

https://dspace.org/

Audio-visual speech perception amongst bilingual speakers

Seedat, Ammaara

2024-03
University of the Witwatersrand, Johannesburg
https://hdl.handle.net/10539/42422
Seedat, Ammaara. (2024). Audio-visual speech perception amongst bilingual speakers [Master's dissertation, University of the Witwatersrand, Johannesburg]. WIReDSpace. https://hdl.handle.net/10539/42422

Downloaded from a DSpace repository. https://dspace.org/



Audio-visual speech perception amongst bilingual speakers

By: Ammaara Seedat

Student Number: 1848552

Dissertation submitted to the University of the Witwatersrand in fulfilment of the degree of Maters in Linguistics

Supervisor

Dr Ramona Kunene Nicolas

Department of Linguistics

School of Literature, Language and Media

March 2024

DECLARATION

I hereby declare that this is entirely my own work and all primary and secondary sources have been appropriately acknowledged. The dissertation has not been submitted to any other institution as part of an academic qualification.

This dissertation is prepared in fulfilment of the requirements of the degree of Master of Arts in African Languages and Linguistics at the University of the Witwatersrand

March 2024

Full Name: Ammaara Seedat Signature: ____A.Seedat_____

ACKNOWLEDGEMENTS

بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ

In the name of God the most Gracious, the most Beneficent.

This thesis would not have been possible without the support and guidance of several individuals who have all contributed and assisted in its completion.

My sincerest thanks to my supervisor Dr Ramona Kunene Nicolas, thank you for introducing me to the world of research. Your valuable time, teachings and support have been invaluable. Thank you for always pushing me to reach my full potential.

To Lerato and Thupetsi, thank you for your continuous support throughout this journey.

Ami and Aba, thank you for always supporting my dreams. Your love and support have been my greatest inspiration.

To Mum, Dad, Safwaanah, Fatima Zahra and Yusuf your love, laughs and motivation made this journey worthwhile.

To my husband- thank you for always pushing me to reach for the stars! -*ily*

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF

ABSTRACT

Why does a face articulating the syllable [ga] presented alongside an auditory /ba/ syllable result in a perceived /da/ syllable/? Language is more than words, and the human face has shown enormous communicatory significance as a mode of nonverbal communication. Multisensory integration is used in audio-visual speech perception when auditory and visual information are integrated at the same time. This integration, however, can be viewed as an involuntary process that occurs automatically. The audio-visual benefit effect occurs when auditory and visual information is synchronized, this is when the visual cue is congruent with its auditory counterpart. Literature on audio-visual speech perception, states that the magnitude of visual influences on audio-visual speech perception varies not only across languages but also across developmental stages. The reasons underlying cross-linguistic and developmental differences in audio-visual speech perception however remain unclear.

With bilingualism becoming the norm rather than the exception around the world (Grosjean & Byers-Heinlein, 2018), strong research foundations for spoken-word comprehension in bilinguals have been carried out. These foundations have been grounded in classical frameworks from monolinguals and formalised in models such as the Bilingual Model of Lexical Access (BIMOLA) (Léwy, 2008) and the Bilingual Language Interaction Network for Comprehension of Speech (BLINCS) (Shook & Marian, 2013). Bilinguals may experience increased audio-visual integration when using their less dominant language because less familiarity with a language creates a greater reliance on the visual channel to make sense of auditory input. This study will therefore examine the extent to which young adult bilinguals benefit from audio-visual speech.

We examine how different listening conditions affect how L2 bilinguals perceive audio-visual speech. Participants in this study were L1 English speakers learning L2 isiZulu between 17-29 years of age. Each participant was introduced to four different conditions. Namely, an audio-only condition, a visual-only condition an and audio-visual condition and an incongruent condition. In the audio-only condition, the stimuli were only auditory, in the visual-only condition the stimuli were perceived without an auditory stimulus. The audio-visual stimulus was made up of both an auditory and visual stimulus whilst the incongruent stimulus was created through dubbing the audio of one word over the visual of another word.

The results of the study highlighted the importance of audio-visual speech in late L2 bilingual acquisition. The differences in the phonetics and phonology of language systems might play an important role in how late L2 bilinguals perceive language in different conditions.

Keywords: Audio-visual; Speech Perception, McGurk Effect, , Bilingualism, L2 learning, L2 learning strategies, Auditory perception, Visual perception

Table of Contents

DECLARATION *2*
ACKNOWLEDGEMENTS *3*
ABSTRACT *4*
List of tables **9**
List of Figures **10**
List of abbreviations **11**
CHAPTER ONE *13*
Background of the Present Study *13*
1.1 Speech perception as a multi-disciplinary field **13**
1.2 Multilingualism in the South African context **14**
1.3 Improving L2 learning strategies **16**
1.4 Rationale **16**
1.5 Problem statement **17**
1.6 Research questions **18**
1.7 The thesis will be written in the following order: **18**
CHAPTER 2 ..... *20*
Literature Review *20*
2.1 Introduction **20**
2.2 The multisensory nature of speech perception **20**
2.3 Bilingualism and speech perception **24**
2.4 Phonological awareness of language properties **26**
2.4.1 Phonemes 26
2.4.2 IsiZulu and English phonetics and phonology 27
2.4.3 isiZulu vowels 28
2.4.4 English vowels 29
2.4.5 IsiZulu consonants 29
2.4.6 English consonants 31
2.4.7 Minimal pairs 33
2.4.8 Models of lexical access 33
2.5 The development of audio-visual perception **35**
2.6 Models of audio-visual integration **37**
2.7 Auditory perception **38**
2.8 What can be understood from visual modality in speech? **39**
2.8.1 Visual cues 39
2.8.2 Effects of visual modality on perception 41

2.8.3 Lip movements in speech perception 42
2.8.4 Variation in lip reading 43
2.8.5 Visual speech in quiet environments 45
2.8.6 Variation of visual modality usage in perception 46

2.9 Audio-visual perception 47
2.9.1 The development of audio-visual integration 49
2.9.2 The perception of connected speech 50
2.9.3 Audio-visual illusions 51
2.9.4 Context and audio-visual integration 51
2.9.5 Non-speech audio-visual illusions 51
2.9.6 Incongruent audio-visual speech in children 52

2.10 The McGurk Effect 53
2.10.1 Stimuli used in previous McGurk studies 55
2.10.2 Differences between McGurk effect studies 56
2.10.3 The McGurk effect across various languages 56
2.10.4 Theoretical explanation of the McGurk effect 57
2.10.5 Limitations of the McGurk Effect 58

2.11 Speech perception in the L2 59
2.11.1 Speech perception in noise 60
2.11.2 Uncertainty in audio and visual cues 61
2.11.3 Audio-visual perception across languages 61
2.11.4 Properties affecting L2 perception 64
2.11.5 Difficulty learning tone in an l2 65

2.12 Theoretical Framework 66

2.13 Chapter summary 68

Methodology 69

3.1 Introduction 69

3.2 Research design 69

3.3 Sampling method 71

3.4 Participants 71
3.4.1 Language background questionnaire 72

3.5 Procedure 74

3.6 Pilot study 75

3.7 The video stimuli 75
3.7.2 Visual-only condition 79
3.7.4 Incongruent stimuli 83

3.8 Coding and transcription 84
3.8.1 Audio-only condition coding 84
3.8.2 Visual-only condition coding 85
3.8.3 Audio-visual condition coding 86
3.8.4 Incongruent condition coding 87

3.9 Reliability and Validity 87

3.10 Ethical considerations 88

3.11 Summary of the chapter 88

CHAPTER 4 89

Analysis *89*
4.1 Introduction **89**
4.2 Audio-only condition findings **92**
4.2.1 Audio-only tonal pairs 92
4.2.2 Audio-only similar word pairs 94
4.3 Visual-only condition findings **96**
4.4 Audio-visual condition **97**
4.4.1 Audio-visual tonal pair 97
4.4.2 Audio-only minimal pair findings 99
4.5 Incongruent condition findings **100**
4.6 Summary of chapter **102**
CHAPTER 5 ..... *103*
Discussion and conclusion *103*
5.1 Introduction **103**
Research Question 1 ..... **104**
5.2 Research question 1: **106**
5.3 Research Question 2 ..... **115**
5.3.1 How do different stimuli affect the way L2 bilinguals perceive audio-visual speech? 115
5.3 Conclusion **119**
5.4 Limitations and future recommendations of the current study **120**
5.4.1 Limitations 120
5.4.2 Future recommendations 121
5.5 Summary of chapter **121**
CHAPTER 6 ..... *123*
References *123*
Appendices *176*
Appendix 1- Participation information sheet **176**
Appendix 2-Consent Form **178**
Appendix 3- Ethics approval **180**
Appendix 4- Audio-only condition coding **181**
Appendix 5-Visual-only condition coding **182**
Appendix 6- Audio-visual only condition coding **183**
Appendix 7- Incongruent condition coding **184**

List of tables

TABLE 1: DATA SAMPLE 72
TABLE 2 SHOWS THE DATA GATHERED FROM THE LINGUISTIC QUESTIONNAIRE 73
TABLE 3: A REPRESENTATION OF THE AUDIO-ONLY TONAL PAIR CODING SCHEMA 85
TABLE 4: A REPRESENTATION OF THE AUDIO-ONLY MINIMAL PAIR CODING SCHEMA 85
TABLE 5: A REPRESENTATION OF THE VISUAL-ONLY CONDITION CODING SCHEMA 86
TABLE 6: A REPRESENTATION OF THE AUDIO-VISUAL TONAL PAIR CODING SCHEMA 86
TABLE 7: A REPRESENTATION OF THE AUDIO-VISUAL MINIMAL PAIR CODING SCHEMA 87
TABLE 8: A REPRESENTATION OF THE INCONGRUENT CONDITION CODING SCHEMA 87
TABLE 9: TABLE SHOWING THE TONAL PAIR STIMULUS IN AUDIO-ONLY CONDITION 92
TABLE 10: TABLE SHOWING THE RESULTS OF MINIMAL PAIRS IN THE AUDIO-ONLY CONDITION 94
TABLE 11: TABLE SHOWING THE RESULTS OF TONAL PAIR IN THE VISUAL-ONLY CONDITION 96
TABLE 12: TABLE SHOWING THE RESULTS OF TONAL PAIRS IN THE AUDIO-VISUAL CONDITION 98
TABLE 13: TABLE SHOWING THE RESULTS OF MINIMAL PAIRS IN THE AUDIO-VISUAL CONDITION 99
TABLE 14: TABLE SHOWING THE RESULTS IN THE INCONGRUENT CONDITION 100

