infancy. The majority of studies on visual speech perception in infants use the matching technique. The visual speech matching technique is where the infants are presented with two video screens and a single auditory source equidistant from the video screens. One video screen has a speaker producing matching speech while the other video is of a speaker producing mismatched speech. A child's preference is measured by the amount of sucking or the amount of time he/she spends watching a given stimulus (Burnham and Dodd, 2004, Kuhl et al., 2006).

Studies on visual face perception have evaluated at what age an infant can discriminate between the mother's and a stranger's face on visual information alone. At the age of 4 to 5 months infants can consistently discriminate between the mother's face and a stranger's face (Burnham, 1993). When the mother's face is

coupled with her voice, the infant can discriminate between the mother's face and a stranger's face by the age of 1 month. The use of auditory-visual information enables the infant to discriminate faces at an age well below that of visual information alone (Burnham, 1993).

Patterson and Werker (2003) investigated whether infants at 2 months of age could match visual speech to auditory speech. Infants were shown two video images in synchrony on computer screens, side by side. One video screen had a face articulating the vowel /**i**/ while the other was articulating the vowel /**a**/. A soundtrack played one of the vowel sounds through a speaker placed between the two computer screens. They found that infants as young as 2 months of age have the ability to match heard vowels with the appropriate lip movements (Patterson and Werker, 2003). The ability to match consonants comes later at the age of 6 months (MacKain et al., 1983). Burnham (1988) found that infants at the age of 4.5 months preferred matching native speech compared to non-native speech. Furthermore, Weikum et al. (2007) found that 4 to 6 month old infants could distinguish native language visual speech cues from non-native visual speech cues, but this ability was not present in 8 month old infants (Weikum et al., 2007).

Other studies investigated how visual cues influence learning of speech. Legerstee (1990) explored the role of visual speech in eliciting imitation of speech sounds. They presented the vowel sounds /**u**/ and /**a**/ to infants 3 to 4 months of age via speaker. Infants were divided into two groups. The first group of infants was represented with an adult who articulated the same vowels silently. An adult in the second group was articulating the opposite vowels, once again silently. The result of this experiment was that children in the first group who were exposed to matching

auditory and visual stimuli were able to reproduce these vowels. The results were taken to suggest that visual speech is useful in stimulating learning and the acquisition of speech (Legerstee, 1990). More recently, Teinonen et al. (2008) investigated the influence of visual cues on learning phonetic discrimination. They tested two groups of 6-month-old infants on a /**ba**/–/**da**/ auditory continuum. One group of infants was simultaneously presented with visual cues for /**ba**/–/**da**/. The second group of infants was exposed to the same /**ba**/–/**da**/ auditory continuum, but they were only given one visual speech cue either /**ba**/ or /**da**/.The results showed that the infants, who were presented with two visual cues, were able to discriminate between the auditory continuum. However, the infants who only had one visual speech cue were not able to discriminate between the auditory continuum. The results were taken to show that the visual speech cues enhance phoneme discrimination and thereby might contribute to the learning of phoneme parameters (Teinonen et al., 2008).

The afore mentioned research demonstrates infants' ability to match auditory-visual speech at a very young age, and their preference to native language visual cues. The meaning of these findings is that visual and auditory speech stimuli are interconnected from a very young age and are crucial in infant's development of speech.

1.5.3 Auditory-Visual Development of Speech

The McGurk effect has been used to evaluate whether auditory-visual integration occurs in infants and to study the development of auditory-visual integration in

children. Burnham and Dodd (2004) investigated the McGurk effect in infants 4 months of age. To assess the presence of the McGurk effect they used a habituation test paradigm. The infants were divided into two groups, an experimental group and a control group. Each group was habituated similarly but to different stimuli. The experimental group was habituated to the McGurk stimuli, which are an auditory /**ba**/ and a visual /**ga**/. On the other hand, the control group was habituated to an auditory /**ba**/ and a visual /**ba**/. After the habituation phase the test phase began. The test included auditory only stimuli /**ba**/ or /**da**/. The auditory stimuli /**ba**/ was chosen, because it was the auditory sound presented both to the control and experimental group. The auditory stimuli /**da**/ was chosen, because it is the perceived auditory response for the McGurk stimuli. Familiarity to the sound was scored based on the infant's visual fixation on a motionless face during the presentation of the auditory stimulus. The results showed that the experimental group showed longer fixation for the /**da**/ that is the McGurk response compared to the control group. These results were interpreted as evidence of the McGurk effect occurring in infants.

Although auditory-visual integration of speech occurs at a young age, there is an abundant amount of evidence for developmental change due to maturation and experience with the native language. Auditory-visual integration as measured by the McGurk effect in English speaking children up to the age of 8 years occurred only half as often as adults (McGurk and MacDonald, 1976). Massaro et al. (1986) showed that children up to the age of 10 years were less affected by visual speech cues compared to adults. From the ages of 5 to 11 years, the effect of visual speech cues on speech perception gradually increased in English speaking children (Hockley and Polka, 1994). This development of auditory-visual integration must be due to an

increased advantage in bimodal speech perception learned over time (Jerger et al., 2009).

However, the susceptibility to the McGurk effect increased with age depended on the native language. Sekiyama and Burnham (2008) evaluated the development of the McGurk effect in a group of Japanese and English speaking children in three age groups (6, 8, and 11 years). They found that visual influence during auditory-visual speech perception improved significantly more for English speaking children compared to the Japanese speaking children. Furthermore, Mugitani et al., (2009) found that the development of auditory-visual matching of vowels was slower in Japanese speaking infants compared to English speaking infants. Their results showed that lip–voice vowel matching in Japanese speaking infants is slower at 8 to 11 months of age compared to English speaking infants at 2 to 4 months of age (Mugitani et al., 2009).

Another method of evaluating auditory-visual integration in children has been the visual fill-in effect. In this method the initial consonant of a word or syllable and the formant transition cues would be removed from the auditory stimuli while the entire word would remain intact in the visual stimuli. For example, the auditory stimulus for the word ‘bag’ would be /**ag**/ while the simultaneous visual stimulus would be the complete word ‘bag’. If auditory-visual integration occurred, then the participant would report hearing the word ‘bag’ that is the visual stimulus would fill in the gap in the auditory stimulus. Jerger et al., (2014) found that children’s auditory-visual integration ability measured by the visual fill-in effect increased between the ages of 4 to 14 years. The reduced ability in younger children to use visual speech cues has

been attributed to linguistic developmental experience and the utilization of sensory information (Jerger et al., 2014).

Maidment et al. (2015) evaluated the benefit that children gained from visual cues in identifying speech in noise. They tested children from the ages of 4 to 11 years in auditory only and auditory-visual conditions. Their results showed that young children compared to older children have lower identification ability for speech in noise. Also, children below the age of 6 years did not gain any significant benefit from visual speech cues (Maidment et al., 2015). Thus, it can be concluded that the children's ability to benefit from visual cues in identifying speech in noise increased with age.

It has been proposed that this preference in children for auditory cues over visual cues might not be specific to speech development. Thus, it can be argued that this preference for auditory cues might be due to a later developmental period for the visual system compared to the auditory system. The auditory system begins to respond to sound between the 25^{th} and 27^{th} week of gestation (Bimholz and Benaceraff, 1983), but the visual system does not reach a similar level of functioning until 6 months post natal (Banks and Salapatek, 1981). To evaluate auditory-visual integration depending on stages of child development Tremblay et al. (2007) tested children aged from 5 to 19 years on both the McGurk effect (auditory-visual speech task) and the Illusory Flash effect (auditory-visual non-speech task). They found that as children got older the percentage of the McGurk effect increased, however the results for the Illusory Flash effect were the same across the age groups (Tremblay et al., 2007). This suggests that the increase of reliance on visual speech cues seen in the development is not due to the maturation of the peripheral visual system. This

implies that the developmental increase in auditory-visual integration of speech is due to increased experience with the native language, which enables the perceptual system to create visual speech mental representations specific to the native language.

The above studies show clearly that the influence of visual speech cues on speech perception while certainly present in infants' auditory-visual integration does not reach maturity until over the age of 10 years. As children mature due to increased experience with the native language, their visual and auditory speech cues become more developed and fine tuned to the native language. This might be represented perceptually as development of auditory and visual native language mental representations. It can be seen that there is a shift in weight between auditory and visual cues in speech perception of children that is to say as children get older their reliance on visual cues during speech perception increases (Jerger et al., 2014, Maidment et al., 2015, Tremblay et al., 2007, Hockley and Polka, 1994, Massaro et al., 1986). The increase in dependency on visual speech cues suggests that the internal process of auditory-visual integration is flexible, and the weight or relevance given to the visual modality is contingent on the development of the visual speech mental representations. This has also been examined in chapter 3 where it was found that the reliance on visual cues was dependent on the native language of the listener (see chapter 3 section 3.5.1).

Children's auditory-visual integration ability reaches maturity during adolescence (Jerger et al., 2014, Maidment et al., 2015, Tremblay et al., 2007), which is much later in life compared to their early ability by 12 months to tune into their native language both for auditory and visual cues separately (Best, 1994, Weikum et al., 2007, Werker and Tess, 2002). This implies that the auditory-visual integration of

speech perception will rely on the linguistics of the native language. Perhaps, even more relevant is that auditory-visual integration develops after the ability to discriminate non-native speech sounds has decreased in children, which would further suggest that auditory-visual integration would develop based on the auditory and visual cues of the native language.

1.5.4 Auditory-Visual Neural Development

The above behavioural studies suggest that auditory-visual development is gradual and based on experience with the native language. Similar results have also been found in the neurophysiological research. For instance the peripheral auditory system is mature at birth however the maturation of the auditory cortex proceeds relatively slowly. Myelinisation of the primary auditory cortex begins around 3 months but is not complete until around 11 years of age (Moore and Guan, 2001). The peripheral visual system is immature at birth and reaches maturity at 5 months of age (Abramov et al., 1982). Auditory and visual peripheral maturity occurs early in development, however higher cortical maturation takes significantly more time to develop.

To better understand cross-modal interactions, multisensory neurons (e.g. auditory-visual) have been studied. Wallace and Stein (2001) have found multisensory neurons in new born monkeys, yet the adult monkey has double the number of multisensory neurons. The integrative abilities of multisensory neurons in adults are more refined compared to infants (Stein and Rowland, 2011). The development of modality specific neurons, such as visual neurones, progresses in an identical manner

to the development of multisensory neurons (Wallace and Stein, 2001). Postnatal experience shapes the development of visual and auditory cortical systems (Bavelier et al., 2001). The existence of neurons that respond to combined auditory-visual input provides evidence that the brain integrates information from the auditory and visual modality.

A recent electrophysiological study using the McGurk effect has shown effects in event-related potentials around 290 ms post-stimulus onset to combination stimuli (auditory /**ga**/ + visual /**ba**/ = /**gba**/) in five-month-old infants, but not for fusion responses (auditory /**ba**/ + visual /**ga**/= /**da**/) (Kushnerenko et al., 2008). This suggests that neural response profiles in the developing infant are indeed sensitive to the most salient auditory-visual discrepancies, but not tuned into more complex auditory-visual integration processes. The maturation of cognitive functions has been investigated using neuroimaging technology. It has been found that the perisylvian language areas in the cortex show a fairly long developmental course, from childhood to adolescence (Sowell et al., 2004). This lengthy maturation period of the perisylvian language areas might be related to the lengthy period for language development. Shaw et al., 2008, suggest that the developmental period for neural cortical areas responsible for auditory-visual speech integration develop late into adolescence, similar to higher order language cortical areas (Shaw et al., 2008).

MacSweeney et al. (2002) tested the effect of auditory-visual experience on the activation of auditory cortex by visual speech by comparing normal hearing participants with deaf individuals (profound hearing loss from birth) and found significantly less auditory cortex activation for visual silent speech for the deaf group than the normal hearing group, suggesting that the development of the auditory-

visual network involved in this response is affected by experience (MacSweeney et al., 2002).

The above neurophysiology studies indicate that even though multisensory neurons are present at birth, the maturation of neural integrating circuits follows a long developmental course postnatally. This implicates the possible role of sensory experience in shaping the final state of these multisensory systems. These neurophysiological studies support the behavioural studies that auditory-visual integration is a process that develops gradually from infancy to adolescence. This development would then be influenced by exposure and experience with the native language.

1.6 Cross-linguistic Studies of Auditory-Visual Integration

Auditory-visual integration has been tested frequently in English; it has only been tested in a few other languages such as Italian, Dutch, Spanish, Chinese, and Japanese (Bovo et al., 2009, Massaro et al., 1995, Sekiyama, 1997). Languages differ in their phonemes and phonotactics. Therefore, it is beneficial to look at auditory-visual integration in different languages to analyse the possible differences in results, which may help in further understanding the processing framework of auditory-visual integration during speech perception.

The McGurk effect in Italian, Dutch and Spanish native listeners was found to be similar in frequency to that of native English listeners (Massaro et al., 1993, Bovo et al., 2009, Massaro et al., 1995). However, this was not the case for Japanese native listeners where the frequency of auditory-visual integration as measured by the

McGurk effect occurrence was lower than English native listeners (Massaro et al., 1993, Sekiyama and Tohkura, 1993). Japanese and English syllables were presented to both Japanese and English listeners. The order of the McGurk effect from largest to smallest was as follows: 1) American English listeners listening to Japanese syllables, 2) American English listeners listening to English syllables, 3) Japanese participants listening to English syllables, and 4) Japanese participants listening to Japanese syllables. These results suggest that the use of visual cues during auditory-visual speech perception might depend upon the native language of the listener. Yet, when auditory masking noise was added to the stimuli the Japanese native listeners showed a high increase in the percentage of the McGurk effect (Sekiyama and Tohkura, 1991). This implies that Japanese native listeners use visual cues in a complementary nature but are less likely to use visual cues for auditory-visual integration compared to English listeners.

Sekiyama (1997) in a later study examined the McGurk effect in Chinese (Cantonese) participants who had lived in Japan from 4 months to 6 years. The stimuli were 10 syllables (/**pa**/, /**ba**/, /**ma**/, /**na**/, /**da**/, /**ta**/, /**ga**/, /**ka**/, /**ra**/, /**wa**/) spoken by two speakers, one American English speaker and one Japanese speaker. The Chinese native listeners had a lower percentage of McGurk effect compared to the Japanese native listeners (Sekiyama, 1997). In another study de Gelder and Vroomen (1992) found that Chinese listeners had poorer visual perception of /**da**/ and /**ba**/ compared to Dutch listeners (De Gelder et al., 1995). Additionally, Hazan et al. (2006) found that Spanish listeners show a much greater sensitivity to visual cues than Japanese listeners when differentiating between a non-native labial/labiodental consonant contrasts (Hazan et al., 2006).

These studies suggest that linguistic factors in the native language might assist in explaining the reduction in the use of visual speech cues during auditory-visual speech perception seen in both the Japanese and Chinese native listeners (Massaro et al., 1993, Sekiyama, 1995, Sekiyama, 1997). Chinese is a tonal language, there are four tones that are used to change the meaning of the word. For example, *mā* with a flat tone means "mother", *má* with a rising tone means "hemp", *mă* with a falling then rising tone means "horse", *mà* with a falling tone means "scold". Since tones are more effectively identified by auditory speech cues than visual speech cues (Chen, 2000) this may lead to a reduction in reliance of visual cues and an increased reliance on auditory speech cues by Chinese listeners (De Gelder et al., 1995, Sekiyama, 1997).

Furthermore, a distinct feature of Japanese is the use of pitch accents. That is in Japanese some syllables can have a high or low pitch, which would change the meaning of the word. For example, the word '*hashi'* can either be '*hashi*^{L}*'* meaning "chopsticks" or '*hashi*^{H}*'* meaning "bridge". Since pitch accent is more readily perceptible in the auditory modality than in the visual modality (Sekiyama and Tohkura, 1991) this may be why Japanese listeners rely less on visual cues compared to English listeners.

Thus, the differences in speech features across languages are relevant to the extent in which visual speech cues are used in the perception of auditory-visual speech. However, Sekiyama reported that the reduction in the McGurk effect percentage seen in Japanese and Chinese might also be due to cultural differences. In the Japanese and Chinese culture direct eye contact is disrespectful and therefore this may lead to a reduction in the use of visual cues during speech perception

(Sekiyama, 1997, Sekiyama and Tohkura, 1991). These discrepancies suggest the need for further cross-linguistic research to further investigate the effect of native language on auditory-visual integration of speech (Rosenblum, 2007).

1.7 Arabic and Auditory-Visual Speech

The choice of Arabic for this research was based on the following analysis of the literature. To evaluate what the parameters of visual speech cues are, some studies evaluated auditory-visual integration of speech across different languages. Comparing results from different languages would assist in determining what is universal in the process of auditory-visual integration of speech and what relies on the visual cue features of the native language.

Auditory-visual integration has been studied mostly in Indo-European languages such as English, Spanish, Dutch, and Italian (Massaro et al., 1993, Bovo et al., 2009, Massaro et al., 1995). These studies have shown similar results, arguing that visual cues in Indo-European languages have a strong influence on speech perception. On the other hand, studies on Chinese and Japanese have shown that there is a reduced reliance on visual cues during speech perception (Massaro et al., 1993, Sekiyama, 1995). Unfortunately, Sekiyama et al. 2008 were unable to determine whether this reduction in use of visual cues during speech perception was due to the visual cues of the language or the admonishment of eye gaze within the culture. This uncertainty in whether the culture is affecting the results is because in Chinese and Japanese cultures, looking directly at the speaker is considered disrespectful (Sekiyama and Tohkura, 1993).

To evaluate novel features of visual cues within the native language which influence speech perception, it is necessary to use a non-European language, which has a culture that does not admonish eye contact and which has ambiguous visual cues that might lead to a reduced reliance on the visual modality during auditory-visual integration of speech. Arabic is a language which fulfils these three requirements. Arabic is a Semitic language; therefore, it does not belong to Indo-European language family. Unlike the Chinese and Japanese culture, in Arabic culture eye contact shows interest and truthfulness during communication (Feghali, 1997). Therefore, any differences found in auditory-visual integration of speech for Arabic individuals as compared to English individuals can be attributed to the visual cues within the native language without any influence of cultural differences in visual contact.

The degree to which visual information is integrated in speech perception might depend on the degree to which visual information is useful in disambiguating close phonetic neighbours. There are certain linguistic features in Arabic which may lead to a reduced use of visual speech cues. Arabic is a Semitic language and, like most modern Semitic languages, it has a series of emphatic phonemes which contrast with plain phonemes. In the case of Arabic there are four emphatic phonemes they are; /**ðˤ**/, /**tˤ**/, /**dˤ** /, and /**sˤ**/. Their corresponding non- emphatic counterparts are /**ð**/, /**t**/, /**d**/, and /**s**/ respectively. However, in the Saudi dialect the emphatic phoneme /**dˤ** / is not produced as a plosive but instead as an emphatic fricative /**ðˤ**/ (Alhammad, 2014, Al-Raba'a, 2015). Emphatic consonants in Arabic are produced with a primary coronal articulation and a secondary articulation in such a manner that the back of the tongue retracts into the pharynx. There is also a sulcalisation of the tongue,

which causes the tongue to be depressed in the centre and lowering of the jaw which helps enlarge the oral cavity. It is the secondary articulation which differentiates between emphatic phonemes and their non-emphatic counterparts. To indicate that a phoneme is emphatic a subscript /ˤ/ is placed after the emphatic phoneme.

For example, /**t**/ and /**tˤ**/ are both alveolar, stop, voiceless consonants, but /**tˤ**/ is an emphatic phoneme. In addition the phoneme /**q**/ is also considered by some as the emphatic counterpart of /**k**/ (Watson, 2002, Heselwood, 1992). The emphatic phoneme **/tˤ/** and **/q/** are considered to have the most emphasis (Laufer and Baer, 1988). Another difference between emphatic and non-emphatic phonemes is the effect they have on the vowels next to them. Vowels next to an emphatic phoneme have a higher F1, and lower F2, than when they are next to non-emphatic phonemes. These differences are caused by the oral cavity enlarging which causes F2 to lower its frequency (larger spaces resonate with lower frequencies). While the pharyngeal cavity becomes smaller causing the F1 to increase its frequency (smaller spaces resonate with higher frequencies). The visual similarity between plain and emphatic phonemes might lead to an increase in visual ambiguity of speech sounds in Arabic.

Additionally, a distinct feature of Arabic is the presence of many guttural phonemes (Heselwood and Al-Tamimi, 2011, Watson, 2002).

Guttural phonemes are phonemes produced in the rear of the oral cavity from the uvula to the glottis. In Arabic there are 7 guttural phonemes **/q, χ, ʁ, ħ, ʕ, h, ʔ/** (see Table 1.3). The 3 guttural uvular phonemes **/q, χ, ʁ/** are produced by a retracted and raised tongue body. For the phoneme **/q/** and **/χ/** there is also a raising and flattening of the soft palate, while for the **/ʁ/** the soft palate is lowered which causes a constriction in the uppermost pharynx. The pharyngeal phonemes **/ħ, ʕ/** are both

Table 1.3 Consonantal Phoneme Inventory for Standard Saudi Arabian Arabic Dialect.

Manner			Place of Articulation								
			Bilabial	Labio-dental	Dental	Alveolar	Palato-Alveolar	Velar	Uvular	Pharyngeal	Glottal
Nasal	Voiced	Non-Emphatic	m			n					
Stop	Voiceless	Non-Emphatic				t		k			ʔ
		Emphatic				tˤ			q		
	Voiced	Non-Emphatic	b			d		g			
		Emphatic				dˤ					
Fricative	Voiceless	Non-Emphatic		f	θ	s	ʃ		χ	ħ	h
		Emphatic				sˤ					
	Voiced	Non-Emphatic			ð	z			ʁ	ʕ	
		Emphatic			ðˤ						
Affricate	Voiced	Non-Emphatic					dʒ				
Approximant	Voiced	Central	w			r	j				
		Lateral				l					

produced by a retraction of the tongue root, the anterior wall of the pharynx, and the epiglottis towards the posterior wall of the pharynx. The phonemes **/h,ʔ/** are both produced at the glottis, **/h**/ is produced with an open glottis while /**ʔ**/ is produced with a constricted glottis. Consequently, visual cues would probably not be very beneficial for differentiating between guttural phonemes. Having many ambiguous visual cues in Arabic may lead to a reduced reliance on the visual modality during auditory-visual integration of speech for Arabic listeners compared to English listeners.

Auditory-visual integration in Arabic native listeners was investigated in one study; (Ali et al., 2005) the participants were ten native bilingual Arabic listeners residing in the United Kingdom. Ali and colleagues found that the percentage of auditory-visual integration in these Arabic native listeners was similar to that of English

native listeners. However, they used Arabic bilingual listeners, and this may have influenced their findings. There is evidence of cross-linguistic influence between the two languages of bilingual individuals in sound perception, word meaning, word formation, and sentence structure (Kohnert et al., 1999, Kovelman et al., 2008, Paradis and Navarro, 2003). Furthermore, it has been found that during speech processing of the same native language a bilingual's brain has a significantly greater increase in activation in the classic language area (i.e. left inferior frontal cortex) compared to a monolingual's brain (Kovelman et al., 2008). In addition, percentage of auditory-visual integration is different when comparing a monolingual and a bilingual (Wang et al., 2008). Sekiyama (1997) found that, the longer Chinese native listeners lived in Japan, the higher the McGurk percentage (Sekiyama, 1997). Sekiyami (1997) suggested that when Chinese native listeners learn a foreign language their reliance on visual speech cues increases. Therefore, it is likely that experience with a second language may influence the process of auditory-visual integration for the first language.

Additionally, there was a difference of auditory-visual integration percentage between participants from the Gulf compared to other Arab countries (Ali et al., 2005). This could be due to the fact that people from Gulf countries usually live in their own country and travel just for a period of study, but Arab people from non-Gulf countries are more likely to immigrate to other countries. Therefore, the non-Gulf participants might have had more experience with English compared to the Gulf participants, which may have led to the non-Gulf participants having a higher percentage of auditory-visual integration compared to the Gulf participants. Furthermore, one of the ten participants in Ali's 2005 study was found to have poor

auditory-visual integration, he was an Arabic teacher. Consequently, his linguistic experience was more focused on Arabic and not English which may have led to his poor auditory-visual integration ability as measured by the McGurk effect (Ali et al., 2005). Therefore, it is essential when investigating auditory-visual integration in a certain language that the participants be monolingual to ensure that the second language does not affect the auditory-visual integration process.

The cross-language differences found in auditory-visual integration (De Gelder et al., 1995, Hazan et al., 2006, Massaro et al., 1995, Sekiyama, 1997, Sekiyama and Tohkura, 1991) imply that speech perception is dependent on mental representations of visual cues within the native language. Although the features of visual cues which shape auditory-visual integration of speech are still under investigation, the influence of native language on the development of auditory and visual cues has been well established (Best et al., 1988, Kuhl et al., 2006, Patterson and Werker, 2003). The inventory of visual mental representations within the native language will influence auditory-visual integration of speech. It is therefore hypothesized that since Arabic has many phonemes which are produced in the back of the oral cavity compared to English this would suggest that during speech perception Arabic listeners will be less reliant on visual cues compared to English listeners.

1.8 Summary

Auditory-visual integration of speech refers to the processing of auditory and visual information to form a unified percept based on mental representations of the native language. Visual lip, jaw, tongue, cheek and facial cues are used in addition to

auditory cues in order to process speech (McGurk and MacDonald, 1976, Desjardins and Werker, 2004). The presentation of auditory-visual speech has been found to be more intelligible than auditory speech only. The improvement gained in the intelligibility of speech perception by the addition of visual speech cues is greatest when the auditory signal is degraded, for example in a noisy environment (MacLeod and Summerfield, 1987). Even in optimal listening environments, a speech perception advantage is observed if accompanied by visual speech (Davis and Kim, 2004). During face to face speech, visual cues influence our perception of speech which helps to enhance our understanding of the listener.

Most studies on the development of speech perception conclude that infants up to the age of six months have the ability to perceive speech sounds in a language-independent manner. However by the end of the first year there is a decrease in infants' ability to perceive sounds in a language-general manner due to increased experience with the native language (Best, 1994, Polka and Bohn, 1996, Polka et al., 2009). The native language shapes the way we categorize speech sounds in a phonologically relevant way. Combining information from auditory and visual cues can affect speech perception even in early postnatal life (Burnham and Dodd, 2004, Rosenblum et al., 1997, Woodhouse et al., 2009). Behavioural studies report age-related differences in multisensory processing (Desjardins and Werker, 2004, Flom and Bahrick, 2007), and neurophysiology studies provide compelling evidence of the role of experience in the development of multisensory processing (Bavelier et al., 2001, Desjardins and Werker, 2004, Flom and Bahrick, 2007, Kushnerenko et al., 2008, Wallace and Stein, 2001). There is also a great deal of evidence that listeners' native language experience may determine the way certain visual cues are

used in speech perception (Hazan et al., 2006, Massaro et al., 1993, Sekiyama, 1995, Sekiyama, 1997, Sekiyama and Tohkura, 1991).

Thus, speech perception theories must account for both auditory and visual speech cues. These theories must try to explain the interaction between the auditory and visual modality. However, it is still not clear how these very different sensory experiences are integrated to form a unitary speech percept. Some speech perception theories propose that auditory and visual signals are integrated automatically as a function of the ability to extract non-modality specific (amodal) cues across the senses at early stages of speech processing (Burnham and Dodd, 2004, Dodd et al., 2008, Green et al., 1990, Rosenblum, 2007). Other theories propose that we analyse the auditory and visual signals and then match them to phonetic templates (mental representations) stored through learned associations in our memory at late stages of speech processing (Altieri et al., 2011, Bernstein et al., 2004a, Massaro et al., 1993). These theories argue that speech should be viewed as a form of pattern recognition in which stimuli are identified and categorized on the basis of previous experience. In order to have a complete theory of speech perception, it is essential to include the weighted function of visual speech cues and how they are integrated with auditory speech cues.

A framework suggested in this thesis is that basic auditory-visual multisensory responses may be present at birth, but that processing matures only after a period of postnatal sensory experience with the native language. Different native languages would have different visual cues and therefore this would lead to a difference in the process of auditory-visual integration during speech perception. In chapter 2 the

literature will be explored further and a framework for auditory-visual speech perception suggested.

Chapter 2

Theories of Auditory-Visual Speech Perception

2.1 Introduction

In this chapter the literature related to auditory-visual speech perception is discussed in order to hypothesize a framework by which auditory and visual cues integrate. For over half a century various speech perception theories have been developed to help understand the process behind perceiving different components of speech. Classic speech perception theories for example the TRACE Model (McClelland and Elman, 1986) and the Cohort Theory (Marslen-Wilson and Tyler, 1980) only include the auditory modality; however, more recently the evidence of the effect of visual cues on speech perception has influenced the development of speech perception theory to include the visual modality. Studies of how auditory-visual integration of speech might operate have helped in understanding the process of speech perception.

Although the literature supports the idea of speech perception as a multimodal process, the underlying framework is still under debate. An important division in the literature identifies two possible classes of theory to explain auditory-visual affects on speech perception; early theories (amodal) and late theories (modal) of auditory-visual speech perception. Researchers have debated whether auditory and visual information is combined early on into a unified code (early integration theories), or instead is processed in separate independent channels before final determination of the linguistic context (late integration theories).

Early integration theories are considered amodal theories, that is auditory-visual integration of speech is a property of the input information itself (Rosenblum, 2008). Hence early integration theories (amodal theories) do not depend on the auditory and visual mental representations of the native language. On the other hand late integration theories are considered modal theories, that is auditory-visual integration of speech depends on the auditory and visual mental representations within the native language (Altieri et al., 2011, Bernstein et al., 2004a, Rosenblum, 2008). In this chapter both early and late integration theories of speech perception are discussed. Furthermore, a framework for auditory-visual integration in speech perception that is dependent on the visual cues of the native language proposed in this thesis is explained.

2.2 Early Integration Theories of Auditory-Visual Speech Perception

Some researchers (Green et al., 1990, Rosenblum, 2007) suggest that the automatic and total integration of auditory-visual speech occurs due to the processing of speech cues without the need for learned mental representations (Campbell and Dodd, 1984). From an early viewpoint, this framework of auditory-visual speech perception does not differentiate between these different modalities and holds that there is a common representation of speech. Thus, supporters of early integration of speech suggest that each auditory and visual unimodal source of information contains inherently amodal information at the most basic level. When the input activates the speech processing regions of the brain, the underlying amodal information from each source is extracted and combined because both sources share a common means of transfer, a “common currency”. The information from the two

modalities could be combined into a single channel before the process of phonetic recognition in which the decision process considers only the totality of the information and not the auditory and visual parameters in the separate modalities (Rosenblum, 2007).

The theoretical basis for some of the theories founded on early speech perception is based on the gestural theories of speech perception. These theories make the assumption that the linguistic representations extracted from the signal are gestures. The most famous, of course, is the first gesture theory by Liberman, which is also known as the motor theory of speech perception. In essence Liberman suggests that the object of speech perception is not the auditory signal, but the representation of the articulatory gesture. By articulatory gesture Liberman meant the invariant configurations of the teeth, tongue, lips, jaw etc. that make up a phonetic segment. In the motor theory visible speech cues are important since they are the vessel through which the articulation gestures of the speaker are reflected (Liberman et al., 1967).
Liberman and his colleagues explored the auditory cues of perception with the means of the sound spectrograph and also pattern playback (Liberman et al., 1967). One of the significant findings of their experiment was that auditory cues for consonants were incredibly sensitive to context, which was conditioned by coarticulation. Lieberman found that in the identification of the synthetic syllables /**di**/ and /**du**/ the transition of the second formant was crucial. Although in the case of /**di**/ transition is high and rising, and in the case of /**du**/ is falling and low, in the context of each syllable, the consonants sounded alike to listeners (Liberman et al., 1967). On the other hand, taken out of context, they sound different. Liberman's conclusion was that except for contextual sensitivity, both syllables were produced identically a

constriction of the tongue tip behind the teeth. Consequently, listeners' perception was based on speaker's articulation (Altieri et al., 2011). Further research demonstrated that stop consonants can be recognised through their formant transitions or "*based on a burst of energy that, in the natural speech, precedes the transitions and occurs as the stop constriction is released*" (Weiner & Freedheim, 2003, p. 255). Based on these findings, Lieberman questioned which stimulus becomes primary in perception articulation or sound. His conclusion was that "*the perception always goes with articulation*" (Lieberman, 1957, p. 121) .

Another representative of this school is Fowler and her Direct Realism theory (Fowler and Smith, 1986). This theory holds that speech perception is not mediated by representations, but it is a property of the input information itself; that is speech is perceived by the signals, for example for visual cues it is the light patterns and for auditory cues it is the patterns of changing air pressure. Like motor theorists, Fowler claimed that the objects of speech perception are not auditory but articulatory phenomena; however, she denied the specific processes necessary for speech perception. Instead, she argued that speech signals contain rich information that listeners can detect irrespective of cognitive processes of inference. The realist nature of this theory is conditioned by the belief that listeners recover the physical properties of the articulated phonetic gestures from the auditory signal (Altieri et al., 2011). Thus, the central idea in early integration theories is that both visual and auditory speech cues carry gestural information in its most elementary level. The input from both the auditory and visual modality is transformed into a common code prior to integration.

The main difference between these two theories is that the motor theory relies on accessing one's own gestural representations as triggered by exposure to someone else's speech, whereas direct realism relies on direct perception of the speaker's gestures through a 'transparent' auditory signal. In other words, the central difference between Lieberman's motor theory and Fowler's direct realism is that while the first one argues that the listener's own vocal gestures are the objects of perception, the second theory suggests that the speaker's gestures that are perceived directly are the objects of perception (Heselwood, 2013).

The main criticism of direct realism theory is that it seems to assume that the perceptual systems have no effect on the representation of the stimulus and that perceptual objects are identical to external objects. In this regard, it is criticised for ignoring the filtering function of the auditory system that is aimed at reshaping the properties of pressure-waves into psychoauditory objects, meaning cognitive images are based on processing of the given information (Heselwood, 2013). Another criticism of direct realism and its immediate perception premise is that perception involves numerous causal series and physical processes which occur with different speed and add different aspects of information for the formation of the final speech perception (Le Morvan, 2004). Thus, direct realism cannot explain the entire spectrum of processes and factors influencing speech perception.

On the other hand, there are other researchers that support early integration but not gestural theories of speech (Burnham and Dodd, 1996). Dodd et al., (2008) compared the percentage of McGurk effect between children with phonological delays and those with phonological disorders. It was suggested that since both groups have the ability to extract gestural information from articulation, any

difference found between the two groups could only be due to deficits in phonological processing. The group with phonological disorders perceived the McGurk effect less than the group with phonological delays. The difference was taken as evidence that speech perception is based on phonological information and not gestural information (Dodd et al., 2008). The Phonetic Plus Post-Categorical Model (Burnham, 1998) proposes an early model of speech perception which is based on phonological information integrated from both modalities at early stages of processing (Burnham and Dodd, 2004, Dodd et al., 2008).

Burnham (1998) proposes that auditory-visual integration of speech occurs initially without any influence of phonological prototypes of the native language. He supports this by evidence of auditory-visual integration being present in young infants. He proposes that any cross-linguistic differences occur due to post-categorical effects based on the native language (Burnham and Dodd, 2004, Dodd et al., 2008). Infants at 10 weeks of age have been found to match auditory-visual speech at a similar percentage for native versus non-native speech. However, by the age of 20 weeks infants have a preference in matching auditory-visual native speech compared to non-native speech (Dodd, 1979, Dodd and Burnham, 1988). These results suggest that we begin with a universal auditory-visual speech perception process. Yet, as we become more experienced with the native language auditory-visual speech perception becomes dependent on the visual mental representations within the native language.

Irrespective of the difference in their explanation of the process of auditory-visual integration, early speech perception theories propose that speech is perceived by deciphering modality independent speech information (a common metric whether

gestural or phonetic) and this occurs at the early stages of speech perception. In the next section behavioural studies supporting early integration of speech are reviewed.

2.3 Support for Early Integration Theories of Speech Perception

The support for early integration theories comes from speech science where it has been suggested that there is no one-to-one association between a phonetic segment and a set of auditory cues, while articulation gestures can more effectively describe phonetic segments (Rosenblum, 2007). The early integration theories are often supported by behavioural studies. One example is the research by Green and Miller (1985), showing that visual cues for percentage of articulation influences the perception of voice onset time (VOT). During the experiment, the participants were shown auditory-visual clips of a speaker saying a syllable in the continuum from /**bi**/ to /**pi**/. The visual information corresponding to this continuum was played at a different pace, either fast or slow. The outcome demonstrated that syllables being articulated rapidly increased the probability of /**bi**/ being perceived as /**pi**/. In terms of the support of early integration theories, the authors suggest that this is evidence of the integration of auditory and visual information in the early stages of phonetic perception (Green and Miller, 1985). However, Bernstein (2005) argues that this is due to a learned predictable association between auditory and visual speech input and that integration occurs later in the process of speech perception (see section 2.4).

Some auditory-visual studies have been used to support early integration theories of speech perception. For example, the McGurk effect has been found to occur even when the sound being dubbed is produced by a man and the visual speech cues is

produced by a woman or vice versa (Green et al., 1990). That is a reduction in cognitive congruency does not reduce the strength of the McGurk effect. This was taken as evidence that higher cognitive properties do not reduce auditory-visual integration, which would support the early integration of speech. In other words, the fact that cognitive differentiation of genders had no impact on the strength of McGurk effect suggests that cognition was not a crucial component in speech perception. Conversely, recent findings have shown that auditory-visual integration of speech is influenced by higher cognitive, semantic, and lexical processes, which will be discussed in detail in section 2.5.

Further evidence for early integration comes from studies on speech perception in infants. Research has demonstrated that infants can match auditory speech to the appropriate visual lip movements at 2 months of age for vowels (Patterson and Werker, 2003) and 6 months of age for consonants (MacKain et al., 1983). These studies on infants demonstrated that the auditory and visual speech streams are entwined in the earliest stage of perception, which precedes even word recognition (see chapter 1 section 1.5.2). The presence of the McGurk effect in young infants (Burnham and Dodd, 2004) was seen as further support for the early and immediate auditory-visual integration of speech (see chapter 1 section 1.5.3). In other words, it is argued that this demonstrates that auditory-visual integration occurs before the development of clear mental representations of speech. In the next section theories based on late auditory-visual integration of speech will be discussed which oppose early auditory-visual integration theories.