List of Figures

Figure 1: The language distribution in South Africa (Census Dissemination, 2022) 14
Figure 2: Grouping of consonants according to the International Phonetic Alphabet (IPA) 27
Figure 3: An illustration of the isiZulu vowel chart (Retrieved from Pascoe & Jeggo, 2019) 28
Figure 4: An illustration of English vowels 29
Figure 5: An illustration of the places of articulation in the vocal tract 31
Figure 6: An illustration of the BIMOLA model (adapted from Grosjean, 2012) 34
Figure 7:Visual representation of the BLINCS model (retrieved from Shook & Marian, 2012) 35
Figure 8: An illustration of audio-visual integration (Peelle & Sommers, 2015) 38
Figure 9: An illustration of visemes and their relevant phonemes (Pattamadilok & Sato, 2022) 45
Figure 10: An illustration of the participant's placement during the laboratory session 74
Figure 13: An illustration of the word" isisu" (stomach) in the audio-only condition 78
Figure 14: An illustration of the word " usisi" (sister) in the audio-only condition 78
Figure 15: An illustration of the audio-only condition where partcipants were seated in front of a blank screen as they observed an auditory stimuls 79
Figure 16: An example of the visual stimulus "isithelo" 80
Figure 17: An example of the option participants encountered during the visual-only condition3.7.3 Audio-visual condition 80
Figure 18: An illustration of the word ùkúsìndà (to smear) in the audio-visual condition 81
Figure 19: An illustration of the word ùkùsíndà (to be heavy; to recover) in the audio-visual condition 81
Figure 20: An illustration of the word isitho (organ) in the audio-visual condition 82
Figure 21: An illustration of the word isipho in the audio-visual condition 82
Figure 22: An illustration of the visual-only stimuli participants were exposed to 90
Figure 23: An illustration of the options participants received after viewing the visual-only stimulus 91
Figure 24: A bar graph displaying the identification of the difference between the audio-only tonal pair words 93
Figure 25: A bar graph displaying the results of minimal pair words in the audio-only condition 95
Figure 26: An illustration of the visual cues in the /s/ and /z/ sound 97
Figure 27: A bar graph displaying the tone detection in the audio-visual condition 99

List of abbreviations

AV- Audio-visual

AVMP- Audio-visual minimal pair

BIMOLA- Bilingual Model of Lexical Access

BLINCS- Bilingual Language Interaction Network for Comprehension of Speech

H- High tone

HLHLT- High, low, high, low tone

HLT- High low tone

HREC- Health Research Ethics Commity

Id of Diff- Identification of difference

IPA- International Phonetic Alphabet

L- Low tone

L1- First Language

L2- Second language

LHLLT- Low, high, low, low tone

LHT- low, high tone

LPHE- Language Policy for Higher Education

MP- Minimal Pair

PAM- perceptual assimilation model

TP- Tonal Pair

n= Number of participants who answered correctly

Clarification of terms

First Language (L1):
This is the language acquired in early childhood (Saville-Troike & Barto, 2017).

Multilingualism-
The ability of a person to speak or be more proficient in two or more languages (Saville-Troike & Barto, 2017).

Second Language (L2):
The language a person acquires in addition to their native language (Saville-Troike & Barto, 2017).

Phonetic Awareness: The understanding that words are comprised of individuals (Hempenstall, 1997)

CHAPTER ONE

Background of the Present Study

The Constitution of the Republic of South Africa, 1996 (the Constitution) is rooted in a framework that supports bilingualism and multilingualism. However, to implement bilingualism and multilingualism effectively, successful implementation, effective learning methods and learning tools need to be adopted.

Higher learning institutions such as the University of the Witwatersrand have implemented language policies enabling students to learn isiZulu or Sesotho as an additional language; this has been done to promote language and linguistic diversity (University of the Witwatersrand, 2024). Because English (a West Germanic language in the Indo-European language family) and isiZulu (a Bantu language in the Nguni branch) belong to different language families their language systems differ, therefore, the pronunciation and perception of certain isiZulu sounds can present a perceptual challenge to late language learners. One of the reasons for, this could be that the learner thinks about their respective L1 (first language) language system, and they substitute these sounds for the sounds of the second language (L2). This study seeks to look at audio-visual aspects that may enhance or hinder second language learning in adult bilinguals.

1.1 Speech perception as a multi-disciplinary field

Speech perception in itself lies within a multi-disciplinary field and includes phonology, phonetics, and psycholinguistics disciplines. Within the field of phonetics, speech perception comprises the acoustic dimensions in speech signals together with acoustic and physiological correlates. Relating to psycholinguistics, speech perception involves the explanation of the processing of speech sounds. Various psycholinguists have argued that people process signals through an intermediate pre-lexical level that reconciles the unprocessed acoustic information and the words in the lexicon (Pitt & McQueen 1998). The phonological field looks at representations of a sound as a contrastive unit as it is made up of characteristics that set it apart from representations of other sounds. Therefore, it can be argued that phonetics, psycholinguistics and phonology all agree that it is essential for sound categories to maintain a level of abstraction.

1.2 Multilingualism in the South African context

South Africa has a rich linguistic landscape with 12 official languages. According to the Census (2022), the twelve official languages include isiZulu, isiXhosa, isiNdebele, Sepedi, siSwati, Setswana, Sesotho, Xitsonga, Tshivenda, Afrikaans, English and the newly added South African Sign Language. Of these 12 official languages, isiZulu is the most spoken language in South Africa accounting for 24,4% of the population (Census Dissemination, 2022). Figure 1 is an illustration of the language distribution in South Africa based on the 2022 Census.

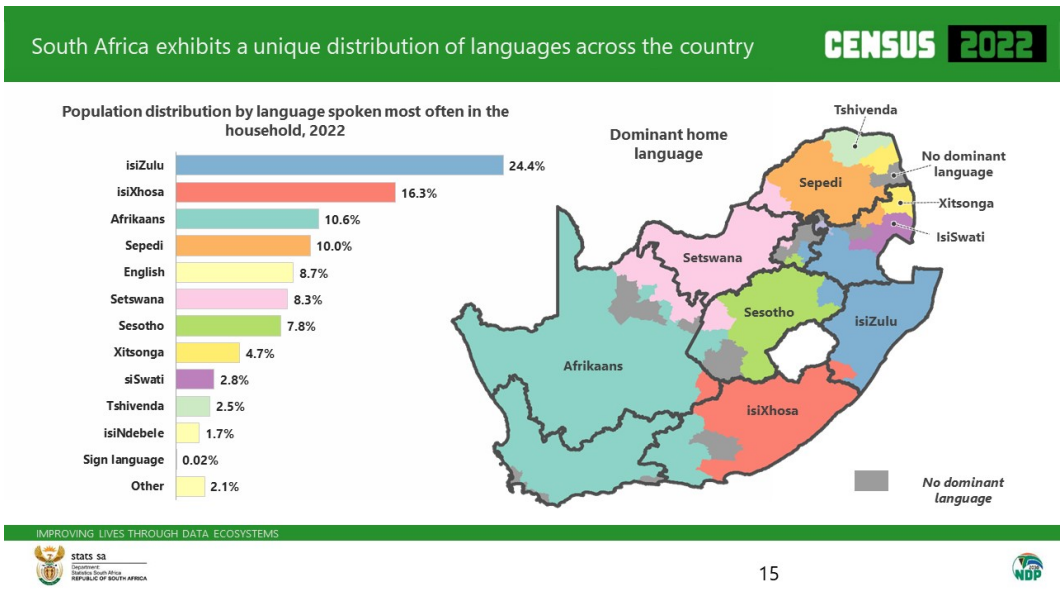


Figure 1: The language distribution in South Africa (Census Dissemination, 2022)

The growing research in bilingual speech is justifiable in today's multicultural world. Research on bilingualism is essential since it provides insight into the theories of language acquisition. This is especially important as the majority of language acquisition theories are situated around monolingual data (Genesee, 2022). Therefore, understanding the various principles and enquiring into the most effective methods of late bilingual acquisition is essential.

New L2 adult speakers, of an additional language, commonly face various challenges when learning an L2. These challenges lie in word learning, the comprehension of sentences and the learning of non-native sounds. Therefore, to alleviate the difficulties concerned with late L2 learning, it is important to optimise the modalities used in L2 learning.

This study seeks to investigate the extent to which young adult early-stage bilinguals benefit from audio-visual speech. The study aims to understand what the effect of modality is on speech perception and how incongruent stimuli affect the perception of audio-visual speech in this population. Learning an L2 is a cognitively demanding process as languages differ in their phonology and speech sounds. Phonological differences are an aspect of language that is challenging to learn (Goswami, 2002). Mastering and articulating new speech sounds and acquiring new concepts of speech are essential processes in learning an L2. For example, L1 English individuals learning isiZulu would have to learn about tone as it is a distinctive feature in the isiZulu language.

The audio-visual modality introduces both auditory and visual tools to individuals. This modality is more nuanced as it is made up of both auditory and visual media. Multisensory integration occurs when both the auditory and visual stimuli are integrated. This integration occurs automatically. However, we need to ask what happens when the auditory and visual stimuli do not match? Do L2 speakers integrate the stimuli from both modalities and do they ignore one modality? This study aims to explore the effect of incongruent speech in L2 speakers in order to understand what modality learners rely on as they learn another language.

Utilising this modality in L2 learning would allow students to synchronise the speech they have heard with the visuals they see to offer a more thorough understanding. The absence of audio-visual material in L2 learning means that students do not gain certain aspects of the L2 (that is, tone) (Cabrera & Werner, 2017). Therefore, the lack of audio-visual material means that L2 speakers may have difficulty with certain phonetic elements, pronunciation and production of non-native elements which could decrease their L2 perception. Audio-visual cues in speech perception can be more challenging for bilingual speakers who have to differentiate between languages particularly when the languages have contrasting linguistic features. Research has shown that L1 listeners gain more from visual cues through lip reading than L2 listeners do (Drijvers & Özyurek, 2019). Similarly, new L2 speakers use visual cues compared to more proficient L2 speakers (Hazan et al., 2006; Sueyoshi & Hardison, 2005). Despite, this information there has been limited research concerning L2 speech perception in languages belonging to the Bantu language family.

Based on the above information it is valid to hypothesise that the audio-visual modality will provide L2 speakers with additional information such as visible articulatory information which would be beneficial in L2 perception and would complement the auditory stimuli.

1.3 Improving L2 learning strategies

With the rise of bilingualism and multilingualism in society it is essential for learning institutes to adopt effective language learning strategies, particularly in late L2 learners. To ease and improve non-native cue learning, language instructors should employ teaching methods that enhance L2 learning processes. Effective learning strategies are a key factor in L2 learning (Oxford, 2003). Various researchers (Kumaravadivelu, 2006; Herrera et al, 2012; Richter, & Lara Herrera, 2016) have highlighted the need for L2 instructors and curriculum planners to design and employ effective L2 teaching methods that will aid in second language learning. Learning strategies are used by curriculum planners to assist in the new cognitive demand of learning a new language system. Therefore, highlighting effective modalities of learning can assist L2 speakers in maximising accurate pronunciation, perception and learning new sound systems. Together with aiding in L2 learning using effective modalities of speech perception can also aid hearing-impaired individuals via improved identification of phonetic detail and sound systems. Thus, examining the most effective modalities for speech perception is essential as it is the foundation of language development.

1.4 Rationale

The following research will explore the effects of modality in young adult bilinguals. The results of this study could contribute to understanding early steps in developing more comprehensive models of speech perception to match the changes in teaching and learning. In the South African context, future planning, reimagined speech research, teaching, and practice all demand evidence-based research that can aid in the facilitation of service delivery models. Such models should be culturally, linguistically, and contextually appropriate to the South African context and population (Khoza-Shangase & Masondo, 2020).

This data could assist theorists in determining whether any additional key component skills should be included in L2 perception models for bilingual speakers. Previous studies on non-

native perception have focused on auditory and visual perception at the phoneme or syllable level; in contrast, this study will focus on lexical stimuli.

Moreover, few studies to date have examined how variations in the perception of lexical cues can add to general listening comprehension, particularly in bilingual speakers (Fricke et al., 2016). General listening comprehension is an important skill in developing L2 learning. The data in this study could assist in decoding additional components such as compensatory strategies used to decode degraded speech. Language instructors have limited knowledge of the necessary training and support for L2 learners particularly in classroom settings (du Plessis, 2020; Webb et al, 2015). A better understanding of the effect of modality on speech perception will allow educators to design more informed listening training that will equip students with the skills they need to overcome the listening challenges they face in their daily lives. As language learning strategies have developed in the last decade, literature promoting language acquisition through an audio-visual stimulus has been on the rise (Perez & Rodgers, 2019; Peters & Muñoz, 2020).