2.4 Late Integration Theories of Auditory-Visual Speech Perception

Late integration of speech states that auditory and visual speech processing result in separate modality specific representations. Late integration theories of auditory-visual speech perception propose that we deconstruct the auditory and visual signals into segments. Then these perceptual segments are matched with templates or mental representations stored through learned associations in our memory (Bernstein et al., 2004a, Massaro, 1987). These theories suggest that speech is categorized based on language specific mental representations for the auditory and visual inputs.

The Fuzzy Logical Model of Speech Perception (Massaro, 1987) is one example of a late integration theory of auditory-visual speech perception. This theory is based on the idea that speech stimuli arriving via the auditory or visual modality are processed separately prior to the integration process. This initial processing creates a summary description for the auditory and visual information individually. These summary descriptions are compared separately to mental representations within the memory in order to define how well these auditory and visual speech cues align with mental representations stored in the memory. The evaluation of speech cues is described as a process in which the sensory systems compare modality specific stimulus features with ideal features that make up category mental representations in the memory. That is integration occurs after labelling occurs, which is referred to as a post-labelling model of speech (Seldran et al., 2011). At the final stage of the Fuzzy Logical Model of Speech Perception the auditory and visual stimuli are integrated together. For example, to explain the McGurk effect, the Fuzzy Logical Model of Speech Perception states that the mental representation /**da**/ is selected based on the phonetic features that the auditory /**ba**/ and visual /**ga**/ signals have in common.

Overall, the main assumptions of the model are that there are four stages: (1) the features of the auditory and visual modality are first evaluated independently (2) the features from both modalities are integrated (3) the result of the integration is compared to the mental representations available in memory (4) perceptual identification is based on the most reliable mental representation to produce a general measure of best fit (Massaro, 1998). In other words in the Fuzzy Logical Model of Speech Perception the selection of a particular perceptual category is chosen based on the mental representation in memory that best matches the phonetic information afforded by the auditory and/or visual signals. The Fuzzy Logical Model of Speech Perception has been able to reliably model human data obtained in many speech perception studies (Massaro and Light, 2004).

Similarly, Braida (1991) proposed a Pre-labelling Model which is a modality specific model. In this model, the auditory and visual speech cues are processed separately which then leads to a multi dimensional vector that characterizes the speech sound. This vector is then mapped to a category label and speech is perceived, that is integration occurs prior to labelling. The Pre-labelling Model suggests that that auditory-visual speech perception optimizes the use of modality specific speech input (Braida, 1991).

The main difference between the Pre-labelling Model and the Fuzzy Logical Model of Speech Perception is their assumption about whether speech integration is continuous or categorical. The pre-labelling model suggests that continuous sensory data is combined across modalities before response labels are assigned (Seldran et al., 2011). Hence integration occurs before a response decision is made for each modality. On the other hand the Fuzzy Logic Model (post-labelling) categorizes the

input from each modality separately before integration occurs. This model suggests that integration occurs after summary descriptions for speech information from each modality has been made.

However, both of these models suggest that auditory and visual inputs are processed separately initially and the features are compared to mental representations which are specific to the native language. Next there is a weighting of the auditory and visual input based on how well they match mental representations within the native language. In other words the greater the predictive power of the auditory or visual input is the greater its influence on the perceived speech. Therefore, auditory-visual integration is suggested to occur at late stages of the process of speech perception.

2.5 Support for Late Integration of Speech Perception

This section will review the research supporting late integration of auditory-visual speech. That is, the following studies suggest that auditory and visual speech cues are integrated not at the initial input level but at a later level in the speech perception process. In order to find out whether there is a late influence on auditory-visual integration of speech, Walker et al. (1995) presented McGurk stimuli to participants who were familiar or unfamiliar with the faces of recorded talkers. The participants who were familiar with the talkers were significantly less susceptible to the McGurk effect then in cases when faces and voices were unknown (Walker et al., 1995). This suggests that there is a cognitive or top-down influence on auditory-visual integration for familiar speakers.

The authors argue that the McGurk stimuli contradict the perceiver's expectations more readily with familiar speakers. The relevance of this research is that familiarity or experience changes the relative importance of different dimensions of visual mental representations, placing emphasis on the recognition of familiarity rather than only early input or amodal information perception of auditory-visual speech. Consequently, in terms of late integration models where mental representations and experience are crucial, Walker's (1995) experiment demonstrates that auditory-visual integration is conditioned by cognitive mental representations of familiarity. Therefore, these findings argue against the notion of auditory-visual integration occurring automatically at an early stage which is not influenced by auditory and visual speech mental representations. This supports the notion that speech perception depends on experience with the native language which forms specific auditory and visual mental representations.

Recently, there has been some research on the lexical modulation of auditory-visual speech perception (Barutchua et al., 2008). It was found that the McGurk effect was elicited in real words when the auditory-visual discrepancy was placed at the beginning of the word. However, when the auditory-visual discrepancy was placed at the end of the word the McGurk effect was not elicited (Barutchua et al., 2008). The explanation given was that at the offset of the word the perception of the word has already been formed; therefore, the auditory-visual discrepancy will not be able to elicit the McGurk effect. Whereas, when the auditory-visual discrepancy was placed at the onset of the word the perception of the word had not yet been formed; thus, the McGurk effect was elicited. In another study, it was found that the McGurk responses were elicited less frequently when the auditory input formed a real word

and the McGurk input was a pseudo-word (Brancazio, 2004). For example, for an auditory input /**bat**/, and a visual input /**gat**/ they predicted that the McGurk effect would elicit /**dat**/, however the participant reported /**bat**/. There seems to be a lexical modulation of speech perception before auditory-visual integration occurs (Brancazio, 2004, Barutchua et al., 2008). Therefore, this is further evidence that auditory-visual integration occurs at a later stage of speech and that it is not automatic.

There also has been research conducted on the effect of semantic cueing on the McGurk effect. The Encyclopaedia of Clinical Neuropsychology gives the following definition of semantic cue: "*is a prompt that contains semantic information, and is given to facilitate word retrieval. Semantic information is knowledge that is related to the meaning of the word. This may include a formal description or definition, word/phrase associations, sentence completion and perceptual information*" (Kreutzer, 2011, p. 2241). Depending on the amount of the provided semantic information, semantic cue can be either strong or weak. Its most common use is in standardized naming tests that focus on one's naming capacity. The relevance of semantic cuing for a working framework of auditory-visual speech perception is in its cognitive function. In other words, semantic cuing is based on one's ability to contextualise a certain word and meaning within the existing language system. Therefore, it is based on mental representations which one acquires learning words and the context associated with it within the native language.

Sharma (1989) conducted an experiment of a positive semantic cueing, meaning that understanding of the word was conditioned by favourable lexicological or contextual information. In other words, the sentence context was structured to favour the

expected fusion this lead to an increase in the McGurk response (Sharma, 1989). For example, with the sentence "Letters are stamped with today's [bait (auditory) - gate (visual)]" the expected McGurk response, (date) was more prevalent. Following Sharma's experiment, further researchers used negative semantic cueing where the sentence context was structured to favour either the auditory input or the visual input, but not the expected McGurk effect (Windmann, 2004). For example, in the sentence 'Two peas in a [pod (auditory) - Todd (visual)]' and the expected McGurk response was (cod), the fusion percentage is likely to be reduced because the sentence context brings semantic bias in favour of the word in the auditory channel (pod) (Ali, 2007).

This type of negative semantic cueing did decrease the McGurk effect percentage compared to isolated words, but in a few sentences the McGurk response actually increased. For example, the sentence 'Where the tongue [slips (auditory) - slicks (visual)] it speaks the truth' the expected McGurk response is (slits), it was found that the McGurk response increased to 63% when compared to isolated words of 40% (Ali, 2007). This suggests that negative semantic cueing can lower the percentage of the McGurk effect, but that it does not block the McGurk effect completely even when the negative semantic cueing is strong. It was also found that semantic cueing was strongest when the McGurk effect word was placed at the end of the sentence than at the beginning of the sentence. The semantic studies highlight that the semantic context and word meaning can influence auditory-visual integration (Windmann, 2004, Sharma, 1989, Ali, 2007). Overall, these speech perception studies suggests that while auditory-visual integration occurs prior to word

identification, there also seems to be a cognitive, lexical, and semantic top-down modulation at this stage of speech perception.

Additional support for late integration of auditory-visual speech comes from developmental studies. It has been found that the ability to integrate auditory-visual speech does not mature until 11 years of age (Hockley and Polka, 1994). That is to say that while the perception of auditory speech signals matures at an early age of 6 years, visual speech cues development matures at 12 years of age (Sekiyama and Burnham, 2008). This implies a separate development process for auditory speech compared to visual speech. These findings are evidence in support of late integration theories where the maturation of each modality might be different due to different cues and developmental characteristics of mental representations of speech.

Furthermore, cross-linguistic studies on auditory-visual speech support the late integration model of speech. For example, children with different native languages have shown a difference in auditory-visual integration development (Sekiyama and Burnham, 2008). The study showed that Japanese children's auditory-visual integration abilities as measured by the McGurk effect while increasing by age were significantly lower than English speaking children. Kuhl and Melzoff (1996) suggest that the representations that combine multimodal information are largely influenced by the early linguistic environment to which an individual is exposed (Kuhl and Meltzoff, 1996). These results suggest that each language has its own correlation of auditory-visual stimuli, meaning that in some languages the auditory modality can be predominant and result in weakening of McGurk effect.

Recently, Fava et al (2014) investigated auditory-visual native and non-native speech for children from the ages of 3 to 14 months. They used an infrared spectroscope to

measure changes in blood flow in the temporal cortex. They found that initially the blood flow in the temporal cortex in infants was the same for native and non-native auditory-visual speech. However, by the age of 12 months the amount of blood flow to the temporal cortex was significantly greater for native speech compared to non-native speech. This suggest that initially infants react the same to native and non-native auditory-visual speech, but by the age of 12 months the children are tuned into native auditory-visual speech compared to non-native auditory-visual speech (Fava et al., 2014). These studies suggest that the perception of auditory-visual speech relies on the native language mental representations for speech. Therefore, these results support the late integration theories of speech perception.

2.6 Factors Influencing the Auditory-Visual Integration Framework during Speech Perception

In this section four factors related to auditory-visual speech perception are discussed in order to propose a working framework by which auditory and visual cues integrate. They are:

1. The effect of native language experience and development on auditory-visual speech perception.
2. Ambiguity of visual speech cues
3. Auditory-visual weighting
4. Speech assimilation

2.6.1 Native Language Experience and Development

There are a number of speech theories that are based on the influence of the native language on mental representations of speech. For example, Kuhl's Native Language Magnet theory (Kuhl, 1991) argues for linguistic conditionality of speech perception. This theory assumes that perceptual space is divided into phonetically conditioned categories; they are represented by category mental representations or also known as "*category's best exemplar*" (Lacerda, 1995, p. 140). These mental representations function as "perceptual magnets" attracting exemplars corresponding to their area of influence. In mathematic calculation this finding corresponds to a formula in which "*discrimination is proportionate to the square or the cube of the auditory distance between the mental representation and the exemplar*" (Lacerda, 1995, p. 140).

In other words, the relevance of this theory is that it outlines a framework through which phonetic perception is altered by native language experience. In this regard, the magnet effect demonstrates that the impact of the native language results in the distortion of the initially perceived distances between stimuli (Thyer et al., 2000). Thus, the native language experience distorts the auditory space. It is argued that speech is perceived and processed through a distorted lens, which depends on the native language. Consequently, the difference between two sounds perceived by an individual in one native language might not even be noticed by an individual with a different native language. For example Arabic listeners find it hard to distinguish between the English phonemes /**b**/ and /**p**/ (see chapter 3, section 3.4).

Kuhl's testing of the magnet effect in adults and infants demonstrated that it was strong both in adults and infants. However, for infants the mental representations

were still in the process of development as discrimination between sounds was not as accurate as in adults. This was further shown in the study of different age groups, with a significant increase in recognition within mental representation groups with the increase of age and correlated increase in the cognitive function of the brain (Kuhl, 1991). Thus, this model also argues in favour of the developed rather than inborn nature of speech perception.

Another relevant theory is Peter Jusczyk's Word Recognition and Phonetic Structure Acquisition model (1997). In terms of speech perception, he argued in favour of innately guided learning. He proposed that the cognitive system of infants uses this innate predisposition in order to perceive and learn to process speech dependent on the language spoken in their environment. This is conducted through warping the perceptual space based on the features of the native language. Perceptual speech facilities gradually become tuned to perceive the native language (Jusczyk, 1997). The model argues that the preliminary speech perception level of the child's brain is limited but can be expanded with experience.

Consequently, the multiple tokens will correspond to multiple representations. One of the components of the model is a weighting scheme that attracts attention to essential language specific features, development of which gradually stimulates the transformation of perceptual space. These language specific features are then stored in memory. These memorised components become the basis of child's lexicon. Finally, comparing new patterns with the memorised ones (traces) takes place. He writes: "*as more tokens of each utterance are collected, more traces will be activated by new tokens, so recognition of patterns will soon become more efficient and will eventually lead to extraction of words*" (Jusczyk, 1997, p. 112).

Although Jusczyk argued that the ability to learn speech is innate, in terms of auditory-visual speech perception he argued for conditionality of a linguistic environment to which an infant was exposed. The research conducted by Polka and colleagues followed the path of lexical context of auditory-visual integration and took into account works of two previous researchers. The obtained findings (Polka et al., 2001, Sundara et al., 2006) proposed the existence of an early perceptual system that is capable of discriminating most contrasts of the world languages. Through the interaction with the language input, infants begin to demonstrate a higher sensitivity to the sounds particularly relevant for their native language and become more ignorant to the contrasts that are linguistically irrelevant for their native language.

Complicated learning is involved in the perception of speech sounds, one theory that tries to explain this process is the exemplar theory. Exemplar theory proposes that we have auditory representations or mental representations of speech within our cognitive space (Johnson, 2006). When we hear a sound it is compared to mental representations already stored from experience with our native language. There are parameters which are still unclear that measure the speech signal to find a match in our cognitive space. For this theory to be comprehensive it needs to incorporate visual mental representations of speech and not only auditory mental representations.

Since languages differ in their visual and auditory speech cues (Paradis and Navarro, 2003), one potential explanation for the variation of auditory-visual integration of speech observed across languages (Hazan et al., 2006, Massaro et al., 1993, Sekiyama, 1997) is that information might be extracted from auditory-visual speech dependent on the specific mental representations of the visual and auditory cues within the native language. Therefore, when a person for example sees and hears the

phoneme /**t**/ they will compare the visual and auditory cues to the mental representations they have within their cognitive space based on their native language repertoire. Establishing mapping from an auditory-visual input space to a perceptual space is a developmental process that depends on language experience (Kuhl et al., 2008).

Ortega-Llebaria et al. (2001) examined the identification of consonants in auditory and auditory-visual conditions among Spanish and English native listeners. They found that the Spanish listeners only benefitted from visual cues which were present in their native language (Ortega-Llebaria et al., 2001). Difficulty in perceiving non-native visual speech categories demonstrates that speech is perceived through the lens of the native language visual categories. The perceptual space changes to reflect the regularities of the native speech input (Kuhl et al., 2008). However, this still does not explain why some languages rely less on visual cues during auditory-visual speech perception compared to others.

2.6.2 Ambiguity of Visual Speech Cues

A bimodal perceptual speech system will increase identification of the speech signal due to redundancy. In the centre of this argument is a redundancy hypothesis, which suggests that when information is represented across two sense modalities this attracts attention and assists perceptual differentiation more productively than if the equivalent information was presented by only one modality. Furthermore, bimodal stimulation can facilitate perceptual learning (Reynolds and Lickliter, 2003).

However, for speech perception to be an efficient system it must rely on patterns or cues that are clear and unambiguous.

One of the best examples is the Cohort speech perception model developed by Marslen-Wilson (1987). According to this model, auditory and visual input corresponds or is mapped to listener's lexicon. Mapping to the lexicon is how speech mental representations might be structured or interrogated. Every time an individual begins to hear a word it activates all elements in the lexicon that start with the same phoneme, with each phoneme added variations from the lexicon are filtered finally ending up with the correct word (see Figure 2.1). Consequently, in terms of this model, words compete for recognition which is determined by how many words share an onset pattern (Marslen-Wilson, 1987). Some words have many competitors,

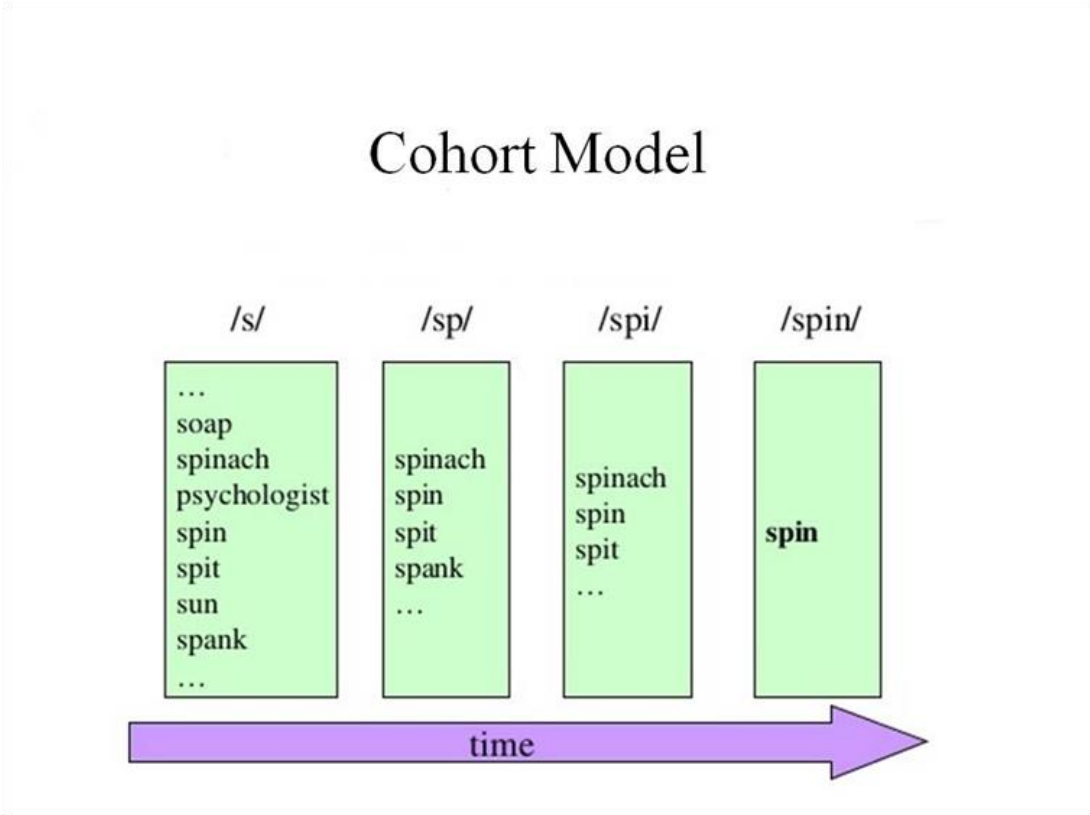

Figure 2.1 Cohort Model (Marslen-Wilson, 1987).

whereas others are subject to much less competition. The moment when only one real word is consistent with all of the input received is called the uniqueness point. Research has shown that word recognition takes place sooner for words with early uniqueness points (Harley, 2009).

Another well known model that discusses ambiguity is the Neighbourhood Activation Model (Luce and Pisoni, 1998). This models spoken word recognition as the identification of a target from among a set of activated candidates. All words phonologically similar to a given word are in the word's neighbourhood. Words that differ by only a single phoneme were considered in the same auditory neighbourhood. The difference could be due to a sound substitution like 'bat' and 'cat', a sound deletion like 'bat' and 'at', or a sound addition like 'bat' and 'bait'. Auditory recognition of a word is based on the probability that the stimulus word was presented compared to the probability that other words in the neighbourhood were in fact presented. A neighbourhood can either be described as being sparse or dense. When there are only a few words that are similar to the target word the neighbourhood is described as sparse for example the word 'song'. However, when a word has many words that sound similar to it then the neighbourhood is described as dense for example the word 'cat' (Grant, 2002). That is probability is influenced by lexical frequency (see Figure 2.2).

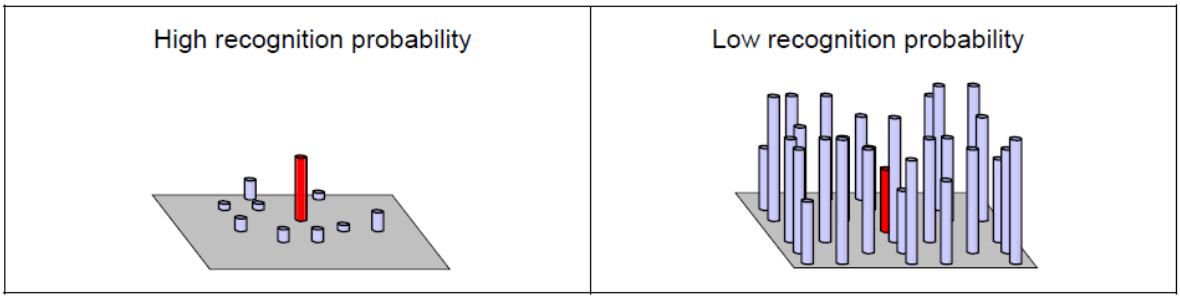

Figure 2.2 Neighbourhood Activation Model (Luce & Pisoni, 1998).

Therefore, non-ambiguous words will have a stronger and faster lexical activation, yielding stronger predicting affects (Chan and Vitevitch, 2009, Goldstein and Vitevitch, 2014). This implies that ambiguity affects speech perception negatively. Furthermore, the effect of visual lexical competition on speech perception has been investigated in a number of studies. Words that differ by only a single viseme (basic unit of visual speech) were considered in the same visual neighbourhood, for example 'fork' and 'ford'. Mattys et al. (2002) showed that the accuracy for lip-reading words varied as a function of the number of words that were visually similar to the stimulus word. That is words with many visual neighbours were harder to identify than words with fewer visual neighbours (Mattys et al., 2002). Auer (2002) found similar results with identification of auditory-visual words, which suggest that speech perception is influenced by the visual neighbourhood density (Auer, 2002). Felt and Sommers (2011) also examined the influence of visual neighbourhood density on consonants and vowels in a phonemic context to minimize co-articulation effect. The results support previous research that visual speech identification is correlated with the visual mental density.

Tye-Murray et al. (2007) investigated the effect of the lexical neighbourhood for auditory, visual, and auditory-visual speech which was referred to as the intersection

density. Words that differ by only a single phoneme were considered in the same auditory neighbourhood. Words that differ by only a single viseme were considered in the same visual neighbourhood. They presented words in the auditory-visual condition which had similar density for auditory and visual neighbourhoods. The results showed that identification of words improved when the target word had sparse intersection neighbourhood density (see Figure 2.3). In this example the word 'fish' had a higher correct identification percentage compared to 'fork' although the auditory and visual neighbourhood densities were similar. They concluded that the difference in identification percentage was due to the difference in intersection density (Tye-Murray et al., 2007). These results show that visual neighbourhood density influences auditory-visual speech perception.

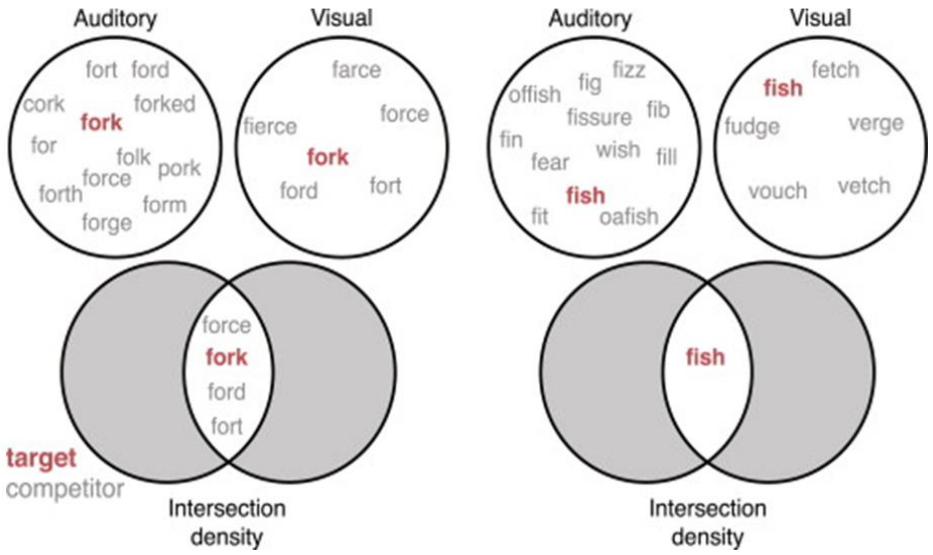

Figure 2.3 Illustrates auditory, visual, and auditory-visual lexical neighbourhood density taken from Tye-Murray et al. (2007).

These studies suggest that the influence of visual speech is determined by the amount of competition in the speech perceptual space among visually similar words within the native language. These models suggest how mental representations might be structured or interrogated and this is an account that can add to the fuzzy logic model component related to how mental representations are weighed.

2.6.3 Auditory-Visual Weighting

It has been proposed that auditory-visual integration of speech might rely on a weighting framework of the visual and auditory cues (Massaro et al., 1993). Hazan and colleagues (2006) also argue for the necessity of studying frameworks of relative weighting of visual and auditory cues in terms of distinctiveness of visual cues. The study explored to which extent learners of a second language are sensitive to information included in a visual stimulus, when asked to identify a non-native phonemic contrast. The study consisted of Spanish and Japanese learners of English. The authors tested the perception of labial consonant contrasts in auditory only, visual only and auditory-visual conditions. While both groups performed best in the auditory-visual condition the Spanish group performed better in general, demonstrating greater sensitivity to visual stimuli than the Japanese group. The findings demonstrate that the weight of visual cues on speech perception is dependent on the participants' native language (Hazan et al., 2006). Similar research on how individuals from different native languages weigh auditory and visual inputs differently include Sekiyama (1997) (see chapter 1, section 1.6) and Ortega-Llebaria et al. (2001) (see section 2.6.1).

Further, evidence for the weighting framework of auditory-visual speech perception can be found in the results of MacDonald et al., (2000). They applied visual degradation filters to the McGurk effect. They presented dubbed stimuli at various visual degradation levels (videotaped images of a speaker's face were quantised by a mosaic transform). They found that coarser visual input caused a reduction in the number of McGurk effects. Interestingly, they also found that as visual degradation increased, the clarity of the auditory stimuli was reported to increase as well. In other words, when the visual stream was more degraded, participants reported the auditory stream as being perceptually clearer. It was concluded that the participants were able to modulate (or weight) their use of visual and auditory information based on whatever modality was clearer (MacDonald et al., 2000).

Support for this also comes from Huyse et al. (2013) they concluded that auditory-visual speech perception is a flexible process which is modulated by the predictive power of visual speech cues (Huyse et al., 2013). In addition, Brunellière et. al, (2013) compared the latency processing speed for N1 evoked potential for words that begin with strong or weak visually salient visemes. They concluded that the facilitation in processing of auditory signals appears to be directly a function of the predictive power of the visual cues (Brunellière et al., 2013). This is further support for the flexibility of the perceptual system, and suggests that auditory-visual speech advantage reflects a complicated interplay of both auditory and visual sensory systems.

2.6.4 Speech Assimilation

The Perceptual Assimilation Model (Best, 1994) suggests that non-native phonemes will be categorized to the closest phonological category based on their native language mental representations (Best, 1991, Harnsberger, 2001, Nagao et al., 2003). In other words, when a person is exposed to a non-native auditory sound, they will categorize the non-native auditory cues to the closest existing auditory speech category based on their mental representations of the native language. For example if an Arabic listener hears /**p**/ they do not have an auditory representation for that phoneme, therefore assimilation will occur to the closest category in this case /**b**/.

A framework hypothesized in this thesis is that this assimilation will also occur for visual speech cues. Thus when a person sees a visual speech cue which is not in their visual repertoire they will assimilate to similar visual mental representation within the native language. For example when English listeners see Arabic /**qa**/ they do not have this visual representation in their native language, therefore visual assimilation occurs to a similar visual phoneme within the native language which might be in this case /**ka**/.

These results provide support for the hypothesis that auditory-visual integration framework depends on auditory and visual native language mental representations, which was demonstrated in analysed literature above (Massaro et al., 1993, Ortega-Llebaria et al., 2001, Sekiyama and Tohkura, 1993, Hazan et al., 2006).

2.7 Working Framework of Speech Perception

Based on the results of the previously mentioned studies, a mental representation framework with late integration of auditory and visual speech signals is proposed in this thesis to explain speech perception. Models of speech perception development incorporate auditory dimensions that map onto cognitive mental representations of speech categories (e.g. phonemes) that depend on the native language (Kuhl et al., 2006, Massaro and Friedman, 1990). Based on the literature review, it is expected that auditory-visual integration of speech will also depend on the native language where visual dimensions augment the auditory mental representations of speech sounds. Furthermore, a framework suggested in this thesis for auditory-visual integration for Arabic listeners is that experience with the visual cues within the native language will fundamentally shape the mental representations of speech. These auditory and visual native language mental representations will influence the perceived auditory-visual speech. In other words, native Arabic listeners' perceptual space will be tuned for the regularities of Arabic visual and auditory cues.

Additionally, a working framework in this thesis is that speech perception is not dominated by either the auditory or visual modality. The dominance is determined by the estimate of how reliable the information in a modality is for a specific stimulus. Therefore, the extent to which visual cues influence speech perception depends on how reliable the information is assessed to be by the perceptual system. Thus, the more ambiguous a visual cue is the less reliable or the less weight it will incur during the auditory-visual integration process of speech. By integrating speech information by this weighting framework, the predictive power of the perceived

speech signal is increased. This weighting framework yields the most reliable unbiased estimate possible.

In other words, the weighting framework of speech perception differs in Arabic compared to English. The dependency of the speech perception process on visual speech cues compared to auditory speech cues might be due to a weighting framework based on the features of the visual cues within the native language. Therefore, for some languages the perception of speech relies more on visual cues than it does for others, depending on the density of the visual neighbourhood, which is the perceptual space populated by visual cues for the phonemes within the native language.

The main hypothesized factors for the auditory-visual native language speech perception framework are that:

1. Auditory-visual integration of speech happens at a late stage in perceptual processing (see section 2.4 and 2.5).

2. Perception relies on auditory and visual native language mental representations (see section 2.6.1).

3. Visual cues are integrated depending on the predictive power and weight they provide (see section 2.6.2 and 2.6.3).

4. Non-native auditory and visual speech cues undergo assimilation to native auditory and visual mental representations of speech (see section 2.6.4).

Figure 2.4 summarizes a working framework for auditory-visual integration of speech perception. A framework includes the analysis of both auditory and visual cues. Speech is recognized in a series of three stages based on the Fuzzy Logic model of perception (Massaro, 1987). In the first stage, speech is analysed in terms

of the general parameters of auditory cues and visual cues that are present in the signal itself. Examples of auditory cues are formant frequency, VOT, and voicing. Visual cues can consist of the movement of the speech articulators such as the lip, jaw and tongue (Munhall et al., 2004).

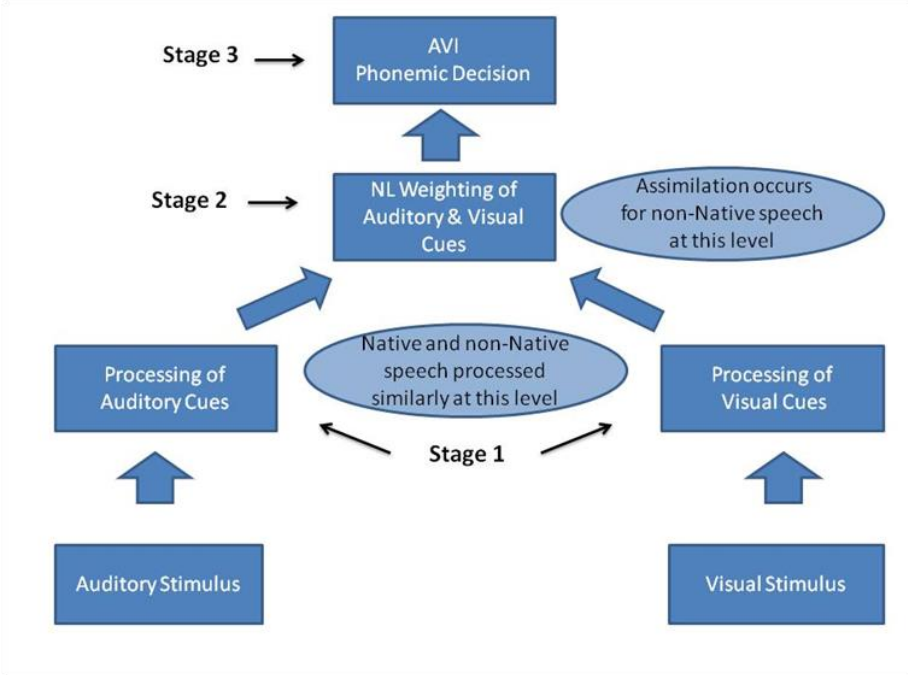

Figure 2.4 A working framework for auditory-visual integration (AVI) of speech for the native language (NL).

During the first stage which is a reception stage of peripheral processing both native and non-native speech are analysed similarly. In the second stage, the visual and auditory cues are compared to the native language mental representations and integration occurs. If the speech signal is non-native then assimilation to native language mental representations occurs at this stage based on Perceptual Assimilation Model (Best, 1994). In the third stage, the overall evidence is used to classify the speech sound based on the native language mental representations.

Figure 2.5 is a depiction of the auditory-visual phonetic perceptual space during speech perception. During stage two the auditory input and the visual input which have undergone peripheral processing in stage one will now be weighed. Visual mental representations which have fewer visual phonetic neighbours within the native language perceptual space will have a greater influence during the integration process. In other words, the less dense the visual neighbourhood is the less number of invisible phonetic contrasts there are, this is based on the Neighbourhood Activation Model (Luce and Pisoni, 1998). Figure 2.5 depicts the auditory-visual perceptual space when the response is weighted towards the visual stimulus. In this case the visual mental representation has a sparse visual neighbourhood and therefore its weight on auditory-visual speech perception is great. Thus the sound perceived in the auditory-visual perceptual space (encircled) is greatly influenced by the visual stimulus input. For example, if the visual stimulus was visually salient such as a labiodental phoneme like /**f**/, it would be very distinct as it is very visually prominent. However, if the visual stimulus was ambiguous such as a velar sound like /**k**/ this would carry less weight during the integration process as it is not visually clear. In other words the more unique (Cohort speech model) the visual cue is, the more influence or weight it will have during the integration stage. Stage three is the final stage where the speech sound is identified.

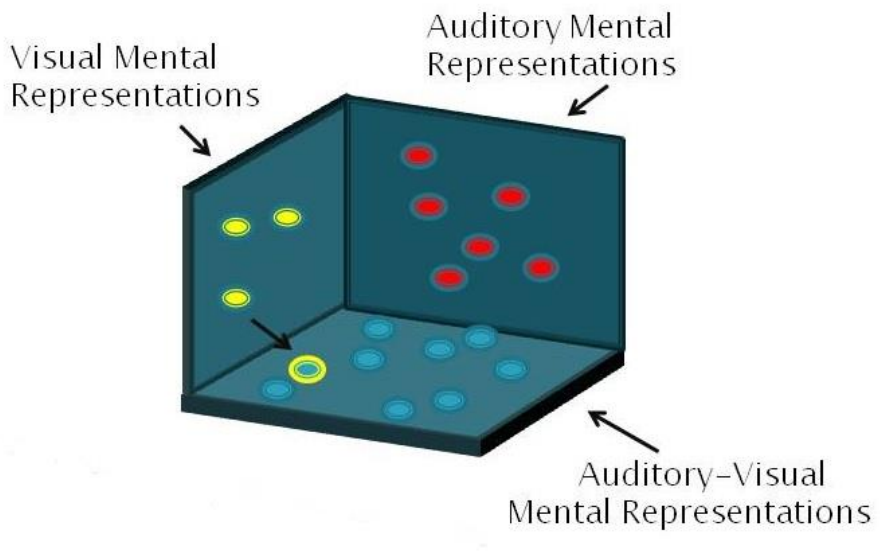

Figure 2.5 Auditory-visual phonetic perceptual space during a visual response.

The hypothesis of this thesis is that the weight that is given to the visual cues during integration is not only dependent on the basic visual parameters of speech; it is also influenced by the native language. In other words a language with more ambiguous visual cues will rely relatively more on the auditory modality than on the visual modality. Therefore auditory-visual integration during speech perception would be modulated by the visual and auditory mental representations of the native language and by the predictive power of visual speech cues within the native language. The auditory-visual integration framework hypothesised in this thesis is flexible enough to account for the different characteristics of Arabic where speech perception might rely more on auditory cues than visual cues, due to the ambiguity of visual cues in Arabic compared to English.

2.8 Purpose of Current Study

The influence of visual information on speech perception processing has been clearly established in the review of the theories within this chapter (Best, 1994, Braida, 1991, Luce and Pisoni, 1998, Marslen-Wilson, 1987, Massaro, 1987, McClelland and Elman, 1986) Thus, it is clear that in order to have a better understanding of speech perception, visual information should be considered a viable information source (Fava et al., 2014, Kushnerenko et al., 2008, Munhall et al., 2009, Shaw et al., 2008, Sommers et al., 2005, Rosenblum, 2007). This thesis aims to investigate the framework underlying the perceptual advantage provided by visual speech information in Arabic. Specific research questions were formulated in the light of the literature reviewed above.

2.8.1 Research Questions

1. Do Arabic and English listeners have a difference in their auditory-visual speech perception process measured by the McGurk effect?
2. Are these differences in auditory-visual integration measured by the McGurk effect due to early integration processing or late integration processing?
3. What are the viseme groups in Arabic and what is the degree of visual ambiguity within each group?
4. Do large viseme groups in Arabic influence the auditory-visual speech perception process less than small viseme groups?

In the next chapter a cross-linguistic experiment is conducted to compare auditory-visual integration as measured by the McGurk effect. This first experiment will investigate how native and non-native visual speech cues are perceived during auditory-visual speech perception. This experiment will explore the first two research questions mentioned above.

Chapter 3

The McGurk Effect in Arabic versus English Native Listeners

3.1 Introduction

In chapter 2 it was suggested that native language influences auditory-visual integration during speech perception. This chapter investigates whether native language influences the use of visual cues in native Arabic listeners during speech perception differently compared to native English listeners. As described in chapter 2 section 2.6, there is evidence which suggests that auditory-visual integration in speech perception might rely on boundaries based on native language visual and auditory cues (Massaro, 1998, van Wassenhove et al., 2007).

In a cross-linguistic study Lisker and Abramson, (1970) showed that when the voice onset time of initial stop consonants was manipulated, the initial stop consonants were perceived in a manner that followed the stop sounds in their native language, such that the voice onset time for the stop categorical boundaries reflected the listener's native language values (Lisker and Abramson, 1967). Furthermore, studies investigating the category boundaries for place of articulation of speech sounds indicate that listeners' perception of a non-native vowel is influenced by the vowel's boundaries within the native language of the participant (Rochet, 1995). Their results showed that native English listeners assimilated the French /**y**/ to their /**u**/ vowel category while native Portuguese participants assimilated the French /**y**/ to their /**i**/ vowel category. Additionally, Best and Strange (1992) performed a study using

English and Japanese native listeners measuring their perception on /**w**/-/**j**/ continuum. Japanese native listeners shifted the boundary towards /**j**/ compared with English native listeners (Best and Strange, 1992). These studies suggest that the boundaries that define speech sounds are dependent on the characteristics of the inventories within the native language.

Furthermore, some researchers have suggested that auditory-visual integration of speech might also be dependent on the visual boundaries of the native language (see chapter 1 section 1.5.2 and 1.5.3). Sekiyami (1997) found that Japanese and Chinese listeners' reliance on visual cues during auditory-visual integration was less than English listeners. As discussed in chapter 2 section 2.7, the dependency of auditory-visual speech perception on visual speech cues compared to auditory speech cues might be due to the linguistic features of the native language. Sekiyami concluded that for some languages, the perception of speech might rely more on visual cues than for others. However, do to eye gaze being disrespectful in the Japanese and Chinese culture Sekiyami (1997) was unable to determine whether these differences in auditory-visual integration during speech perception were due to the linguistics of the native language or due to the factor of eye gaze within the cultural (Rosenblum, 2007). The balance of evidence discussed in chapter 1 section 1.6, suggests an involvement of native visual speech cues features during auditory-visual speech perception.

The place of articulation for phonemes greatly affects how visually distinctive they are in the speech perception process (Jiang and Bernstein, 2011). The phonemes produced in the anterior portion of the vocal tract are more visually prominent than those produced posteriorly. As discussed in chapter 1 section 1.7, Arabic has fewer

visually prominent phonemes compared to English which might lead to a reduction in use of visual cues by Arabic listeners. A question that emerges is whether there is a difference between Arabic and English native listeners in the use of visual cues during speech perception? In this experiment the auditory-visual integration of speech for Arabic listeners was compared to that of English listeners. If auditory-visual integration in speech perception is reliant on native language visual speech cues then (since Arabic has more visually ambiguous phonemes than English), Arabic listeners should rely less on visual speech cues compared to English listeners during auditory-visual speech perception. Furthermore, if visual speech cues features are language specific; listeners would not be able to efficiently use non-native visual speech information as they are not familiar with the visual cues features specific to the non-native language.

In this chapter it is proposed to investigate this premise by using the McGurk effect to measure the influence of native language on auditory-visual integration during speech perception. The McGurk effect is an auditory illusion created by dubbing an auditory phoneme onto a visual phoneme, or viseme (McGurk and MacDonald, 1976). As discussed in chapter 1 section 1.4.1 the McGurk effect is thought to only occur through successful integration of the auditory and visual modalities, hence the McGurk effect may be used to examine factors that influence the integration of auditory-visual speech.

The percentage and response types of the McGurk effect were measured in monolingual native Arabic and English listeners. It is hypothesised that the perception of auditory-visual integration of speech is based on native language visual

speech cues. (Refer to chapter 2 section 2.7 for a discussion about an auditory-visual speech perception framework).

3.2 Rationale

Experiment 1 was an identification task using the McGurk effect. The percentage of the McGurk response was measured to estimate the strength of auditory-visual speech integration of the participant (McGurk and MacDonald, 1976). This experiment employed a cross linguistic design which included monolingual native Arabic and English speaking adults. The consonant vowel (CV) stimulus set used were Arabic syllables (/**ba**/, /**qa**/) and English syllables (/**pa**/, /**ka**/). Both Arabic and English stimuli sets were chosen following Summerfield's rules of categories that are most likely to induce a McGurk effect (Summerfield, 1987). Summerfield (1987) found that the stimulus set that is mostly likely to lead to a McGurk response is a bilabial auditory sound and a velar visual sound.

A distinct feature of Arabic is the presence of many guttural phonemes (Elgendy and Pols, 2001). Guttural phonemes are sounds produced in the back of the mouth such as uvular, pharyngeal and glottal sounds. Therefore visual speech cues would not be very beneficial for identifying guttural sounds due to their perceptual dense neighbourhood of visual speech mental representations. The Arabic stimulus /**qa**/ was chosen because it is a uvular sound and therefore might have a perceptually dense visual neighbourhood for Arabic listeners. This could then lead to less reliance on the visual input and thus a reduced McGurk response percentage compared to the English listeners.

1. If auditory-visual integration is dependent on the linguistic structure of the native language; Arabic native listeners will have a reduced percentage of McGurk responses compared to English native listeners because Arabic has more sounds that are visually ambiguous as they are produced in the back of the mouth compared to English. Thus Arabic listeners are expected to use visual cues less for guttural phonemes during auditory-visual integration compared to English listeners as measured by the McGurk effect.

2. There will be different McGurk responses between Arabic and English native listeners because there are different visual speech mental representations in Arabic and English. The /**q**/ phoneme in Arabic is an emphatic sound (Heselwood, 1992, Watson, 2002). Consequently, Arabic listeners will be able to pick up on emphatic visual cues and choose a fusion response that is also an emphatic phoneme. However since English listeners do not have emphatic phonemes within their native language repertoire they will not pick up on this visual cue.

This experiment will investigate auditory-visual integration of speech in Arabic and the influence of visual cues during speech perception in native versus non-native language. Furthermore, this will enable the investigation of whether visual cues to emphatic phonemes are identified and incorporated in the perception of speech. According to the auditory-visual speech perception framework proposed in this thesis auditory-visual integration of speech is based on the auditory and visual mental representations developed due to experience with the native language (see chapter 2, section 2.7).

3.3 Method

One English stimulus set that is likely to induce a McGurk response is the bilabial auditory syllable /**pa**/ coupled with the velar visual syllable /**ka**/. From previous research the expected McGurk response for an English native listener for auditory /**pa**/ and visual /**ka**/ is /**ta**/ (van Wassenhove et al., 2007). An Arabic stimulus set that is likely to induce a McGurk response is the bilabial auditory syllable /**ba**/ coupled with the uvular visual syllable /**qa**/. Since experiments on the McGurk effect using these Arabic syllables have not been previously published, the response expected for this Arabic stimulus set is unknown. To investigate non-native phonemes the Arabic consonant /**q**/ was used because it is not used by native English listeners (Giegerich, 1992) while English consonant /**p**/ was used because it is not used by native Arabic listeners (Al-Ani, 1970). Also the Arabic phoneme /**q**/ is an emphatic sound; an emphatic sound is produced with a secondary articulation involving retraction of the root of the tongue towards the pharyngeal wall (see chapter 1, section 1.7).

3.3.1 Participants

For a power of 80% and a significance level of 5 % with a medium effect size 0.25 (Cohen, 1988) a sample size of 17 was estimated for each group. The participants ages were between 20 to 50 years, for the Arabic group the mean age was 34 years (7 male, 10 female) and for the English group the mean age was 37 years (8 male, 9 female). For these experiments monolingual participants were defined as those who do not speak, read, or write a second language. Arabic speaking participants were

recruited from Riyadh, Saudi Arabia and all English speaking participants were recruited from Leeds, United Kingdom. Participants were staff and students from King Saud University and the University of Leeds.

The participants all had normal hearing at octave frequencies and self-reported normal or corrected vision and wore correction if needed. All participants gave their written informed consent to take part in the study, and the study was approved by the School of Healthcare Research Ethics Committee, University of Leeds and by the local committee at the Department of Rehabilitation Sciences, in the Applied Medical Sciences College, King Saud University.

3.3.2 Stimuli

3.3.2.1 Stimulus Generation

Stimuli were recorded from four individuals, to control for speaker effect. Furthermore to obtain the same dialect as the listeners; materials were recorded from two native Arabic Saudi adults living in Riyadh, and two native English adults living in Leeds (one woman and one man in each group). The speakers were videotaped in a well-lit, sound proof room with a plain background. The speakers were recorded looking directly into a camera and their face filled the frame. To make the video recordings a Canon Legria-HFS200 digital video recorder was used and a directional external broadcast quality microphone (Sennheiser- K6) connected to a mixer (Phonic- MM1002a) which was connected directly to the computer. The mixer had lights which indicated the intensity level of the stimulus being recorded. All

recordings were made at an average conversational level, indicated by green lights on the mixer. The Sound Pressure Level (SPL) in A-weighting (dBA) for each stimulus was measured through the headphones using a circumaural plate coupled to an artificial ear (Bruel and Kjaer- 4153) connected to a sound level meter (Bruel and Kjaer- 2250). The mean SPL for the test stimuli was 70.44 dBA (SD=1.52 dB), a one way analysis of variance (ANOVA) indicated that there was no significant difference between the SPL values of all the stimuli spoken by the 4 speakers [$F(3,12) = 0.31$, $p= 0.817$]. A sound calibrator (Bruel and Kjaer-4231) which conforms to EN/IEC 60942 Class LS and Class 1, and ANSI S1.40-1984 was used to calibrate the measurement system.

3.3.2.2 Auditory-Visual Stimulus Alignment

Two sets of auditory-visual stimuli were generated that comprised two experimental conditions (Congruent auditory-visual and incongruent auditory-visual). Congruent auditory-visual stimuli are when the visual and auditory stimuli match. Incongruent auditory-visual stimuli are when the visual and auditory stimuli do not match (see Table 3.1). Adobe Premiere Elements 9 Software (Adobe, 2010) was used, to create all of the stimuli. All the stimuli were edited to begin and end with a neutral facial expression and each stimulus lasted about 5 seconds.

Table 3.1 Experimental Speech Stimulus Native and Non-native conditions

Condition	Speech Material	Listener	Congruent Stimuli		Incongruent Stimuli		Expected McGurk Response
			Heard	***Seen***	***Heard***	***Seen***	
Native	***Arabic***	***Arabic***	/ba/	/ba/	/ba/	/qa/	unknown
			/qa/	/qa/	/qa/	/ba/	
	English	***English***	/pa/	/pa/	/pa/	/ka/	/ta/
			/ka/	/ka/	/ka/	/pa/	
Non-Native	***Arabic***	***English***	/ba/	/ba/	/ba/	/qa/	unknown
			/qa/	/qa/	/qa/	/ba/	
	English	***Arabic***	/pa/	/pa/	/pa/	/ka/	/ta/
			/ka/	/ka/	/ka/	/pa/	

To generate the incongruent stimuli a consonant onset alignment method was adopted, which is the most commonly used method (Grant et al., 2004, Jiang and Bernstein, 2011, Munhall et al., 1996). The incongruent auditory stimulus was aligned with the original auditory signal at the start of the production of the consonant using Adobe Premiere Elements 9 Software (Adobe, 2010). After coarse alignment the original auditory signal was erased and replaced by the incongruent auditory signal. Further fine alignment was performed by visually viewing the video clip frame by frame and aligning the auditory signal to the visual signal by visual inspection of the acoustic waveform by the experimenter.

Each stimulus block consisted of 2 congruent stimuli and 2 incongruent stimuli spoken by 2 native speakers for Arabic and English. Therefore, each stimulus block contained 16 trials consisting of 8 Arabic stimuli and 8 English stimuli ([4 stimuli * 2 native speakers] * 2 languages= 16 trials).There were 20 blocks of stimuli; therefore each participant was required to respond to 320 trials in total (16 CV syllables x 20 blocks= 320 trials).

All stimuli were saved in MPEG file format with; 720 x 480 pixels, frame rate of 29.97 frames/s, and audio sampling rate of 48 kHz. The stimuli were displayed in random order using SuperLab presentation software (Version 4.5, Cedrus Corporation, 2009). For baseline measurements auditory only trials were conducted after testing in the auditory-visual condition was completed. The auditory only trials were conducted last so as not to influence the participants' auditory-visual responses.

3.3.3 Procedure

Each participant took part in one session which lasted about one hour. Participants were given a 5 minute break after every 4 blocks (about every 10 minutes). Participants were tested individually in a sound proof audiology test room. For Arabic listeners the research was undertaken within the audiology suite situated within the School of Rehabilitation Sciences, King Saud University, Riyadh Saudi Arabia. For English listeners the research was undertaken within the audiology suite situated within the School of Healthcare, University of Leeds, United Kingdom. All participants gave their written informed consent to take part in the study. Participants were seated about 70 cm from a 15 inch laptop screen and listened to the speech stimuli through Circumaural headphones (Sennheiser HD438). The volume control of the laptop was set at a level that produced 70dB SPL that is normal conversational level. Each trial consisted of a short video clip (5 sec) of a person producing the speech stimuli described in section 3.3.2.2.

On each trial participants were asked to watch the face of the talker on the laptop screen whilst listening to the output from the headphones. The response required was

to report the CV heard. Following verbal instructions, participants were given a short practice session of 5 trials to familiarize themselves with the protocol. SuperLab presentation software (Version 4.5, Cedrus Corporation, 2009) was used to present the stimuli in a random order and record the participants' free-form response. Both participant and experimenter were blind to stimulus presentation order. The experimenter ensured throughout the session that the participant was looking directly at the screen. After each stimulus a response box was displayed on the laptop monitor and the participant typed in his/her free-form response using the laptop keyboard, so if they heard /**ba**/ they would type "ba" using the keyboard in the response screen. After the participant pressed the "Enter" key a new trial was presented, the testing was self-paced.

3.4 Results

For baseline measurements auditory only trials were conducted in free-form response. Confusion matrices for the auditory only condition are shown in Table 3.2 and 3.3. The number in each cell is the percentage of responses for each of the four auditory signals /**b**, **q**, **p**, **k**/. There were 170 observations for each stimulus, 10 repetitions x 17 participants in each native language group. The two phonemes with the lowest correct identification percentage were the phonemes not present in the listeners native language. This finding was expected as /**p**/ is not present in Arabic and /**q**/ is not present in English and is the most likely reason for incorrect identification by non-native listeners.

Table 3.2 Confusion matrices in the auditory-only condition for Arabic Native Listeners (a) and English Native Listeners (b) using Arabic stimulus. Each number indicates the percentage of responses.

(a)

Stimulus	Response	
	/b/	/q/
/b/	100	
/q/		100

(b)

Stimulus	Response			
	/b/	/q/	/k/	/g/
/b/	100			
/q/			15	85

Table 3.3 Confusion matrices in the auditory-only condition for Arabic Native Listeners (a) and English Native Listeners (b) for English stimulus. Each number indicates the percentage of responses.

(a)

Stimulus	Response	
	/k/	/b/
/p/		100
/k/	100	

(b)

Stimulus	Response	
	/p/	/k/
/p/	100	
/k/		100

Figures 3.1-3.4 show the responses by Arabic and English listeners for Arabic and English stimuli as proportions of four response categories; auditory (e.g., the response to A/**pa**/ + V/**ka**/ was /**pa**/), visual (e.g., the response to A/**pa**/ + V/**ka**/ was /**ka**/), fusion (e.g., the response to A/**pa**/ + V/**ka**/ was /**ta**/), and combination (e.g., the response to A/**pa**/ + V/**ka**/ was /**pka**/). Although a combination response is evidence of auditory-visual integration, it is not considered a McGurk response, as a McGurk response is a fusion of the visual and auditory stimuli producing a new response (McGurk and MacDonald, 1976) and not a combination of the two. It can be seen from Figures 3.1 and 3.3 that the stimuli with auditory bilabials /**b**/ and /**p**/ and a visual uvular or velar /**q**/ and /**k**/ were the only stimuli pairs which produced McGurk responses for Arabic and English listeners. Figures 3.2 and 3.4 illustrate that the stimuli with an auditory uvular or velar /**q**/ and /**k**/ and visual bilabials /**b**/ and /**p**/ were the only stimulus pairs which produced combination responses for Arabic and English listeners.

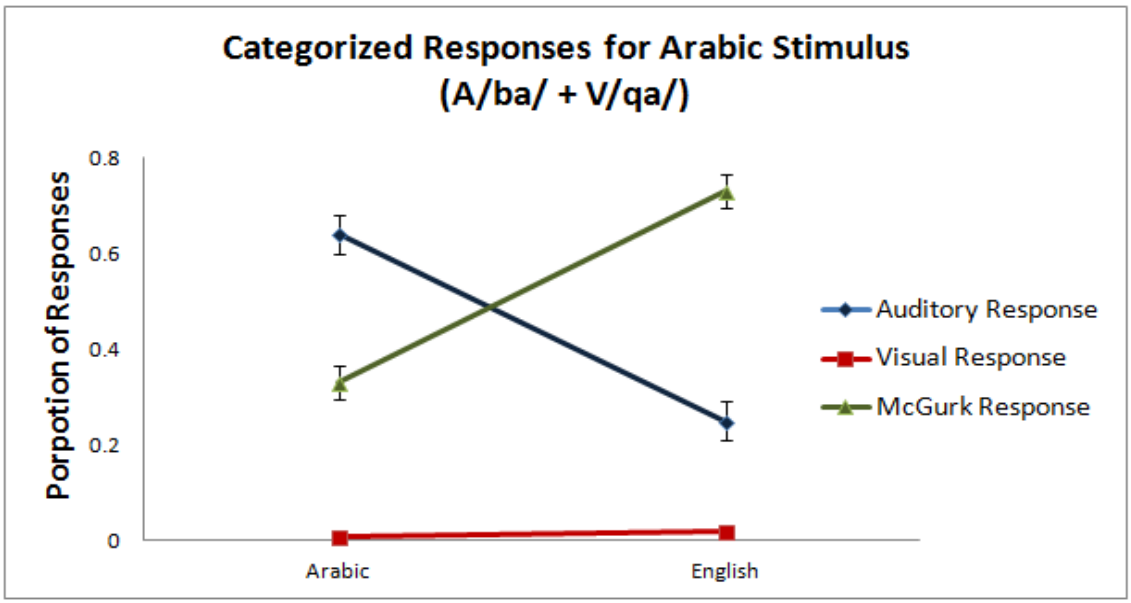

Figure 3.1 Categorized responses (auditory, visual, and McGurk response) for Arabic stimulus (A/ba/ +V/qa/) shown as proportions for Arabic native listeners and English native listeners.

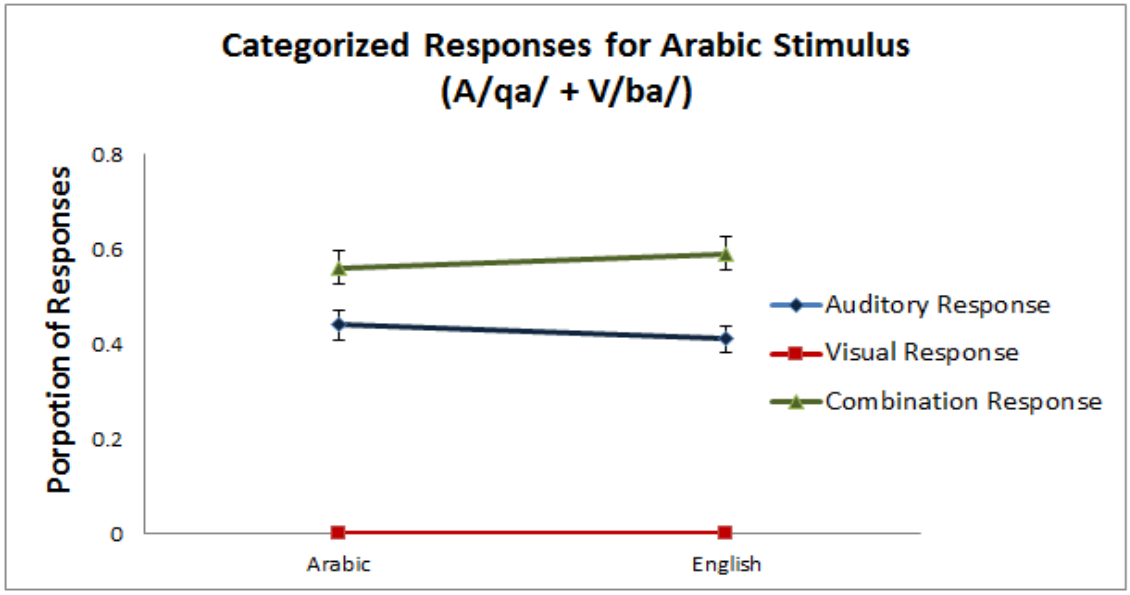

Figure 3.2 Categorized responses (auditory, visual, and combination response) for Arabic stimulus (A/qa/ + V/ba/) shown as proportions for Arabic native listeners and English native listeners.

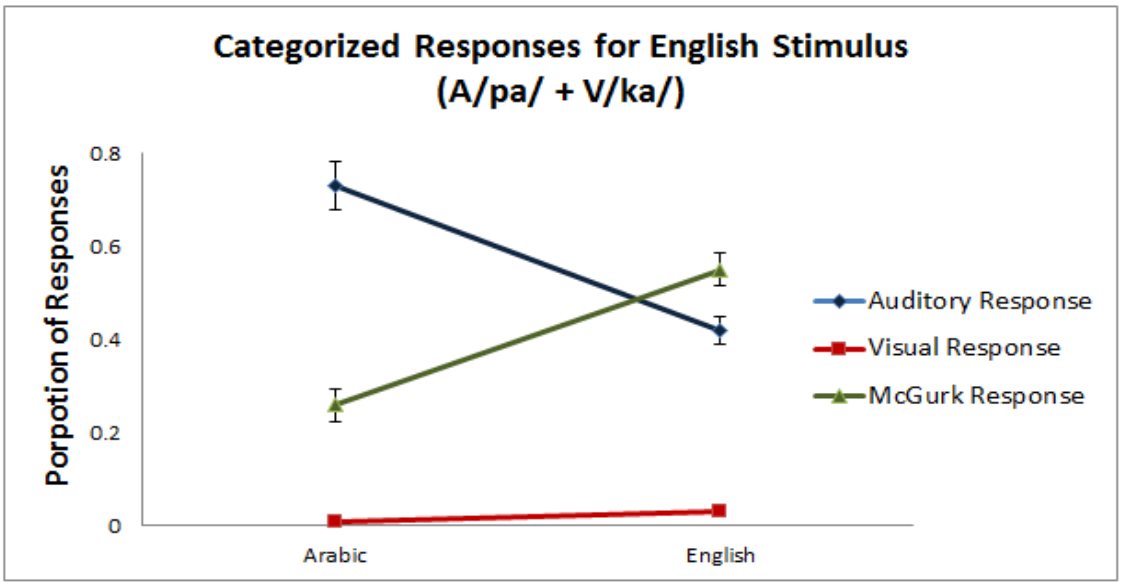

Figure 3.3 Categorized responses (auditory, visual, and McGurk Response) for English stimulus (A/pa/ + V/ka/) shown as proportions for Arabic native listeners and English native listeners.

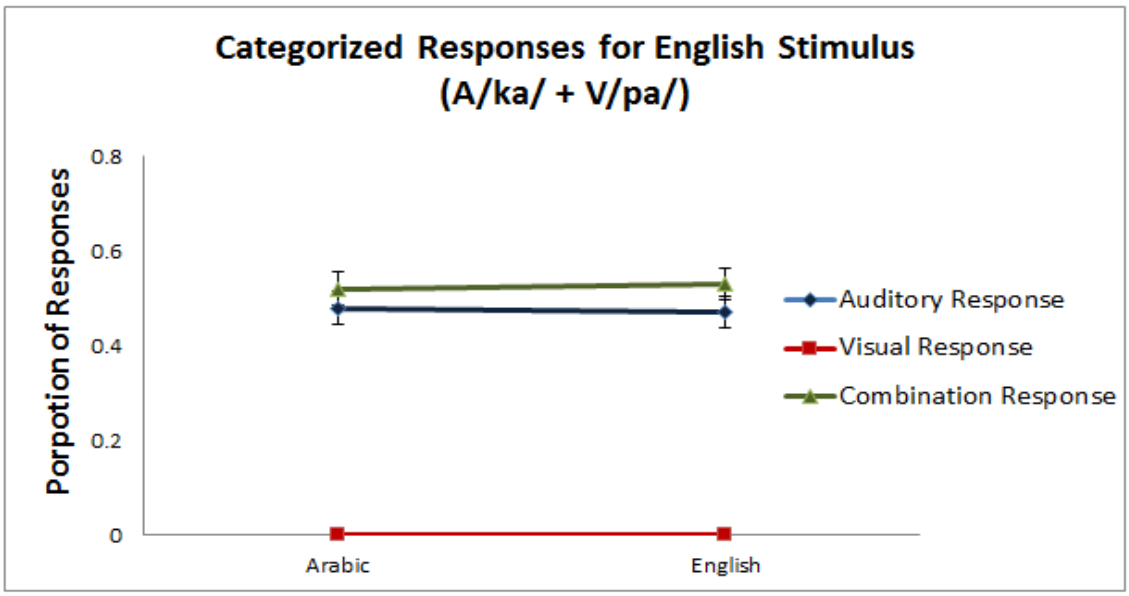

Figure 3.4 Categorized responses (auditory, visual, and combination response) for English stimulus (A/ka/ + V/pa/) shown as proportions for Arabic native listeners and English native listeners.

The effect of language of listener and stimulus type on the percentage of McGurk response (see Figure 3.1 and 3.3) was investigated using a two-way ANOVA for repeated measures. Levene's test verified the equality of variances in the samples (homogeneity of variance) (***p*** > .05) (Martin and Bridgmon, 2012). The within group effect of stimulus type (Arabic or English) was significant [***F*** (1, 32) = 29.82; ***p*** < .0001)], that means that an Arabic stimulus (A/**ba**/ + V/**qa**/) produced a significant larger percentage of McGurk effect compared to an English stimulus (A/**pa**/ + V/**ka**/). The between group effect of native language was also significant [***F*** (1, 32) = 167.29; ***p*** < .0001)] this was due to the significantly larger McGurk response percentage by English listeners compared to Arabic listeners. The interaction between stimulus and native language of listener was also significant [***F*** (1, 32) = 4.95; ***p*** < .03)]. This interaction occurred because for the Arabic listeners both stimulus sets sound the same. In Arabic there is no /**p**/ phoneme and so the phoneme is assimilated to /**b**/ for the Arabic listeners. As can be seen in the auditory only condition Arabic listeners perceived the /**pa**/ auditory stimulus as /**ba**/ (see Table 3.3). While for the English listeners the auditory stimulus /**pa**/ and /**ba**/ are perceived as different since they are both distinct phonemes within the English language. Therefore, for the English listeners there was a difference in the effect of stimulus type, while for the Arabic listeners there was no difference caused by the stimulus type.

The combination response percentage (see Figure 3.2 and 3.4) was also evaluated using a repeated measures two-way ANOVA. A Levene's test verified the equality of variances in the samples (homogeneity of variance) (***p*** >.05) (Martin and Bridgmon, 2012). The within group effect of stimulus type (Arabic or English) was

not significant [F (1, 32) = 1.29; p = .20)], that means that there was no statistical difference in the percentage of combination responses between an Arabic stimulus (A/**qa**/+V/**ba**/) and an English stimulus (A/**ka**/+V/**pa**/). The between group effect of native language was also not significant [F (1, 32) = .96; p = .30)] this was due to similar combination response rates by English and Arabic listeners. The interaction between stimulus and native language of listener was also not significant [F (1, 32) = .11; p = .70)].

The open–set McGurk responses for the incongruent stimulus sets Arabic (A/**ba**/+V/**qa**/) and English (A/**pa**/+V/**ka**/) were tallied. Figures 3.5 and 3.6 show the consonant identification responses by Arabic and English native listeners to the Arabic and English stimuli. These figures show that the phonetic responses varied across the different types of incongruent stimuli. For example, in Figure 3.5 the Arabic stimulus (A/**ba**/+V/**qa**/) resulted in /**d**/ and /**tˀ**/ responses while in Figure 3.6 the English stimulus (A/**pa**/+V/**ka**/) resulted in only /**t**/ responses for both groups of listeners. Furthermore, the fusion response that occurred most frequently for the Arabic stimulus (A/**ba**/+V/**qa**/) was /**tˀ**/ for Arabic listeners, while the English listeners' response was only /**d**/. The visual phoneme in the Arabic stimulus was an emphatic phoneme /**q**/ and Arabic listeners' McGurk response was also an emphatic phoneme /**tˀ**/.

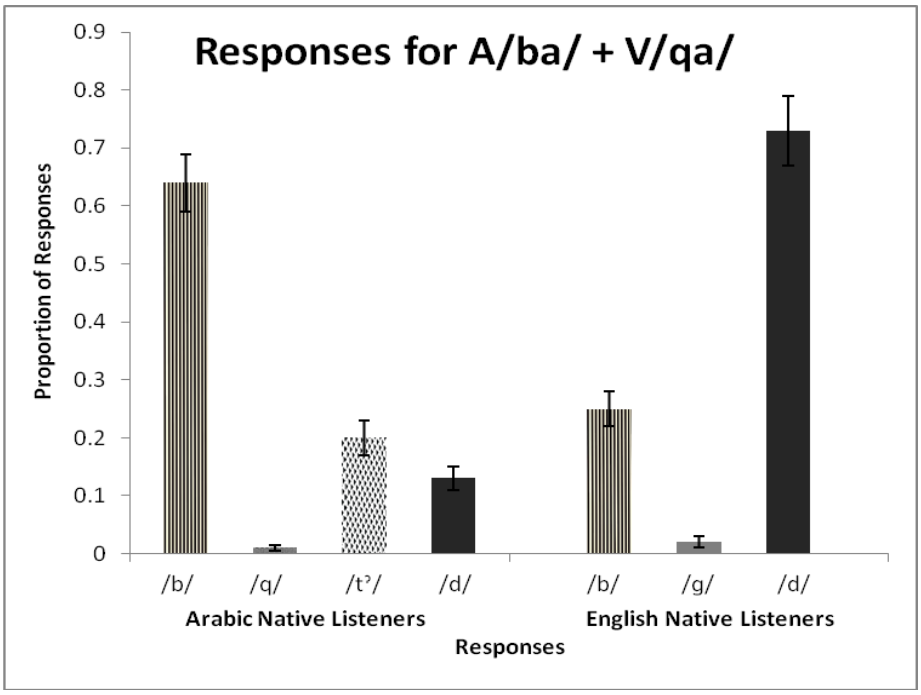

Figure 3.5 Shows the response proportions for consonant identification for Arabic stimuli (A/ba/+V/qa/) by Arabic and English native listeners.

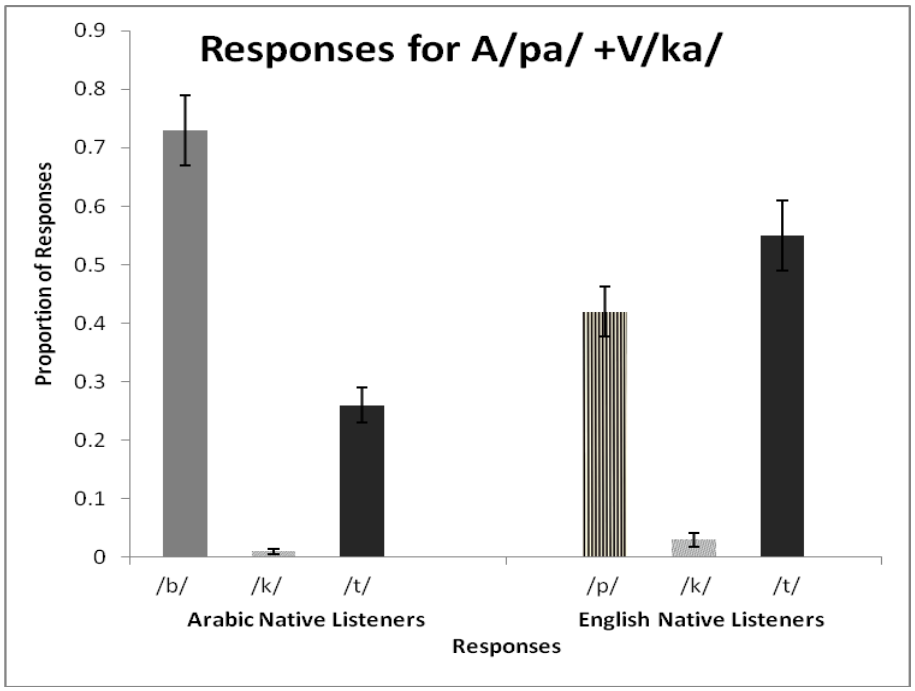

Figure 3.6 Shows the response proportions for consonant identification for English stimuli (A/pa/+V/ka/) by Arabic and English native listeners.

3.5 Discussion

3.5.1 McGurk Response and Combination Response

In the present study the McGurk effect only occurred for the Arabic (A/**ba**/+V/**qa**/) and the English (A/**pa**/ +V/**ka**/) stimuli sets, which had a uvular and velar visual component respectively. The visual cues during the McGurk effect change what is perceived to something different from both the auditory and visual components. For example in the English stimulus the visual /**ka**/ changed the perception of the auditory /**pa**/ into the McGurk response /**ta**/. Van Wassenhove et al. (2005) demonstrated that the perceptual outcome of auditory-visual integration may depend on the ease of perceptual categorization of the visual stimulus. Since velar and uvular phonemes have ambiguous visual cues as they are produced in the back of the mouth they may have a weak visual weight in the auditory-visual integration process explaining why these visual cues lead to a McGurk response.

On the other hand, when the visual cues were prominent in the case of bilabials there was no McGurk fusion, but there was a combination response. The only stimulus sets that produced a combination response were the Arabic (A/**qa**/ + V/ **ba**/) and the English (A/**ka**/+V/**pa**/) which both had a bilabial visual component. McGurk and MacDonald (1976) demonstrated that bilabial phonemes have very prominent visual cues as they are produced at the lips. Consequently bilabial phonemes may have a strong visual weight in auditory-visual integration process. During the combination response an individual perceives to hear both the visual component as well as the auditory component. For example in the English stimulus the visual /**pa**/ changed the perception of the auditory /**ka**/ into the combination response /**pka**/.

The percentage of use of visual cues by Arabic and English listeners for the combination response was similar; this might be due to the place of articulation of the visual cue. This suggests that when the visual phoneme is a bilabial sound Arabic listeners are facilitated by visual information when perceiving speech stimuli at a similar percentage to English listeners. For English listeners visual cues from bilabial, velar and uvular phonemes have a strong influence on auditory-visual integration of speech as measured by the McGurk effect. However, for Arabic listeners visual cues for bilabial consonants carry more influence during auditory-visual speech perception when compared to visual cues for velar and uvular phonemes.

In the present study, when the visual cue was a clearly identifiable bilabial phoneme it increased the Arabic listeners' auditory-visual integration percentage measured by the combination response. Yet, when the visual stimulus was a velar or uvular phoneme the Arabic listeners' auditory-visual integration ability was reduced measured by the lower percentage of the McGurk response. This lends support to the hypothesis that auditory-visual integration relies on the predictive power of visual speech cues within the native language. As there are a greater number of phonemes produced in the back of the mouth in Arabic compared to English this might lead to a difference in visual ambiguity between the two languages (see chapter 2 section 2.7).

3.5.2 Assimilation

The concept of assimilation might also explain why Arabic listeners had a similar McGurk response percentage for the Arabic (A/**ba**/+V/**qa**/) and the English

(A/**pa**/+V/**ka**/). If auditory-visual integration of speech depends on the auditory and visual native language mental representations; since Arabic listeners do not have an auditory /**p**/ in their native language then the auditory-visual native language framework would predict that they frequently misidentify the phoneme /**p**/ as /**b**/ (see Table 3.2). According to Best's (1994) Perceptual Assimilation Model, when people are exposed to non-native language, they will categorize the new non-native speech sounds to the closest existing phonological category based on their mental representations of the respective native language (Best et al., 2001, Harnsberger, 2001, Nagao et al., 2003). Consequently, Arabic listeners' percentage of McGurk response for both stimulus sets would be similar, since Arabic listeners do not perceive a distinct difference between auditory /**b**/ and /**p**/.

On the other hand English listeners do have an auditory /**p**/ in their auditory repertoire and clearly distinguish between an auditory /**p**/ and /**b**/ (see Table 3.3). For English listeners the stimulus set with an auditory /**b**/ causes greater McGurk response percentage than an auditory /**p**/, because the auditory confusion between /**b**/ (auditory stimulus) and /**d**/ (the McGurk response) is higher than the auditory confusion between /**p**/ (auditory stimulus) and /**t**/ (the McGurk response) as noted by Summerfield, (1987). However for Arabic listeners the auditory /**ba**/ and /**pa**/ stimuli sound similar and therefore their auditory-visual integration effect is similar.

Similarly, English listeners assimilated Arabic visual /**q**/ to the /**g**/ which is the closest visual cue within their language. The results are consistent with assimilation, but in this case assimilation occurred for the visual speech stimulus instead of the traditional auditory speech assimilation. That is when English listeners were exposed to a non-native visual component of the auditory-visual stimulus, in this case

emphatic uvular /**q**/, they categorized the non-native visual component of the auditory-visual stimulus to the closest existing visual speech category based on their visual mental representations of the native language in this case /**g**/ and auditory-visual integration still occurs.

Assimilation for both auditory and visual components of auditory-visual speech has been accounted for within the hypothesized auditory-visual native language framework described in chapter 2 section 2.7. These results show that assimilation can occur for non-native visual component of the auditory-visual stimulus in the same way as assimilation occurs for non-native auditory component of the auditory-visual stimulus. This implies that auditory-visual integration of speech is a resilient process which will still occur even if the auditory or visual components of the auditory-visual stimulus are non-native.