The results of this research study will also contribute to the development of L2 isiZulu perception and could aid the educational sector in adopting effective L2 learning models particularly to improve L2 speech perception.

1.5 Problem statement

The perception of L2 speech is essential in the process of mastering an L2. Although there has been a steady stream of research regarding speech perception (Aller et al, 2022; Alsius et al, 2018; Bernstein et al, 2000) there has been limited focus on the speech perception of Bantu languages (Van Der Merwe & Le Roux, 2014). Understanding the extent to which young adult bilinguals perceive audio-visual speech is essential for L2 listeners' perception of sounds. Similarly, it is important to document how L2 speakers can better perceive L2 speech within classroom settings. The limited research on the perception of Bantu languages means that language instructors are forced to rely on intuition instead of evidence in L2 language learning.

1.6 Research questions

To accomplish the aims and objectives of this research this study aimed to examine the extent to which young adult bilinguals benefit from audio-visual speech this study aimed to answer the following questions:

- What is the effect of modality on audio-visual perception?
- How do different conditions affect the way L2 bilinguals perceive audio-visual speech?

1.7 The thesis will be written in the following order:

Chapter One: Introduction

This chapter highlighted the prevalence of multilingualism in South Africa. It also explored the main framework underpinning audio-visual perception and the need for studying speech perception in L2 speakers. Thereafter, the chapter examined the different L2 learning strategies involved with speech perception and introduced the research question and the motivation behind the study.

Chapter 2: Literature Review

In this chapter, the theoretical frameworks of the study are explained. It then explores previous research on speech perception including audio perception, visual perception and audio-visual perception. The language properties of English and isiZulu will also be discussed.

Chapter 3: Methodology

This chapter examines the data collection method utilised in the study. Thereafter, the description of the materials, equipment and procedure is investigated. Lastly, this chapter will explain how the data was coded.

Chapter 4: Analysis

The analysis chapter presents the research findings, the results gathered from this data collection studied through the use of tables and graphs.

Chapter 5: Discussion and Conclusion

In this chapter, the findings of the present research are explored and the most important findings are explained. Moreover, the results are interpreted with reference to other research studies. The conclusion underlines the principal findings of the study. Thereafter, the limitations of the current study are addressed. A list of references established by the research framework will also be included.

Appendices to the manuscript contain ethical clearance certificates and the coding manual used in the methodology.

CHAPTER 2

Literature Review

2.1 Introduction

In this thesis, we explore the extent to which young adult bilinguals benefit from audio-visual speech. In doing so we will investigate the effects of modality on audio-visual perception and how different listening conditions affect how L2 speakers perceive speech. The present chapter highlights the literature that foregrounds this study in aligning with my research questions:

1. What is the effect of modality on audio-visual perception?
2. How do different conditions affect the way L2 bilinguals perceive audio-visual speech?

This chapter presents the literature for this study, and it provides a description of audio-visual modalities. An outline is provided of the languages under review including bilingualism, the McGurk Effect and the theoretical framework to be used in responding to the research questions are explored.

2.2 The multisensory nature of speech perception

The environment around us is multisensory- human beings process information from their surroundings using their senses which they then code and use in various sensory modalities. This provides a cognitive challenge to the human brain as it is forced to distinguish between information that originates from the same object or event and should be connected and information that does not and should be separated. This process is known as sensory binding (Desantis & Haggard, 2016). To formulate coherent perceptual representations and create appropriate response behaviours this process needs to be rapidly resolved.

Speech is more than just a collection of acoustic sounds. The process of creating speech requires the coordinated movement of several articulators (including the jaw, teeth, and lips). To aid the perceiver's language comprehension face-to-face multimodal communication involves hearing the speaker's auditory speech while also seeing the speaker's articulatory movements. Auditory speech accompanied by the speakers mouth together with the auditory stream requires a combination of temporal and phonetic cues (Lalonde & Werner, 2019). As a

result, the integration of auditory and visual stimuli becomes pivotal for optimal speech comprehension (Brancazio et al, 2003). Several studies (Hirata & Kelly, 2010; Holle et al., 2012; Nagels et al., 2015; Wu & Coulson, 2015) relating to audio-visual perception have revealed that people's responses to audio-visual stimuli are considerably quicker compared to auditory-only conditions. It has been shown that the processing of audio-visual language is bimodal and that the act of visual modalities facilitating speech perception is known as the "audio-visual benefit effect" (Bernstein & Liebenthal 2014; Jesse & Janse, 2012).

The process of synthesising several modalities simultaneously and in synchrony is known as "multisensory integration" (Venezia et al; 2016). Multisensory integration is used in audio-visual speech perception when auditory and visual information are integrated at the same time. This integration, however, can be viewed as a natural process that occurs automatically. The audio-visual benefit effect occurs when auditory and visual information is synchronised, this is when the visual cue is congruent with its auditory counterpart. Additionally, another type of audio-visual effect is known as "the McGurk effect" (McGurk & MacDonald, 1976) which is discussed later in this study. The two major strands of research on audio-visual speech perception are the McGurk effect and the audio-visual benefit effect.

According to research on speech perception (which includes both the McGurk effect and the audio-visual benefit effect), the immensity of visual influences on audio-visual speech perception differs not only across languages but also across developmental stages (Burnham & Dodd, 2004; Chen & Hazan, 2009; Chen & Massaro, 2004; Dupont et al., 2005; Lalonde & McCreery, 2020; Lalonde & Werner, 2019; Magnotti et al., 2015; Sekiyama, 1997; Sekiyama & Burnham, 2008; Tremblay et al., 2007). However, the underlying reasons for cross-linguistic and developmental differences in audio-visual speech perception remain unclear.

Spoken word recognition is a critical component of language comprehension. A large number of studies that have been conducted over the last four decades lay a strong foundation for spoken word recognition in monolingual speakers (McClelland & Elman, 1986). With bilingualism becoming the norm rather than the exception around the world (Grosjean & Byers-Heinlein, 2018), strong research foundations for spoken-word comprehension in bilinguals have been carried out. These foundations have been grounded in classical frameworks from monolinguals and formalised in models such as the Bilingual Interactive

Model of Lexical Access (BIMOLA) (Grosjean, 2008) and the Bilingual Language Interaction Network for Comprehension of Speech (BLINCS) (Shook & Marian, 2012).

When the auditory channel is unclear (such as when the audio is degraded or noisy) we rely more on visual input. Visual lip-reading, for example, can aid in the clarification of degraded speech sounds. As a result, when auditory input is less comprehensible, people are more likely to experience incongruent speech. When using their less dominant language, bilinguals may experience increased audio-visual integration to make sense of auditory input because less familiarity with a language creates a greater reliance on the visual channel. Bilingualism has been proven to have long-term effects on a variety of cognitive processes. The bilingual experience, for example, has been shown to improve cognitive control (Bialystok & Herman, 1999), increase meta-linguistic awareness (Cummins, 1978), delay the effects of Alzheimer's disease in older adults (Yew et al., 2012), and improve perspective-taking ability in children (Goetz, 2003). While such examples demonstrate the various advantages of bilingualism, there are some drawbacks. Previous studies have found that bilinguals who use their second language have more difficulty understanding speech in noisy environments than monolinguals (Smiljanic & Gilbert, 2017). The challenge of learning multiple languages may thus shape bilinguals' speech comprehension strategies. According to Navarra and Soto-Faraco (2007), attending to corresponding visual stimuli improves bilinguals' speech comprehension. This could lead to a general preference for the visual channel in speech perception.

Even after achieving language competence, early challenges of learning and managing multiple languages may alter how people perceive and integrate audio-visual stimuli. Poor classroom acoustics have negatively impacted academic achievement and psycho-educational and psychosocial development (Crandell & Smaldino, 2000). High levels of reverberation and noise (unwanted sound inside or outside a room) impair speech perception, reading and spelling ability, classroom attention, concentration, and academic achievement. Because learning multiple languages is a relatively challenging task, bilingual speakers may rely more on visual information than monolinguals, resulting in increased audio-visual integration. Greater reliance on the visual channel during bilingual speech comprehension, and an increase in audio-visual integration as a result of greater difficulty comprehending speech listening conditions increase the likelihood of experiencing the McGurk effect (Sekiyama et al., 2003). This study will further explore the impacts of audio-visual integration and the McGurk effect in bilingual

speakers. Although there have been a plethora of studies focusing on audio-visual aspects, they are largely limited to monolingual speakers.

In the educational setting, media can be described as utilising technology to aid in learning (Clark et al., 2003). In today's world, there are various types of media and technology being used for learning and have been shown to increase a student's motivation, learning and engagement. Within the learning space, media can be introduced through different modalities including the audio-visual modality. Mishra and Koehler (2006) state that media in learning is used to aid in the learning process and can be introduced through visual, video or audio means. The media used in teaching and learning an L2 has to be intentional and appropriate to elicit the best learning results (Noesgaard & Ørngree, 2015). In the learning effective environment, media aid in processing information- particularly through effective visual and auditory modalities. Utilising the best modalities in teaching and learning can aid teachers in providing materials that will best facilitate L2 learning.

With improvements in science and technology, L2 teaching is changing from being predominantly audio towards a more nuanced way of audio-visual teaching (Mayyas & Farhan, 2023; Namaziandost & Nasri, 2019; Zhyrun, 2016). The audio-visual modality elicits both visual and auditory perceptions, while auditory learning elicits auditory perceptions (Campbell, 2007; Surguladze et al., 2001) therefore, L2 learning should head towards multi-modal learning that incorporates both auditory and visual modalities. Various studies have shown that gaining information from a multi-modal perspective is more constructive than information gained from a single modality as multimodal information lasts longer in the memory (Xu et al; 2021; Zhang et al; 2020). The significance of this study understanding the extent to which audio-visual perception aids in L2 perception- will add to the influence of L2 cues in audio-visual perception.

Lei and Gupta (2010) revealed that technology has become more in reach of people than ever before. With increased use, technology will now have to be integrated into places of learning and education settings. Similarly, students have displayed positive responses towards the integration of technology in learning (Peters et al., 2009). Various studies have highlighted the benefit of using technology in L2 learning (Roever, 2006; Zhao, 2010). The study by Taylor and Gitsaki (2004) for instance highlighted that the use of technology in their English classes was beneficial and rewarding. These studies highlight that effective technology models can be

valuable and can aid in face-to-face L2 learning; this is especially helpful, given time constraints limiting the opportunity for learners to exercise their language production. In addition to face-to-face teaching, technology can be utilised to improve language production. Studies (Wang & Munro, 2004) have also shown that technological help can aid in language perception. This was seen in a Cantonese group whose English vowel contrasts improved drastically with the help of computer-based perception training.

2.3 Bilingualism and speech perception

The term "bilingualism", which is a major theme of this manuscript, is derived from two Latin words: "bi" (two) and "lingua" (language). Thus, bilingualism is defined as a speaker's ability to communicate with himself or herself in at least two languages with equal or native-like proficiency (Tsitsipis, 2006). More broadly, bilinguals are now defined as any individual with early or late onset of a second language (Tanner et al; 2012). The term "bilingualism" does not have set criteria rather, it is said to exist within a continuum (González, 1989). The ends of the continuum range from a monolingual speaker to an L2 speaker with native-like acquisition. Although there is no set definition of bilingualism, researchers have described different classes of bilinguals.