3.5.3 Emphatic Visual Cues

The fusion response that occurred most frequently for the Arabic stimulus (A/**ba**/+V/**qa**/) was /**tˀ**/ for Arabic listeners, while English listeners' response was /**d**/. The visual phoneme in the Arabic stimulus (A/**ba**/+V/**qa**/) was the emphatic phoneme /**q**/ and Arabic listeners' McGurk response was also an emphatic phoneme /**tˀ**/ (see chapter 1 section 1.7). This implies that Arabic listeners picked up on the visual cues for an emphatic phoneme, a finding also reported by Ouni and Ouni (2008). Furthermore the results of this experiment have suggested that when the visual component of the auditory-visual stimulus is an emphatic phoneme this affects the perception of the auditory-visual speech. Thus, Arabic listeners incorporated the

emphatic category in their choice of the McGurk response. The auditory stimulus /**b** / is a voiced phoneme however the McGurk response was /**tˤ**/ which is not voiced. The /**dˤ**/ is a voiced plosive emphatic, but it was not perceived by the Saudi listeners and this could be due to dialect. In the Saudi dialect the emphatic phoneme /**dˤ** / is not produced and is substituted with the emphatic fricative /**ðˤ**/ (Alhammad, 2014, Al-Raba'a, 2015). Therefore, the mental representation for the phoneme /**tˤ**/ was the best fit for the Arabic stimulus (A/**ba**/+V/**qa**/) for Saudi listeners.

Since there are no emphatic phonemes in English, the English listeners were not able to recognize the visual cues of the auditory-visual stimulus for an emphatic phoneme. Similarly, Hazan et al. (2006) found that Japanese native listeners' ability to contrast non-native /**r**/-/**l**/ did not improve in the auditory-visual condition compared to the auditory only condition, despite the visible differences between these sounds. Likewise, Han-Gyol Yi et al. (2013) found that English native listeners benefitted more during auditory-visual speech perception when the visual cues were native (Yi et al., 2013). These results imply that listeners reduce their reliance on non-native visual cues and that they are unable to correctly use the visual cues of the auditory-visual stimulus that are not present within their native language (Hazan et al., 2006). These results can be explained by the hypothesized auditory-visual native language framework which states that auditory-visual speech perception relies on the saliency of visual speech cues within the native language repertoire.

3.6 Conclusion

Although Arabic native listeners may be assisted by visual information when perceiving speech stimuli, they do not seem to rely on them as much as English listeners. In the literature there is a division between whether auditory-visual integration is an early or late integration process (see chapter 2 section 2.2 and 2.4). One possible explanation is that the difference in McGurk effect between Arabic and English listeners seen in this chapter is due to early integration differences. It could have been hypothesized that a framework of auditory-visual integration can be explained by early integration of speech through a process of predicting, that is visual speech primes the auditory system to what is about to happen (Buchwald et al., 2009, Kim et al., 2004, van Wassenhove et al., 2005). Could auditory-visual speech perception be an early integration process that relies on pre-phonological representations? That would mean that visual speech cues only represent low-level spatiotemporal correlations of facial movements.

In the following chapter a second experiment was conducted to investigate whether the difference of the McGurk response percentage between Arabic and English listeners could be attributed to a difference in early integration due to bottom-up visual processing speed of visual speech cues. Experiment two will help to determine whether Arabic listeners would have a slower visual processing speed compared to English listeners. That is to say that since Arabic has many phonemes produced in the back of the vocal tract that could lead to a slower visual processing speed time for Arabic listeners compared to English listeners.

Chapter 4

Temporal Constraints on the McGurk Effect in Arabic versus English Listeners

4.1 Introduction

The data reported in chapter 3 showed a significant difference in auditory-visual integration percentage, as measured by the McGurk effect, between Arabic and English native listeners. This was interpreted as meaning that auditory-visual integration of speech is assisted by the visual mental representations of the native language. This suggested that the native English listeners put more weight on visual cues than native Arabic listeners during the perception of auditory-visual speech. However, it could be argued as pointed out in the conclusion in chapter 3 this difference could have been due to early integration due to bottom-up differences in visual processing speeds. In view of this the study reported in this chapter aimed to evaluate whether the difference in McGurk percentage between the Arabic and English native listeners was caused by a difference in early integration due to visual processing speed.

In the human natural environment, the propagation speeds affect the relative timing correlation between a visual and auditory signal. In terms of a human body, the relative timing is also influenced by times of sensory transduction and neural conduction. Another relevant consideration to take into account is that conduction times of their corresponding media are different for auditory and visual signals. As a

result, in the course of an auditory-visual event the auditory component reaches observer's sensory receptors much later than the visual component (Spence and Squire, 2003). In other words, a propagation speed of a visual signal is 300 *106 m/s, which suggest that the signal arrives almost instantly. On the other hand, a propagation speed of an auditory signal is approximately 340 m/s, which results in its delay. This difference in propagation and transduction speed causes visual cues to precede the auditory cues by 100 to 200 ms (Chandrasekaran et al., 2009). This temporal dynamic implies that visual information provides strong predictive cues for the auditory information (Besle et al., 2008, Hertrich et al., 2007, Peelle and Davis, 2012).

An experimental method used to evaluate processing speed, the relationship between auditory and visual cues, is measuring and manipulating the temporal synchrony between the two modalities. Temporal synchrony or timing is a sensory attribute critical for binding auditory-visual stimuli (Calvert, 2001). Van Wassenhove et al., (2002) investigated differences between English listeners' ability in auditory-visual integration of speech measured by the McGurk effect. They divided the participants into two groups; one group had high McGurk response percentage and the second with average McGurk response percentage. They found that by increasing the stimuli's visual lead time the McGurk response percentage for the average McGurk response group could increase to the levels of the high McGurk response group (van Wassenhove et al., 2002). They concluded the difference in auditory-visual integration percentage was due to a difference in bottom-up processing speed of visual speech cues.

In the research of Conrey and Pisoni (2006) it was outlined that, in terms of auditory-visual asynchrony, individuals' capacity for estimating items as synchronous is dependent on the visual cues of the initial phonetic segment of the word. There were certain words that were demonstrated to have a lesser degree of tolerance to audio-visual asynchrony. For instance, the word *theme* was resistant to 300 ms of visual lead while *back* was resistant to only 200 ms of visual lead. The reason given for the difference in temporal tolerance was that it depends on the place of articulation and voice onset time of the initial phonetic segment. These results illustrate the notion of relative and variable cue-weighting.

In the above example *back* begins with the phoneme /**b**/ which is a voiced bilabial phoneme therefore very prominent visually and has a short voice onset time. Whereas in the case of the word *theme* it begins with the phoneme /**θ**/ which is less visually prominent compared to /**b**/. Also /**θ**/ has a longer voice onset time compared to /**b**/ (Conrey and Pisoni, 2006). Conrey and Pisoni (2006) demonstrated that listeners' tolerance of auditory-visual asynchrony depends on the visual saliency of the word. This suggests that phonemes that are visually salient are processed faster since their visual processing speeds are shorter than less salient visual speech phonemes. The above studies suggest that both saliency and familiarity of visual cues can influence the temporal aspects of auditory-visual integration of speech.

It could be suggested that the McGurk response percentage difference found in chapter 3 between native English listeners and native Arabic listeners was due to early integration differences of visual processing speed. Since Arabic has more visual ambiguous phonemes than English this could lead to a slower visual processing speed for these phonemes. In order to determine whether the visual

processing speed is the cause of the difference in McGurk response percentage between Arabic and English listeners an experiment was conducted. This experiment used auditory-visual alignment in temporal asynchronous conditions. That is the temporal relationship between the auditory and visual stimuli was changed by +/- 300 msec in 30 msec steps. This enabled the measurement of a threshold (in msec) where the McGurk response was at the highest percentage for Arabic native listeners versus English native listeners.

4.2 Method

This experiment was a cross linguistic design which included monolingual native Arabic and English speaking adults. The experiment was an identification task measuring the percentage of the McGurk response in two test conditions (temporal synchronous auditory-visual and temporal asynchronous auditory-visual) using English and Arabic syllables. Synchronous auditory-visual stimuli are when the visual and auditory stimuli temporally match. Asynchronous auditory-visual stimuli are when the visual and auditory stimuli do not temporally match. The temporal alignment between the auditory and visual stimuli was manipulated to measure the temporal window which resulted in the highest McGurk response percentage.

4.2.1 Participants

For a power of 80% and a significance level of 5 % with a medium effect size 0.25 (Cohen, 1988) a sample size of 17 was estimated for each group. The participants

were 30 adults ages between 20 to 50 years, for the Arabic group the mean age was 29 years (5 male and 12 female) and for the English group the mean age was 33 (6 male and 11 female). For these experiments monolingual participants were defined as those who do not speak, read, or write a second language. All Arabic listeners were from Saudi Arabia and all English speaking participants were from the United Kingdom. Participants were staff and students from King Saud University and the University of Leeds.

Before the experiment started each participant was given a routine hearing screen in the form of a pure tone audiometric test at 20dB HL (frequencies tested were 500Hz, 1000Hz, 2000Hz, and 4000Hz). All participants reported normal vision and wore correction if needed. None of the participants had visual or hearing problems; therefore none were excluded from the experiment. All participants gave written informed consent to take part in the study, and the study was approved by the School of Healthcare Research Ethics Committee, University of Leeds and by the local committee at the Department of Rehabilitation Sciences, in the Applied Medical Sciences College, King Saud University.

4.2.2 Stimuli

4.2.2.1 Stimulus Generation

The consonant vowel (CV) stimuli set used for both the congruent and incongruent conditions were Arabic syllables auditory /**ba**/ and visual /**qa**/ and English syllables auditory /**pa**/ and visual /**ka**/. These stimuli sets were chosen, because in chapter 3

these stimuli sets were the only ones that produced a McGurk response. To enable the comparison of the effect of native versus non-native stimuli both stimuli that are native and non-native to each group were used. The Arabic consonant **/q/** was used because it is not used by native English listeners, while English consonant /**p**/ was used because it is not used by native Arabic listeners. Therefore the participants would not have mental representations for these non-native phonemes. The stimuli were taken from the recordings made in chapter 3 (see chapter 3 section 3.3.2.1)

4.2.2.2 Auditory-Visual Alignment

Adobe Premiere Elements 9 Software (Adobe, 2010) was used to create all of the stimuli by displacing the auditory file in 30 ms increments with respect to the video file. Negative values are used for an auditory component occurring before its visual counterpart, while positive auditory delays indicate that the auditory component trails the visual component. The physical synchrony of the auditory and visual stimulus components is at 0 ms.

The temporal asynchrony ranged from (−) 300 ms of auditory lead to (+) 300 ms of auditory lag. Auditory-visual integration of speech falls within this time frame (Munhall et al., 1996, van Wassenhove et al., 2007). Thus, a total of 21 stimulus conditions (20 asynchronous conditions and 1 synchronous condition) were used in the study for each of the four speakers. Hence there were 4 video clips with synchronous conditions and 80 video clips with asynchronous conditions between the auditory and visual stimuli. There were 84 trials for English and Arabic stimuli randomized within a block. There were 6 blocks and after each block there was a 5

minute break. All stimuli were saved in MPEG file format with; 720 x 480 pixels, frame rate of 29.97 frames/s, and audio sampling rate of 48 kHz. The stimuli were displayed in random order using SuperLab presentation software (Version 4.5, Cedrus Corporation, 2009).

4.2.3 Procedure

Each participant took part in one session which lasted about an hour and a half. Participants were given a 5 minute break after each block. Participants were tested individually in a sound proof audiology test room. For Arabic listeners the research was undertaken within the audiology suite situated within the School of Rehabilitation Sciences, King Saud University, Riyadh Saudi Arabia. For English listeners the research was undertaken within the audiology suite situated within the School of Healthcare, University of Leeds, United Kingdom.

Participants were seated about 70 cm from a 15 inch laptop screen and listened to the speech stimuli through Circumaural headphones (Sennheiser HD438) at normal conversational level of 70dB SPL. Each trial consisted of a short video clip (5 sec) of a person producing the speech stimuli. There were 4 trials of video clips with temporally synchronous sounds, and 80 trials of video clips with temporally asynchronous sounds. Each of the 84 trials was repeated in 6 blocks; each participant was required to respond to 504 clips in total (84*6=504). Randomized within each of the 6 blocks there were synchronous and asynchronous Arabic and English stimuli. Both Arabic and English native listeners heard both sets of stimuli Arabic and English.

SuperLab presentation software (Version 4.5, Cedrus Corporation, 2009) was used to present the stimuli in a random order and record the participants' free-form response. Both experimenter and participant were blind to stimulus presentation order. The experimenter ensured throughout the session that the participant was looking directly at the screen. After each stimulus a response box was displayed on the laptop monitor and the participant typed in his/her response using the laptop keyboard, so if they heard /**ba**/ they would type "ba" using the keyboard in the response screen. After the participant pressed the "Enter" key a new trial was presented, the testing was self-paced.

Following verbal instructions, participants were given a short practice session of 5 trials to familiarize themselves with the protocol. On each trial participants were asked to watch the face of the talker on the laptop screen whilst listening to the output from the headphones. The response required was to report the CV that was heard. Following each stimulus a response box was displayed on the laptop monitor, then the participant typed his/her response by using the laptop keyboard. After the participant pressed the "Enter" key a new trial was presented. The testing was self-paced that is there was no time restriction on responding.

4.3 Results

4.3.1 Response Categories to the Arabic stimulus (A/**ba**/ + V/**qa**/)

Figures 4.1 and 4.2 show the responses for Arabic stimulus of Arabic and English listeners as proportions of three response categories (auditory, visual, and McGurk

response). Negative numbers on the x-axis indicate that the auditory stimulus preceded the visual stimulus. Figure 4.1 illustrates that for Arabic listeners listening to the Arabic stimulus (A/**ba**/ + V/**qa**/) the McGurk response never reached 50% of the responses; consequently the temporal window could not be measured. The temporal window of integration is defined as the temporal range in ms where the McGurk response percentage is greater than 50% (van Wassenhove et al., 2002). However, the visual lead (auditory lag) of 60ms was the most favourable condition for the McGurk response at a percentage of 45%. As the visual lead increased greater than 60 ms the McGurk response percentage decreased gradually to the percentage of 25% at 300ms. The auditory lead led to a more pronounced decrease in the McGurk response at a percentage of 16% at -300ms. The percentage of decrease in the McGurk response was more pronounced for the auditory lead then for the visual

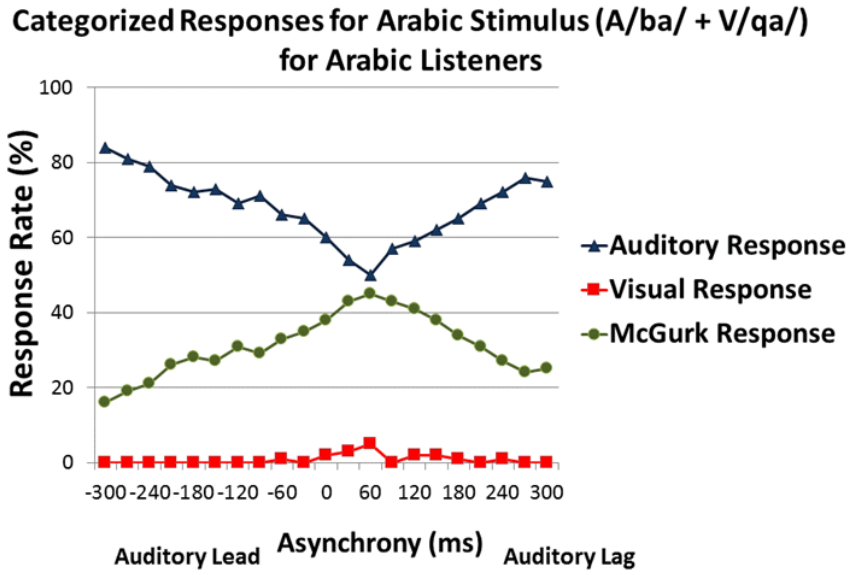

Figure 4.1 Categorized responses (auditory, visual, and McGurk Response) for Arabic stimulus (A/ba/ + V/qa/) shown as proportions for Arabic listeners.

lead. The visual response percentage was overall low, but the range of 30-60ms of visual lead was the most favourable for visual response at a percentage of 3% (see Figure 4.1).

For English listeners listening to Arabic stimulus (A/**ba**/ + V/**qa**/) the McGurk percentage was the greatest at the visual lead of 60ms at a percentage of 81% after that the McGurk percentage gradually decreased reaching 42% at 300ms and 26% at -300ms. The temporal window of integration for English listeners listening to Arabic stimulus was -120ms to 240ms. It can be seen in Figure 4.2 that similar to Arabic listeners the visual response percentage was overall low for English listeners listening to Arabic stimulus, but there was some visual response in the visual lead range which peaked at 60ms at a percentage of 4%.

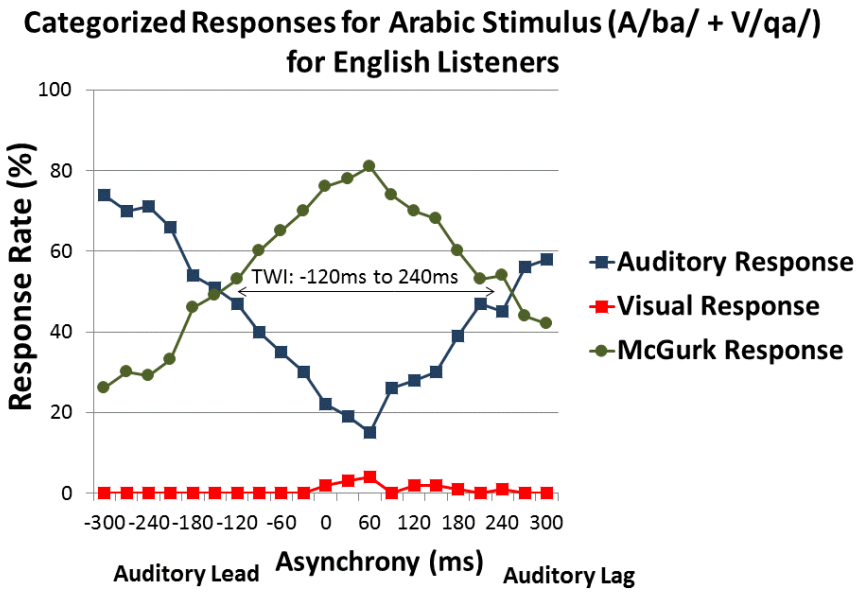

Figure 4.2 Categorized responses (auditory, visual, and McGurk Response) for Arabic stimulus (A/ba/ + V/qa/) shown as proportions for English listeners and the temporal window of integration.

4.3.2 Response Categories to English stimulus (A/pa/ + V/ka/)

Figures 4.3 and 4.4 show the responses for English stimulus of Arabic and English listeners as proportions of three response categories (auditory, visual, and McGurk response). For Arabic listeners listening to an English stimulus (A/**pa**/ + V/**ka**/) the McGurk response never reached 50% of the responses. The highest McGurk response percentage 29% was at a visual lead (auditory lag) of 30ms. As the visual lead increased greater than 30 ms the McGurk response percentage decreased gradually to 15% of the responses at 300ms. The auditory lead led to a rapid decrease in the McGurk response and an increase in auditory response. At -300ms auditory lead McGurk response percentage was 0% and the auditory response was at 100%. Figure 4.3 illustrates that similar to Arabic stimulus the visual response percentage for English stimulus was overall low, but 60ms of visual lead was the most favourable for visual response at a percentage of 4%.

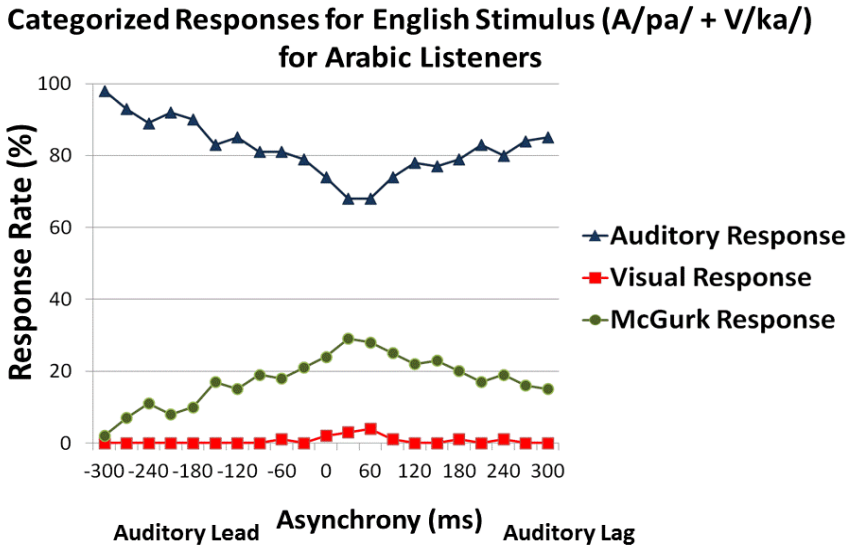

Figure 4.3 Categorized responses (auditory, visual, and McGurk Response) for English stimulus (A/pa/ + V/ka/) shown as proportions for Arabic native listeners.

For English listeners listening to an English stimulus (A/**pa**/ + V/**ka**/) the McGurk percentage was the greatest at the visual lead of 30ms at a percentage of 60%. After that the McGurk percentage gradually decreased to 34% of the responses at 300ms visual lead. The temporal window of integration for English listeners listening to English stimulus was -90ms to 210ms. Beyond the temporal window of integration range the auditory response was greater than the McGurk response and this difference was greater for the auditory lead than the visual lead. Similar to Arabic listeners the visual response percentage was overall low for English listeners listening to English stimulus, but there was some visual response in the visual lead range which peaked at 30 to 60ms at a percentage of 5% (see Figure 4.4).

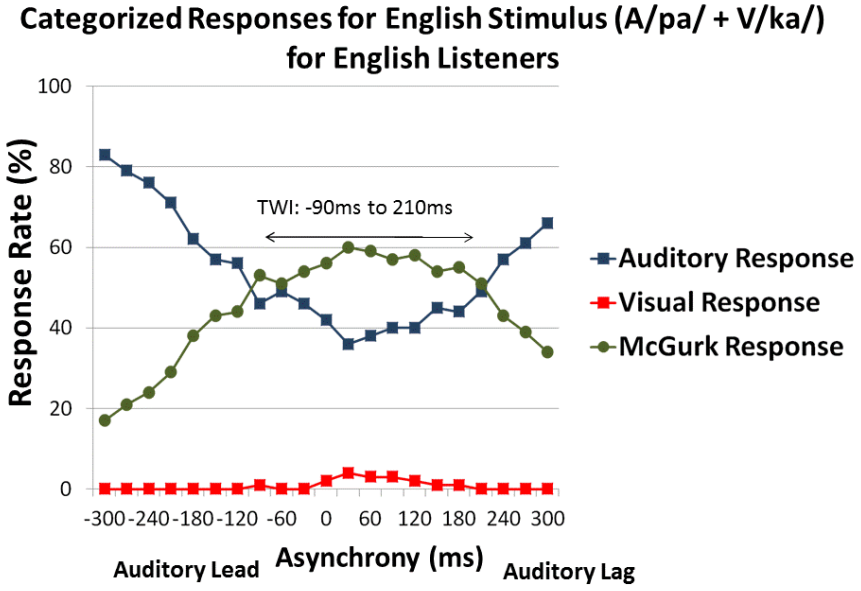

Figure 4.4 Categorized responses (auditory, visual, and McGurk Response) for English stimulus (A/pa/ + V/ka/) shown as proportions for English native listeners.

4.3.3 ANOVA for Arabic and English Stimulus

The effect of language of listener and stimulus pair on the percentage of McGurk response was investigated using a two-way analysis of variance (ANOVA) for repeated measures. Levene's test verified the equality of variances in the samples (homogeneity of variance) (***p***>.05) (Martin and Bridgmon, 2012). The within group effect of stimulus type (Arabic or English) was significant [***F*** (1,32)= 37.25; ***p*** <.0001)], that means that the Arabic stimulus (A/**ba**/ + V/**qa**/) produced a significant larger percentage of McGurk response compared to the English stimulus (A/**pa**/ + V/**ka**/). The between group effect of native language was also significant [***F*** (1,32)= 183.89; ***p*** <.0001)] this was due to the significantly larger McGurk response percentage by the English listeners compared to the Arabic listeners. The interaction between stimulus and native language of listener was not significant [***F*** (1,32)= 1.57; ***p***=.219)].

4.3.4 Open Set Responses for Arabic and English Stimulus

The open–set McGurk responses for the asynchronous stimuli pairs (A/**ba**/ + V/**qa**/ and A/**pa**/ + V/**ka**/) were tallied and Figures 4.5 and 4.6 show the consonant identification responses by Arabic and English native listeners for Arabic and English stimuli. These figures show that the phonetic responses varied across the different types of asynchronous stimuli. It can be seen in Figure 4.5 that for Arabic stimulus (A/**ba**/ + V/**qa**/) the auditory response was **/ba/** for both English and Arabic listeners. However, the visual response for Arabic listeners was **/qa/** while for English listeners it was **/ga/** Likewise the two groups had different responses for the

McGurk response, Arabic listeners' response was /**da**/ and /**taˀ**/ while English listeners' responses was only /**da**/.

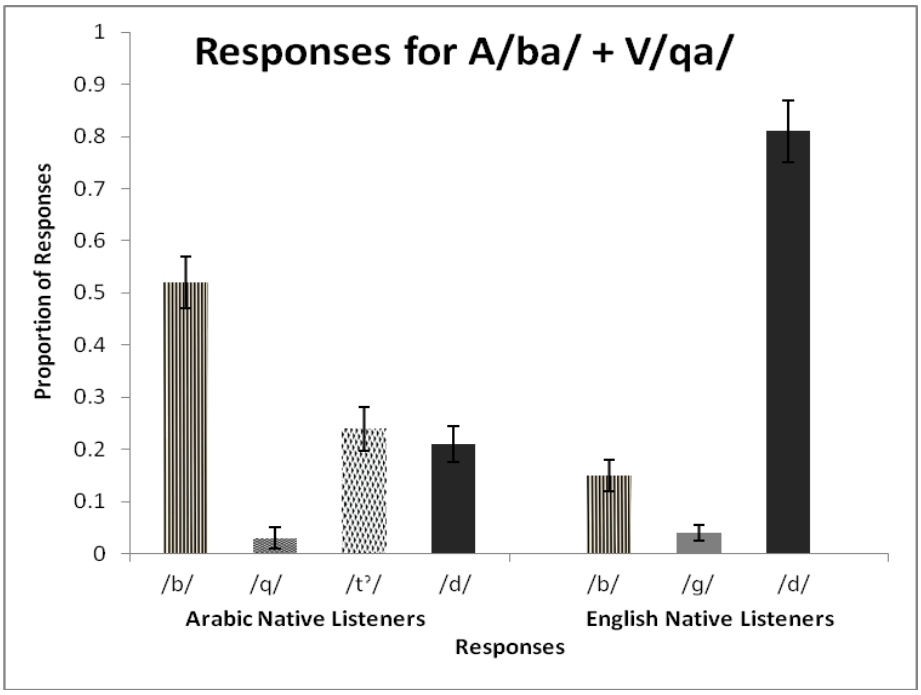

Figure 4.5 Shows the response proportions for consonant identification for Arabic stimuli (A/ba/ + V/qa) by Arabic and English native listeners.

Figure 4.6 illustrates that for English stimulus (A/**pa**/ + V/**ka**/) both groups visual response was /**ka**/ and the McGurk response was /**ta**/. However, their auditory response was /**ba**/ for Arabic listeners and /**pa**/ for English listeners. These responses are similar to the results in experiment 1(chapter 3). The responses are based on the visual and auditory native language mental representations of the listener, in other words the Arabic listeners do not have /**p**/ as an auditory mental representation in their native language therefore auditory assimilation occurs and they perceive /**b**/. The same occurs for the visual stimuli where the English listeners do not have /**q**/ as

a mental representation in their visual mental repertoire therefore visual assimilation occurs and they perceive /**g**/.

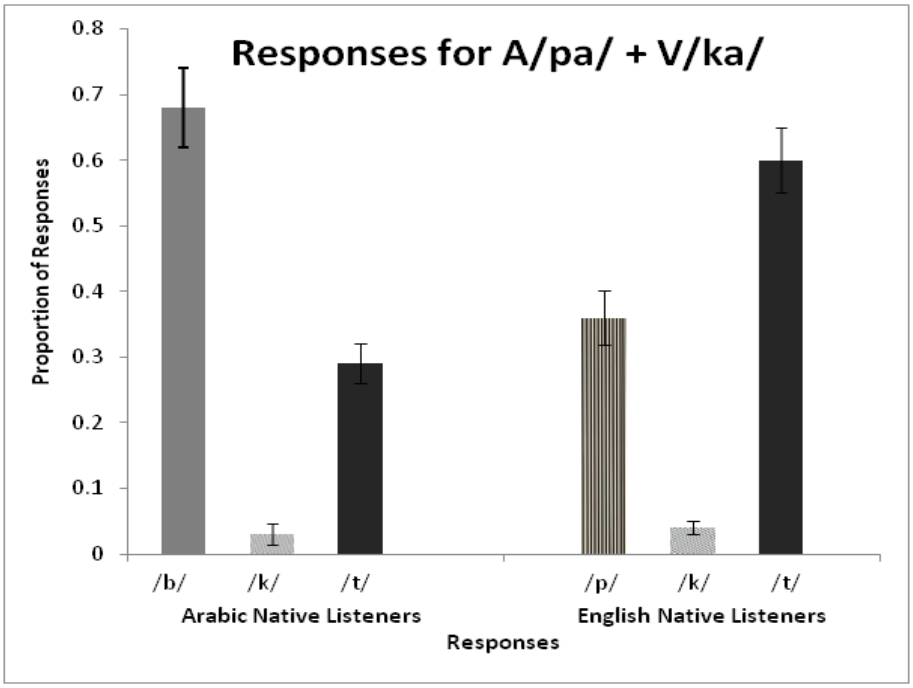

Figure 4.6 Shows the response proportions for consonant identification for English stimuli (A/pa/ + V/ka/) by Arabic and English native listeners.

4.4 Discussion

The results of this experiment do not support early integration of auditory-visual speech as even at the optimal visual lead time the Arabic listeners had a significantly lower McGurk response percentage compared to English listeners. If auditory-visual integration was based on early integration then the Arabic listeners' McGurk percentage at optimal visual lead time should have not been significantly lower than the English listeners. Therefore, these results suggest that the lower McGurk percentage found in Arabic listeners compared to English listeners demonstrated in

experiment 1 (chapter 3) is not due to bottom-up processing speed of visual speech cues. These differences in auditory-visual integration measured by the percentage of McGurk response could possibly be due to differences in the development of mental representations based on the native language. These abstract auditory and visual mental representations are organized around auditory and visual categories of the native language. This would mean that there is a difference in the predictive coding strength of visual cues during auditory-visual integration which would depend on the native language mental representations (see chapter 2 section 2.6).

Further support for dependency of auditory-visual speech on native language mental representations comes from a study by Sánchez-Garcia et al. (2013). They evaluated visual speech lead time for Spanish and English native speakers. They found that visual lead time was more beneficial for the listener's native language. They suggest that native visual speech cues predict native auditory speech cues faster than that for non-native visual speech cues. They concluded that auditory-visual integration is dependent on phonological mental representations (Sánchez-García et al., 2013). These results can be explained by late integration that occurs for auditory-visual speech which is dependent on native language mental representations (see chapter 2 section 2.7).

In this chapter the extent to which temporal incongruence of the visual and auditory information influences the McGurk effect across two different native languages was explored. The findings suggest that for both Arabic and English native listeners auditory-visual integration of speech tolerates temporal asynchrony in the presentation of auditory and visual events, integrating the two streams into a single synchronous event. The McGurk response for both sets of stimuli and listeners

favours a visual lead of 30 to 60ms; this finding is similar to that previously reported (Conrey and Pisoni, 2006). There is a marked asymmetry in the integration of auditory and visual information for both Arabic and English listeners. When the auditory signal leads the visual signal both Arabic and English listeners exhibit an appreciable decline in the percentage of McGurk response relative to the synchronous condition. When the visual signal leads the auditory signal, a different pattern emerges. There is first an increase in the McGurk response percentage and then a gradual decline.

This large tolerance for asynchronous auditory-visual speech stimuli for both Arabic and English stimuli sets seen in both Arabic and English native listeners implies that separate modality specific representations are maintained for the speech stimuli over long stimulus onset asynchronies. This implies that auditory-visual integration of speech probably does not occur at an early stage in the process of speech perception. Temporal overlaps of auditory and visual stimuli are important to convergence (Calvert, 2001). Convergence refers to a framework for combining diverse information in the early stages of perception. Therefore, since auditory-visual integration of speech occurs even when the stimuli are asynchronous this implies that convergence does not occur for speech perception.

The results presented in this experiment suggest that strict timing of visual and auditory speech information is not the major determinant of auditory-visual integration in speech. Participants' perceptions were influenced by the visual stimuli even when the auditory information lagged the visual information by as much as 240 msec. When the auditory stimuli led the visual stimuli, participants showed less tolerance for the lack of synchrony. This preference for visual lead time is consistent

within the research (Munhall et al., 1996, van Wassenhove et al., 2007). This could be due to different propagation speeds of light and sound, humans are accustomed to perceiving everyday life events with auditory signals arriving later than the corresponding visual signals. Consequently, observers are more tolerant of a delay in the auditory signal than a delay in the visual signal for the integration of the auditory-visual stimuli. As such the results that the listeners had a higher McGurk percentage in the visual lead condition compared to the auditory lead condition was expected. This research, similarly, shows that temporal congruency of information from the auditory and visual modality is not essential. However, the results show that the auditory-visual stimuli do show some limits on the range over which the signals from the two modalities are treated as synchronous. The results are consistent with previous research on the temporal constraints of the McGurk effect (Munhall et al., 1996, van Wassenhove et al., 2007).

From a theoretical perspective, the delays cause further considerations on the particular conditions of auditory-visual integration in speech and at which stage the information from both modalities combine. For an early model of auditory-visual integration, perceptual unity for information from both modalities will be required. However, the research of Green et al. (1991) has demonstrated that the knowledge regarding the correspondence between the two modalities, meaning a single factor of perceptual unity, was not a prerequisite for one's perception of the McGurk effect. In this research by Green et al., stimuli to which participants were exposed consisted of voices and faces differed by genders. That is male voices were combined with female faces and visa versa, however no decrease in the degree of the McGurk effect

was demonstrated irrespective of the fact that incompatibility of faces and voices was obvious (Green et al., 1990).

Further research in support of the variable strength of visual speech cues has been carried out by van Wassenhove et al. (2005). They have stated that components of the evoked potential N1 and P2 (see chapter 1 section 1.4.3.1).demonstrated a significant reduction in processing time or latency for auditory-visual syllable presentations compared to auditory presentations alone. Additionally, they found that the size of latency shifts demonstrated to be proportionate to the phoneme's visual saliency. For example /**p**/ is more visually salient then /**k**/ consequently /**p**/ visual stimuli induced a greater latency shift in N1 and P2. This finding suggests that information received from visual speech can be of benefit in processing the consequent following auditory input. This is achieved through the application of predictive mechanisms dependent on the phonemes' saliency or ambiguity.

Interestingly for the English listeners the temporal window of integration was larger for the Arabic stimuli (-120 to +240) compared to the English stimuli (-90 to +210). This could be interpreted using the auditory-visual native language framework which proposes that integration of speech relies on the features of native language mental representations of auditory-visual speech cues. One of the features of these mental representations would need to be the alignment of auditory and visual dimensions. This would suggest that one would be more sensitive to native auditory-visual asynchronous stimuli compared to non-native stimuli. This difference might be due to experience with the native language which increases sensitivity to temporal changes in the stimuli. Changes can occur to the temporal window of integration due

to learning and previous experience (Fujisaki et al., 2004, Navarra et al., 2010, Powers et al., 2009).

In this experiment bottom-up processing speed of visual speech cues did not account for the significant difference in the McGurk percentage between Arabic and English listeners. Even at optimal visual lead the McGurk percentage of Arabic listeners was significantly less than English listeners. This finding suggests that the reduced use of visual speech cues by Arabic listeners compared to English listeners is not due to differences in bottom-up processing. Thus top-down processing differences need to be investigated. In terms of the hypothesis of this thesis, it is argued that in the brain the information flows not only forward according to a hierarchy of processing levels, but that during certain processing stages information also meets a top-down 'prediction' (Altieri, 2014, Barutchua et al., 2008, Brancazio, 2004). Thus, auditory-visual native language mental representations assist in reducing the ambiguity during the process of speech perception. Visual cues add to the auditory information received during speech perception. If the visual cues of one language are more ambiguous than another language this may lead to an auditory-visual integration process that relies less on the visual cues compared to the other language. Accordingly, it is suggested that the difference between Arabic and English listeners in the percentage of McGurk response might be due to a difference in the saliency of the visual cues of the native language.

To be able to investigate the predictive power of visual speech cues in Arabic it was first necessary to identify the viseme categories in Arabic. Visemes are the basic unit of visual speech and each viseme category can have a many to one mapping. Experiment three reported in the following chapter aimed to categorize the visual

cues of consonants in Arabic into viseme categories. This will enable the comparison of the viseme categories of Arabic to published viseme categories of English. The results will permit the evaluation of whether Arabic has a greater number of ambiguous visual cues, which might then lead to a decrease in reliance on visual cues during speech perception. The classification of the Arabic visemes is essential for planning the final experiment in chapter 6 which will evaluate the weight of Arabic visual cues during speech perception.

Chapter 5

Viseme Categories in Arabic

5.1 Introduction

The data from experiment 1 in chapter 3 showed a significant difference in auditory-visual integration percentage, as measured by the McGurk effect, between monolingual Arabic and English native listeners. In a second experiment described in chapter 4 bottom-up visual processing speeds was investigated to determine whether that may influence the difference in auditory-visual integration between Arabic and English listeners. However, visual processing speed did not significantly increase the auditory-visual integration percentage as measured by the McGurk effect for Arabic native listeners compared to English native listeners. That suggests that the difference in the way the visual speech cues are used during auditory-visual integration are not due to a lower level functional change in visual processing speed.

This leads to the question of whether the cross-linguistic difference in auditory-visual integration of speech demonstrated so far could be due to increased visual speech ambiguity in Arabic which makes the predictive power of the visual cues in Arabic less than the visual speech cues in English. For example the results of experiment one and two might be that the mental representations of speech signals across the visual and auditory mental repertoire in Arabic native listeners for guttural phonemes are weighted more in favour of auditory cues rather than visual cues compared to English listeners. This would suggest that there are a greater number of ambiguous visual cues in Arabic. Thus this would cause a greater density of visual

mental representation for guttural phonemes in Arabic leading to a reduced percentage of McGurk responses in Arabic native listeners compared to English native listeners.

In experiment 1 (chapter 3) it was also found that Arabic and English listeners picked up on different visual cues which resulted in a difference in the perceived auditory-visual speech token. For example the fusion response that occurred the greatest for Arabic stimuli (A/**ba**/+V/**qa**/) was /**t ˁ**/ for Arabic listeners, while English listeners' response was only /**d**/. The visual phoneme in the Arabic stimuli was an emphatic phoneme /**q**/ and Arabic listeners' McGurk response was also an emphatic phoneme /**t ˁ**/. Emphatic consonants are pronounced with the back of the tongue approaching the pharynx. This implies that Arabic listeners picked up on the visual cues for an emphatic phoneme. The ability of Arabic listeners to visually differentiate between an emphatic and non-emphatic phoneme has also been found by Ouni and Ouni (2007). In addition the results of experiment 1 (chapter 3) have suggested that when the visual cue is an emphatic sound this can affect the perceived auditory-visual speech token.

Thus Arabic listeners incorporated the emphatic category in their choice of the McGurk response. However, since there are no emphatic phonemes in English, the English listeners were not able to recognize the visual cues for an emphatic phoneme and assimilated the visual stimulus to the closest category within their native language (see chapter 3 section 3.5.2). It has been suggested that the perceptual process might direct which auditory and/or visual features are bound to the speech stimulus (Massaro, 1987). This view is consistent with the hypothesis in this thesis that the role of the speech processing framework might be to weight speech feature

representations from both the visual and auditory modalities. The features that are integrated from the visual and auditory input depend on the predictive power of the speech stimuli provided from both modalities.

5.1.1 Visual Speech

Visual speech information has been shown to be beneficial in processing the auditory input through the means of predictive mechanisms (Bubic et al., 2010, Enns and Lleras, 2008). These predictive mechanisms are thought to depend on the saliency or ambiguity of the phoneme's visual cues (van Wassenhove et al., 2005) (see chapter 2 section 2.6.2). One of the first studies investigating visual speech ambiguity was conducted by Massaro (1987) which demonstrated that visual information influences categorisation of heard phonemes among adults and children. His results showed that during auditory-visual speech perception as the ambiguity of the auditory syllable increased so did the reliance on the visual syllable. In terms of the Fuzzy Logical Model of Perception (Massaro, 1998), this finding was considered to be a characteristic tendency for the visual speech information with the lowest visual ambiguity, to be more influential and reliable with ambiguous auditory speech stimuli (see chapter 2 section 2.4). Therefore if the visual cues of one language are more ambiguous than another language this may influence the auditory-visual integration process so that it relies less on the visual cues compared to the other language.

A distinct feature of Arabic is the presence of many guttural phonemes (Elgendy and Pols, 2001). Guttural phonemes are sounds produced in the back of the mouth such

as uvular, pharyngeal and glottal sounds. Consequently visual speech cues would not be very beneficial for identifying guttural sounds. There are also many phonemes in Arabic which have an emphatic counterpart (see Table 5.1).

Table 5.1 Arabic Consonants used as Visual Stimuli (ˤ emphatic)

Manner			Place of Articulation								
			Bilabial	Labio-dental	Dental	Alveolar	Palato-Alveolar	Velar	Uvular	Pharyngeal	Glottal
Nasal	Voiced	Non-Emphatic	m			n					
Stop	Voiceless	Non-Emphatic				t		k			ʔ
		Emphatic				tˤ			q		
	Voiced	Non-Emphatic	b			d		g			
		Emphatic				dˤ					
Fricative	Voiceless	Non-Emphatic		f	θ	s	ʃ		χ	ħ	h
		Emphatic				sˤ					
	Voiced	Non-Emphatic			ð	z			ʁ	ʕ	
		Emphatic			ðˤ						
Affricate	Voiced	Non-Emphatic					dʒ				
Approximant	Voiced	Central	w			r	j				
		Lateral				l					

For example **/t/** and **/tˤ/** are both alveolar, voiceless, stop consonants, but **/tˤ/** is an emphatic sound. The visual similarity between phonemes and their emphatic phoneme counterpart might lead to an increase in visual ambiguity of speech sounds in Arabic. Therefore, it is suggested that the difference between Arabic and English listeners in the use of visual cues during auditory-visual integration of speech might be due to a difference in the ambiguity of the visual cues of the native language and a different repertoire of visual and auditory mental representations of the native language.

5.1.2 Visemes

Many studies of English visual speech have examined the visual discrimination of consonants. Visemes are visually based categories of contrast, similar to phonemes in the auditory modality. Table 5.2 shows the 10 viseme categories that have been defined for English (Bozkurt et al., 2007). Studies in other European languages have found essentially similar viseme groups compared to English; Welsh (Meredith et al., 1990) and Swedish (Martony, 1974) .

Table 5.2 Viseme categories for English consonants (Bozkurt et al., 2007).

Viseme Category	Phonemes
1	/ p, b, m/
2	/f, v/
3	/w/
4	/ θ, ð /
5	/t, d, n, l/
6	/s, z/
7	/ʃ, dʒ/
8	/r/
9	/j/
10	/k, g, h/

However no investigation using human speech-reading has been conducted on viseme groups covering the entire range of consonants used in Arabic. This experiment was conducted to determine the confusability of all Arabic consonants grouped into their viseme classes. In this thesis it is suggested that speech perception relies on the native language mental representations for both visual and auditory cues. Depending on the visual speech cues of the native language, a different set of features may be at the focus of attention for speech processing.

This chapter reports an experiment conducted to identify viseme categories for all consonants in Arabic and to compare the results to published results for viseme categories within English. Since Arabic has many guttural phonemes (see chapter 1 section 1.7) it is hypothesised that Arabic will have more consonants represented within a viseme category compared to English. This could then lead to increased visual ambiguity during speech perception in Arabic compared to English, explaining why Arabic listeners in experiment one and two relied less on the visual cues compared to English listeners. It is also expected that Arabic listeners will at times distinguish visually emphatic from non-emphatic phonemes supporting the view that mental representation of speech are dependent on the parameters of visual cues within the native language.

The rationale behind this experiment is that by establishing the interclass confusion for Arabic phonemes in their viseme class, a better understanding can be obtained on the weighting framework of the separate auditory and visual speech cues. This can be subsequently applied to enhance our understanding of the fusion stage of auditory-visual integration during speech perception. Which will be evaluated further in chapter 6 of this thesis.

5.2 Aim and Objectives

The aim of this experiment is to determine whether Arabic has more ambiguity within viseme categories compared to published identifications of English visemes.

Objectives

1. To identify the number of Viseme categories for the 29 Arabic consonants
2. To identify the number of consonants within each viseme category
3. To determine if the emphatic consonant can be visually discriminated from their non-emphatic counterparts
4. To analyse whether the guttural consonants can be visually discriminated

5.3 Method

This third experiment was a visual only task. The stimuli included all 29 Arabic consonants spoken in consonant vowel (CV) syllables. Twenty eight of the consonants are traditional Arabic consonants that are both spoken and written, however one consonant /**g**/ is only spoken. In order not to direct the participants in their perception, the experiment required a free-form response.

5.3.1 Participants

For a power of 80% and a significance level of 5 % with a medium effect size 0.25 (Cohen, 1988) a sample size of 36 was estimated. The participants were 36 adults, ages between 20 to 50 years with a mean age of 28 years (SD= 8.6 years; 23 women and 13 men) native Arabic speakers. To control for dialect differences, all the participants were from Riyadh, Saudi Arabia. All participants reported normal or

corrected-to-normal vision. A routine hearing screening was performed on all the participants to ensure normal hearing at 20 dBHL for octave frequencies between 500Hz to 4000Hz. None of the participants had a hearing problem thus all participants were included in the analysis. All participants gave their written informed consent to take part in the study, and the study was approved by the School of Healthcare Research Ethics Committee, University of Leeds, United Kingdom and by the local committee at the Department of Rehabilitation Sciences, in the Applied Medical Sciences College, King Saud University, Saudi Arabia.

5.3.2 Stimuli

Participants were tested on the speech reading stimuli in Arabic. To control for co-articulation affects the stimuli were in the form of a CV syllable. Stimuli used were all 29 Arabic consonants (see Table 5.1). The vowel used was the /**a**/ vowel as it leads to the greatest visual impact in Arabic (Ouni and Ouni, 2007). To control for speaker effect, stimuli were recorded from two individuals from Riyadh (one woman and one man) to ensure consistency in dialect. The speakers were videotaped in a well-lit, sound proof room with a plain background with a Canon Legria-HFS200 video camera onto a memory card. Speakers were instructed to start and end from a neutral mouth position and to avoid blinking.

Recordings were chosen that avoided low level cues which might provide non-linguistic predictive information. The selection criteria for choosing the recordings that would be included were as follows: no blinks or other eye movements occurred during the production of the syllable, the lips were in a neutral position before

articulation began and the lips returned to a resting position after the syllable was produced. Stimuli were transferred to an Intel laptop running Windows 7 for video editing. All video editing was performed in Adobe Premiere Elements 9 Software (Adobe, 2010). The auditory signal was deleted from all stimuli using Adobe Premiere Elements 9 software. The start of each stimulus file was selected by visually inspecting each stimulus for the first visible lip movement and then placing a marker 2 seconds before that point. The end of each stimulus file was chosen by finding the frame at which the speaker's mouth returned to a neutral lip position and then placing a marker 2 seconds after that point. The average stimulus duration was 5 seconds.

In each block there were 58 stimuli randomized within a block (29 CV x 2 speakers= 58 stimuli). There were 10 blocks therefore the session consisted of 580 trials (58 stimuli x 10 blocks). Hence each stimulus was repeated 20 times (580 trials/29 CV syllables= 20 repetitions for each CV syllable).

5.3.3 Procedure

Each participant had one session lasting approximately one hour and a half. After every two blocks the participants had a 5 minute break. The participants were tested individually in a soundproof booth within the audiology suite situated within the School of Rehabilitation Sciences, King Saud University, Riyadh Saudi Arabia. Participants were seated about 70 cm from a 15 inch laptop screen and listened to the speech stimuli through Circumaural headphones (Sennheiser HD438) at normal conversational level of 70dB SPL. Each trial consisted of a short video clip (5 sec) of

a person producing the speech stimuli. The participants were at a 0° angle to the laptop screen while the experimenter was at a 90° angle to the screen. The participants were asked to watch the face of the talker on the screen and then identify the consonant that they think the person is saying. The experimenter ensured throughout the session that the participant was looking directly at the screen. Following the presentation of verbal instructions, the participants were given a short practice session of 5 video clips to familiarize themselves with the protocol.

SuperLab software (Version 4.5, Cedrus Corporation, 2009) (Abboud et al., 2010) was used to present the stimuli in a random order and record the participants' free-form response. Consequently, both experimenter and participant were blind to stimulus presentation order. After each stimulus a response box was displayed on the laptop monitor and the participant typed in his/her response using the laptop keyboard, so if they heard /**ba**/ they would type "ba" using the keyboard in the response screen. After the participant pressed the "Enter" key a new trial was presented, the testing was self-paced.

5.3.4 Analysis

The data was analysed using Hierarchical Cluster Analysis to define the viseme categories for the 29 consonants in Arabic. Hierarchical clustering has become the standard method in the literature for defining viseme categories at a correlation percentage of 75% (Chen and Rao, 1998, Goldschen et al., 1994, Owens and Blazek, 1985). Responses of all participants were combined to form a complete confusion matrix. The stimuli were then defined as variables and the data analysed using the

SPSS package (IBM SPSS). Hierarchical clustering organizes observations in a tree structure based on similarity or dissimilarity between clusters. The algorithm starts with each observation as its own cluster, and successively combines the two or more most similar objects into one cluster. This cluster is then redefined as a single object and the process is repeated until all of the objects are combined into one group. The results of hierarchical clustering are presented in a dendrogram (tree diagram). The advantage of displaying the cluster analysis in dendrogram form is that the closeness of association between the viseme groups is easily seen.

Two important properties of the algorithm are a) the distance measure and b) the linkage method. The distance between pairs is defined by the cosine method, which is a pattern similarity measure. This method was chosen because its function is to group together those variables (which in this case are consonant stimuli) that elicit the most similar responses. The function of the average linkage technique is to define the distance between two clusters as the average of the distances between all pairs of the two clusters' members. Examples of distance and linkage are given in the results section 5.4.

These date were compared to the most recent viseme categories for all of the English consonants (Bozkurt et al., 2007) to investigate whether there are more consonants in the viseme groups of Arabic compared to English.

5.4 Results

The mean correct consonant identification percentage was 42.78%, ranging from 0% for /**g**/ to 99% for /**f**, **w**/. The consonants with identification greater than 90% were

bilabial or labiodental phonemes /**b**, **f**, **w**/. This was expected as these phonemes are visually prominent as their place of articulation includes the lips. The phoneme /**l**/ had the next highest identification percentage at 80% followed by; /**dʒ**/ at 73%, /**j**/ at 72% and /**r**/ at 70%. The phoneme /**q**/ had the highest identification percentage between the guttural consonants at 45%. The remainder of the phonemes had an identification percentage below 45%. Table 5.3 shows the complete confusion matrix for all participants in this experiment.

Table 5.3 Confusion Matrix for all participants

Responses	Stimulus																												
	m	b	f	w	θ	ð	ðˤ	n	t	d	tˤ	dˤ	s	z	sˤ	l	r	ʃ	dʒ	j	k	g	q	χ	ʁ	ħ	ʔ	h
m	0.21	0.38																											
b	0.79	0.62																											
f			0.99																										
w				0.99																									
θ					0.39	0.44	0.31			0.05																			
ð					0.32	0.27	0.28			0.04																			
ðˤ					0.27	0.25	0.39					0.01																	
n								0.32	0.14	0.11							0.07								0.01			
t								0.29	0.33	0.27	0.26	0.18	0.12	0.11	0.04					0.07	0.03							
d						0.01		0.26	0.3	0.32	0.15	0.22	0.07	0.1	0.05	0.09					0.04	0.03		0.01		0.01		
tˤ							0.01		0.16	0.14	0.56	0.51	0.04		0.23	0.06	0.02						0.02					
dˤ																													
s									0.04				0.35	0.41	0.25			0.03						0.01				
z													0.24	0.3	0.02													
sˤ													0.17	0.08	0.39			0.01	0.02									
l					0.01	0.02		0.05	0.03	0.06		0.04				0.81	0.19		0.01	0.06	0.02	0.01						
r								0.03		0.01						0.04	0.7			0.04	0.02	0.01	0.01		0.02			
ʃ																		0.33	0.24									
dʒ															0.01			0.63	0.73	0.03								
j								0.02					0.01							0.72	0.34	0.22	0.01		0.01			
k								0.02									0.01			0.06	0.27	0.25	0.03	0.04		0.02	0.15	0.22
ɡ																				.01	.03	0.04						
q											0.02	0.02									0.06	0.06	0.45	0.25	0.18	0.18	0.08	0.13
χ																							0.15	0.16	0.12	0.15	0.07	0.1
ʁ												0.01									0.08	0.03		0.07	0.08	0.07		0.01
ħ																						0.11	0.09	0.09	0.08	0.17	0.15	0.01
ʕ											0.01										0.06	0.06	0.13	0.06	0.1	0.15	0.06	0.22
ʔ												0.01										0.08	0.04	0.13	0.26	0.09	0.26	0.02
h																				0.02	0.05	0.1	0.07	0.18	0.14	0.16	0.23	0.29

Table 5.4 indicates the actual distance in correlation between the different consonants and at what stages the clusters are combining. Each consonant initially is its own cluster, for example in stage two in Table 5.4 cluster /**θ**/ combines with cluster /**ð**/ at a correlation percentage of 98.5%. The cluster at a later stage can add more consonants, but at a lower correlation percentage. So for the same cluster that appeared in stage 2 composed of /**θ, ð**/ at stage 6 it adds /**ðˤ**/ at a correlation percentage of 93.6%.

Table 5.4 Agglomeration Schedule

Stage	Cluster Combined		Correlation
	Cluster 1	Cluster 2	
1	/m/	/b/	0.999
2	/θ/	/ð/	0.985
3	/dʒ/	/ʃ/	0.983
4	/s/	/z/	0.975
5	/ħ/	/ʕ/	0.958
6	/θ, ð/	/ðˤ/	0.936
7	/tˤ/	/dˤ/	0.922
8	/χ/	/ʁ/	0.902
9	/k/	/g/	0.9
10	/t/	/d/	0.894
11	/ χ, ʁ/	/ħ, ʕ/	0.881
12	/t, d/	/n/	0.854
13	/ χ, ʁ, ħ,ʕ/	/q/	0.816
14	/q, χ,ʁ,ħ,ʕ/	/h, ʔ /	0.759
15	/k, g/	/j/	0.746
16	/s, z/	/ sˤ/	0.623

Figure 5.1 shows a dendrogram which is an application of the cluster analysis technique, indicating the possible viseme grouping for the 29 consonants in Arabic.

The dendrogram is a visual depiction of the correlation between the consonants. The dendrogram is read from left to right, consonants and clusters of consonants are joined together by vertical lines. The distance at which this merger takes place indicates how correlated the consonants in each category are. The distances shown in the dendrogram are 0–100%; that is the highest correlation is 100% and the lowest correlation is at 0%. The first step is at 100% correlation, at this step each consonant is its own cluster. The last merging step cluster solution takes place at 0% correlation, where all the consonants are merged into one cluster.

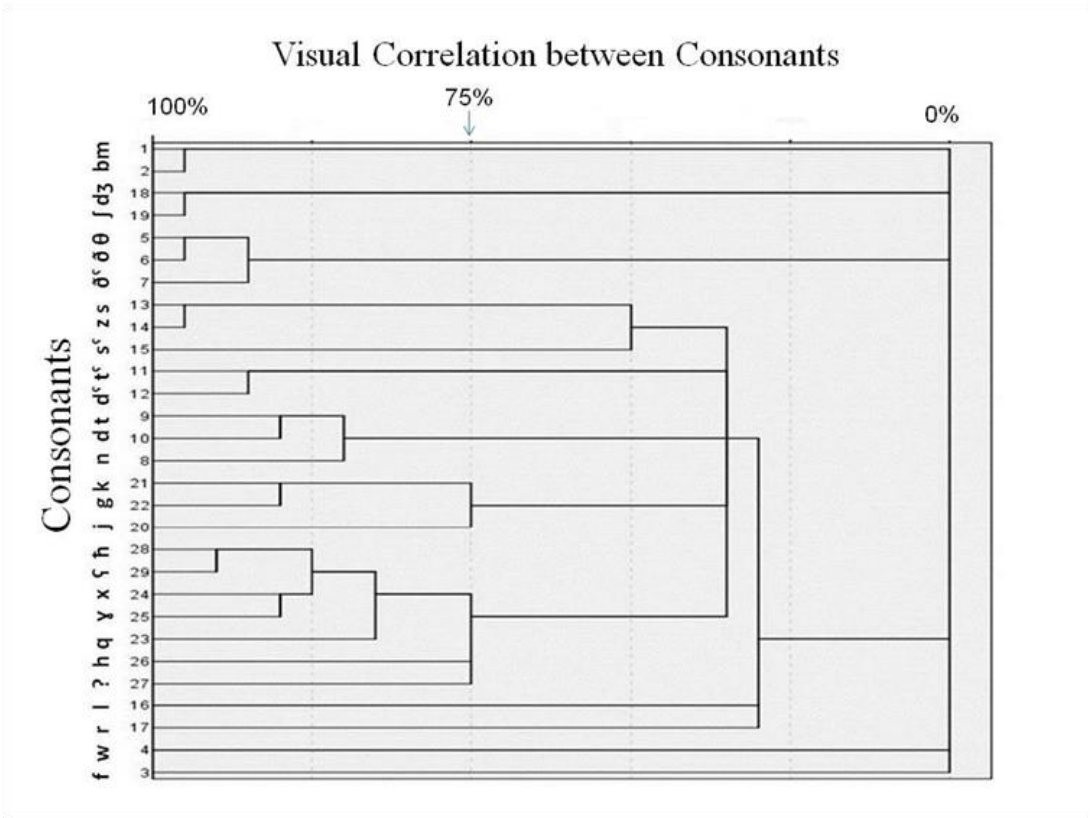

Figure 5.1 Dendrogram for Correlation between the 29 Arabic Consonants.

For example the correlation between /**b**/ and /**m**/ is very high the actual correlation is 0.999 between the two consonants this can be seen in stage 1 in Table 5.4. The vertical line connecting them on the dendrogram is near 100% on the correlation scale. While the /**j**/ consonant combines with the /**g**, **k**/ cluster at 0.716 correlation, this can be seen in stage 14 in Table 5.4.

Viseme groups are defined where the consonants in a cluster have a correlation which exceeds a given threshold, the thresholds assigned to define viseme groups in the literature is 75% correlation (Alothman, 2009, Owens and Blazek, 1985, Xue et al., 2004). Based on the data 13 viseme groups are formed at 75% or greater correlation (see Table 5.5). When analysing the viseme groups it can be seen that the emphatic phonemes, which are consonants that are pronounced in such a manner that the back of the tongue retracts into the pharynx, are not always visually differentiated from their non-emphatic counterpart.

Table 5.5 Viseme Categories for Arabic Consonants

Viseme Category	Phonemes
1	**/b,m/**
2	**/f/**
3	**/w/**
4	**/ θ, ð, ðˤ /**
5	**/t, d, n/**
6	**/ tˤ, dˤ/**
7	**/s, z/**
8	**/ sˤ/**
9	**/ʃ, dʒ/**
10	**/l/**
11	**/r/**
12	**/k, g, j/**
13	**/q, χ, ʁ , ħ, ʕ, h, ʔ/**

For instance, the emphatic consonant is not distinguished from the non-emphatic counterpart in the dental place of articulation. That is the dental emphatic **/ðˤ/** was not distinguished from the non-emphatic **/ð/.** However for the alveolar place of articulation the emphatic consonants were distinguished from the non-emphatic counterpart. That is / **tˤ, dˤ**/ from /**t, d**/ were distinguished from one another so were /**sˤ**/ from /**s**/. In the guttural place of articulation the emphatic /**ħ, ʕ**/ was not distinguished from the non-emphatic /**h, ʔ** /. The number of phonemes within a viseme category ranged from 1 phoneme to 7 phonemes. Viseme group 13 had the greatest number of phonemes which consisted of the guttural phonemes /**q, χ, ʁ, h, ʔ, ħ, ʕ**/, which are uvular, pharyngeal, and glottal consonants. The place of articulation of a uvular consonant is at the uvula as for the pharyngeal consonant it is in the pharynx, and the glottal consonant it is at the glottis. Since all of these articulations occur in the back of the mouth and are therefore not visible from the outside, then the visual cues are ambiguous and are not appropriate to differentiate between them, as was hypothesized.

5.5 Discussion

In this chapter the viseme categories containing the 29 consonants in Arabic were determined. This experiment is the only study that has identified viseme groups covering the entire range of Arabic consonants using human speech-reading. The phonemes in Arabic were classified into 13 viseme groups via speech reading by native Arabic speakers. In some viseme groups there are more than one phoneme in the same group. Although there is more than one phoneme within some viseme groups visual speech cues are still necessary for clarification of auditory confusion.

The additional information present in the visual modality can be applied to the improvement of speech perception. If a phoneme has high visual ambiguity with many other phonemes in the same viseme class, they provide less useful information about identification of the sound. Therefore it is not beneficial to use visual speech information since these sounds look the same in the visual domain. On the other hand, in the case a phoneme possesses very low visual ambiguity with only one or two phonemes within the same viseme class, in this case the visual modality would contain useful and additional information which would complement the auditory speech information.

The viseme category with the greatest number of phonemes was that comprised of the guttural sounds. They are produced in the back of the mouth and therefore difficult to visually distinguish from one another. Jiang et al. (2002) showed that among the different facial regions, the lip area (55%) was the most informative, although the cheeks (26%) and the chin (19%) also contributed significantly to visual intelligibility (Jiang et al., 2002). The variance accounted for in the visual perceptual results by the physical measures demonstrated that visual speech stimulus structure drives visual speech perception.

The results also suggest that Arabic emphatic consonants are sometimes distinguished from their non-emphatic counterparts based on place of articulation. When the place of articulation allows a greater sulcalisation of the tongue and lowering of the jaw then the observer can more readily visually distinguish an emphatic phoneme from its non-emphatic counterpart (see chapter 1 section 1.7). For the dental place of articulation the emphatic /**ðˤ**/ and non-emphatic /**ð**/ phoneme were not distinguished. The place of articulation for a dental phoneme requires that the

tongue remains between the teeth. Therefore, the jaw can only be lowered slightly when producing the emphatic dental phoneme /**ðˤ**/. Also to produce a dental phoneme the tongue has to remain flat to form a 'slit-fricative' rather than a 'grooved fricative'. Dental fricatives are always slit rather than grooved, meaning that the air exits across the width of the tongue, not down a groove in the centre. Therefore since the /**ðˤ**/ is a dental phoneme then it cannot easily be sulcalised and the jaw can only be lowered slightly this would lead to the emphatic visual cues to not be prominent for this phoneme. However, for the emphatic alveolar phonemes / **tˤ, dˤ, sˤ**/ the movement of the jaw is more visually accessible and clear therefore the emphatic / **tˤ, dˤ, sˤ**/ and non-emphatic / **t, d, s**/ phonemes were distinguished from one another. Conversely, the uvular, pharyngeal, and glottal place of articulation is in the back of the mouth which is not visually clear and therefore it is difficult to distinguish between emphatic /**ħ, ʕ**/ and non-emphatic /**h, ʔ** / phonemes.

Damien et al (2009 and 2011), classified Arabic phonemes into viseme categories based on computer analysis of geometric features of the lips. There were many similarities between the viseme categories found in this experiment and the ones Damien et al found. They also found that there was no visual differentiation between uvular, pharyngeal and glottal phonemes (Damien, 2011, Damien et al., 2009). However, they categorized Arabic constants into 10 viseme groups, while results of this experiment showed 13 viseme groups. The viseme groups 1, 2, 3, 4 and 9 (see Table 5.5) are identical to Damien et al (2011). Yet, Damien et al 2011, found no visual difference between emphatic and non emphatic counterparts. This might be due to Damien et al using a different type of dialect Lebanese Arabic where the participants of this study were from Saudi Arabia. Like any language; Arabic

dialect varies from one country to another. The differences in results may be explained further by their method of grouping the visemes. They did not use any listeners but instead they grouped the phonemes based on calculations using four geometric measures of the lips. Using computer analysis of the visual cues might have overlooked certain visual parameters used by humans such as the cheeks and chin (Jiang et al., 2002). Moreover in computer analysis of visual speech Abry and Boë (1986) recommend a set of eight parameters (Abry and Boë, 1986), while Damien et al (2011) only used 4. Ouni and Ouni (2007) also investigated visual speech for some Arabic consonants. They found that Arabic native speakers could visually distinguish between some pairs of emphatic and non-emphatic Arabic phonemes. Similar to the results of this experiment they found that the emphatic consonants at the alveolar place of articulation were distinguished from the non-emphatic counterpart. That is /**sˁ**/ from /**s**/ were distinguished from one another so where / **tˁ**/ from /**t** /.

5. 5.1 Developmental Issues

Teinonen et al. (2008) have explored the significance of visual speech components in speech development. They investigated whether phoneme discrimination can be enhanced by seen articulations. In other words, they examined whether seen articulations play any role in learning of phonetic categories. In their experiment, 6-month-old infants were exposed to speech sounds within the continuum between /**ba**/ and /**da** /. The first group were presented with auditory-visual articulation of a /**ba** / or /**da** /. The second group was presented with auditory only speech sounds.

Their results showed that the infants who were presented with both auditory and visual cues were significantly better in discriminating the /**ba**/ - /**da**/ contrast compared to the infants who were only presented auditory cues. Their conclusion was that visual speech cues improve phoneme discrimination and visual speech cues might also contribute to the learning of phoneme boundaries during infancy (Teinonen et al., 2008).

Additionally, some studies have shown that in Arabic emphatic and guttural phonemes are acquired later in speech development (Amayreh, 2003, Amayreh and Dyson, 1998). Since speech is auditory and visual, the phoneme with the greater number of cues would be more readily accessible to the child during speech development. The results of this experiment suggest that guttural and emphatic phonemes have visual cues that are more ambiguous and therefore this could explain why they are acquired later in development. It would seem that visual saliency of phonemes influences the age of acquisition.

5.5.2 Crosslinguistic Issues

Due to there being different phonemes in Arabic compared to English there was a different number of phonemes in some of the groups (see Table 5.6). For the viseme group 1 **/b, m/** Arabic had less number of phonemes compared to English which has **/b, p, m/**. Also for the viseme group 2 Arabic had **/f/** while English had **/f, v/.** However, for viseme group 4 Arabic had **/θ, ð, ðˤ/** while English has only **/θ, ð/.** Furthermore, the largest number of phonemes within a viseme group in Arabic is 7 phonemes, while in English the largest number of phonemes within a viseme group

is only 4. When Arabic viseme categories are compared to English viseme categories at 75% correlation Arabic was found to have more categories, 13 compared to 10 in English (Bozkurt et al., 2007). These results support the hypothesis that Arabic has more visual ambiguity for guttural phonemes compared to English.

Table 5.6 Viseme categories for English consonants (Bozkurt et al., 2007) and Arabic consonants.

English		Arabic	
Viseme Category	**Phonemes**	**Viseme Category**	**Phonemes**
1	**/ p,b,m/**	**1**	**/b,m/**
2	**/f,v/**	**2**	**/f/**
3	**/w/**	**3**	**/w/**
4	**/ θ,ð /**	**4**	**/ θ, ð, ðˤ /**
5	**/t,d,n,l/**	**5**	**/t, d, n/**
6	**/s,z/**	**6**	**/ tˤ, dˤ/**
7	**/ʃ, dʒ/**	**7**	**/s, z/**
8	**/r/**	**8**	**/ sˤ/**
9	**/j/**	**9**	**/ʃ, dʒ/**
10	**/k, g, h/**	**10**	**/l/**
		11	**/r/**
		12	**/k, g, j/**
		13	**/q, χ, ʁ , ħ, ʕ, h, ʔ/**

The results also confirm that Arabic listeners can at times detect emphatic from non-emphatic counterparts. These findings might help to explain the results from the first experiment where Arabic listeners had a decrease in auditory-visual integration as measured by the percentage of the McGurk effect compared to English listeners. The increase in visual ambiguity for guttural phonemes in Arabic might then lead to a decrease in predictive power of visual mental representations for guttural phonemes

due to there being more phonemes within this group in Arabic compared to English. This would lead to a more difficult visual perception of guttural phonemes for Arabic listeners compared to English listeners. This decrease in auditory-visual integration might suggest a shift in weighting between Arabic and English listeners; where for Arabic listeners less weight is put on visual cues for guttural phonemes due to an increase in visual ambiguity within their native language as compared to English listeners. Furthermore some Arabic emphatic phonemes were found to be visual distinguishable from their non-emphatic counterpart; this would lead to emphatic visual cues to be a feature within the visual mental representations of Arabic listeners. This would explain why in experiment one Arabic listeners sometimes perceived an emphatic phoneme during auditory-visual integration, while English listeners never did.

The present study determined the viseme groups of Arabic consonants and that emphatic and guttural phonemes lead to an increase in the number of phonemes within some viseme groups. To further analyse this, in the following chapter a fourth experiment was conducted to compare the percentage of visual influence in a McGurk paradigm across the 13 viseme groups of Arabic. This would help to evaluate whether visual ambiguity within the viseme group leads to a decrease in visual influence during auditory-visual speech perception.

Chapter 6

Visual Speech Effect in Arabic

6.1 Introduction

In experiment 3 (chapter 5), the viseme categories for all 29 consonants in Arabic were identified (see Table 6.1). Results indicated that Arabic has 13 viseme categories compared to 10 in English and that the largest number of phonemes within a viseme group in Arabic is 7 phonemes, while in English the largest number of phonemes within a viseme group is only 4. The question that arose was whether this increase in visual ambiguity found in Arabic compared to English could assist in explaining the results found in the first experiment in chapter 3. The findings from that experiment indicated a decrease in auditory-visual integration measured by the percentage of the McGurk effect for Arabic participants compared to English participant. This reduction in the McGurk percentage in Arabic participants seemed to be due to a decrease reliance on visual speech cues compared to English participants.

Although the results of the third experiment in chapter 5 showed comparatively large viseme categories which suggest more visual ambiguity in Arabic it is now necessary to evaluate if this visual ambiguity specifically affects auditory-visual integration during speech perception. Consequently, the main rationale for conducting the next experiment is based on the following considerations. First of all, since auditory cues appear to be dominant in speech perception for Arabic, it is assumed that the weighting between auditory and visual cues in speech perception

Table 6.1 Viseme Categories for 29 Arabic Consonants

Viseme Group	Phonemes
1	/b,m/
2	/f/
3	/w/
4	/ θ, ð, ðˤ /
5	/t, d, n/
6	/ tˤ, dˤ/
7	/s, z/
8	/ sˤ/
9	/ʃ, dʒ/
10	/l/
11	/r/
12	/k, g, j/
13	/q, χ, ʁ , ħ, ʕ, h, ʔ/

could have an effect on auditory-visual integration of speech. Since some of the visual cues have been demonstrated to be more ambiguous than others in Arabic, the consequent question was what their role in auditory-visual integration is. Thus, experiment four was conducted using the McGurk effect to investigate whether visual ambiguity of phonemes affects auditory-visual integration of speech. The hypothesis proposed in this thesis is that visual ambiguity of a phoneme will lead to less auditory-visual integration during speech perception, measured by a reduced McGurk response percentage. Resulting in greater reliance on auditory cues and auditory-visual speech perception is influenced by visual speech features specific to the native language.

The hypothesis will be tested by comparing the auditory-visual integration responses across the different viseme groups in Arabic. In this experiment the effect of visemes

with low levels of visual ambiguity (4 or less phonemes within the group) was compared to phonemes from groups with a high level of visual ambiguity (more than 4 phonemes within the group) on the McGurk effect. Additionally, emphatic visual cues in Arabic were evaluated to investigate if they influence auditory-visual speech perception. Moreover, place of the auditory stimulus were evaluated to investigate whether it affected auditory-visual integration of speech.

6.2 Aim and Objectives

The aim of this experiment is to assess the effect of visual cues of phonemes across the 13 viseme categories in Arabic on the McGurk effect.

The specific objectives are:

1. To evaluate whether the visual ambiguity of phonemes in Arabic affects the percentage of visually influenced responses (visual correct, McGurk, combination). The rationale for this is that if visual ambiguity of speech cues influences the weighting between auditory and visual cues during speech perception, then the role of highly ambiguous visemes in speech perception would be expected to be quite low, because the mental representations will be more tuned to the auditory cues rather than these highly ambiguous visual phonemes. On the other hand, viseme groups which are unambiguous have distinct visual mental representations that are more dominant during speech

perception. Consequently, it is expected that the phonemes in the largest viseme group will have the smallest visually influenced response percentage.

2. To compare the influence of the auditory stimulus bilabial /**b**/ and alveolar /**l**/ on auditory-visual integration of speech. It is expected that the place of articulation of the auditory stimulus would influence the type of response category during auditory-visual integration of speech (whether McGurk or combination response). However, it is expected that for both auditory /**b**/ and /**l**/ the more ambiguous the visual stimulus is the less impact it will have on auditory-visual speech perception.

6.3 Method

In this experiment the effect of visual phonemes from groups with a low level of visual ambiguity (4 or less phonemes within the group) was compared to phonemes from groups with a high level of visual ambiguity (more than 4 phonemes within the group) on auditory-visual integration during speech perception. This experiment is a within participant design with an auditory-visual identification task using the McGurk effect. The auditory consonants that were used are /**b**/ and /**l**/ as they produce the largest McGurk effect (Jiang and Bernstein, 2011). The visual consonants were all 29 consonants of Arabic (see Table 6.2). The vowel used was the /**a**/ vowel as it leads to the greatest visual impact in Arabic (Ouni and Ouni,

2007). The stimuli were in the form of a consonant vowel (CV) syllable to control for co-articulation affects.

Table 6.2 Arabic Consonants used as Visual Stimuli (ˤ emphatic)

Manner			Place of Articulation								
			Bilabial	Labio-dental	Dental	Alveolar	Palato-Alveolar	Velar	Uvular	Pharyngeal	Glottal
Nasal	Voiced	Non-Emphatic	m			n					
Stop	Voiceless	Non-Emphatic				t		k			ʔ
		Emphatic				tˤ			q		
	Voiced	Non-Emphatic	b			d		g			
		Emphatic				dˤ					
Fricative	Voiceless	Non-Emphatic		f	θ	s	ʃ		χ	ħ	h
		Emphatic				sˤ					
	Voiced	Non-Emphatic			ð	z			ʁ	ʕ	
		Emphatic			ðˤ						
Affricate	Voiced	Non-Emphatic					ʤ				
Approximant	Voiced	Central	w			r	j				
		Lateral				l					

6.3.1 Participants

For a power of 80% and a significance level of 5 % with a medium effect size 0.25 (Cohen, 1988) a sample size of 46 was estimated. The participants were 46 adults, ages between 20 to 50 years with a mean age of 32 years (SD= 6.8 years; 27 women and 19 men) native listeners of Arabic. To control for dialect differences, all the participants were from Riyadh, Saudi Arabia. All participants reported normal or corrected-to-normal vision. A routine hearing screening was performed on all the participants to ensure normal hearing within a 20 dBHL for octave frequencies

between 500Hz to 4000Hz. None of the participants had a hearing problem thus all participants were included in the analysis. All participants gave their written informed consent to take part in the study, and the study was approved by the School of Healthcare Research Ethics Committee, University of Leeds, United Kingdom and by the local committee at the Department of Rehabilitation Sciences, in the Applied Medical Sciences College, King Saud University, Saudi Arabia.