The first classification is based on the stage of life of becoming bilingual early and late. This is determined according to the age of exposure of the L2 (Berken et al; 2016). The acquisition of a second language during the pre-adolescent phase is known as early bilingualism while late bilingualism refers to the acquisition of another language after eight years of study (Berken et al; 2016). The other classification is described as balanced and dominant bilinguals (Bhattacharjee, 2014). Balanced bilinguals are those who obtain native-like proficiency and mastery in both their L1 and L2 while dominant bilinguals have a higher proficiency in one language (Bhattacharjee, 2014). Additive and subtractive bilinguals are another classification-additive bilinguals are those who can acquire an L2 without losing their L1 while a subtractive bilingual acquires an L2 at the expense of the L1 (Bhattacharjee, 2014). Another classification is simultaneous and sequential bilingualism (Amengual, 2017). Simultaneous bilinguals do not have a distinct L1 or L2 as both languages were learnt simultaneously (Silva-Corvalán, 2014; Thordardottir, 2017). For the purpose of this study, the term bilingualism will refer to sequential bilinguals. Sequential bilingualism is defined as a person who first acquires an L1 and then acquires an L2 later in development (Castilla et al; 2009; Kohnert, 2008). A sequential

bilingual could be any individual who acquires an L2 after 3 years of age because at this age their L1 is intact (Castilla et al; 2009; Kohnert, 2008).

Bilingualism has become the norm around the world, although the general tone of bilingualism is either official or unofficial, depending on the country. Official bilingualism is generally assertively entrenched in a country's constitution and is practically applied in sectors such as education, and national, international, and official engagements. Teaching and learning in primary, secondary, and tertiary institutions fall within these sectors together with parliamentary proceedings and judicial activities.

South Africa has twelve official languages and promotes a multilingual policy. The languages that fall under the twelve official languages are Sepedi, Sesotho, Setswana, siSwati, Tshivenda, Xitsonga, Afrikaans, English, isiNdebele, isiXhosa, South African Sign Language and isiZulu. To put the South African constitutional multilingual provision into action, the government enacted the Language Policy for Higher Education in 2002. This policy requires universities to develop bilingual language policies and to incorporate bilingualism into their institutional environment as well as their teaching and learning programmes. Although the country's language policy has been critiqued for a wide range of reasons (Nkomo & Madiba, 2012), it has undoubtedly created an ideological space for the adoption of bilingual language policies in most universities. The policy has resulted in the implementation of bilingual education language policies in approximately 20 of the 23 universities in the country.

Although these practices are stated, the reality is that changes (particularly related to the COVID-19 pandemic) have meant that there has been a shift away from traditional in-person models of teaching and learning. Although several educational establishments have returned to in-person teaching a larger number have continued with distanced learning. Distanced learning may hinder an individual's speech perception particularly, owing to inadequate access to effective technology together with maintaining a stable power supply. Therefore, exploring the effects of speech perception is important in developing adequate teaching methods for teaching and learning particularly for new L2 speakers. Similarly, this will aid in creating effective methods and models for distance learning.

There have been various hypotheses regarding the methods which L2 speakers use to understand the linguistic input in each specific language. The language-selective hypothesis and the nonselective hypothesis proposed by Dijkstra and Kroll (2005) sought to explain the

bilingual language selection process. The first (language-selective) hypothesis, demonstrates that, depending on linguistic input, bilinguals have two distinct selectively accessed lexicon systems. According to this hypothesis, the bilingual mind chooses and employs one language like a monolingual, since the two languages are stored and processed independently (Kroll & Bialystok, 2013).

The alternative (nonselective) hypothesis, argues that bilinguals have an integrated lexicon, in which lexical structures from both languages are simultaneously activated during the word recognition and selection process. This is backed by evidence from neuroimaging studies, which suggests that bilinguals who read, talk, and listen to speech in one language alone are co-activating their linguistic knowledge rather than choosing one particular language at a time

2.4 Phonological awareness of language properties

Infants develop the perception of speech sounds from birth (Moon et al; 2013; Webb et al; 2015). It is from birth that infants develop the awareness of phonemes from various languages: however, increased exposure to one particular language over time sees a decline in the perception of speech sounds of other languages (Beattie & Manis, 2012; Cantiani et al; 2023). As children develop and reach new milestones their phoneme-grapheme knowledge is essential for phonemic awareness (Ouellette & Haley, 2011). Phonological awareness thus contributes to other systems such as phonological representations and phonemic specifications which are all essential for speech perception (Anthony et al; 2005; Lewis et al., 2012; Preston & Edwards, 2010). Therefore, the perception of speech at the segmental and suprasegmental levels moderate auditory sensitivity of phonemes through phonological awareness.

2.4.1 Phonemes

Phonemes are the smallest sound units were substituting a phoneme segment with another might alter the meaning (Zwicky, 2017; Zwicky &, Pullum, 1983). The International Phonetic Alphabet (IPA) was created in 1887 and comprises a set of phonemes that can be used to describe speech across languages. Phonemes are made up of groups of vowels and consonants and appear in the IPA chart as shown in Figure 2. This shows that as children reach new milestones their phonological awareness improves (Ambrose et al; 2012). As their

phonological awareness increases in the L1 it becomes more challenging for an individual to develop the same perception in their L2.

THE INTERNATIONAL PHONETIC ALPHABET (revised to 2020)

CONSONANTS (PULMONIC) ©①◎ 2020 IPA

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

Figure 2: Grouping of consonants according to the International Phonetic Alphabet (IPA)

2.4.2 IsiZulu and English phonetics and phonology

IsiZulu is one of the 11 recognised languages of South Africa. The Bantu language, isiZulu falls under the greater Niger-Congo family (Nicolas & Ahmed, 2016). It is part of the Nguni group of languages (Kunene-Nicolas, 2015; Pacoe & Jeggo, 2019; Posthumus, 2016). Owing to the multicultural society in South African the universities have adopted language policies that enable young adults to learn one of the recognised languages as an L2. This study chose to look at early L2 isiZulu speakers because of the scarcity of the scarcity of research regarding the L2 perception of young and adult bilinguals in Bantu languages.

A few key factors differentiate English and isiZulu; importantly isiZulu has a larger phonetic inventory than English has. According to Naidoo et al. (2005), there are approximately 24 consonants and 14 vowels found in English whilst isiZulu accounts for 59 consonants and seven vowels (Naidoo, 2002; Naidoo et al; 2005). Further, isiZulu contains clicks which do not appear in English (Land, 2015). In contrast, the English language allows consonant clusters (Stockman, 2006) whilst isiZulu does not permit them (Taljaard & Snyman, 1991). A further difference is that isiZulu allows only open syllables (Poulos & Msimang, 2001) whilst English contains both open and closed syllables (Borowsky, 1989). Moreover, aspiration is not a characteristic in English while aspirated sounds in isiZulu can change meaning (Van Der Merwe & Le Roux, 2014).

In Bantu languages, words commonly adhere to the CVCV pattern (Mosaka, 2000). IsiZulu may include V, C and CV (Van Der Merwe & Le Roux, 2014). IsiZulu also uses tone which could affect the meaning of the word. The tone could be presented as H (high tone), L (low tone) and could include HH, LL, HL or LH structure (Van Der Merwe & Le Roux, 2014).

2.4.3 isiZulu vowels

A vowel can be identified based on vertical modification, horizontal modification, the extent of lip rounding and the position of the cardinal vowel. The horizontal modification of the oral cavity is related based on the front or back of the mouth. Therefore, vowels are categorised as front, central or back vowels. Similarly, the vertical modification is based on the height of the tongue in the oral cavity and can be categorised as low, mid and high.

The [i] is a high, front vowel, unrounded which occurs at cardinal positions 1 and 2. The [i] sound can be found in words such as [usi.si] (sister) while the [ɛ] is a mid-low front, unrounded vowel (Ncobela, 2015). The [ɔ] is the mid low, back, rounded vowel while the [u] is a high, back, rounded vowel. The mid-high front, unrounded vowel is the [e] while [o] is the mid-high, back, rounded vowel (Ncobela, 2015) (See Figure 3). The next section looks at the phonemic structure of English.

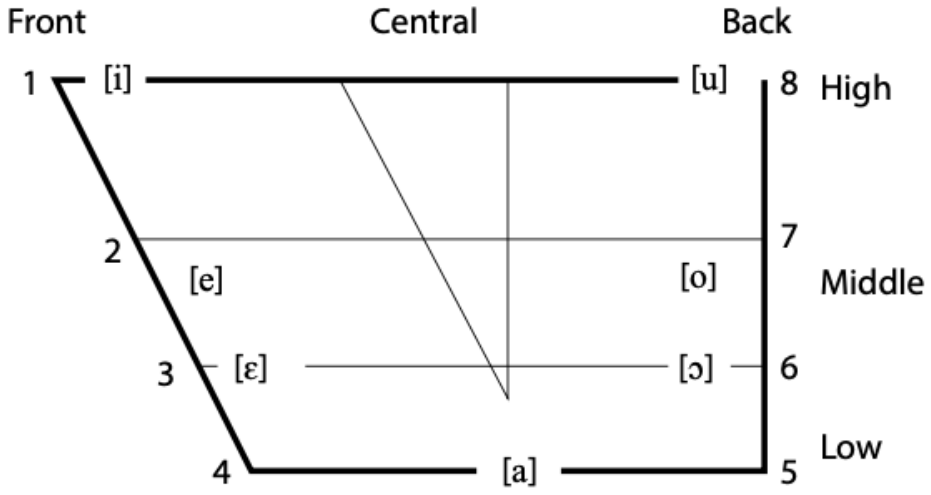


Figure 3: An illustration of the isiZulu vowel chart (Retrieved from Pascoe & Jeggo, 2019)

2.4.4 English vowels

The English language has a comparatively high number of vowel combinations and variations; Figure 4 provides a relatively comprehensive list of the vowels in the language. The English language vowels make use of the descriptions “open” and “close” to describe position in the mouth. The front vowels include the [i] which is unrounded. The /ʌ/ is a central vowel whilst /ə/ is a central, unrounded vowel.

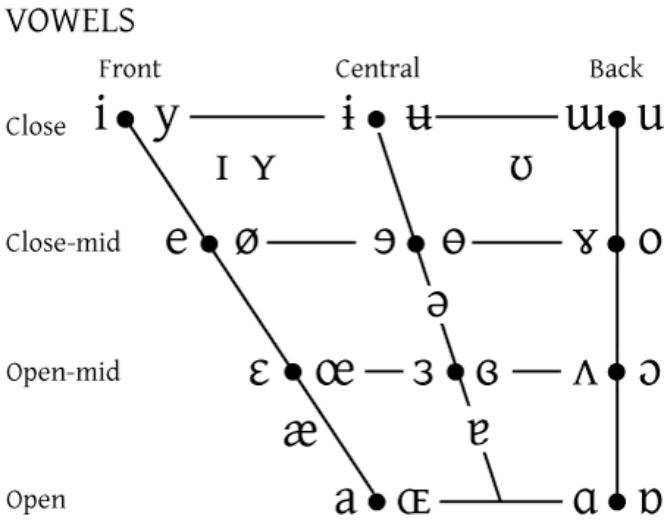


Figure 4: An illustration of English vowels

2.4.5 IsiZulu consonants

Consonants in isiZulu are based on 6 specifications.

a.) Manner of articulation

This applies to the modification of the airstream during articulation. Based on whether the airstream is obstructed or modified, four speech sounds are recognised, namely, explosives [p’],[t’], [k’], implosives [ɓ], fricatives [f], [s], [h], resonants [w], [m] and affricates [/t͡ʃ/].

b.) Direction of the airstream

The airstream moves either outward or inwards into the oral cavity. The speech sound that occurs when the airstream moves into the cavity is the click [!], [/], [/ /] and the implosive sound [ɓ]. The remaining sounds are eggressive.

c.) Origin of airstream

The origin of airstream is either an oral [l] or glottal [t'].

d.) Channel of air release

The two channels of air release are oral [l] and nasal [n]. Nasal sounds are articulated when the airstream moves through the nose.

e.) State of the glottis

This refers to the opening between the two vocal cords during articulation. The common types of sound that can be produced are voiced, voiceless, devocalised and aspirated consonants.

f.) Place of articulation

This refers to the tract where articulation occurs. Place of articulation consists of velars (occurs by raising the back of the tongue to touch the soft palate), bilabials (sounds which use both lips), alveolar (occur when the tip of the tongue is raised to touch the tooth ridge), palatals (the sound takes place at the oral roof region found behind the alveolar ridge), labiodental (produced by raising the lower lip to touch the upper part of the front teeth) and alveolar laterals (See Figure 5).