6.3.2 Stimuli

6.3.2.1 Stimulus Generation

To control for speaker effects, stimuli were recorded from two individuals. Furthermore, to obtain the same dialect as the participants; materials were recorded from native Arabic Saudi adults living in Riyadh (one woman and one man). The video and auditory recordings were made using the same procedure described in experiment one (see chapter 3 section 3.3.2.1). The mean SPL was 70.61 dB (SD=1.34 dB), a t-test indicated that there was no significant difference between the SPL values of all the stimuli by the 2 speakers ($p= 0.86$, $t = .03$, $df = 162$). A sound calibrator (Bruel and Kjaer-4231) which conforms to EN/IEC 60942 Class LS and Class 1, and ANSI S1.40-1984 was used to calibrate the measurement system.

6.3.2.2 Auditory-Visual Stimulus Alignment

Adobe Premiere Elements 9 Software (Adobe, 2010) was used to generate congruent and incongruent auditory-visual stimuli following the same method as in experiment one (see chapter 3 section 3.3.2.2). For each of the two speakers 29 congruent stimuli were generated as controls and 56 incongruent stimuli (see Table 6.3- 6.5); therefore there were 85 stimuli generated for each of the two speakers.

In each block there were 170 auditory-visual CV syllables ([29 congruent stimuli + 56 incongruent stimuli] x 2 native speakers= 170 stimuli). There were 3 blocks of stimuli with 170 CV syllables including both of the two speakers randomized within a block. Each session consisted of 510 trials (170 stimuli x 3 blocks) therefore there were 6 presentations for each stimulus.

Table 6.3 Congruent stimuli

Number	Auditory Stimuli	Visual Stimuli
1	/ma/	/ma/
2	/ba/	/ba/
3	/wa/	/wa/
4	/fa/	/fa/
5	/θa/	/θa/
6	/ða/	/ða/
7	/ðˤa/	/ðˤa/
8	/na/	/na/
9	/ta/	/ta/
10	/tˤa/	/tˤa/
11	/da/	/da/
12	/dˤa/	/dˤa/
13	/sa/	/sa/
14	/sˤa/	/sˤa/
15	/za/	/za/
16	/la/	/la/
17	/ra/	/ra/
18	/ʃa/	/ʃa/
19	/dʒa/	/dʒa/
20	/ja/	/ja/
21	/ka/	/ka/
22	/ga/	/ga/
23	/qˤa/	/qˤa/
24	/ χa/	/ χa/
25	/ʁa/	/ʁa/
26	/ħa/	/ħa/
27	/ʕa/	/ʕa/
28	/ʔa/	/ʔa/
29	/ha/	/ha/

Table 6.4 Incongruent stimuli auditory /ba/

Number	Auditory Stimuli	Visual Stimuli
1	/ba/	/ma/
2	/ba/	/wa/
3	/ba/	/fa/
4	/ba/	/θa/
5	/ba/	/ða/
6	/ba/	/ðˤa/
7	/ba/	/na/
8	/ba/	/ta/
9	/ba/	/tˤa/
10	/ba/	/da/
11	/ba/	/dˤa/
12	/ba/	/sa/
13	/ba/	/sˤa/
14	/ba/	/za/
15	/ba/	/la/
16	/ba/	/ra/
17	/ba/	/ʃa/
18	/ba/	/dʒa/
19	/ba/	/ja/
20	/ba/	/ka/
21	/ba/	/ga/
22	/ba/	/qˤa/
23	/ba/	/χa/
24	/ba/	/ʁa/
25	/ba/	/ħa/
26	/ba/	/ʕa/
27	/ba/	/ʔa/
28	/ba/	/ha/

Table 6.5 Incongruent stimuli auditory /la/

Number	Auditory Stimuli	Visual Stimuli
29	/la/	/ma/
30	/la/	/ba/
31	/la/	/wa/
32	/la/	/fa/
33	/la/	/θa/
34	/la/	/ða/
35	/la/	/ðˤa/
36	/la/	/na/
37	/la/	/ta/
38	/la/	/tˤa/
39	/la/	/da/
40	/la/	/dˤa/
41	/la/	/sa/
42	/la/	/sˤa/
43	/la/	/za/
44	/la/	/ra/
45	/la/	/ʃa/
46	/la/	/dʒa/
47	/la/	/ja/
48	/la/	/ka/
49	/la/	/ga/
50	/la/	/qˤa/
51	/la/	/χa/
52	/la/	/ʁa/
53	/la/	/ħa/
54	/la/	/ʕa/
55	/la/	/ʔa/
56	/la/	/ha/

6.3.3 Procedure

Participants were tested individually in a sound proof room situated at the School of Rehabilitation Sciences, King Saud University, Riyadh Saudi Arabia. Each participant took part in one session which lasted about one hour and a half. Participants were given a 5 minute break after each block (approximately every 15 minutes). Participants were seated about 70 cm from a 15 inch laptop screen and listened to the speech stimuli through Circumaural headphones (Sennheiser HD438) at normal conversational level of 70dB SPL. The participants were at a 0° angle to the laptop screen while the experimenter was at a 90° angle to the laptop screen. On each trial participants were asked to watch the face of the talker on the laptop screen whilst listening to the output from the headphones and then identify the consonant or consonants that were heard. The researcher ensured throughout the session that the participant was looking directly at the screen. Following the presentation of verbal instructions, the participants were given a short practice session of 5 trials to familiarize themselves with the protocol.

SuperLab presentation software (Version 4.5, Cedrus Corporation, 2009) was used to present the stimuli in a random order and record the participants' response, so if they heard /**ba**/ they would type "ba" using the keyboard in the response screen. The two speakers and stimuli were randomized within each block where, neither experimenter nor participant knew which stimuli were incongruent and which were congruent. Each trial consisted of a short video clip (5 sec) of a person saying the experimental stimuli. After the participant presses the "Enter" key a new trial was presented, the testing was self-paced.

6.3.4 Analysis

In chapter 5, 13 Arabic viseme categories were identified. Phonemes within a viseme category are visually indistinguishable from one another, therefore the responses within each of the 13 Arabic viseme categories were averaged together. To evaluate the visual effect of the phoneme on speech perception the responses were categorized into four categories; auditory (e.g., the response to A/**ba**/ + V /**ka**/ was /**ba**/), visual (e.g., the response to A/**ba**/ + V /**ka**/ was /**ka**/), combination (e.g., the response to A/**ba**/ + V /**ka**/ was /**bka**/) and fusion (e.g., the response to A/**ba**/ + V /**ka**/ was /**da**/). Any response other than auditory indicated influence of the visual phoneme. For a response to be considered a McGurk response it must be a response that is not the same as the auditory or any of the visual signals within the viseme category. For example if the response to A/**ba**/ + V /**ka**/ was /**ga**/ it would not be considered a McGurk response since /**g**/ and /**k**/ are within the same viseme category (i.e. they are not visually distinguishable).

The effect of viseme group on the presence of a visual influenced response during speech perception was investigated using a binary logistic multiple regression model. Visual, McGurk, and combination responses were considered a visual influenced response (Jiang and Bernstein, 2011). Only an auditory response was considered unaffected by the visual stimuli. Binary logistic multiple regression is a statistical technique which measures the relationship between a categorical dependent variable (visual influenced response) and several independent variables (13 viseme groups). The dependent variable for the model was visual influenced response (1=present and 0 not present) and predictors were the 13 Arabic viseme groups.

The odds ratio was measured for each of the 13 viseme groups. The odds ratio in logistic regression can be interpreted as the measure of a ratio of effect size of the predictor on an outcome compared to the effect size of the other predictors. In this experiment, odds ratio is measuring the ratio of effect size of each viseme group on the percentage of visual influenced responses compared to the effect size of the other viseme groups. The odds ratio can range from 0 to infinity, the higher the odds ratio is for a viseme group the greater the effect of the visual stimuli is in that group on auditory-visual speech perception. Therefore if one viseme group has an odds ratio of 3 and the other of 2, then the phonemes in the latter viseme group has less visual effect during auditory-visual speech perception compared to the phonemes in the other viseme group.

6.4 Results

6.4.1 Auditory /ba/

Tables 6.6, 6.7 and 6.8 show the proportion of responses for auditory /**ba**/ across the 13 viseme categories. The mean response category proportions for auditory /**ba**/ were auditory correct 0.4, visual correct 0.24, and McGurk response 0.36. For the auditory /**ba**/stimulus set there were no combination responses. The guttural viseme group /**q, χ, ʁ, ħ, ʕ, h, ʔ**/ had the largest auditory response (64%) and viseme group /**f**/ had the smallest auditory response (26%).

Table 6.6 Viseme groups ordered from largest to smallest for Auditory responses for Auditory /ba/.

Viseme Group	Auditory Response
/q, χ, ʁ, ħ, ʕ, h, ʔ/	0.64
/w/	0.58
/k, g, j/	0.54
/ʃ, dʒ/	0.47
/r/	0.46
/ sˤ/	0.37
/s, z/	0.35
/tˤ, dˤ/	0.32
/ l/	0.31
/ θ, ð, ðˤ /	0.28
/t, d, n/	0.27
/f/	0.26
Average	0.4

Table 6.7 Viseme groups ordered from largest to smallest for Visual responses for Auditory /ba/.

Viseme Group	Visual Response
/f/	0.74
/t, d, n/	0.63
/tˤ, dˤ/	0.54
/θ, ð, ðˤ /	0.33
/w/	0.19
/ʃ, dʒ/	0.13
/l/	0.11
/r/	0.09
/q, χ, ʁ, ħ, ʕ, h, ʔ/	0.06
/s, z/	0
/ sˤ/	0
/k, g, j/	0

Table 6.8 Viseme groups ordered from largest to smallest for McGurk responses for Auditory /ba/.

Viseme Group	McGurk Response
/s, z/	**0.65**
/ sˤ/	**0.63**
/l/	**0.58**
/k, g, j/	**0.46**
/r/	**0.45**
/θ, ð, ðˤ /	**0.39**
/ʃ, dʒ/	**0.4**
/q, χ, ʁ, ħ, ʕ, h, ʔ/	**0.3**
/w/	**0.23**
/tˤ, dˤ/	**0.14**
/t, d, n/	**0.1**
/f/	**0**
Average	**0.36**

On inspection it can be seen that in Figure 6.1 and 6.2 that the viseme group /**f**/ had the greatest visual responses (74%) and the greatest number of McGurk responses was the viseme group /**s,z**/ (65%). The McGurk responses were /**da**/at 84%, **ða**/ at 8.7% and /**tˤa/** at 7.3%. Only four viseme groups had McGurk responses other than /**da**/ they are viseme groups /**t, d, n**/, /**sˤ**/ , /**ʃ, dʒ**/ and / **q, χ, ʁ, ħ, ʕ, h, ʔ**/ (see Figure 6.3).

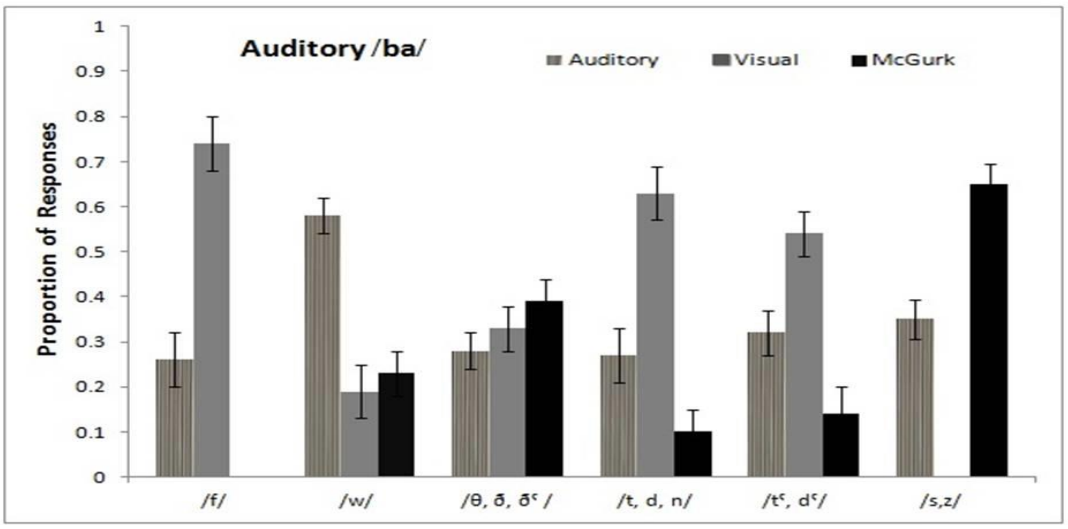

Figure 6.1 Categorized responses (auditory correct, visual correct, and McGurk) shown proportions by Arabic viseme groups 2-7 (y-axis) for Auditory /ba/.

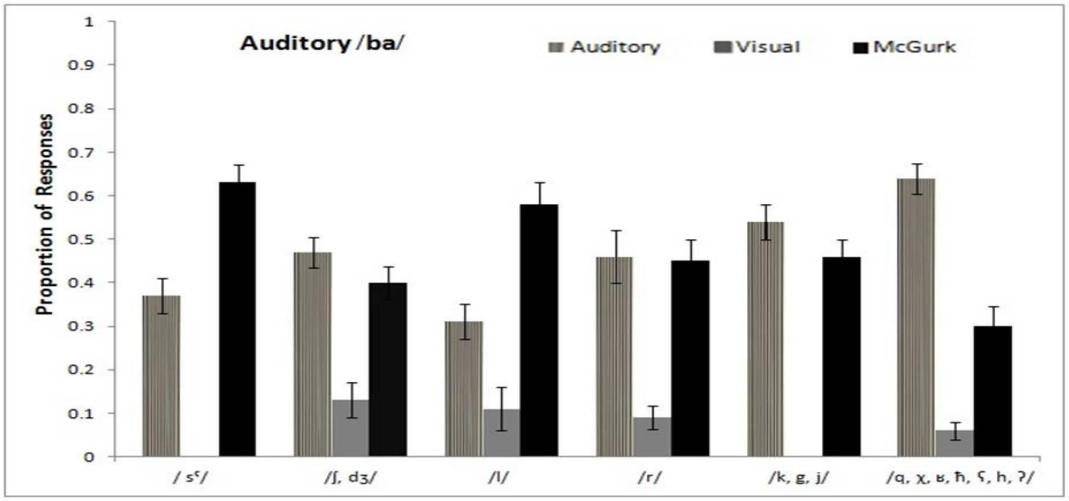

Figure 6.2 Categorized responses (auditory correct, visual correct, and McGurk) shown proportions by Arabic viseme groups 8-13 (y-axis) for Auditory /ba/.

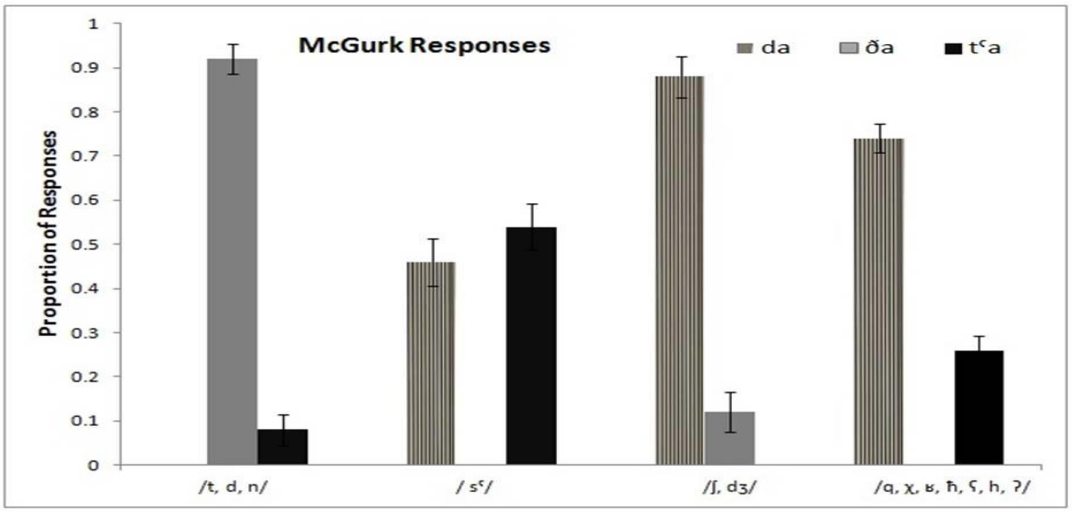

Figure 6.3 Categorized McGurk responses (/da/, /dˤa/, and /ða/) shown as proportions by viseme group (y-axis) for Auditory /ba/.

Table 6.9 displays the odds ratio for each of the viseme groups with the significance level. For auditory /**ba**/, the visual influence of the viseme groups was significant ($p<0.001$). The largest odds ratio (6.64) for producing a visual influenced response, that is sum of visual and McGurk responses, was for viseme group /**f**/ and the smallest odds ratio (1.00) for producing a visual response was for viseme group /**q, χ, ʁ, ħ, ʕ, h, ʔ**/. This means the phoneme in viseme group /**f**/ is 6.64 times more likely to influence a visual response during auditory-visual speech perception than phonemes in viseme group /**q, χ, ʁ, ħ, ʕ, h, ʔ**/.

Table 6.9 Odds Ratio for the Viseme groups with auditory /ba/.

Viseme	Phonemes	Significance	Odds Ratio
2	/f/	<0.001	6.64
3	/w/	0.02	1.99
4	/θ, ð, ðˤ /	<0.001	6.2
5	/t, d, n/	<0.001	6.5
6	/tˤ, dˤ/	<0.001	4.52
7	/s, z/	<0.001	4.19
8	/ sˤ/	<0.001	4.01
9	/ʃ, dʒ/	<0.001	3.37
10	/l/	<0.001	5.97
11	/r/	<0.001	3.43
12	/k, g, j/	0.001	2.74
13	/ q, χ, ʁ, ħ, ʕ, h, ʔ/	<0.001	1

6.4.2 Auditory /la/

Table 6.10 shows the proportion of responses for auditory /**la**/ across the 13 viseme categories. The overall response category proportions for auditory /**la**/ were auditory correct 0.75 and combination 0.25. For the auditory /**la**/ stimulus set there were no visual or McGurk responses. The guttural viseme group /**q, χ, ʁ, ħ, ʕ, h, ʔ**/ had the largest auditory response percentage (96%) and viseme group **/f/** had the smallest auditory response percentage (49%).

Table 6.10 Viseme groups ordered from largest to smallest for Auditory and Combination responses for Auditory /la/.

Viseme Group	Auditory Response	Viseme Group	Combination Response
/q, χ, ʁ, ħ, ʕ, h, ʔ/	**0.96**	**/b,m/**	**0.47**
/k, g, j/	**0.88**	**/f/**	**0.51**
/r/	**0.88**	**/w/**	**0.43**
/t, d, n/	**0.87**	**/ʃ, dʒ/**	**0.39**
/tˤ, dˤ/	**0.86**	**/θ, ð, ðˤ /**	**0.28**
/ sˤ/	**0.85**	**/s, z/**	**0.17**
/s, z/	**0.83**	**/ sˤ/**	**0.15**
/θ, ð, ðˤ /	**0.72**	**/tˤ, dˤ/**	**0.14**
/ʃ, dʒ/	**0.61**	**/t, d, n/**	**0.13**
/w/	**0.57**	**/r/**	**0.12**
/b,m/	**0.53**	**/k, g, j/**	**0.12**
/f/	**0.49**	**/q, χ, ʁ, ħ, ʕ, h, ʔ/**	**0.04**
Average	**0.75**	**Average**	**0.25**

On inspection it can be seen that in Figure 6.4 and 6.5 viseme group /**f**/ had the smallest auditory response percentage at 49% while viseme group **/q, χ, ʁ, ħ, ʕ, h, ʔ/** had the largest auditory response percentage at 96%. The largest combination response rates were for viseme group /**f**/ at 51%, group /**b,m**/ at 47% and group /**w**/ at 43% . The smallest combination response percentage was for viseme group **/q, χ, ʁ, ħ, ʕ, h, ʔ/** at 4%.

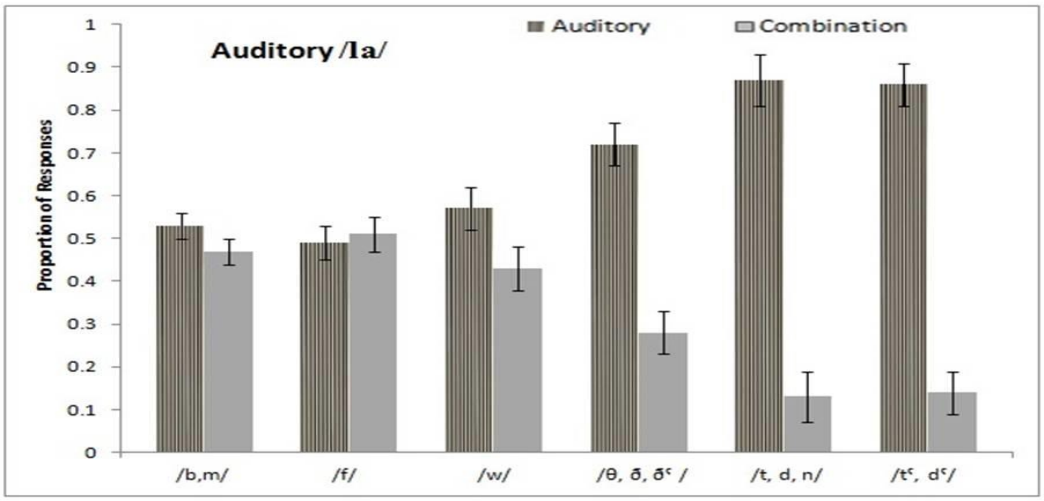

Figure 6.4 Categorized responses (auditory correct and combination) shown as proportions by Arabic viseme groups 1-6 (y-axis) for Auditory /la/.

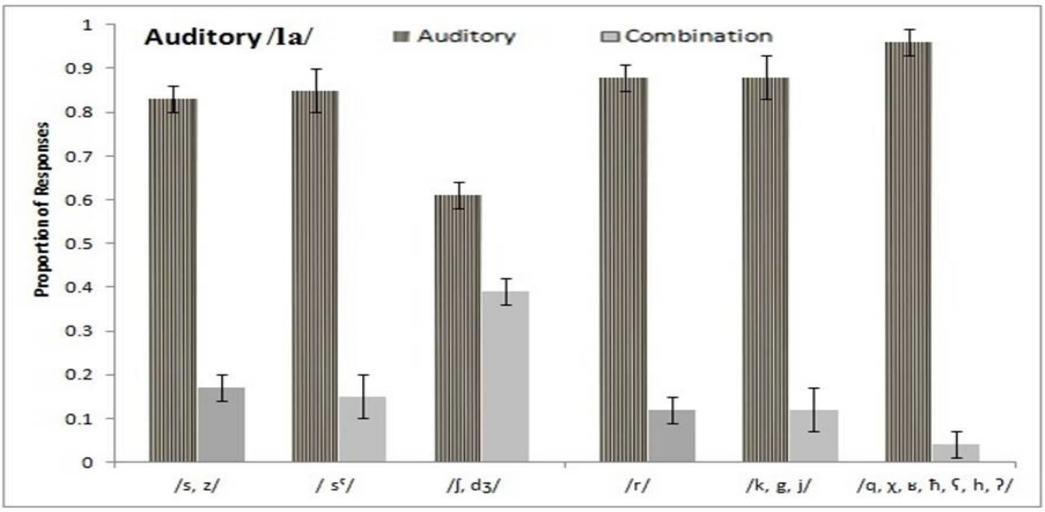

Figure 6.5 Categorized responses (auditory correct and combination) shown as proportions by Arabic viseme groups 7-13 (y-axis) for Auditory /la/.

For auditory /**la**/, the visual influence of the viseme groups was significant ($p<0.001$). Table 6.11 displays the odds ratio for each of the viseme groups. The greatest odds ratio (25) for producing a combination response was for viseme group /**f**/. The least odds ratio (1.00) for producing a combination response was for viseme group /**q, χ, ʁ, ħ, ʕ, h, ʔ**/. This means that when the stimulus was auditory /**la**/ a combination response is 25 times more likely to occur when the visual stimulus is the phoneme /**f**/ rather than the phonemes /**q, χ, ʁ, ħ, ʕ, h, ʔ**/.

Table 6.11 Odds Ratio for the Viseme groups with auditory /la/.

Viseme	Phonemes	Significance	Odds Ratio
1	/b,m/	<0.001	23.06
2	/f/	<0.001	25
3	/w/	<0.001	18.1
4	/θ, ð, ðˤ /	<0.001	9.33
5	/t, d, n/	<0.001	3.59
6	/tˤ, dˤ/	<0.001	3.98
7	/s, z/	<0.001	4.92
8	/ sˤ/	<0.001	4.24
9	/ʃ, dʒ/	<0.001	15.34
11	/r/	0.02	2.97
12	/k, g, j/	0.02	2.67
13	/ q, χ, ʁ, ħ, ʕ, h, ʔ/	<0.001	1

6.5 Discussion

Experiment 4 was performed to evaluate the visual influence of different viseme categories on auditory-visual speech perception. To test the hypothesis that greater visual ambiguity leads to a reduced visual effect, the auditory-visual responses (auditory correct, visual correct, McGurk, and combination) across all 13 viseme categories in Arabic were compared (see chapter 5). In this regard, this experiment

has demonstrated that the objectives outlined above were fulfilled (see section 6.2). The objectives were to evaluate whether the number of phonemes within a viseme group influenced the percentage of visually influenced responses (visual correct, McGurk, and combination). Second the influence of place of articulation for the auditory stimulus on the type of auditory-visual influenced response was evaluated. This was achieved by comparing the auditory stimulus bilabial /**b**/ to the alveolar /**l**/.

First of all, it was shown that phonemes belonging to large viseme groups had a reduced percentage of visually influenced responses (visual correct, McGurk, and combination) than phonemes in smaller viseme groups. Phonemes in larger viseme groups would have a larger visual phonetic density which would increase the number of invisible phonetic contrasts. Visual phonetic density depends on the density of the visual neighbourhood, which is the perceptual space populated by visual cues for the phonemes within the native language (see chapter 2 section 2.6.2.). Thus the density of the visual neighbourhood probably increased the importance of auditory cues compared to highly ambiguous visual cues for speech perception. Secondly, the results suggest that emphatic visual cues influenced auditory-visual speech perception. Arabic listeners picked up on the visual cues for an emphatic phoneme and incorporated the emphatic category in their choice of the McGurk response. Furthermore the results of this experiment have suggested that when the visual component of the auditory-visual stimulus is an emphatic phoneme this affects the perception of the auditory-visual speech. Moreover, the perception of an emphatic phoneme when the visual cue was an emphatic phoneme and the auditory stimulus a non-emphatic phoneme is suggested to be due to the specifics of visual mental representations particular to Arabic.

Finally, in terms of the McGurk paradigm, two auditory speech tokens were paired with each of 28 visual tokens across all 13 viseme categories in Arabic in order to confirm that the effect was conditioned by the ambiguity of the viseme categories and not due to the specific auditory stimulus. The auditory stimuli /**ba**/ and /**la**/ were used, because they are the most likely to induce a visual influenced response (Jiang and Bernstein, 2011). Responses were categorized as either being visually influenced which included visual, McGurk and combination response or not visually influenced that is to say an auditory response.

The results indicate that the auditory stimulus differentially influences the proportion of responses in each of the four response categories of auditory, visual, combination or McGurk. Place of articulation of the auditory stimulus influences the type of visually induced phoneme that is being perceived. Overall, the responses to the stimuli with auditory /**ba**/ were more susceptible to visual influences than those with auditory /**la**/. 60% of the responses to auditory /**ba**/ were visually influenced (i.e. a visual, combination or McGurk response was produced), while auditory /**la**/ produced only 25% visually influenced responses.

Bilabial phonemes in McGurk experiments have been found to be the auditory stimulus most likely to produce a visually induced response (Jiang and Bernstein, 2011). The place of articulation for the /**b**/ phoneme is bilabial, produced by the closure of the lips; therefore the visual cue is very visually distinct. As a result when the auditory /**b**/ phoneme is combined with a visual speech stimulus of a phoneme without closure of the lips there is a clear contradiction to the visual mental representation or mental representation of what the articulation of a /**b**/ phoneme should appear like. Hence, identification of the auditory-visual speech token was

more likely to be influenced by the visual stimulus, resulting in a McGurk response or a visual response. This might explain why the McGurk response at 36% and the visual response at 24% were the visually influenced responses for the auditory /**ba**/. In the case of the /**l**/ phoneme, the place of articulation is alveolar that is it is produced with the tip of the tongue touching the alveolar ridge. The visual stimuli for the articulation of an alveolar phoneme is not as easily distinguished as a bilabial phoneme (Jiang and Bernstein, 2011, McGurk and MacDonald, 1976). Therefore, the visual mental representation for the /**l**/ phoneme is more ambiguous compared to the /**b**/ phoneme. Consequently the majority of responses to the /**la**/ stimulus were auditory responses. The only visually influenced responses for auditory /**la**/ were the combination responses at a percentage of 25%. In a combination response both the auditory stimuli and visual stimuli are perceived, for example auditory /**la**/ and visual /**fa**/ would be perceived as /**fla**/. A combination response is considered the weakest visually influenced category as the auditory stimulus is still perceived. These results suggest that place of articulation of the auditory stimulus influences the type of visually influenced response (visual correct, McGurk, and combination).

Through this experiment, the influence of visemes in Arabic on auditory-visual integration was analysed. Phonemes from large viseme groups demonstrated a lesser degree of impact on auditory-visual integration because visual mental representations are tuned to auditory cues rather than phonemes belonging to viseme groups with many invisible phonetic contrasts. In other words, phonemes with a visually dense phonetic neighbourhood carry less information and thus are less relevant in auditory-visual integration of speech perception. Finally, the ambiguity of

the visual mental representations influences the weighting frameworks of auditory-visual integration.

The response type depended on the auditory and visual stimuli presented. In the following sections, the four response categories visual, McGurk, combination and auditory will be discussed.

6.5.1 Visual Response

The viseme groups which had the greatest visual response percentage for auditory /**ba**/ was viseme group /**f**/. That is to say, when the participants were presented with auditory /**ba**/ and visually the phoneme /**f**/ 74% of the time their response was the phoneme /**f**/ (visual stimulus). For auditory /**ba**/ the following viseme groups also had high visual response rates; viseme group /**t,d,n**/ at 63%, viseme group /**tˠ, dˠ**/ at 54%, and viseme group /**θ, ð, ðˠ**/ at 33%. The reason why viseme group /**f**/ had the highest percentage of visual responses could be due to low visual ambiguity. There is only one phoneme in this group, additionally in experiment 3 (chapter 5) the phoneme /**f**/ had a 99% visually correct identification percentage. However, this is also true for viseme group /**w**/ it only has one phoneme in the group and had 99% correct visual identification (see chapter 5 section 5.4), but it only had 19% visual response percentage for auditory /**ba**/. Therefore, there must be an additional factor other than the visual ambiguity of the phoneme that influences the percentage of visual response such as auditory-visual confusion discussed below.

First, the contradiction between the visual mental representations for the auditory and visual components of the presented stimuli must be considered. For example,

when the stimulus was an auditory /**b**/ and a visual /**f**/ the visual mental representations of these two phonemes are very distinct. A /**b**/ phoneme is a bilabial which requires a closure of the lips while the /**f**/ phoneme is a labiodental which is articulated by the lower lip against the upper teeth. This is also true for the dental phonemes in viseme group /**θ, ð, ðˤ**/ and the alveolar phonemes in viseme group /**t,d,n**/ and /**tˤ, dˤ**/ they are visually distinct from the bilabial /**b**/ phoneme. However, the phoneme /**w**/ is a bilabial and therefore the visual mental representation would then be similar to that of bilabial /**b**/.

Second, the similarity between the auditory-visual mental representations for the auditory and visual components of the presented stimuli must be evaluated. Mesgarani et al., (2008) compared the confusion of English consonants in auditory-visual condition to the auditory only condition (see Figure 6.6). The grey scale demonstrates the probability of reporting a certain phoneme (in a column) for a certain input phoneme (in a row). Consequently, the colour demonstrates the intensity or confusion percentage. The stimuli were congruent phonemes that is when the auditory stimulus was /**b**/ the visual stimulus was also /**b**/. They found that in the auditory-visual condition the phoneme /**b**/ was most likely to be confused with /**v**/. In English /**v**/ and /**f**/ are in the same viseme group that is they are not distinguished visually from one another.

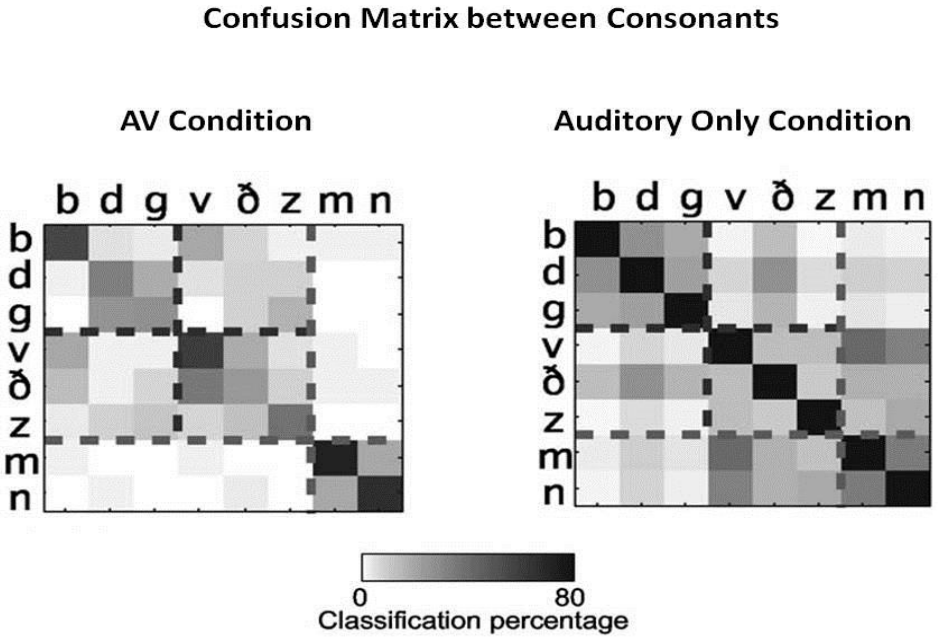

Figure 6.6 The confusion matrix for consonants in the auditory-visual condition and the Auditory only condition (Mesgarani et al., 2008).

Consistent with the results of Mesgarani et al., (2008), the results in experiment 4 (chapter 6) showed that the viseme group with the highest visual response was viseme group /**f**/ when the auditory stimulus was /**ba**/. As can be seen in Figure 6.5 the phonemes /**d**/ and /**ð**/ have a high auditory confusion percentage with the phoneme /**b**/, in other words the boundaries of the auditory mental representations are close. That can help in explaining why the viseme groups /**t,d,n**/ and /**θ, ð, ðˤ**/ had high visual response rates. Since visually they are distinct to the auditory stimulus /**b**/ and they have a high auditory confusion percentage with the /**b**/ phoneme, therefore their visual influence during speech perception of an auditory /**b**/ would be strong. Interestingly Mesgarani et al., (2008) found a low confusion percentage between phoneme **/b/** and **/v/** in the auditory only condition. This

indicates that visual input can alter speech perception when compared to the auditory only condition.

A framework proposed in this thesis is based on the processing of visual and auditory cues and then predicting which native language mental representations match these features. At this final step the prediction value between the auditory and visual mental representations are compared or weighed. Depending on these prediction values between the two modalities a decision is made on the perceived speech (see Figure 6.7).

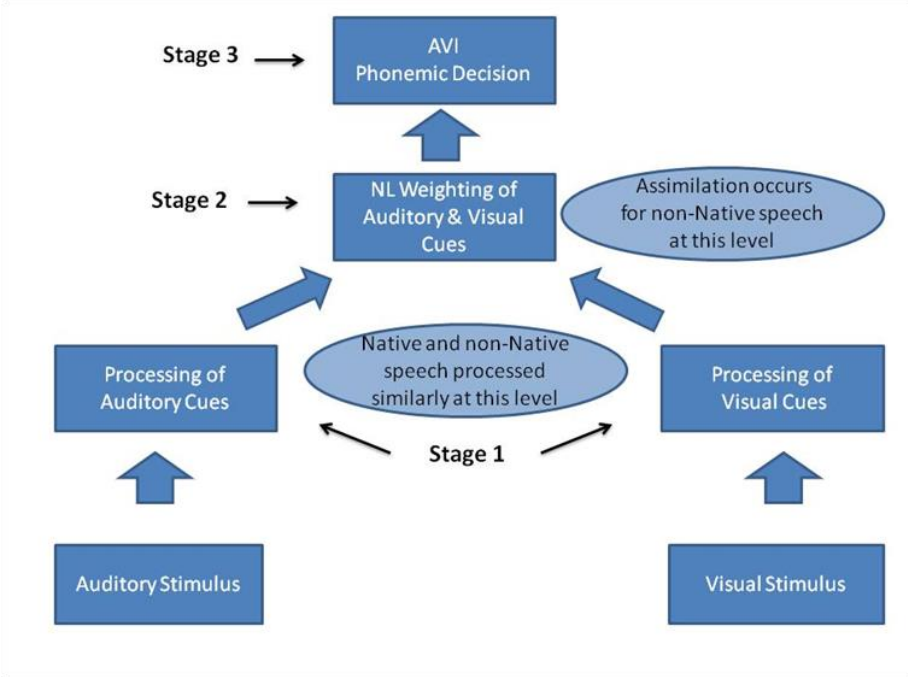

Figure 6.7 A working framework for auditory-visual integration (AVI) of speech for the native language (NL).

Thus, for a visual stimulus to have a greater influence or weight during auditory-visual speech perception it must be visually unambiguous and have a high auditory or auditory-visual confusion percentage with the auditory stimulus. Also the visual mental representations of the visual and auditory components must contrast. In other words the phoneme /**f**/ is visually unambiguous, and visually distinct from the phoneme /**b**/. Furthermore, in the auditory-visual condition it has a high confusion percentage with /**b**/ (Jiang and Bernstein, 2011, Mesgarani et al., 2008) therefore its visual influence on auditory /**b**/ during auditory-visual speech perception is strong (see Figure 6.8).