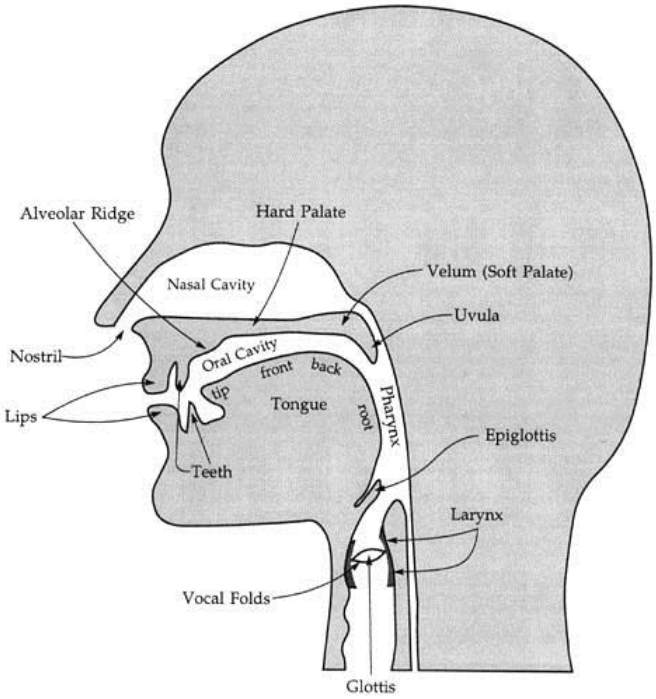


Figure 5: An illustration of the places of articulation in the vocal tract

Phonemes can be either voiced or voiceless. Voicing takes place in the larynx where the vocal cords exist. The meaning of an isiZulu word can easily be altered or be regarded meaningless by simply changing one sound to another. For example: “ukuthula” [to be quiet] and “ukuthola” [to find] (Ncobela, 2015). When an isiZulu sound is uncommon to an L1 English speaker, individuals often replace the sound with a similar English sound. This is often noted with the implosive [ɓ] that is uncommon in English. L2 speakers will likely replace the implosive with a voiced bilabial explosive [b]. Therefore, incorrect pronunciation of an isiZulu word could result in a change of meaning to the word or could mean that the word is meaningless. These incorrect pronunciations most commonly occur in L2 learners.

2.4.6 English consonants

English consonants are similar to isiZulu consonants, except there are some additional challenging sounds. English consonants consist of velar stops such as [k] which is a voiceless, velar, and plosive. Similarly, [g] is a voiced, velar, plosive. Alveolar stops are also present in sounds such as [t] which is a voiceless, alveolar, plosive. There is also the bilabial stop which is found in sounds such as [p] a voiceless, bilabial, stop and [b] which is a voiced, bilabial,

plosive. We also have alveolar which is found in the voiced, alveolar, fricative [z] and in [f] a voiceless, labiodental fricative.

The size of a language's phoneme inventory may also affect perception. When comparing English vowels between two L2 cohorts with different L1 studies showed that English learners with more limited L1 vowel inventories such as Spanish and French resulted in poorer recognition and performance compared to L1 speakers with bigger vowel inventories such as German (Iverson & Evans, 2009). The results of this study contradict previous work which stated that individuals with a smaller L1 phoneme inventory might be at an advantage as they have space to learn new L2 categories (Flege et al., 1995).

There are certain sounds which do not appear in English and are therefore replaced by an equivalent English sound in the pronunciation of isiZulu words. One of these sounds which does not appear in English- and hence results in mispronunciations is the implosive bilabial consonant [ɓ] (Naidoo, 2002). The implosive is, therefore, often compensated by a [b] in isiZulu (Ncobela, 2015). Similarly, the alveo-lateral fricative [ɬ] is unfamiliar in English, L2 isiZulu speakers often replace this sound with a voiceless, alveolar, fricative [s] (Ncobela, 2015). These incorrect pronunciations can therefore result in meaning changes. Shedding light on the phonetic comparison between isiZulu and English allows us to understand the difficulties that may arise for L2 speakers in their perception of word sounds. When a particular word arises in the L2 that does not occur in a person's L1, this may result in mispronunciation or auditory perception of the word.

When looking at phonemes in the phonological environment, it is important to consider the coarticulation feature. When phonemes appear next to each other, they are influenced on both the visual and acoustic level by the previous and forthcoming phonemes (Benguerel & Pichora-Fuller, 1982). Therefore, the prior viseme can hint at the next viseme. Coarticulation, however, can hinder lip reading in visual-only conditions particularly when the articulatory gestures are less salient (Benguerel & Pichora-Fuller, 1982). However, more recent studies have revealed that audio-visual input enhances an individual's potential to predict approaching sounds (Campbell, 2007; Howson et al; 2020;). Therefore, it can be said that audio-visual facilitation on the phoneme level impacts lexical activation revealing visual phonemic cues in speech to assist in retrieving lexical entries (Fort et al., 2013).

2.4.7 Minimal pairs

A minimal pair can be described as a set of two words which have different lexical meanings owing to a difference in one sound (Pennington, 1999). Minimal pairs are often used to demonstrate the meaning confusion of words. A study referenced by Nunes (2015) used minimal pairs in a study comparing English and Portuguese vowel and consonant sound systems. Minimal pairs were used to illustrate meaning confusion relating to the Brazilian speaker's mispronunciation of vowels. A study looking at the sound systems most difficult for Brazilian learners to grasp centred around the consonants /t/ and /d/ as there was an expected confusion in words such as "deep" and "jeep" (Baptista, 2001).

2.4.8 Models of lexical access

It is essential to examine the research underpinning bilingual perception. The more localist model BIMOLA (Grosjean, 2012) has a foundation in earlier activation models such as the activation model of (McClelland & Rumelhart, 1981) and the TRACE model of (McClelland & Elman, 1986). The objective of the BIMOLA model (Grosjean, 2012) deals with handling the recognition of spoken words. According to BIMOLA, there are no distinct levels of language nodes. Rather, the lexicons of various languages are stored in a variety of language networks. This model is, therefore, based on word recognition in bilinguals. BIMOLA states that there are three node levels, namely: features, phonemes, and words (McClelland & Rumelhart, 1981). According to this model, the feature level is a commonality between both languages however, the phoneme and word levels are broken into subsets based on the language. This highlights that features activate words.

The activation that occurs between phonemes and words is thus bidirectional however, the activation of the feature level and phoneme level are bottom-up (McClelland & Rumelhart, 1981). In this model, language is contained in distinct networks from the phoneme level through to the word level. Languages are thus arranged separately within the phoneme and word levels and can be hindered between levels of one language only. Therefore, languages do not correspond with each other. In this model words and consonants belonging to a particular language will increase the overall activation of the language network and hasten word recognition in that language (McClelland & Rumelhart, 1981). Figure 6 below illustrates the

BIMOLA model; as can be seen in this figure, BIMOLA maintains both a bottom-up and top-down processing. Overall, BIMOLA is a traditional approach to bilingual speech processing.

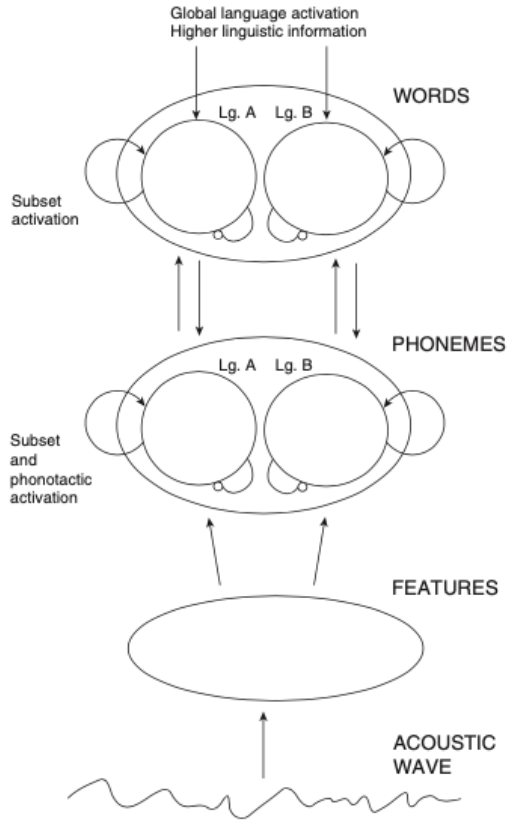


Figure 6: An illustration of the BIMOLA model (adapted from Grosjean, 2012)

The BLINCS model concerns interconnected levels of representations which are phonological, phono-lexical, ortho-lexical and semantic. This model states that the influence of visual stimuli on language processes occurs through the phonological and semantic levels. Exposure to spoken language automatically involves the integration of audio and visual information (Gentilucci & Cattaneo, 2005). It is at the lexical level that the BLINCS model presumes that a person's two languages are both separated and integrated. The phono-lexical level is where words are separated into language regions however the words are still able to interact between themselves. The distinction between the two language orthographies is highlighted at the ortho-lexical level (Cornut, et al., 2021; Shook & Marian, 2012). Therefore it can be seen that the BLINCS model was not created to account for visual recognition rather it was established to centre around the effects that information gathered through the visual systems might have on age language processing. Figure 7 is a visual illustration of the BLINCS model.

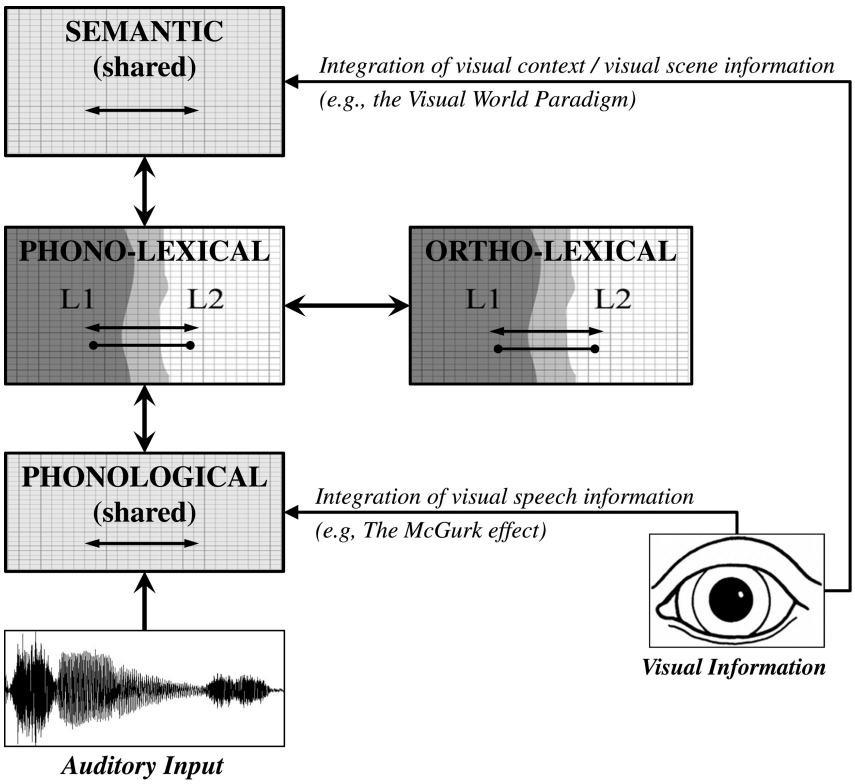


Figure 7: Visual representation of the BLINCS model (retrieved from Shook & Marian, 2012)

These models are important to understand, as they describe how individuals recognise lexical item meaning. This provides insight to how people access words and word meanings.

2.5 The development of audio-visual perception

At the age of 2.5 months, infants are sensitive to the congruency of facial articulation and voice, according to research by Dodd (1979) and Burnham and Dodd (2004). Infants as young as two months old can match auditory and visual cues. Infants are presented with a silent video of two identical faces side-by-side, (one with the visual articulation of one syllable and the other with the visual articulation of a different syllable). Later when the infants are presented with the same video of two faces adding the sound of one of the syllables, infants preferentially look at the side where the visual articulation is matched with the sound (Patterson & Werker, 2002, 2003). In addition to cross-modal matching. Prince and Hollich (2005) discovered an audio-visual benefit effect in 7.5-month-old infants, indicating that visual information aids infants' auditory speech perception. Infants, like adults, demonstrated the McGurk effect, according to

research carried out by Rosenblum et al. (2007) and Burnham and Dodd (2004). In the Burnham and Dodd (2004) study, it was discovered that in the test phase, infants in the experimental group fixated on the auditory [da] or [ða] for a substantially longer time than auditory [ba], suggesting that participants in the experimental group recognised the McGurk stimulus (auditory [ba] accompanied by visual [ga]) as auditory [da] or [ða] rather than [ba] (Burnham & Dodd, 2004). In summary, infant audio-visual speech perception research discovered that: (1) infants are sensitive to auditory and visual synchrony; (2) infants can use visual cues to aid auditory speech perception, and (3) infants exhibit the McGurk effect.

Developmental studies of audio-visual speech perception have repeatedly found that children have a weaker visual influence than adolescents and adults. Dupont et al. (2005) recruited eight 4- and 5-year-old French Canadian children and ten adults as participants, who were shown congruent and incongruent visual, auditory, and audio-visual stimuli. Adults were required to write down the perceived utterances, while children were instructed to repeat what they believed the speaker had just said (Dupont et al., 2005). The results showed that approximately 90% of children's responses to audio-visual incongruent stimuli were identical with the auditory component of the audio-visual incongruent stimuli, while adults' audio percepts accounted for only about 40% of responses, implying that the adults were much more biased towards visual information than the children were (Dupont et al., 2005).

Sekiyama and Burnham (2008) enlisted native Japanese and English speakers as participants, who were divided into four age groups in both languages, namely groups of six-years-old, eight-year-old, eleven-year-olds and. After being presented with syllables in audio-visual (with congruent or mismatched auditory and visual components), audio-only, and video-only trials at various signal-to-noise levels, participants were asked to press one of three buttons for a “ba”, “da”, or “ga” response accurately and without delay (Sekiyama & Burnham, 2008). In summary, the strength of the McGurk effect was discovered to be adult-like in children over the age of ten. English-speaking children and Canadian French-speaking children under the age of ten have a weaker McGurk effect than adolescents and adults, and the degree of the McGurk effect increased gradually between the ages of four and nine. This increase may be most noticeable between the ages of six and eight years of age. However, the McGurk effect remains strong for Japanese speakers even after six years of age.

Visual articulators, primarily the lips, teeth, and tongue, aid in relaying important aspects of speech, such as the location of consonant and vowel articulation. Lip-reading combined with audition allows the listener with normal to comprehend speech more accurately, even when the acoustic information is undistorted (Summerfield, 1992). Visually impaired infants with normal hearing, for example, are frequently delayed in acquiring the phonetic differences that are easy to see but challenging to hear (such as /b/ and /d/). The visibility of the lips allows the listener with normal hearing to perceive speech more accurately (Summerfield, 1992). In circumstances of noise or hearing impairments, visual modalities play a complementary role. This allows individuals with normal-hearing to comprehend speech in conditions with a degraded signal.

Although the auditory modality is the primary modality for speech perception, the visual modality also assists in infant language development and improves speech perception accuracy in children and adults (Bergeson, et al., 2010). Based on the above information we can deduce that children have a good foundation of exposure from infancy to suprasegmental and segmental elements of auditory speech which allows them to acquire any language they are exposed to (Watson, et al., 2014).

2.6 Models of audio-visual integration

The exact stage in which audio and visual information are integrated is vague. The fuzzy logical model and the Revised Motor theory form the basis of the ideas of how audio and visual information integrate.

These models can be explained and broken down into three classes. The late integration model suggests that audio and visual cues are processed individually before integration. The early integration model however states that audio and visual signals are integrated during perception and they do not occur separately (Schroeder et al., 2008). van Wassenhove (2013) claims that the biggest limitation of the early and late models of integration is that they characterize audio-visual integration as unidirectional and one-dimensional. Following this critique, Zhang et al. (2018) recommended a hybrid multistage model which argues that audio and visual signals are processed at both the earlier and later stages. Multistage integration permits the early effects of visual cues on auditory processing and the integration of lower-level processing outputs at a

later stage (Zhang et al., 2018). Peelle and Sommers (2015) illustrated the models shown in Figure 8:

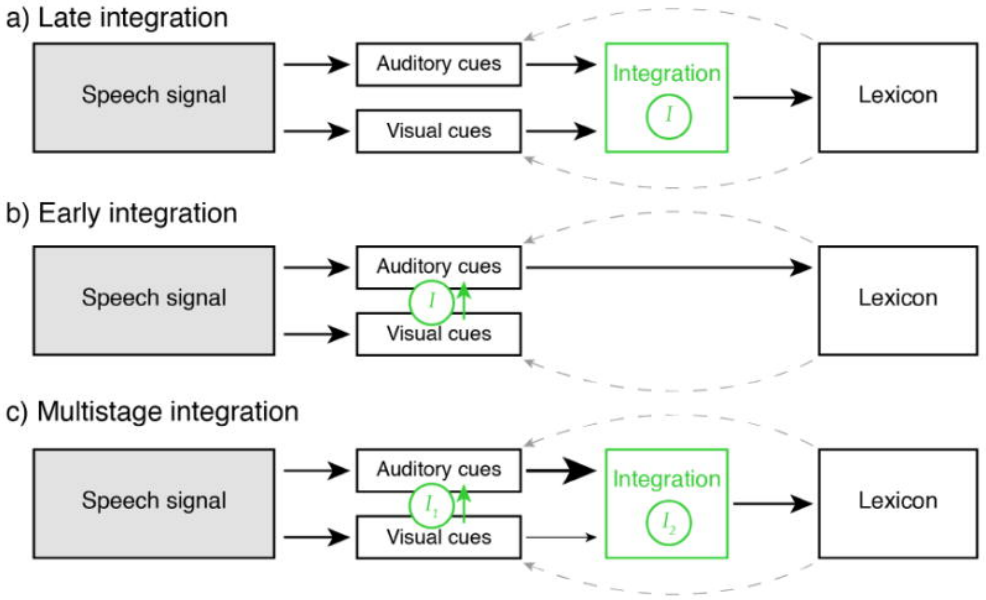


Figure 8: An illustration of audio-visual integration (Peelle & Sommers, 2015)

The above illustration outlines the modes of audio-visual integration. In summary, the early integration model states that integration takes place during perception, while the late integration model argues that audio-visual integration takes place after audio and visual cues are processed. In contrast, the hybrid, multistage poses that integration can take place at an earlier and or late stage (Peelle & Sommers, 2015).

2.7 Auditory perception

To understand the processes that utilise auditory and visual speech it is important to understand the use of each modality. The auditory perception is a fundamental aspect of how we perceive speech. Auditory perception is made up of listening and hearing and includes the detection, analysis and comprehension of sound. The auditory system modifies acoustic and mechanical signals into neural signals which results in the perception of a sound (Fuchs et al., 2016; Litovsky et al., 2010; Moore, 2012). The key elements of auditory stimuli are loudness, pitch and timbre (Fastl & Zwicker, 2007). Pitch perception occurs because of tonality, harmonicity and periodicity in acoustic signals. The timbre highlights different properties of sound which include onset and offset transients and the strength of harmonics. The basic unit of audio speech is phonemes. Phonemes are the smallest linguistic units that carry meaning.

With the help of Gestalt principles (Klapp & Jagacinski, 2011) the auditory system breaks down these units into individual streams. Therefore, when we hear a stimulus it is not the individual properties such as timbre that we hear but rather the auditory object.

The auditory modality offers L2 listeners with consonant identification cues. The identification cues include the place and manner of articulation, and voicing. The place of articulation provides information about the place in the vocal tract where the sound is produced. The manner of articulation concerns the method in which the airstream is changed when producing the sound. The voicing relates to vocal fold vibration during speech production. These cues are the information provided by auditory stimuli.

2.8 What can be understood from visual modality in speech?

2.8.1 Visual cues

The key link between auditory speech and visual cues is the change of shape of the mouth based on the basal structure of the uttered word (Bahal, 2012). The central purpose of visual cues is to complement and cross-verify the auditory signal from the speaker (Altieri, 2014; Peelle and Sommers, 2015). Therefore, visual cues are said to have three essential roles in the perception of speech. According to Cusak et al. (2004), the first role is that visual signal aid in localising the speaker, which allows the listener to situate their attention. Second, these cues play a role in supplying contextual knowledge about the place of articulation (Peelle & Sommers, 2015). Thirdly, it aids in temporal information about the speech signal, increasing the accuracy of anticipating the auditory signal (Peelle & Davies, 2012). Lip-reading is beneficial to the majority of people ranging from those with normal hearing to hearing-impaired people as it compensates for insufficient auditory cues (Summerfield, 1992).

The benefit of audio-visual modality, however, is largely due to the perceptual weighting of audio and visual modalities (Hazan et al., 2006). A study that assessed this weighting was conducted by McGurk and McDonald (1976). Their study explored audio-visual perception consisting of incongruent stimuli. The stimuli were made up of CV syllables with a stop consonant (/p/,/g/,/k/,/b/) followed by a vowel /a/. The stimuli were then made to be incongruent which was represented as an audio [ba] with a visual [ga], an audio [ga] with a visual [ba], an audio [pa] with a visual [ka] and an audio ka with a visual pa. The study included preschool, primary school and adult participants. The results revealed when the audio was

[baba] and the visual was [gaga], in the incongruent condition, some stimuli, [dada] was perceived while with other stimuli such as audio [gaga] and a visual [baba] saw participants perceive [gabga] (McGurk & McDonald, 1976). These results suggests that audio and visual modalities of speech influence speech perception.

Using the McGurk effect, various other studies looked into the effect of visual cues on L1 and L2 contrasts. Studies looking into L1 contrasts (Traunmüller & Öhström, 2007), explored Swedish vowels, while exploration into Japanese syllables was undertaken by Sekiyama and Tohkura (1991). Their study concerned the perception of L2 contrasts, of Chinese and Japanese syllables by Chinese and Japanese speakers (Sekiyama, 1997), Hazan et al. (2006) explored English consonants by Spanish and Japanese speakers while Hazan et al. (2005) looked at English consonants Spanish learners of English. These studies have shown that L1 and L2 listeners depend on visual information in order to distinguish L1 and L2 speech contrasts.