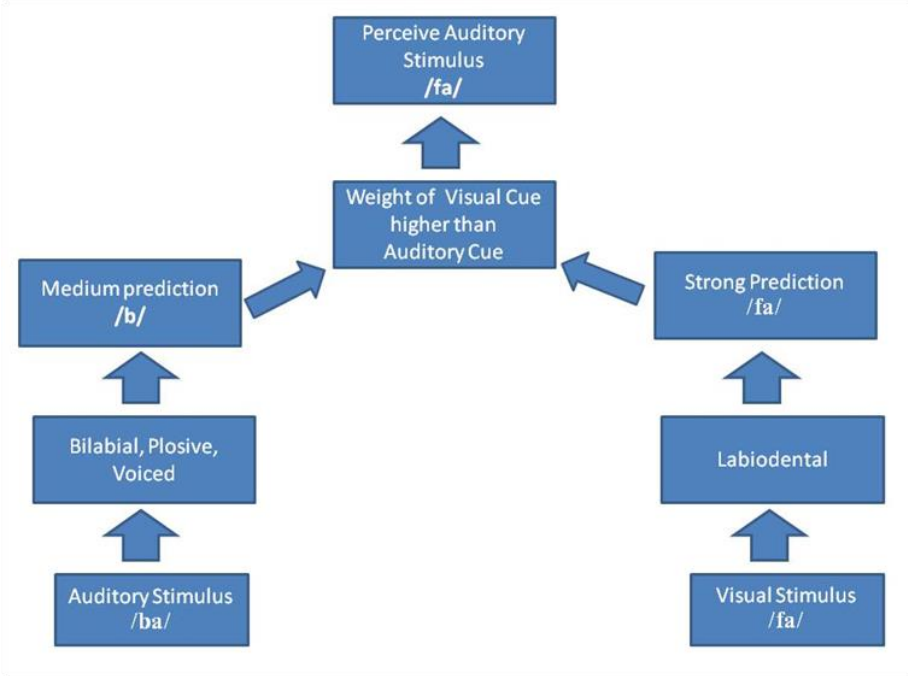

Figure 6.8 An example of the hypothesized auditory-visual native language framework for auditory-visual integration with a highly predictive visual speech cue /f/.

For a visual response to be perceived, the mental representation of the visual stimulus needs to be visually distinct. The above outlined analysis of visual response data provides a substantial argument in favour of the posed hypothesis. Further evidence that supports the hypothesis is provided in the following section describing the McGurk response data.

6.5.2 McGurk Response

Stimulus sets with auditory /**ba**/ were the only ones which led to McGurk responses and the majority of the McGurk responses were /**da**/. When the auditory stimulus is /**ba**/ the phoneme that is most likely to be perceived during a McGurk response is /**da**/ (Rosenblum, 2008). Mesgarani et al., (2008) found that in the auditory only condition the highest confusability with phoneme /**b**/ is the phoneme /**d**/. Consequently, the more similar the visual stimulus is to the perceived /**da**/ response the more likely that a fusion response will occur. The viseme group which had the greatest McGurk response percentage for auditory /**ba**/ was viseme group 7 /**s, z**/ at 65%. The main information that the visual stimulus carries is place of articulation and both /**d**/ and /**s, z**/ are alveolar phonemes, this might be why the viseme group 7 /**s, z**/ had the highest McGurk response percentage. Jiang and Bernstein (2011) also found that auditory /**ba**/ paired visually with an alveolar phoneme produced high McGurk response rates. A framework proposed here includes a weighting mechanism between the auditory and visual mental representations. Hence, when the visual stimulus for example /**s, z**/ has a visual mental representation that is similar to

the perceived auditory response, in this case /**d**/ this will increase its influence or weight during the auditory-visual integration process and thus fusion may occur.

Figure 6.9 depicts the auditory-visual phonetic perceptual space when both the auditory stimulus and visual stimulus have a *similar* influence on the perceived speech. In this case the visual stimulus has a medium density visual phonetic neighbourhood. Also the visual stimulus mental representation must contrast that of the auditory stimulus therefore neither the auditory or visual stimulus is perceived, instead a fusion of the two stimuli is perceived.

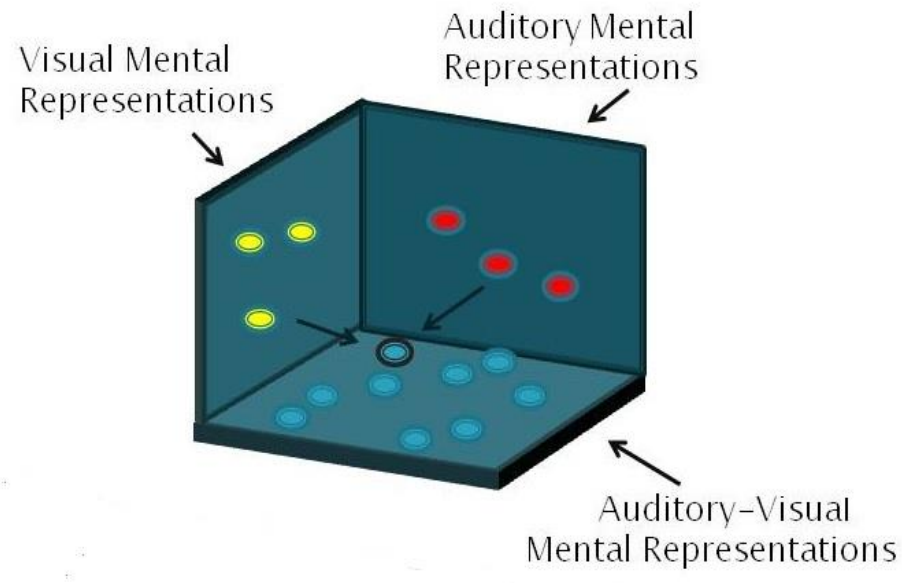

Figure 6.9 Auditory-visual phonetic perceptual space during a McGurk response.

The second viseme group with the highest McGurk response percentage was group 8 /**sˤ**/ at 63%. Interestingly this viseme group was the only group to have the majority of the perceived fusion response as /**tˤa**/. Both the visual cue /**sˤ**/ and the fusion response /**tˤ**/ are emphatic phonemes. As was reported in experiment 1 (chapter 3),

for Arabic listeners the emphatic visual cues are features that are sometimes incorporated into the McGurk response when the visual stimulus is an emphatic phoneme. This is evidence that auditory-visual integration of speech depends not only on the auditory mental representations but also on the visual mental representations in the native language. In the following section combination responses will be discussed.

6.5.3 Combination Response

Stimulus sets with auditory **/la/** were the only ones which had combination responses. A combination response is when the participant perceives both the auditory and visual stimulus (e.g., the response to A/**la**/ + V/**ka**/ was /**lka**/). The viseme groups with the highest combination responses were; group 2 **/f/** at 51%, group 1**/b, m/** at 47%, group 3 **/w/** at 43%, and group /**ʃ, dʒ**/ at 39%. These viseme groups all involve the movement of the lips to produce the phoneme, therefore they are visually prominent and unambiguous. Furthermore, in experiment 3 (chapter 5) the correct visual identification within these viseme groups was very high group 2 **/f/** at 99%, group 1 **/b, m/** at 99%, group 3 **/w/** at 99%, and group /**ʃ, dʒ**/ at 96%. On the other hand viseme group 13 **/q, x, ʁ, ħ, ʕ, h, ʔ/** had the least combination response percentage at 4%. This viseme group is the largest viseme group in Arabic therefore it is highly ambiguous compared to the others. These findings may be explained by the proposed framework, the viseme groups that are clearly distinguishable visually would lead to them having more weight or influence during auditory-visual integration of speech leading to a higher probability of a combination response.

Conversely those viseme groups which are ambiguous will have less weight during the auditory-visual integration process as can be seen in the above data.

However, viseme group 11 **/r/** had a low percentage of combination responses of 12%. This group has only one phoneme but in experiment 3 (chapter 5) visual **/r/** and **/l/** were visually similar (See chapter 5 section 5.4). For a combination response to occur the visual stimulus must contradict the visual mental representation of the auditory stimulus. These results are similar to the results found above for the influence on a visual response when the stimulus was auditory **/ba/**.

Figure 6.10 depicts the auditory-visual perceptual phonetic space when a combination response is made. In this case the visual stimulus is in a medium density visual perceptual neighbourhood (see chapter 2 section 2.6.2) and it is in contrast to the auditory stimulus. These are similar conditions to producing a McGurk response, however a combination response is perceived. The reason is not due to the visual stimulus, but the auditory stimulus. In the above data an auditory **/la/** produced the combination responses but not an auditory **/ba/**.

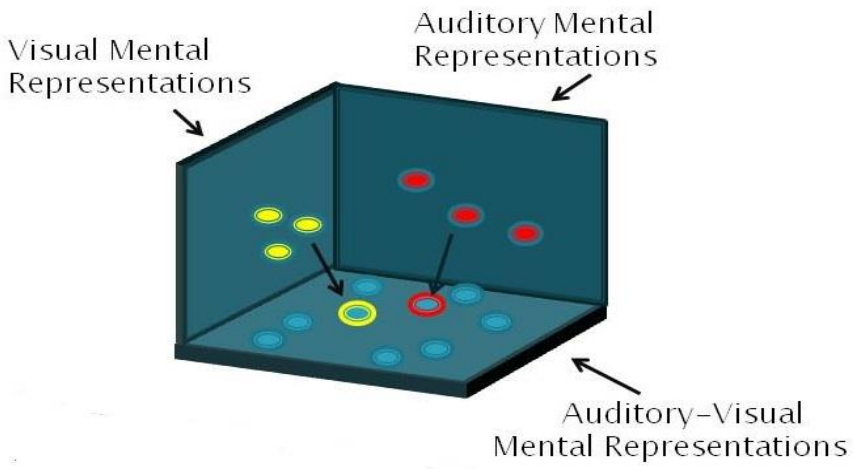

Figure 6.10 Auditory-visual phonetic perceptual space during combination response.

To explain these results the visual mental representations of the auditory stimulus must be evaluated. In the case of the phoneme **/b/** it is very visual distinct and therefore the listener can clearly perceive if it was presented or not. However, in the case of the phoneme **/l/** since it is an alveolar phoneme it is less visually distinct compared to the phoneme **/b/**. When the visual stimulus is a visually distinct phoneme like **/f/**, the listener perceives **/f/** however since the auditory stimulus is **/la/** the listener also perceives **/la/**. The auditory-visual phonetic perceptual space allow for the perception of **/l/** since it is not visually distinct, so in the perceptual process A **/la/** + V **/fa/** can be perceived as **/fla/**. The visual stimulus was always perceived first followed by the auditory stimulus. This can be explained by the processing speed of vision and hearing, it is well established in temporal experiments that there is a preference for visual speech stimuli preceding auditory speech stimuli (Grant et al., 2004, Munhall et al., 1996, Navarra et al., 2010, van Wassenhove et al., 2007). In the following section auditory responses will be reviewed.

6.5.4 Auditory Response

Although the auditory stimuli /**ba**/ and /**la**/ differed in the amount of influence the visual stimuli had during auditory-visual speech perception, the viseme group that produced the greatest proportion of auditory responses or the least visually influenced responses was the guttural viseme group for both the auditory /**ba**/ and /**la**/ stimuli. Massaro (2005) reported that visual cues provide information mostly for phonemes produced in the front of the mouth rather than the back of the mouth. The guttural viseme group includes 7 phonemes (viseme group 13 /**q, χ, ʁ, ħ, ʕ, h, ʔ**/).

This viseme group had the greatest number of phonemes compared to the other 12 viseme groups in Arabic; therefore it has the greatest amount of visual ambiguity. These results support the hypothesis that greater visual ambiguity will lead to a reduction in the predictive value of the visual cues during speech perception. The predictive value will influence the weight of these visual cues during the process of auditory-visual speech integration (see Figure 6.11). In this case the visual stimulus has a highly dense visual neighbourhood (large viseme group) with many invisible contrasts. Hence, the visual stimulus has little perceptual influence and the auditory stimulus is perceived.

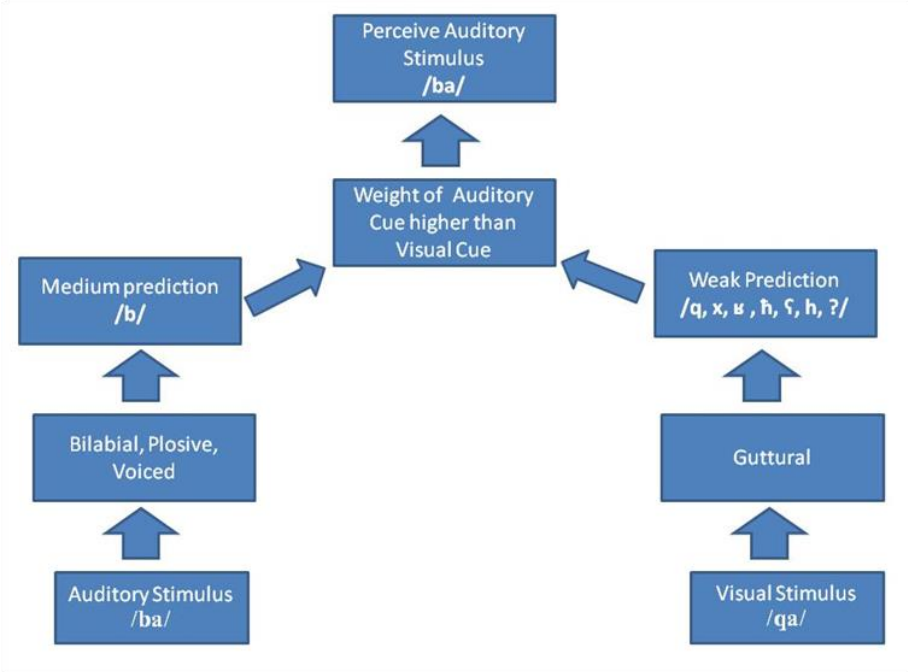

Figure 6.11 An example of the hypothesized auditory-visual native language framework for auditory-visual integration with an ambiguous visual speech cue /qa/.

Arnal (2009) found that the facilitation in processing of auditory signals appears to be directly a function of the predictability of visual cues(Arnal et al., 2009, Vroomen and Stekelenburg, 2010). Furthermore the results of van Wassenhove et al. (2005) support the findings of experiment four; they reported that the speed of processing auditory-visual speech is influenced by the ambiguity of visual speech cues. That is when the visual cue was unambiguous for example /**b**/ the speed of processing measured by the latency of N1/P2 was less compared to an ambiguous visual speech cue for example /**k**/ (see chapter 1 section 1.4.3.1). van Wassenhove et al. (2005) results also showed that speech processing is faster when the stimuli are auditory-visual compared to auditory only. They measured a reduction in N1/P2 latencies for native English listeners when the speech stimuli were auditory-visual (van Wassenhove et al., 2005). On the other hand, Hisanaga et al. (2009) only found a reduction for N1latencies but not for P2 latencies in native Japanese listeners (Hisanaga et al., 2009). This suggests that the influence or the weight of visual speech cues on auditory-visual speech perception for native Japanese listeners is less than that for native English listeners.

These studies are consistent with the current findings that due to the increased visual ambiguity of guttural phonemes in Arabic the information value of the mental representations of these visual cues for Arabic listeners might be less than for English listeners. Therefore the perceptual space of Arabic listeners for guttural phonemes might be tuned into the mental representations for auditory cues more than the mental representations for visual cues compared to English listeners. These results can be explained by the working framework which proposes that auditory-

visual integration of speech is modulated by visual and auditory mental representations of the native language and by the degree of the predictive value of visual cues within the native language.

Chapter 7

Discussion and Conclusions

7.1 Introduction

In this chapter, the results obtained from experiments reported in chapters 3, 4, 5 and 6 are discussed in relation to the research questions posed in chapter 1 and 2. The implications for the research in advancing our understanding of auditory-visual speech perception are also discussed. Finally, the limitations of the research are discussed and recommendations made regarding the direction of future research.

7.2 What is the cross-linguistic difference in the McGurk effect between Arabic and English Listeners?

To investigate this question, auditory-visual speech perception was examined using the McGurk technique for Arabic and English listeners using native and non-native stimuli. Previous work has suggested that the process of speech perception seems to rely on the prediction value between the auditory and visual mental representations within the native language (Hazan et al., 2006, Massaro et al., 1995). Depending on these prediction values between the two modalities a decision is made on the perceived speech. The native language mental representation which best matches the auditory-visual speech input is the one which is perceived.

An important finding in experiment 1 and 2 (chapter 3 and 4) was that the influence of visual speech information reflected by the McGurk response percentage was found to be significantly lower in the Arabic listeners in comparison to English

listeners. This was observed when the visual cues were velar /**k**/ and uvular /**q**/ which are produced in the back of the mouth. However when the visual speech cues were /**b**/ and /**p**/ (both bilabial phonemes), the Arabic listeners incorporated the visual cues in their perception of speech at the same percentage as the English listeners measured by the percentage of combination responses. This is new evidence for the influence of native visual cues in the perception of speech. This suggests that Arabic listeners rely less on visual speech cues compared to English listeners during the perception of speech when the visual speech cues have reduced saliency due to being produced in the rear of the oral cavity. An explanation for this cross-linguistic difference was suggested in experiment 1 (chapter 3), that for guttural phonemes the perceptual space of Arabic listeners might be tuned into the mental representations for auditory cues more than the mental representations for visual cues compared to English listeners.

These results are consistent with research evaluating native language differences in the use of visual cues during speech perception that found differences in the percentage of the McGurk effect between Chinese (Cantonese), Japanese and English listeners (Sekiyama, 1997, Sekiyama and Burnham, 2008, Sekiyama and Tohkura, 1993). Massaro et al. (1993) also found differences in the McGurk effect between Japanese, Spanish, and English listeners (Massaro et al., 1993). Similarly, Hazan et al. (2006) found that Japanese listeners relied less on visual cues compared to Spanish listeners. The explanation given by Sekiyama (1997) for the reduced reliance on visual cues in Chinese is because it relies on tones and for Japanese it uses pitch accents (see chapter 1 section 1.6). However, these reasons can not be applied to Arabic since it neither relies on tones nor pitch accents. Therefore there

must be a novel feature in Arabic which reduces the reliance on visual cues during speech perception.

The findings in Arabic noted above raised the question that one of the factors influencing the reliance on visual speech cues is the ambiguity of the visual speech gestures within the native language. To investigate the visual ambiguity of Arabic phonemes, the visemes in Arabic were identified in experiment 3 (chapter 5). The results showed that the guttural phonemes comprised the largest viseme category which included 7 phonemes (see chapter 5 section 5.4). It has also been reported in previous research (Damien, 2011, Damien et al., 2009) that the guttural phonemes in Arabic are visually ambiguous. Novel evidence was found in experiment 4 (chapter 6) which showed that the viseme category with the smallest percentage of McGurk response and combination response was the guttural viseme group **/q, χ, ʁ, ħ, ʕ, h, ʔ/**. These findings suggest that the mental representations of speech signals across the visual and auditory mental repertoire in Arabic native listeners for guttural phonemes is more tuned into auditory cues rather than visual cues. This seems to be due to the large number of guttural phonemes within the viseme group in Arabic.

These data could be explained by the Fuzzy Logical Model of Perception suggested by Massaro (1998) in terms of a perceptual bias which is dependent on the visual mental representation of the native language. The speech perception process seems to be flexible where there is a shift in weighting from auditory or visual input depending on the relevant information obtained from both modalities. Additionally the Neighbourhood Activation Model (Luce and Pisoni, 1998) implies that the perception of auditory-visual speech is reliant on the predictive power of the native

language visual cues and the density of the phonetic visual neighbourhood (see chapter 2 section 2.6.2). The results in experiment 2 and 4 (chapter 3 and 6) can be explained by a native language framework of auditory-visual speech perception. As the predictive power of a visual speech cue increases so does the visual weight it will incur during the auditory-visual integration process of speech (see chapter 2 section 2.7). This is new evidence to explain the cross-linguistic differences in the percentage of McGurk response between Arabic and English listeners.

7.3 What are some cross-linguistic differences in visual speech cues between Arabic and English Listeners?

It was hypothesized in chapter 3 that experience with native visual cues fundamentally alters auditory-visual speech perception as measured by the McGurk effect. This was evidenced by different McGurk responses found for native Arabic and English listeners (see chapter 3 section 3.5.3) for the same speech stimuli. This can be explained by the ideas represented in a working framework which predicts, based on Kuhl et al. (2006), that there are different visual speech mental representations in Arabic and English this leads to different McGurk responses (see chapter 2 section 2.7). For example when the visual stimulus was **/qa/** and the auditory stimulus was **/ba/** the majority of McGurk responses for the Arabic listeners was **/tʿa/.** These results add support to the concept that the phoneme **/qa/** is an emphatic phoneme (Watson, 2002, Heselwood, 1992). The Arabic listeners were able to recognize the emphatic visual cues and chose a fusion response that is also an emphatic sound. However, since English listeners do not have emphatic phonemes

within their native language repertoire they were not able to pick up on the emphatic visual cues and their McGurk response was always /**da**/ a non-emphatic phoneme. This is new evidence of the influence of emphatic visual cues in Arabic on auditory-visual speech perception. This suggests that for Arabic listeners there are mental representations for visual emphatic cues based on their native language which English listeners do not have.

The results of experiment 3 and 4 (chapter 5 and 6) also confirm that Arabic listeners can visually distinguish emphatic phonemes from non-emphatic counterparts. Ouni and Ouni (2009) also found in a group of ten participants that emphatic phonemes in Arabic can be visually distinguished from non-emphatic phonemes. However, Damien (2009) investigated the visual cues for the emphatic Arabic consonants. He did not find them to be visually distinct from their non-emphatic counterpart. Damien evaluated the visual cues by using an algorithm computer analysis of lip movement. This suggests that the distinguishing visual cue for emphatic phonemes is not the movement of the lips, it could be the lowering of the jaw. The results of experiment 3 (chapter 5) provide some support for this notion, for example it was found that when the place of articulation allows a greater visual component in identifying the emphatic movement then the participants could distinguish an emphatic phoneme from the non-emphatic counterpart. A distinct visual feature for emphatic phonemes seems to be sulcalisation of the tongue and the lowering of the jaw which assists in increasing the size of the oral cavity compared to the non-emphatic counterpart. For example in the alveolar place of articulation the movement of the jaw is more visually accessible and clear therefore the emphatic /**tˤ, dˤ,sˤ**/ and non-emphatic /**t, d,s**/ phonemes were distinguished from one another. Also there

was a distinction between the non-emphatic /**k**/ and the emphatic /**q**/. Although the /**q**/ is a uvular phoneme the cue for an emphatic phoneme seems to be the lowering of the jaw so this can still be observed even for a guttural phoneme. These results show that emphatic visual cues are used in speech perception by Arabic listeners however these visual cues are not available to listeners who do not speak Arabic.

Another example of the importance of visual native language mental representations is that English listeners in experiment 1 (chapter 3) assimilated Arabic visual /**q**/ to the /**g**/ being a close visual cue within their language. It was also found that Arabic listeners assimilated the auditory cue /**p**/ which is not present in Arabic to the auditory /**b**/ (see chapter 3 section 3.5.2). The findings in experiment 1 and 2 (chapter 3 and 4) support the notion that assimilation occurs for the visual characteristics of internal mental representations as well as auditory ones. That is when English listeners were exposed to a non-native visual category, in this case emphatic, they categorized the non-native visual cues to the closest existing visual speech category based on their visual mental representations of their native language and auditory-visual integration still occurred. Assimilation for both auditory and visual speech cues has been accounted for within this framework and it seems to occur for non-native visual speech cues in the same way as assimilation occurs for non-native auditory cues.

Evidence in the literature consistent with this idea was reported by Werker and Tees (1992) who found that for the McGurk effect the stimuli A/**ba**/ + V/**ða**/ produced a /**ða**/ response for English listeners but not for French listeners (Werker et al., 1992). The phoneme /**ð**/ is not used in French therefore they substituted /**da**/ for /**ða**/. Burnham and Keane (1997) also found that Japanese participants substituted /**da**/ for

/**ða**/ because /**ða**/ is not present in Japanese . In both studies the French and Japanese listeners' /**ða**/ response increased as a function of experience with the English language.

In summary the results from experiment 1 and 2 (chapter 3 and 4) are consistent with the notion that the auditory-visual integration process accesses phonemic mental representations of sounds, the form and relative weighting of which are dependent on the phonological features of the native language. Due to a different repertoire of visual and auditory cues in Arabic compared to English the net result of auditory-visual integration produces a different response in Arabic listeners compared to English listeners.

7.4 Can bottom-up visual processing speed explain the difference found in McGurk response percentage between Arabic and English listeners?

An experimental method used in experiment 2 (chapter 4) to evaluate the relationship between auditory and visual cues was measuring and manipulating the temporal synchrony between the auditory-visual stimuli. Consistent with the literature it was found that the highest percentage of McGurk effect occurred when the visual speech input led the auditory speech input (Conrey and Pisoni, 2006, van Wassenhove et al., 2007). However, the percentage of the McGurk effect at optimal visual lead time was still significantly greater for English listeners compared to Arabic listeners. Hence visual processing speed did not account for the differences in auditory-visual integration found between native Arabic and English listeners. This

suggests that the difference between Arabic and English listeners in auditory-visual integration of speech is not due to differences in visual processing speed. Also the results of this experiment suggest that integration does not occur at a pre-phonetic stage as processing speed did not explain the cross-linguistic differences in the McGurk response percentage. These results are consistent with findings from other chapters that can be explained by a native language framework of auditory-visual speech perception.

7.5 Does predictive power of native visual speech cues affect the percentage of auditory-visual integration of speech?

It is proposed that the difference between the Arabic and English listeners in their use of visual speech cues might be due to the density of the phonetic visual neighbourhood within the native language (see chapter 2 section 2.6.2). This can be explained by a framework of auditory-visual speech perception described in chapter 2 section 2.7. That is to say that the use of visual speech cues during speech perception is dependent on how useful these visual cues are in disambiguating close phonetic visual neighbours within the native language. Therefore, if the visual speech cues are unambiguous this will yield strong auditory-visual integration. However, when visual speech cues are ambiguous more weight or reliance is then focused on the auditory domain during the speech perception process. Hence phonemes that have a large viseme group, in other words many invisible phonetic contrasts, are less effective for speech discrimination and would not be expected to have a large influence on auditory-visual speech perception.

To examine this question it was first necessary to quantify Arabic visemes. Visemes are visually based categories of contrast, similar to phonemes in the auditory modality. Many studies of English visemes have examined the visual discrimination of consonants. However no investigation has been conducted on viseme groups covering the entire range of the 29 consonants used in Arabic. In experiment 3 (chapter 5) the confusability of all Arabic consonants grouped into their viseme classes was determined. The rationale behind this approach is that by establishing the interclass confusion for a group of phonemes in their viseme class, a better understanding can be obtained of the complementary nature of the separate auditory and visual information sources and this can be subsequently applied to understand the fusion stage of auditory-visual speech perception, explained in chapter 6.

This enabled a comparison of the viseme categories of Arabic with the published viseme groups of English. A distinct feature of Arabic is the presence of emphatic and guttural phonemes (Elgendy and Pols, 2001). Guttural sounds are produced in the back of the mouth; consequently visual cues are not very beneficial for identifying guttural phonemes. Also in Arabic, there are four emphatic phonemes; they are /**tˤ**/, /**dˤ** /, /**sˤ**/and /**ðˤ**/. Emphatic consonants are pronounced in such a manner that the back of the tongue retracts into the pharynx (see chapter 1 section 1.7). For example, /**d** / and /**dˤ**/ are both voiced, alveolar, stop consonants, but /**dˤ**/ is an emphatic sound. The visual similarity between plain and emphatic phonemes might lead to an increase in visual ambiguity of speech sounds in Arabic. Visual cues add to the auditory information received during speech perception. If the visual cues of one language are less reliable compared to another language this may lead to an auditory-visual integration process that relies less on the visual cues. The results

permitted an evaluation of whether Arabic has more ambiguous visual cues, which might then lead to a difference in predictive coding of the visual cues across the native language. The results showed that Arabic consonants were grouped into 13 viseme categories. The viseme category with the greatest number of phonemes was the guttural phonemes. This viseme category has high visual ambiguity since it has many phonemes in the same viseme category. This is new evidence which confirms that Arabic has more visually ambiguous phonemes compared to English.

In experiment 4 (chapter 6) the influence of visual cues across the 13 viseme categories in Arabic was investigated. It was found that predictive power of visual cues is a factor which influences the degree in which visual cues are integrated into the process of speech perception as shown by the McGurk results. For example the **/qa/** visual cue had less visual influence than **/fa/**. In experiment 3 (chapter 5) it was shown that the /**f**/ phoneme was in a viseme category on its own; however /**q**/ was in a viseme category with six other phonemes. It would then not be beneficial to use visual speech information when many of the phonemes look the same in the visual domain. As a result, these visually ambiguous phonemes would provide very little useful information about identification of those phonemes. This suggests that visual bias is influenced by the lack of visible phonetic contrasts within the native language.

Support for this finding comes from developmental studies in Arabic which have shown that emphatic and guttural phonemes are acquired at a later age during speech development (Amayreh, 2003, Amayreh and Dyson, 1998). Since speech is auditory and visual it would be expected that phonemes with a greater number of cues would be more readily accessible to the child during speech development. Conversely

visually ambiguous phonemes do not have many visual cues therefore they develop at a later age compared to visually salient phonemes. Experiment 3 (chapter 5) determined the viseme groups of Arabic consonants and that emphatic and guttural phonemes lead to an increase in the number of phonemes within some viseme groups. Experiment 4 (chapter 6) determined that guttural phonemes in Arabic have a decreased influence on speech perception which would help to explain why these phonemes would have a later age of acquisition compared to the visually salient phonemes in Arabic.

The findings noted above help to explain the results in experiment 1 and 2 (chapter 3 and 4) where Arabic listeners had a decrease in auditory-visual integration as measured by the percentage of the McGurk effect compared to English listeners. The results in experiment 4 (chapter 6) suggest that the increase in invisible phonetic contrasts among the guttural phonemes in Arabic led to a decrease in visual bias elicited by these phonemes. This decrease in auditory-visual integration suggests a shift in weighting between Arabic and English listeners; where for Arabic listeners less weight is put on visual cues within the guttural viseme group due to an decrease in the predictive power of visual speech cues within the native language as compared to English listeners. Therefore, the density of phonetic visual neighbourhood within the native language influences the weight given to visual cues during the process of speech perception.

A number of studies suggest that ambiguity of visual cues affects speech perception negatively (Brunellière et al., 2013, Nielsen, 2002, Huyse et al., 2013, Kawase et al., 2014). Brunellière et al (2013) showed that the greater the predictive power of visual

cues the greater the influence on auditory-visual speech processing. Nielsen (2004) showed that although visual information improves speech perception, the amount of visual contribution to the perception of speech was unequal between consonants. Huyse et al (2013) found that the auditory modality influence on the process of speech perception increases when visual information is degraded (see chapter 2 section 2.6.2).

The results within this thesis provide evidence for the importance of ambiguity of visual cues dependent on native language. Experiment 3 and 4 (chapters 5 and 6) suggest that auditory-visual integration of speech perception relies on the predictive power of visual phonetic mental representations of sounds, which are dependent on the visible contrasts within the native language.

7.6 Relevance within the literature

The thesis contributes to our understanding of auditory-visual speech processing in Arabic. An important finding was the reduced percentage of McGurk effect in Arabic listeners compared to English listeners for guttural phonemes. This is suggested to be due to a larger proportion of invisible phonetic contrasts in Arabic compared to English, which leads to a decrease in importance of visual information during speech perception for Arabic listeners compared to English listeners.

Auditory-visual integration in Arabic appears to depend on the characteristics of the language where visual dimensions augment the development of mental representations of speech sounds. In other words, native Arabic listeners' speech perceptual space will be tuned for the regularities of Arabic visual speech cues.

Novel findings were seen in experiment 1 (chapter 3) where the Arabic listeners incorporated the visual stimulus's emphatic cues in their auditory-visual response and reported hearing an emphatic phoneme. This was also supported by the findings in chapter 5 where in the visual only condition emphatic phonemes where visually distinguished from their non-emphatic counterparts for alveolar and velar/guttural place of articulation. Additionally, in experiment 1 (chapter 3) assimilation of non-native visual and auditory speech stimuli occurred during the McGurk effect.

In experiment 4 (chapter 6) it was shown that the perception of auditory-visual speech is a complex process that is dependent on an interaction between the auditory and visual speech cues. Based on the suggested framework the mental representation matching process is proposed to be dependent on the native language auditory and visual mental representations. If there is a clear match between the auditory and visual speech input then the matching native language mental representation will be robust (Peelle and Sommers, 2015). However, when the input between the visual and auditory speech cues do not match, as in the case of the McGurk effect, then the speech input which has greater predictive power will dominate the perception process. The matching mental representation will be perceived regardless of modality. This weighting process for the auditory and visual speech input is dependent on the auditory and visual speech mental representations within the native language.

Additionally, an important finding was that visemes had a variable influence that was not only dependent on predictive power of the visual cues. For there to be a McGurk effect there needs to be a clear conflict between the visual mental

representations of the stimuli presented in the visual modality compared to the stimuli presented in the auditory modality. This suggests that the weighting system is more complex than initially supposed and there seems to be a more complex analysis of the predictive power of the information received from both modalities (chapter 6 section 6.6). These findings can be explained by the framework which suggests that the speech perception process seems to analyse the auditory and visual input and compare it to the visual and auditory native language mental representations. In other words a hypothesis is formed on what was most likely said based on the auditory and visual native language mental representations. By integrating speech information by this weighting framework, the predictive power of the perceived speech signal is increased. The weighting framework yields the most reliable speech estimate possible. The findings in experiment 4 (chapter 6) support speech perception as a flexible system that reflects a complicated interplay of both auditory and visual native language mental representations (MacDonald et al., 2000, Massaro, 1998, Kuhl et al., 2008, Kuhl et al., 2006).

New evidence in this thesis has shown cross-linguistic differences on the influence of visual speech cues during auditory-visual integration of speech between Arabic and English native speakers. Additionally, in experiment 3 (chapter 5) it was found that the Arabic consonants can be categorized into 13 viseme groups. The largest viseme group in Arabic was found to be for the guttural phonemes **/q, χ, ʁ, ħ, ʕ, h, ʔ/**. Also the results of this thesis support the finding that the alveolar emphatic phonemes are visually distinguished from their non-emphatic counterpart (Ouni and Ouni, 2007). Furthermore, the /**q/** phoneme was found to be visually identified as an emphatic phoneme supporting the concept that it should be categorized as the

emphatic form of /**k**/ (Watson, 2002, Heselwood, 1992). Finally, it can be concluded that this thesis provides support for the notion that auditory-visual integration of spoken Arabic is conditioned by the consequent level of predictive power of the visual cues characteristic in Arabic and the assessment and comparison of the mental representations between the auditory and visual speech cues.

7.7 Future research

Based on the findings of this thesis, suggested future research includes investigating the development of auditory-visual integration during the first-language acquisition of Arabic. This will allow a more systemic investigation of the role of visual cues in the development of speech, thus allowing a more detailed evaluation of whether the later development of emphatic and guttural phonemes is due to the increase in visual ambiguity of these phonemes.

One important finding was that emphatic visual cues had a reduced visual bias compared to their non-emphatic counterparts. Therefore a more detailed investigation of the influence of emphatic phonemes on auditory-visual speech perception can be performed. An experiment can be conducted which uses emphatic auditory cues to evaluate whether emphatic visual cues have a similar visual influence to their non-emphatic counterpart. It would also be beneficial to investigate auditory-visual integration in an Arabic dialect that produces the /**dˤ** / phoneme in order to investigate if there are differences in the auditory-visual integration compared to those found in this thesis for Saudi native speakers.

References

ABBOUD, H. A., HELLER, K., SCHULTZ, H. & ZEITLIN, V. 2010. SuperLab 4.5. San Pedro, CA: Cedrus Corporation.

ABRAMOV, L., GORDON, J., HENDRICKSON, A., HAINLINE, L., DOBSON, V. & LABOSSIERE, E. 1982. The retina of the newborn human infant. *Science,* 217**,** 265-267.

ABRY, C. & BOË, L. 1986. Laws for lips. *Speech Communication,* 5**,** 97-104.

ADOBE, S. I. 2010. Adobe Premiere Elements 9. San Jose.

AL-ANI, S. H. 1970. *Arabic phonology: An acoustical and physiological investigation.* Hague, Netherlands: Mouton.

AL-RABA'A, B. I. 2015. The Manner of Articulation of the emphatic /dˤ/in both Saudi and Palestinian dialects. *International Journal of Language and Linguistics,* 3**,** 1-7.

ALHAMMAD, R. 2014. *Emphasis spread in najdi Arabic.* M.A. thesis., California State University.

ALI, A. N. 2007. Exploring semantic cueing effects using McGurk fusion. *Audio-Visual Speech Processing, 31 August,* Hilvarenbeek, Netherlands.

ALI, A. N., HASSAN-HAJ, A., INGLEBY, M. & IDRISSI, A. 2005. McGurk fusion effects in Arabic words. *Auditory-Visual Speech Processing, 24 July,* British Columbia, Canada.

ALOTHMAN, N. 2009. *Classification of Visemes using Visual Cues.* Ph.D. thesis, University of Pittsburgh.

ALTIERI, N. 2014. Multisensory integration, learning, and the predictive coding hypothesis. *Frontiers in Psychology,* 5**,** 257-261.

ALTIERI, N., PISONI, D. & TOWNSEND, J. 2011. Some behavioral and neurobiological constraints on theories of audiovisual speech integration: A review and suggestions for new directions. *Seeing and Perceiving,* 24**,** 513-517.

AMAYREH, M. M. 2003. Completion of the consonant inventory of Arabic. *Journal of Speech, Language and Hearing Research,* 46**,** 517-529.

AMAYREH, M. M. & DYSON, A. T. 1998. The acquisition of Arabic consonants. *Journal of Speech and Hearing Research,* 41**,** 642-653.

ARNAL, L. H., MORILLON, B., KELL, C. A. & GIRAUD, A. L. 2009. Dual neural routing of visual facilitation in speech processing. *Journal of Neuroscience,* 29**,** 13445-13453.

ARNOLD, P. & HILL, F. 2001. Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology,* 92**,** 339-355.

ASCHENBERNER, B. & WEISS, C. 2005. *Phoneme-viseme mapping for German audio-visual speech synthesis. Ph.D. thesis,* Universität Bonn.

AUER, E. T. 2002. The influence of the lexicon on speech read word recognition: Contrasting segmental and lexical distinctiveness. *Psychonomic Bulletin Review,* 9**,** 341-347.

BAART, M. & VROOMEN, J. 2010. Do you see what you are hearing? Cross-modal effects of speech sounds on lipreading. *Neuroscience Letters,* 471**,** 100-103.

BANKS, M. S. & SALAPATEK, P. 1981. Infant pattern vision: A new approach based on the contrast sensitivity function. *Journal of Experimental Child Psychology,* 31**,** 1-45.

BARNARD, M., HOLDEN, E. & OWENS, R. 2002. Lip tracking using pattern matching snakes. *The 5th Asian Conference on Computer Vision, 10 January,* Melbourne, Australia.

BARUTCHUA, A., CREWTHERA, S., KIELYA, P., MURPHYA, M. & CREWTHERAB, D. 2008. When /b/ill with /g/ill becomes /d/ill: Evidence for a lexical effect in audiovisual speech perception. *European Journal of Cognitive Psychology,* 20**,** 1-11.

BAVELIER, D., BROZINSKY, C., TOMANN, A., MITCHELL, T., NEVILLE, H. & LIU, G. 2001. Impact of early deafness and early exposure to sign language on the cerebral organization for motion processing. *Journal of Neuroscience,* 21**,** 8931-8942.

BERNSTEIN, L., AUER, E. & J., M. 2004a. Modality specific perception of auditory and visual speech. *In:* G. A CALVERT, C. S., B. E STEIN. ed. *The Handbook of Multisensory Processing.* Cambridge, MA: MIT Press**,** 203-223.

BERNSTEIN, L. E., AUER, E. T. & TAKAYANAGI, S. 2004b. Auditory speech detection in noise enhanced by lipreading. *Speech Communication,* 44**,** 5-18.

BESLE, J., FISCHER, C., BIDET-CAULET, A., LECAIGNARD, F., BERTRAND, O. & GIARD, M. H. 2008. Visual activation and audiovisual interactions in the auditory cortex during speech perception: Intracranial recordings in humans. *Journal of Neuroscience,* 28**,** 143-148.

BEST, C. T. 1991. Phonetic influences on the perception of nonnative speech contrasts by 6–8 and 10–12 month olds. *Society for Research in Child Development, 16 February,* Seattle, WA.

BEST, C. T. 1994. The emergence of native-language phonological influences in infants: A perceptual assimilation model. *In:* GOODMAN, J. C. & NUSBAUM, H. C. eds. *The development of speech perception: The transition from speech sounds to spoken words.* Cambridge, MA: MIT Press**,** 167-224.

BEST, C. T., MCROBERTS, G. W. & GOODELL, E. 2001. Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *Journal of the Acoustical Society of America,* 109**,** 775-794.

BEST, C. T., MCROBERTS, G. W. & SITHOLE, N. M. 1988. Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance,* 14**,** 345-360.

BEST, C. T. & STRANGE, W. 1992. Effects of phonological and phonetic factors on cross-language perception on approximants. *Journal of Phonetics,* 20**,** 305-330.

BIMHOLZ, J. C. & BENACERAFF, B. B. 1983. The development of the fetal human hearing. *Science,* 222**,** 516-518.

BOVO, R., CIORBA, A., PROSSER, S. & MARTINI, A. 2009. The McGurk phenomenon in Italian listeners. *Acta Otorhinolaryngologica Italica,* 29**,** 203-208.

BOZKURT, EROGLU, ERZIN, ERDEM & OZKAN 2007. Comparison of phoneme and viseme based acoustic units for speech driven realistic lip animation. *Signal Processing and Communications Application, 11 June,*. Eskisehir, Turkey.

BRAIDA, L. D. 1991. Crossmodal integration in the identification of consonant segments. *The Quarterly Journal of Experimental Psychology,* 43**,** 647-677.

BRANCAZIO, L. 2004. Lexical influences in audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance,* 30**,** 445-463.

BRUNELLIÈRE, A., SÁNCHEZ-GARCÍA, C., IKUMI, N. & SOTO-FARACO, S. 2013. Visual information constrains early and late stages of spoken-word recognition in sentence context. *International Journal of Psychophysiology,* 89**,** 136-147.

BUBIC, A., VON CRAMON, D. Y. & SCHUBOTZ, R. I. 2010. Prediction, cognition and the brain. *Frontiers in Human Neuroscience,* 22**,** 4-25.

BUCHWALD, A., WINTERS, S. & PISONI, D. 2009. Visual speech primes open-set recognition of spoken words. *Language and Cognitive Processes,* 24**,** 580-610.

BURNHAM, D. 1993. Visual recognition of mother by young infants: Facilitation by speech. *Perception,* 22**,** 1133-1153.

BURNHAM, D. 1998. Language specificity in the development of auditory–visual speech perception. *In:* CAMPBELL, R., DODD, B. & BURNHAM, D. eds. *Hearing by eye II: The psychology of speechreading.* London: Psychology Press**,** 27-60.

BURNHAM, D. & DODD, B. 1996. Auditory–visual speech perception as a direct process: The McGurk effect in human infants and across languages. *In:* STORK, D. G. & HENNECKE, M. E. eds. *Speechreading by humans and machines.* Berlin: Springer-Verlag**,** 103-114.

BURNHAM, D. & DODD, B. 2004. Auditory-visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology,* 45**,** 204-220.

BURNHAM, D. & KEANE, S. 1997. The Japanese McGurk effect: The role of linguistic and cultural factors an auditory-visual speech perception. *Audio-Visual Speech Processing, 59,* 93-96.

CALVERT, G. A. 2001. Crossmodel processing in the human brain: Insights from functional neuroimaging studies. *Cerebral Cortex,* 11**,** 1110-1123.

CALVERT, G. A., BULLMORE, E. T., BRAMMER, M. J., CAMPBELL, R., WILLIAMS, S. C., MCGUIRE, P. K., WOODRUFF, P. W., IVERSEN, S. D. & DAVID, A. S. 1997. Activation of auditory cortex during silent lipreading. *Science,* 276**,** 593-596.

CAMPBELL, R. 2008. The processing of audio-visual speech: Empirical and neural bases. *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences,* 363**,** 1001-1010.

CAMPBELL, R. & DODD, B. 1984. Aspects of hearing by eye. *In:* H BOUMA, D. B. ed. *Attention and performance, x, control of language processes.* Hillsdale, NJ: Lawrence Erlbaum Associates**,** 299-314.

CARLYON, R. P., CUSACK, R., FOXTON, J. M. & ROBERTSON, I. H. 2001. Effects of attention and unilateral neglect on auditory stream segregation. *Journal of Experimental Psychology,* 27**,** 115-127.

CHAN, K. Y. & VITEVITCH, M. S. 2009. The influence of the phonological neighborhood clustering coefficient on spoken word recognition. *Journal of Experimental Psychology,* 35**,** 1934-1949.

CHANDRASEKARAN, C., TRUBANOVA, A., STILLITTANO, S., CAPLIER, A. & GHAZANFAR, A. A. 2009. The natural statistics of audiovisual speech. *PLoS Computational Biology,* 5, 21-25.

CHEN, M. Y. 2000. *Tone Sandhi: Patterns across Chinese dialects.* Cambridge: Cambridge University Press.

CHEN, T. 1998. Audio-visual integration in multimodal communication. *Proceedings of the IEEE, 80, 25-33.*

CHEN, T. 2001. Audiovisual speech processing. *IEEE Signal Processing,* 18**,** 9-21.

CHEN, T. & RAO, R. R. 1998. Audio-visual integration in multimodal communication. *Proceedings of the IEEE* 86**,** 837-852.

COHEN, J. 1988. *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Earlbaum Associates.

CONREY, B. & PISONI, D. B. 2006. Auditory-visual speech perception and synchrony detection for speech and nonspeech signals. *Journal of the Acoustical Society of America,* 119**,** 4065-4073.

DAMIEN, P. 2011. Visual speech recognition of modern classic Arabic language. *International Symposium on Humanities, Science and Engineering Research,* 11 April, Lebanon.

DAMIEN, P., WAKIM, N. & EGÉA, M. 2009. Phoneme-viseme mapping for modern, classical Arabic language. *Advances in Computational Tools for Engineering Application, 25 May,* Lebanon.

DAVIS, C. & KIM, J. 2004. Audio-visual interactions with intact clearly audible speech. *Quarterly Journal of Experimental Psychology,* 57**,** 1103-1121.

DE GELDER, B., BERTELSON, P., VROOMEN, J. & CHEN, H. 1995. Inter-language differences in the McGurk effects for Dutch and Cantonese listeners. *Proceedings of the Fourth European Conference on Speech Communication and Technoloy, 27 October,* Madrid, Spain.

DEKLE, D. J., FOWLER, C. A. & FUNNELL, M. G. 1992. Audiovisual integration in perception of real words. *Perception & Psychophysics,* 51**,** 355-362.

DESJARDINS, R. N. & WERKER, J. F. 2004. Is the integration of heard and seen speech mandatory for infants? *Developmental Psychobiology,* 45**,** 187-203.

DODD, B. 1979. Lip reading in infants: attention to speech presented in- and out-of-synchrony. *Cognitive Psychology,* 4**,** 478-484.

DODD, B. & BURNHAM, D. K. 1988. Processing speechread information. *The Volta Review: New Reflections on Speechreading,* 90**,** 45-60.

DODD, B., MCINTOSH, B., ERDENER, D. & BURNHAM, D. 2008. Perception of the auditory-visual illusion in speech perception by children with phonological disorders. *Clinical Linguistics & Phonetics,* 22**,** 69-82.

ELGENDY, A. M. & POLS, L. C. W. 2001. Mechanical versus peceptual constraints as determinants of articulatory strategy. *Institute of Phonetic Sciences, 16 July,* Amsterdam, Netherlands.

ENGSTRÖM, C. 2003. Articulatory Analysis of Swedish Visemes. *Speech, music and hearing,* 40**,** 231-240.

ENNS, J. T. & LLERAS, A. 2008. What's next? New evidence for prediction in human vision. *Trends in Cognitive Science,* 9**,** 327-333.

ERICKSON, L., ZIELINSKI, B., ZIELINSKI, J., LIU, G., TURKELTAUB, P., LEAVER, A. & RAUSCHECKER, J. 2014. Distinct cortical locations for integration of audiovisual speech and the McGurk effect. *Frontiers in Psychology,* 5**,** 534- 538.

FAVA, E., HULL, R. & BORTFELD, H. 2014. Dissociating cortical activity during processing of native and non-native audiovisual speech from early to late infancy. *Brain Sciences,* 4**,** 471-487.

FEGHALI, E. 1997. Arab cultural communication patterns. *International Journal of Intercultural Relations,* 21**,** 345-378.

FELD, J. & SOMMERS, M. 2011. There goes the neighborhood: Lipreading and the structure of the mental lexicon. *Speech Communication,* 53**,** 220-228.

FISHER, C. G. 1968. Confusions among visually perceived consonants. *Journal of Speech and Hearing Research,* 11**,** 781- 796.

FLOM, R. & BAHRICK, L. E. 2007. The development of infant discrimination of affect in multimodal and unimodal stimulation: The role of intersensory redundancy. *Developmental Psychology,* 43**,** 238-252.

FOWLER, C. A. & SMITH, M. 1986. *Speech perception as "vector analysis": An approach to the problems of segmentation and invariance.* Hillsdale, NJ: Erlbraum.

FUJISAKI, W., SHIMOJO, S., KASHINO, M. & NISHIDA, S. 2004. Recalibration of audiovisual simultaneity. *Nature Neuroscience* 7**,** 773-778.

GERVAIN, J. & MEHLER, J. 2010. Speech perception and language acquisition in the first year of life. *Annual Review of Psychology,* 61**,** 191-218.

GIEGERICH, H. 1992. *English phonology: An introduction.* Cambridge: Cambridge University Press.

GOLDSCHEN, A. J., GARCIA, O. N. & PETAJAN, E. 1994. Continuous optical automatic speech recognition by lipreading. *Systems and Computers,* 1**,** 572-577.

GOLDSTEIN, R. & VITEVITCH, M. S. 2014. The influence of clustering coefficient on word-learning: How groups of similar sounding words facilitate acquisition. *Frontiers in Psychology,* 18**,** 1307- 1309.

GRANT, K. 2002. Measures of auditory-visual integration for speech understanding: A theoretical perspective. *Journal of the Acoustical Society of America,* 112**,** 30-33.

GRANT, K. & SEITZ, P. 1998. Measures of auditory-visual integration in nonsense syllables and sentences. *Journal of the Acoustical Society of America,* 104**,** 2438-2450.

GRANT, K., WALDEN, B. & SEITZ, P. 1998. Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *Journal of the Acoustical Society of America,* 103**,** 2677-2690.

GRANT, K. W., GREENBERG, S., POEPPEL, D. & WASSENHOVE, V. 2004. Effects of spectro-temporal asynchrony in auditory and auditory-visual speech processing. *Seminars in Hearing,* 25**,** 241-255.

GREEN, K. P. & MILLER, J. L. 1985. On the role of visual rate information in phonetic perception. *Perception and Psychophysics,* 38**,** 269-276.

GREEN, P., STEVENS, B., KUHL, K. & MELTZOFF, M. 1990. Exploring the basis of the "McGurk effect": Can perceivers combine information from a

female face and a male voice. *Journal of the Acoustical Society of America,* 87**,** 125- 133.

HARLEY, T. A. 2009. *Psychology of language, from data to theory.* New York: Psychology Press.

HARNSBERGER, J. 2001. On the relationship between identification and discrimination of non-native nasal consonants. *Journal of the Acoustical Society of America,* 110**,** 489-503.

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. *Journal of the Acoustical Society of America,* 119**,** 1740-1751.

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**,** 360-378.

HERTRICH, I., MATHIAK, K., LUTZENBERGER, W., MENNING, H. & ACKERMANN, H. 2007. Sequential audiovisual interactions during speech perception: A whole-head MEG study. *Neuropsychologia,* 45**,** 1342-1354.

HESELWOOD, B. 1992. *Extended axiomatic-functionalist phonology: An exposition with application to modern standard Arabic.* Ph.D. thesis, University of Ulster.

HESELWOOD, B. 2013. *Phonetic transcription in theory and practice* Edinburgh. Edinburg, UK: University Press.

HESELWOOD, B. & AL-TAMIMI, F. 2011. A study of the laryngeal and pharyngeal consonants in Jordanian Arabic using nasoendoscopy, videofluoroscopy and spectrography. *In:* HASSAN, Z. M. & HESELWOOD, B. eds. *Instrumental studies in Arabic phonetics.* Amsterdam: John Benjamins**,** 163-192.

HISANAGA, S., SEKIYAMA, K., IGASAKI, T. & MURAYAMA, N. 2009. Audiovisual speech perception in Japanese and English: Inter-language differences examined by event-related potentials. *Auditory-Visual Speech Processing Workshop, 14 August,* Norwich.

HOCKLEY, N. S. & POLKA, L. 1994. A developmental study of audiovisual speech perception using the McGurk Paradigm. Poster presented at the *12th Meeting of the Acoustical Society of America, 10 January,* Austin, Texas.

HOLDEN, E. J. & OWENS, R. 2000. Visual speech recognition using cepstral images. *Signal and Image Processing,* 5**,** 331-336.

HUYSE, A., BERTHOMMIER, F. & LEYBAERT, J. 2013. Degradation of labial information modifies audiovisual speech perception in cochlear-implanted children. *Ear and Hearing,* 34**,** 110-121.

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

JERGER, S., DAMIAN, M., TYE-MURRAY, N. & ABDI, H. 2014. Children use visual speech to compensate for non-intact auditory speech. *Journal of Experimental Child Psychology,* 126**,** 295-312.

JERGER, S., DAMIAN, M. F., SPENCE, M. J., TYE-MURRAY, N. & ABDI, H. 2009. Developmental shifts in children's sensitivity to visual speech: A new multimodal picture-word task. *Journal of Experimental Child Psychology,* 102**,** 40-59.

JIANG, J., ALWAN, A., KEATING, P. A., AUER, E. T. J. & BERNSTEIN, L. E. 2002. On the relationship between face movements, tongue movements and speech acoustics. *Journal on Applied Signal Processing,* 11**,** 1174–1188.

JIANG, J. & BERNSTEIN, L. E. 2011. Psychophysics of the McGurk and other audio-visual speech integration effects. *Journal of Experimental Psychology,* 37**,** 1193-1209.

JOHNSON, K. 2006. Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics* 34**,** 485-499.

JUSCZYK, P. W. 1997. *The discovery of spoken language.* New York: MIT Press.

JUSCZYK, P. W., CUTLER, A. & REDANZ, N. J. 1993. Infants' preference for the predominant stress patterns of English words. *Child Development,* 64**,** 675-687.

KAURAMÄKI, J., JÄÄSKELÄINEN, I. P., HARI, R., MÖTTÖNEN, R., RAUSCHECKER, J. P. & SAMS, M. 2010. Lipreading and covert speech production similarly modulate human auditory-cortex responses to pure tones. *Journal of Neuroscience,* 30**,** 1314-1321.

KAWASE, S., HANNAH, B. & WANG, Y. 2014. The influence of visual speech information on the intelligibility of English consonants produced by non-native speakers. *Journal of the Acoustical Society of America,* 136**,** 1352-1362.

KHALIL, I. 2013. *Contemporary challenges and solutions for mobile and multimedia technologies.* Pennsylvania: IGI Global.

KIM, J., DAVIS, C. & KRINS, P. 2004. Amodal processing of visual speech as revealed by priming. *Cognition,* 93**,** 39-47.

KOHNERT, K. J., BATES, E. & HERNANDEZ, A. E. 1999. Balancing bilinguals: Lexical-semantic production and cognitive processing in children learning Spanish and English. *Journal of Speech, Language, & Hearing Research,* 42**,** 1400-1413.

KOVELMAN, I., BAKER, S. A. & PETITTO, L. A. 2008. Bilingual and monolingual brains compared: A functional magnetic resonance imaging investigation of syntactic processing and a possible "neural signature" of bilingualism. *Journal of Cognitive Neuroscience,* 20**,** 153-169.

KREUTZER, J., DELUCA, J. & B., C. 2011. *Encyclopaedia of clinical neuropsychology.* New York: Springer.

KRYTER, K. D. 1970. *The effects of noise on man.* New York: Academic Press.

KUHL, P. K. 1991. Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics,* 50**,** 93-107.

KUHL, P. K., CONBOY, B. T., COFFEY-CORINA, S., PADDEN, D., RIVERA-GAXIOLA, M. & NELSON, T. 2008. Phonetic learning as a pathway to language: New data and native language magnet theory expanded. *Philosophical Transactions of the Royal Society of London Series B,* 363**,** 979-1000.

KUHL, P. K. & MELTZOFF, A. N. 1996. Infant vocalizations in response to speech: Vocal imitation and developmental change. *Journal of the Acoustical Society of America,* 100**,** 2425-2438.

KUHL, P. K., STEVENS, E., HAYASHI, A., DEGUCHI, T., KIRITANI, S. & IVERSON, P. 2006. Infants show a facilitation effect for native language

phonetic perception between 6 and 12 months. *Developmental Science,* 9**,** F13-F21.

KUSHNERENKO, E., TEINONEN, T., VOLEIN, A. & CSIBRA, G. 2008. Electrophysiological evidence of illusory audiovisual speech percept in human infants. *Proceedings of the National Academy of Sciences of the United States of America,* 105**,** 11442-11445.

LACERDA, F. 1995. The perceptual-magnet effect: An emergent consequence of exemplar-based phonetic memory. *Proceedings of 13th International Congress of Phonetic Sciences, 17 May,* Stockholm, Sweden.

LAUFER, A. & BAER, T. 1988. The emphatic and pharyngeal sounds in Hebrew and in Arabic. *Language and Speech,* 31**,** 181-205.

LE MORVAN, P. 2004. Arguments against direct realism and how to counter them. *American Philosophical Quarterly,* 41**,** 221-234.

LEGERSTEE, M. 1990. Infants use multimodal information to imitate speech sounds. *Infant Behavior and Development,* 17**,** 829-840.

LIBERMAN, A. M. 1957. Some results of research on speech perception. *Journal of the Acoustical Society of America,* 29**,** 117-123.

LIBERMAN, A. M., COOPER, F. S., SHANKWEILER, D. P. & STUDDERT-KENNEDY, M. 1967. Perception of the speech code. *Psychological Review,* 74**,** 431-461.

LISKER, L. & ABRAMSON, A. 1967. The voicing dimension: Some experIments in comparitive phonetics. *Procedings 6th International Congres of Phonetic Sciences, 5 August,* Prague, Czech Republic.

LUCE, P. A. & PISONI, D. B. 1998. Recognizing spoken words: The neighborhood activation model. *Ear and Hearing,* 19**,** 1-36.

MACDONALD, J., ANDERSEN, S. & BACHMANN, T. 2000. Hearing by eye: How much spatial degradation can be tolerated? *Perception,* 29**,** 1155-1168.

MACKAIN, K., STUDDERT-KENNEDY, M., SPIEKER, S. & STERN, D. 1983. Infant intermodal speech perception is a left-hemisphere function. *Science,* 219**,** 1347-1349.

MACLEOD, A. & SUMMERFIELD, Q. 1987. Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology,* 21**,** 131-141.

MACSWEENEY, M., CALVERT, G. A., CAMPBELL, R., MCGUIRE, P. K., DAVID, A. S., WILLIAMS, S. C., WOLL, B. & BRAMMER, M. J. 2002. Speechreading circuits in people born deaf. *Neuropsychologia,* 40**,** 801-807.

MAGNO CALDOGNETTO, E., ZMARICH, C., COSI, P. & FERRERO, F. 1997. Italian consonantal visemes: Relationship between spatial/ temporal articulatory characteristics and coproduced acoustic signal. *Audio-Visual Speech Processing, 7 May,* Rhodes, Greece.

MAIDMENT, D. W., KANG, H. J., STEWART, H. J. & AMITAY, S. 2015. Audiovisual integration in children listening to spectrally degraded speech. *Journal of Speech and Hearing Research,* 58**,** 61-68.

MARSLEN-WILSON, W. 1987. Functional parallelism in spoken word recognition. *Cognition,* 25**,** 71-102.

MARSLEN-WILSON, W. & TYLER, L. K. 1980. The temporal structure of spoken language understanding. *Cognition,* 8**,** 1-71.

MARTIN, W. E. & BRIDGMON, K. D. 2012. *Quantitative and statistical research methods.* Colorado: Jossey-Bass.

MARTONY, J. 1974. On speechreading of Swedish consonants and vowels. *Speech Transmission Laboratory,* 15**,** 11-33.
MASSARO, D. 1987. *Speech perception by ear and eye.* Hillsdale, NJ: Lawrence Erlbaum Associates
MASSARO, D., THOMPSON, L., BARRON, B. & LAREN, E. 1986. Developmental changes in visual and auditory contributions to speech perception. *Journal of Experimental Child Psychology,* 41**,** 93-113.
MASSARO, D. W. 1998. *Perceiving talking faces: From speech perception to a behavioral principle.* Cambridge: MIT Press.
MASSARO, D. W., COHEN, M. M. & SMEELE, P. M. 1995. Cross-linguistic comparisons in the integration of visual and auditory speech. *Memory & Cognition,* 23**,** 113-131.
MASSARO, D. W. & FRIEDMAN, D. 1990. Models of integration given multiple sources of information. *Psychological Review,* 97**,** 225-252.
MASSARO, D. W. & LIGHT, J. 2004. Using visible speech to train perception and production of speech for individuals with hearing loss. *Journal of Speech and Hearing Research,* 47**,** 304-320.
MASSARO, D. W., TSUZAKI, M., COHEN, M. M., GESI, A. & HEREDIA, R. 1993. Bimodal speech perception: An examination across languages. *Journal of Phonetics,* 21**,** 445-478.
MATTYS, S. L., BERNSTEIN, L. E. & AUER, E. T. J. 2002. Stimulus-based lexical distinctiveness as a general word-recognition mechanism. *Perceptual Psychophysics,* 64**,** 667-679.
MCCLELLAND, J. L. & ELMAN, J. L. 1986. The TRACE model of speech perception. *Cognitive Psychology,* 18**,** 1-86.
MCGURK, H. & MACDONALD, J. 1976. Hearing lips and seeing voices. *Nature,* 264**,** 746-748.
MEREDITH, R., STEPHENS, S. D. & JONES, G. E. 1990. Investigations on viseme groups in Welsh. *Clinical Linguistics and Phonetics,* 4**,** 253-265.
MESGARANI, N., DAVID, S. V., FRITZ, J. B. & SHAMMA, S. A. 2008. Phoneme representation and classification in primary auditory cortex. *Journal of the Acoustical Society of America,* 123**,** 899-909.
MILLER, G. A. & NICELY, P. E. 1955. An analysis of perceptual confusion among some English consonants. *Journal of the Acoustical Society of America,* 27**,** 329-335.
MOORE, J. K. & GUAN, Y. L. 2001. Cytoarchitectural and axonal maturation in human auditory cortex. *Journal of the Association of Research in Otolaryngology,* 2**,** 297-311.
MUGITANI, R., PONS, F., FAIS, L., DIETRICH, C., WERKER, J. F. & AMANO, S. 2009. Perception of vowel length by Japanese- and English-learning infants. *Developmental Psychology,* 45**,** 236-247.
MUNHALL, K., HOVE, M., BRAMMER, M. & PARÉ, M. 2009. Audiovisual integration of speech in a bistable illusion. Current Biology, 19**,** 735-739.
MUNHALL, K., KROOS, C., JOZAN, G. & VATIKIOTIS-BATESON, E. 2004. Spatial frequency requirements for audiovisual speech perception. *Perception & Psychophysics,* 66**,** 574-583.
MUNHALL, K. G., GRIBBLE, P., SACCO, L. & WARD, M. 1996. Temporal constraints on the McGurk effect. *Perception & Psychophysics,* 58**,** 351-362.

NAGAO, K., LIM, B.-J. & DE JONG, K. 2003. Perceptual acquisitions of non-native syllable structures by native listeners of Japanese. *The 15th International Congress of Phonetic Sciences, 22 April,* Barcelona, Spain.

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,* 6**,** 84-93.

NIELSEN, K. 2002. Segmental differences in the visual contribution to speech intelligibility. *Journal of the Acoustical Society of America,* 115**,** 2533-2536.

OGAWA, S., LEE, T. M., KAY, A. R. & W., T. D. 1990. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Science,* 87**,** 9868–9872.

OKADA, K. & HICKOK, G. 2009. Two cortical mechanisms support the integration of visual and auditory speech: A hypothesis and preliminary data. *Neuroscience Letters,* 452**,** 219-223.

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. *International Conference on Auditory-Visual Speech Processing, 7 June,* London.

OUNI, S. & OUNI, K. 2007. Aspects of Visual Speech in Arabic. *Interspeech, 20 May,* Antwerp, Belgium.

OWENS, E. & BLA ZEK, B. 1985. Visemes observed by hearing-impaired and normal-hearing adult viewers. *Journal of Speech and Hearing Research,* 28**,** 381-393.

PARADIS, J. & NAVARRO, S. 2003. Subject realization and crosslinguistic interference in the bilingual acquisition of Spanish and English: What is the role of the input. *Journal of Child Language* 371-390.

PATTERSON, M. L. & WERKER, J. F. 2003. Two-month-old infants match phonetic information in lips and voice. *Developmental Science,* 6**,** 191-196.

PEELLE, J. E. & DAVIS, M. H. 2012. Neural oscillations carry speech rhythm through to comprehension. *Frontiers in Psychology,* 3**,** 320-322.

PEELLE, J. E. & SOMMERS, M. S. 2015. Prediction and constraint in audiovisual speech perception. *Cortex,* 68**,** 169-181.

POLKA, L. & BOHN, O. S. 1996. A cross-language comparison of vowel perception in English-learning and German-learning infants. *Journal of the Acoustical Society of America,* 100**,** 577-592.

POLKA, L., COLANTONIO, C. & SUNDARA, M. 2001. A cross-language comparison of /d/-/th/ perception: Evidence for a new developmental pattern. *Journal of the Acoustical Society of America,* 109**,** 2190-2201.

POLKA, L., VALJI, A. & MATTOCK, K. 2009. Language preference in monolingual and bilingual infants. *Journal of the Acoustical Society of America,* 125**,** 277-279.

POTAMIANOS, G., NETI, C., LUETTIN, J. & MATTHEWS, I. 2004. Audio-visual automatic speech recognition: An overview. *In:* BAILLY, G., VATIKIOTIS-BATESON, E. & PERRIER, P. eds. *Issues in visual and audio-visual speech processing.* Cambridge: MIT Press**,** 106-134.

POWERS, R., HILLOCK, A. R. & WALLACE, M. T. 2009. Perceptual Training Narrows the Temporal Window of Multisensory Binding. *Journal of Neuroscience,* 29**,** 165–174.

RABINOWITZ, W., EDDINGTON, D., DELHORNE, L. & CUNEO, P. 1992. Relations among different measures of speech reception in subjects using a cochlear implant. *Journal of the Acoustical Society of America,* 92**,** 1869-1881.

REISBERG, D., MCLEAN, J. & GOLDFIELD, A. 1987. Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli *In:* DODD, B. & CAMPBELL, R. eds. *Hearing by eye: The psychology of lip-reading.* Hillsdale, NJ: Lawrence Erlbaum Associates**,** 97-113.

REYNOLDS, G. D. & LICKLITER, R. 2003. Effects of redundant and nonredundant bimodal sensory stimulation on heart rate in bobwhite quail embryos. *Developmental Psychobiology,* 43**,** 304-310.

ROCHET, B. L. 1995. *Perception and production of second-language speech sounds by adults.* Baltimore, MD: York Press.

ROSENBLUM, L., JOHNSON, J. & SALDAÑA, H. 1996. Point-light facial displays enhance comprehension of speech in noise. *Journal of Speech and Hearing Research,* 39**,** 1159-1170.

ROSENBLUM, L. D. 2007. The primacy of multimodel speech perception. *In:* PISONI, D. ed. *Handbook of speech perception.* Oxford: Blackwell.

ROSENBLUM, L. D. 2008. Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science,* 17**,** 405-409.

ROSENBLUM, L. D., SCHMUCKLER, M. A. & JOHNSON, J. A. 1997. The McGurk effect in infants. *Perception and Psychophysics,* 59**,** 347-357.

SAMS, M., AULANKO, R., HÄMÄLÄINEN, M., HARI, R., LOUNASMAA, O., LU, S. & SIMOLA, J. 1991. Seeing speech: Visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience Letters,* 127**,** 141-145.

SÁNCHEZ-GARCÍA, C., ENNS, J. T. & SOTO-FARACO, S. 2013. Cross-modal prediction in speech depends on prior linguistic experience. *Experimental Brain Research,* 225**,** 499-511.

SCHWARTZ, J., BERTHOMMIER, F. & SAVARIAUX, C. 2004. Seeing to hear better: Evidence for early audio-visual interactions in speech identification. *Cognition,* 93**,** B69-78.

SEKIYAMA, K. 1995. Differences in auditory-visual speech perception between Japanese and Americans: McGurk effect as a function of incompatibility. *Journal of the Acoustical Society of Japan,* 15**,** 143-158.

SEKIYAMA, K. 1997. Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects. *Perception & Psychophysics,* 59**,** 73-80.

SEKIYAMA, K. & BURNHAM, D. 2008. Impact of language on development of auditory-visual speech perception. *Developmental Science,* 11**,** 306-320.

SEKIYAMA, K. & TOHKURA, Y. 1991. McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *Journal of the Acoustical Society of America,* 90**,** 1797-1805.

SEKIYAMA, K. & TOHKURA, Y. 1993. Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics,* 21**,** 427-444.

SELDRAN, F., MICHEYL, C., TRUY, E., BERGER-VACHON, C., THAI-VAN, H. & GALLEGO, S. 2011. A model-based analysis of the "combined-stimulation advantage". *Hearing Research,* 282**,** 252-264.

SHARMA, D. 1989. *Audio-visual integration and perceived location.* Ph.D. thesis, University of Reading.

SHAW, P., KABANI, N. J., LERCH, J. P., ECKSTRAND, K., LENROOT, R., GOGTAY, N., GREENSTEIN, D., CLASEN, L., EVANS, A., RAPOPORT, J. L., GIEDD, J. N. & WISE, S. P. 2008. Neurodevelopmental trajectories of the human cerebral cortex. *Journal of Neuroscience,* 28**,** 3586-3594.

SOMMERS, M. S., TYE-MURRAY, N. & SPEHAR, B. 2005. Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear & Hearing,* 26**,** 263-275.

SOWELL, E., THOMPSON, P., LEONARD, C., WELCOME, S., KAN, E. & TOGA, A. 2004. Longitudinal mapping of cortical thickness and brain growth in normal children. *Journal of Neuroscience,* 24**,** 8223-8231.

SPENCE, C. & SQUIRE, S. 2003. Multisensory integration: Maintaining the perception of synchrony. *Current Biology,* 13**,** 519-521.

STEIN, B. E. & ROWLAND, B. A. 2011. Organization and plasticity in multisensory integration: Early and late experience affects its governing principles. *Progress in Brain Research,* 191**,** 145-163.

SUMBY, W. & POLLACK, I. 1954. Visual Contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America,* 26**,** 212-215.

SUMMERFIELD, Q. 1987. Some preliminaries to a comprehensive account of audio-visual speech perception. *In:* DODD, B. & CAMPBELL, R. eds. *Hearing by eye: The psychology of lip-reading.* London: Lawrence Erlbaum Associates**,** 3-52.

SUMMERFIELD, Q. & MCGRATH, M. 1984. Detection and resolution of audio-visual incompatibility in the perception of vowels. *Quarterly Journal of Experimental Psychology,* 36**,** 51-74.

SUNDARA, M., POLKA, L. & GENESEE, F. 2006. Language-experience facilitates discrimination of /d-th/ in monolingual and bilingual acquisition of English. *Cognition,* 100**,** 369-388.

TEINONEN, T., ASLIN, R., ALKU, P. & CSIBRA, G. 2008. Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition,* 108**,** 850-855.

THYER, N., HICKSON, L. & DODD, B. 2000. The perceptual magnet effect in Australian English vowels. *Perception & Psychophysics,* 62**,** 1-20.

TREMBLAY, C., CHAMPOUX, F., VOSS, P., BACON, B. A., LEPORE, F. & THEORET, H. 2007. Speech and non-speech audio-visual illusions: A developmental study. *Plos One,* 2**,** e742.

TYE-MURRAY, N., SOMMERS, M. & SPEHAR, B. 2007. Auditory and visual lexical neighborhoods in audiovisual speech perception. *Trends in Amplification,* 11**,** 233-241.

VAN WASSENHOVE, V., GRANT, K. & POEPPEL, D. 2007. Temporal window of integration in auditory-visual speech perception. *Neuropsychologia,* 45**,** 598-607.

VAN WASSENHOVE, V., GRANT, K. W. & POEPPEL, D. 2002. Temporal Integration in the McGurk Effect. *Cognitive Neuroscience Annual Meeting, 9 May,* San Francisco.

VAN WASSENHOVE, V., GRANT, K. W. & POEPPEL, D. 2005. Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences,* 102**,** 1181-1186.

VROOMEN, J. & STEKELENBURG, J. J. 2010. Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *Journal of Cognitive Neuroscience,* 22**,** 1583-1596.

WALDEN, B. E., PROSEK, R. A., MONTGOMERY, A. A., SCHERR, C. K. & JONES, C. J. 1977. Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research,* 20**,** 130-145.

WALKER, S., BRUCE, V. & O'MALLEY, C. 1995. Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect. *Perception & Psychophysics,* 57**,** 1124-1133.

WALLACE, M. T. & STEIN, B. E. 2001. Sensory and multisensory responses in the newborn monkey superior colliculus. *Journal of Neuroscience,* 22**,** 8886-8894.

WANG, Y., BEHNE, D. & JIANG, H. 2008. Linguistic experience and audio-visual perception of non-native fricatives. *Journal of the Acoustical Society of America,* 124**,** 1716-1726.

WATSON, J. C. E. 2002. *The Phonology and Morphology of Arabic.* Oxford: Oxford University Press.

WEIKUM, W. M., VOULOUMANOS, A., NAVARRA, J., SOTO-FARACO, S., SEBASTIÁN-GALLÉS, N. & WERKER, J. F. 2007. Visual language discrimination in infancy. *Science,* 316**,** 1159.

WEINER, I. B. & FREEDHEIM, D. K. 2003. *Handbook of psychology.* New Jersey: John Wiley & Sons.

WERDA, S., MAHDI, W. & HAMADOU, A. 2007. Lip localization and viseme classification for visual speech recognition *International. Journal of Computing & Information Sciences,* 5**,** 62-75.

WERKER, J. F., FROST, P. E. & MCGURK, H. 1992. Cross-language influences on bimodal speech perception. *Canadian Journal of Psychology,* 46**,** 551-568.

WERKER, J. F. & TESS, R. C. 2002. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development,* 25**,** 121-133.

WHALEN, D. H., IRWIN, J. & FOWLER, C. A. 1996. Audiovisual integration of speech based on minimal visual information. *Journal of the Acoustical Society of America,* 100**,** 25-39.

WINDMANN, S. 2004. Effects of sentence context and expectation on the McGurk illusion. *Journal of Memory and Language,* 50**,** 212-230.

WOODHOUSE, L., HICKSON, L. & DODD, B. 2009. Review of visual speech perception by hearing and hearing-impaired people: Clinical implications. *International Journal of Language & Communication Disorders,* 44**,** 253-270.

XUE, J., ALWAN, A., AUER JR , E. T. & BERNSTEIN, L. 2004. On audio-visual synchronization for viseme-based speech synthesis. *Journal of the Acoustical Society of America,* 116**,** 2480-2481.

YI, H., PHELPS, J., SMILJANIC, R. & CHANDRASEKARAN, B. 2013. Reduced efficiency of audiovisual integration for nonnative speech. *Journal of the Acoustical Society of America,* 134**,** 387-393.