Hazan and colleagues (2006) looked at the effects of visual cues on the perception of non-native consonants. In the first experiment, they used labial (/b/-/v/) and labiodental (/v/) to understand the perception of English phonemic contrasts. In the second experiment, they made use of a less visual contrast with /r/ and /l/. Both experiments were tested on Japanese, Spanish and Korean learners of English. The test took place across three conditions (1) video-only, (2) audio-only and (3) audio-visual. The results of Experiment 1 showed that participants attained better scores in the audio-visual condition highlighting the audio-visual benefit while in Experiment 2, results showed there was no audio-visual benefit for either group (Hazan et al., 2006).

Based on neuroimaging research, the temporal lobes Herschl's gyrus is located in the temporal lobe and is responsible for processing acoustic signals together with visual processing areas which are found in the occipital lobe (Kyle et al., 2013). The posterior superior temporal sulcus (pSTS) has been linked with the integration of multimodal input such as audio-visual speech (Kyle et al., 2013). Seeing a talking face would stimulate the visual movement cortex in the occipitotemporal lobe whilst watching someone talk would stimulate left-lateralized regions which indicates that articulatory movements are processed similarly to language-specific movements (Kyle et al., 2013). Spatial resolution does not have to be high to extract useful visual cues for a face to be seen (Munhall et al., 2004).

Rosenblum and Saldaña (1996) established that speech perception can be facilitated simply by a face marked with a single illuminated point at anchor places on the lips and bottom half of the face. At the suprasegmental level audio-visual speech also facilitates the processing. At the acoustic level, lexical tones are systemic variations of the fundamental frequency that have the ability to alter the meaning of a word (Ong et al., 2017). In a study carried out by Smith and Burnham (2012) with Mandarin Chinese lexical tones, both native and non-native individuals were able to tell apart lexical tones in both clear and degraded speech conditions. However, the study revealed that visual cues enabled durational information of the tones. This shows that together with improving speech perception on various levels (Smith & Burnham, 2012) visual cues in speech are not confined to language-specific elements and could be useful in L2 acquisition. The present study will make use of a visual-only condition to investigate L2 speakers perception of non-native visual cues.

2.8.2 Effects of visual modality on perception

Unlike the abundance of studies carried out on the auditory modality of speech perception, the visual modality of audio-visual speech perception has been limited. Adults can tell two languages apart based on the visual cues available to them alone even in instances when the two languages share similar properties such as Spanish and Catalan (Mazzaro & Calantoni, 2016). This could be because adults have years of exposure in their native language/es and are attuned to the visual and acoustic properties of speech. In instances of being exposed to badly dubbed acoustics, the irregularities between the audio modality and visual modality are evident to adult speech perceivers, however, long-term exposure to asynchronous speech could result in adult speakers becoming less sensitive to this irregularity (Navarra et al; 2005). The use of audio and visual modalities to perceive speech is not restricted to pre-recorded stimuli only. Rather, adults utilise both modalities to understand auditory speech events (Grant & Seitz, 2000; Ross & Tremblay, 2009). This is particularly the case in noisy environments (Sumby & Pollack, 1954), when exposed to foreign language (Navarra & Soto-Franco, 2005) and complex semantic circumstances (Reisberg et al., 1987).

According to Bahal (2012), the main distinction between visual cues and auditory speech is that the shape of the mouth differs depending on the word's basal structure. It has been suggested that the primary purpose of visual cues is to cross-check and complement auditory information (Peelle & Sommers, 2015). It is thought to serve three key roles in speech

perception Firstly, it aids the process of localising a speaker, which provides an opportunity to modify listening to correspond with the speaker's location (Carlyon et al., 2001). Furthermore, it presents extra background, contextual, and environmental information concerning the place of articulation (Peelle & Sommers, 2015). Lastly, it improves the accuracy of predicting the acoustic signal by supplying temporal information about the speech signal (Peelle & Davis, 2012).

Munhall and colleagues (2004) reveal that the visual modalities in speech perception provide supplementary information that allows individuals to recognise the place of articulation of phonemes such as /b/ and /d/. In instances where auditory recognition might be ambiguous the visual modality assists in differentiating phonemes such as /b/ and /w/ in their articulation (Bernstein et al., 2010). When phonemes are articulated they correlate with particular changes in the mouth shape which cover the movement of articulators such as the jaw, lip, and tongue and which account for the visual cues. Some phonemes, however, are difficult to tell apart visually and can only be distinguished vocally; such phonemes in English are /p, b, m/ and /d, t/ (Liedstam et al., 2006; Màrtony, 1974). These phonemes are all bilabial consonants and share different properties acoustically. However, they look similar visually (Brahme & Bhadade, 2016; Mattheyses et al., 2013).

Phonemes that share similar visual movements are known as visemes (Fisher, 1968). A viseme can be described as the visual representation of a group of phonemes. Visemes have been studied extensively in various languages such as English (Bozkurt et al., 2011), German (Mattheyses et al., 2013), Italian (Cosi et al., 1996), French (Mahdi et al., 2008) and Swedish (Nasir et al., 2020). However, only limited research into isiZulu viseme classification has been carried out. An isiZulu viseme classification would be able to highlight speech sounds that look the same visually, such lip reading. The above information shows that visual information in speech perception can be complementary to the perception of speech.

2.8.3 Lip movements in speech perception

A key element of audio-visual speech concerns the time of articulatory cues and the equivalent auditory cue. Articulatory movements dominate the acoustic cues for a listener. Therefore, it can be said that visual information is seen before the auditory output is heard even though both modalities are integrated into one occurrence.

All elements of speech (including smaller units of units of speech such as phonemes or large units such as words) can be recognised visually (Bernstein & Liebenthal, 2014). However, the occurrence of using visual modality alone to identify speech may differ greatly among individuals (Bernstein & Liebenthal, 2014). Phonemes which sound similar can be identified in the audio modality owing to the different acoustic properties in voicing, place and manner or articulation. An example of this can be seen in the syllables /pa:/ and /ba/ (Peelle & Sommers, 2015). These two syllables share a manner of articulation as they are both plosive. However, the difference in these two syllables is noted when /pa:/ is voiceless but /ba:/ is voiced this indicates that the syllables sound different. However they belong to the same viseme group indicating that from a visual perspective, them might seem very similar to each other. Therefore we can identify a disadvantage in the visual-only modality as it cannot transfer important properties such as the vibration of the vocal cords. Rather it only provides information on the place of articulation (Sekiyama, 1997).

2.8.4 Variation in lip reading

An individual's lip reading ability may vary owing to a range of circumstances. People with congenital hearing loss may possess better lip-reading abilities because of their dependence on visual information. Information in speech including phoneme units and words can be recognized visually (Bernstein & Liebenthal, 2014).

Bernstein and colleagues (2000) carried out a study which contrasted the ability of people with normal hearing and people with impaired hearing to recognise words and sentences (Bernstein et al., 2000). Adults with hearing impairment outperformed adults with normal hearing in both tasks. This reveals that there is a variation in lip reading ability in adults with normal hearing versus those with a hearing impairment. Although the above study (Bernstein et al., 2000) reported that adults with hearing impairment were competent lip-readers the majority of participants were not able to accurately recognise phonemes. Using the foundations of the latter study Auer and Bernstein (2007) used a large sample size and concluded that hearing-impaired individuals recognized 43% of words in sentences compared to normal-hearing individuals who only recognised 18% of words in sentences. However, it should be noted that the lip-reading ability depends on the complexity of the stimuli presented (Altieri et al., 2011; Tye-Murray et al., 2014).

Lip-reading assists people in word recognition. A study comparing three different conditions- an audio-only condition, an audio-visual condition and a condition, where the speaker's mouth was covered revealed that English L1 speakers benefited the most in the audio-visual condition whilst the lowest word identification was in the audio-only condition (Aubanel et al., 2016). Studies have shown that visual cues increase word identification. Similarly, the study of Drijvers and Özyürek (2017) study concluded that articulatory gestures assist with word comprehension in L1 listeners. In addition to the above-mentioned advantages, gazing at the articulatory gestures from speakers decreases lexical competition which contributes to word recognition. Cunillera et al. (2010) state that audio-visual input enhances segmentation when the visual and auditory information components are provided simultaneously. However perfect synchronisation is not necessary.

Being able to recognise speech from a visual modality alone differs greatly between individuals (Bernstein & Liebenthal, 2014). Studies showed that listeners with normal hearing and hearing impairment were able to recognise speech sounds in quiet conditions (Gatehouse & Gordon, 1990; Liu & Zeng, 2006; Ross et al., 2011). However, the extent to which individuals can recognise speech sounds from a visual alone is where the variation lies (Bernstein & Liebenthal, 2014).

Numerous cognitive skills have been linked with lip-reading including working memory and processing speed of verbal and spatial information (Feld & Sommers, 2009). The study carried out by Feld and Sommers (2009) concluded that younger adults were more proficient in lip-reading sentences compared to older adults as they had quicker processing speed and better working memory. Similarly, the effect of gender on lip-reading abilities has also been studied and has yielded inconsistent evidence. A study carried out by Dancer et al. (1994) as cited by Tye Murray et al., 2007 stated that females had better lip-reading abilities than males however further research revealed that this evidence was conflicting as gender did not greatly affect lip-reading capabilities (Auer et al., 2007). This shows that the true capabilities of visual information on speech perception have not been fully explored. Understanding the true effect of visual modalities on speech perception is essential, particularly in non-native speech speakers.

Although visual cues are important, it is important to realize that visual modality alone has limited use in speech perception (Stacey et al., 2016). The study carried out by Grant et al.

(1998) showed that in a visual-only condition, the identification of sentences ranged between 0% to 20% while audio-visual scores ranged between 23% and 94%. For the audio-only condition, scores ranged between 5% and 70% in the audio-only condition. This shows that in order to access the visual advantages of speech a significant amount of auditory information is needed to enable understanding of the visual cues (Grant & Seitz, 2000).

It is understood that the L1 of a speaker will affect the visual interpretation of English vowels (McGuire & Babel, 2012). Visemes are classified based on the phonemes they portray. Pattamadilok and Sato (2022), provided the following visual to illustrate the visemes and their relevant phoneme (See Figure 9).

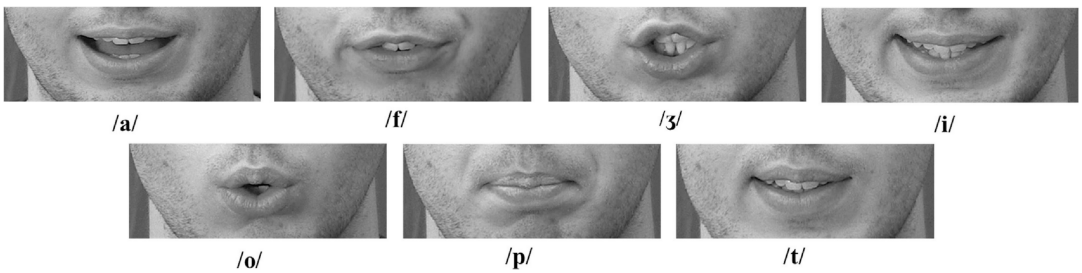


Figure 9: An illustration of visemes and their relevant phonemes (Pattamadilok & Sato, 2022)

2.8.5 Visual speech in quiet environments

Research has shown that in a quiet environment people have a performance advantage when introduced to congruent auditory and visual information. Studies have also revealed that people's performance shows improvements when they can see the talker's face (Reisberg et al., 1987). The results of other studies (Arnold & Hill, 2001) corresponded and show that a person is advantaged when they can see a person can see a speaker's face rather than simply hearing the auditory voice. This verifies the above studies in showing that listeners do indeed benefit from visual cues presented to them particularly when the need for language comprehension is more strenuous. Therefore, visual cues serve to aid in the perception of speech as aids in processes such as phoneme recognition. Other studies, however, have yielded results showing that visual cues do not greatly impact the perception of speech. In the study by Jesse et al. (2000), although participants had better results in the identification of sentences in the audio-visual modality compared to the auditory-only modality there were no crucial differences (as the perception was generally high). However, Jesse et al. (2000) stated that no

crucial difference could be due to task difficulty, increasing the task difficulty could have yielded different results.

A study by Erdener and Burnham (2005) also showed that speakers of different languages such as English, French and Spanish were better at recognising syllables in an audio-visual condition compared to an auditory-only condition. Some studies have demonstrated that (compared to audio-only stimulus) the multi-modal use of audio and visual modalities improves the ease of listening (Arbona et al., 2023; Zhyrun, 2017). Other studies have revealed that visual modalities can supplement audio speech cues (Ginther, 2002). This shows that visual cues provide supplementary information to listeners when perceiving language and that visual cues do not go unacknowledged in the perception of speech.

2.8.6 Variation of visual modality usage in perception

The native language of a speaker may also affect the extent to which they make use of visual cues. It has been argued that when adults converse with one another they appear to concentrate on their social partner's eyes which provides them with social cues (Birmingham et al., 2008). However, when speakers are met with obstacles in their audio-visual speech (such as ambiguities) they tend to focus on the talker's mouth (Lansing & McConkie, 2003). A study by Sekiyama and Tohkura (1993) revealed that Japanese speakers made limited use of visual cues. This is despite Hardison (2003) concluding that Japanese speakers were affected by the McGurk effect. Similarly, Yehia et al. (2002) maintained in their study that obstacles such as noise resulted in listeners fixating on the speaker's mouth which promoted speech processing. Lansing and McConkie (2003) influenced the auditory cues in their study through changes in volume and studied the listener's comprehension while recording their eye gaze. When the speech was silent, the listener gazed at the speaker's eyes however, when the speech was influenced by noise they gazed at the mouth of the speaker. In a study carried out by Thompson and Malloy (2004), it was found that older adults (aged between 65 and 75 years) had a greater dependence on visual features as it compensated for slower perception owing to factors such as hearing loss.

Hazan and colleagues (2006) looked at Japanese and Korean participants who were learners of English on the /l/-/r/visually salient contrast. There was no major visual benefit noted for either group. However, the Korean participants' audio and audio-visual results were better than the visual results while Japanese participants noted poor results across all three conditions (audio,

visual and audio-visual). These results reflected the L1 of the speaker and the salience of visual cues in L2 contrasts.

More recently, studies have shown that people alter their attention based on the task they need to carry out. For example, for the need to read emotions the speaker's eyes will be the main point of fixation, while to read noise, a speaker's mouth will be the main point of fixation (Buchan et al., 2008). Lusk and Mitchel's (2016) study exposed participants to an artificial language through an audio-visual modality. Their study concluded that as familiarity grew with new words the listener decreased their attention to the mouth. Barenholtz et al. (2016) looked to see whether auditory sentences matched with audio-visual sentences participants heard in both their native and non-native languages. The results revealed that attention was overall placed on the speakers mouth and the fixation on the mouth increased when in the non-native stimulus (Barenholtz et al., 2016). The information above (2.7.6) discusses how individuals vary their fixation on the visual cues presented to them during different speech conditions.

2.9 Audio-visual perception

Lewkowicz & Ghazanfar (2009) researched the development of audio-visual speech progressing from lower- to higher-level cues. They found that as children develop, their perception and processing of perceptual cues improve. At the lowest level, audio-visual speech is recognised by concurrent on and offsets of auditory and visual attributes. It is at the higher levels where categorical amodal characteristics exist. Examples of these include the talker's gender.

Communication of speech most frequently occurs face-to-face. In this way, visual cues are paired with speech, which is then available to the listener, these visual cues carry several benefits (Girin et al., 2001) as they are resistant to acoustic noise. Similarly, these visual cues notify the listener of the precise temporal onset of speech sounds. This is helpful in informational masking as it permits the listener to centere on the speech sound that is matched with the onset of visual speech production (Macalusoet al., 2004; Vatakis & Spence, 2006). Lastly, visual signals supply basic phonemic clues of speech sounds. Various authors have reported that a large number of consonants and vowels can be identified based on the shapes of speech articulators that are needed to produce the sound (Schwartz et al., 2004).

Audio and visual aspects of speech, such as rhythm and intensity share certain attributes. However, other aspects also link them. This means that some visual articulatory gestures correlate with certain acoustic features. The glimpse of the mouth opening the vocal articulators (tongue, teeth, lips, palate) and the narrowing of the vocal tract, are linked to the acoustic structures of vowels. When the mouth closes and the articulators limit the flow of air in the vocal tract can be linked to another acoustic property of consonants. However, this link extends beyond the above explanation. Mouth aperture for example affects aspects including the range frequencies, with high spectral peaks giving vowels with their specific attributes. However, the capability of accurately perceiving speech sounds through visual stimuli differs according to the environment that the perception occurs. A study revealed that young adults with no known hearing impairment had roughly 50 % accuracy of perception (Summerfield, 1992). Nevertheless, these young adults continue to utilise contextual information together with speech reading which enables improved speech perception (Matthews et al., 2002),

The process of understanding which modality is perceived first and at which state the perception occurs saw the development of various theoretical frameworks (Peelle & Sommers, 2015). Massaro (1989), introduced the fuzzy logical model of perception. This model reveals that audio and visual modalities at the beginning are perceived individually at the beginning and are then combined at a later stage. Previous integration models highlighted that the auditory and visual signals link at the beginning processing stage and that perception occurs from premature experience in a language (van Wassenhove et al., 2007). However, recent studies proceeded to put forward a fused approach combining various integration stages (Peelle & Sommers, 2015).

The analysis-by-synthesis method was produced through the work of Hallé and Stevens (1962). This method argues that incoming speech input is diligently analysed through synthesising the production process based on stored phonemic representations. Another model was built from the natural dynamics of speech stating that the visible articulatory gesture comes before the acoustic signaling (van Wassenhove et al, 2007). This model states that visual input forms a projection which is examined against the acoustic cue. When the visual is highly salient, it authorises a secure prediction concerning the forthcoming acoustic cue. Therefore, visual input speeds up the perception process and makes it easier to recognise phonemes in speech when

the visual and auditory information in the speech signal are congruent and within the acceptable desynchronisation time range.

2.9.1 The development of audio-visual integration

Audio-visual integration occurs at the merging of auditory signals and visual cues such as lip movement which form a well-organized representation. The time at which the integration of audio-visual cues is learnt is divided into two perspectives. The first perspective accounts for audio-visual integration that is present in an infant (Aldridge et al., 1999; Bahrick et al., 2005; Kuhl & Meltzoff, 1996; Patterson & Werker, 2003), while the second perspective refers to audio-visual integration, evolves throughout an individual's life through learning and experience (Bejjanki et al., 2011; Jansen et al., 2013; Ross et al., 2011; Sowell et al., 2004). While the majority of studies substantiate for the latter perspective, this evolution, however, remains unclear.

The vagueness of the development of audio-visual perception has resulted in the development of behavioural and neural studies (including electrophysiological methods and functional neuroimaging). Evidence supporting the idea that audio-visual perception is present in infancy has been based on both behavioural and neurological research. Patterson and Werker (2003) made use of isolated vowels to show a link between auditory and visual structures and established that children as young as two months of age can match up phonetic vowel details with the accuracy of articulation through facial presentation. Contrasting evidence shows that visual articulatory speech has a limited effect on speech perception in school-aged children (Jerger et al., 2009).

To add to this research, neural studies were conducted which observed the cortical anatomy in the perisylvian language area and found that this cortical area went through an extended developmental path suggesting that audio and visual systems evolve over time (Sowell et al, 2004). The findings from Sowell et al., (2004) were expanded by Jansen et al. (2013) who found that a completely developed audio-visual integration relies on factors such as vision and cognition. The findings from the author's study (Jansen et al., 2013), disclosed that in developing adults these factors are fully developed. In comparison, observations on developing children showed that even though their visual and auditory modalities are accounted for their brains continue to develop revealing that the fusion of modalities is incomplete. This evidence

shows that the neural connections of audio and visual pathways do follow a developmental pathway.

To our knowledge, the only study that examined the attention of native and non-native speakers was carried out by Barenholtz et al., (2016). Their study highlighted that people concentrated most of their attention on a speaker's mouth during the perception of their native speech however, the fixation time on the mouth increased when participants were exposed to non-native speech cues.

2.9.2 The perception of connected speech

Owing to various reduction procedures that occur frequently in speech, casual speech with pronunciation variations and ambiguities presents a greater listening challenge to non-native listeners than precisely pronounced speech (Broersma & Scharenborg, 2010). The authors (Broersma & Scharenborg, 2010) assert that native English speakers adhere to the simplicity of articulation principles in which distinctions between segments (consonants or vowels) are kept to a minimum. As a result, connected speech is mostly composed of reduced forms in which sounds are deleted or changed.

Examples of connected speech include the phonological process of assimilation (changing sounds to complement neighbouring sounds, elision (the deletion of sounds), and contraction (the compression of word forms). Non-native learners face difficulties with these reduced forms, particularly when their first and second languages do not share similar properties in reduced variants of pronunciation (Mitterer & Tuinman 2012). This difficulty has however mostly been demonstrated in a noise-free laboratory setting (Gaskell & Snoeren, 2008; Mitterer & Tuinman, 2012) and has received minimal attention in less ideal listening conditions such as multi-talker noise.

Listeners frequently face unfavourable listening conditions in everyday situations, in which auditory, acoustic, or linguistic signals relevant to speech recognition are deteriorated or are completely lost. These settings impair speech perception in native English speakers, but they are even more challenging for non-native English speakers. Identifying the impact of these unfavourable listening conditions on speech perception is critical for improving the effectiveness of second language-connected speech perception (Vlasenk et al., 2012).

2.9.3 Audio-visual illusions

To gain a better idea of how our senses work alongside each other, it is useful to determine what influences the perception of audio-visual illusions. In their review study, Shams and Kim (2010) observed in a review that vision is frequently regarded as the dominant sense. This was evidenced by various academic articles as outlined by (Shimojo, 2001). However, the factors that determine which senses outperforms other senses are not fully understood. Whether audio vision predominates is context-dependent (Walker & Scott, 1981) and can vary depending on the task demands (Robinson et al., 2016). Walker and Scott (1981) discovered that when an auditory stimulus (tone) was presented separately from a visual stimulus (light), the tone was perceived for longer than the light. When the tone and light were presented simultaneously, the light was perceived for the same duration as the tone when it was presented alone. This suggests that the duration of auditory stimuli influenced the temporal decisions of the visual stimuli. Similarly, it was found that if one modality was degraded during the perception process, a person could shift the modality to the more reliable one which would potentially influence perception.

2.9.4 Context and audio-visual integration

Another factor that may influence audio-visual speech perception is the situation in which perception occurs. When presented with additional contextual information, the perception of spoken signals improves phonologically, lexically, and grammatically. Listeners, for example, use visual information more effectively when it is presented alongside a meaningful utterance than when it is presented together with a nonsensical one. Furthermore, Van Engen et al. (2014) revealed that when the recognition of semantically normal and anomalous sentences was studied with four types of auditory masking, the semantic context was beneficial across all types of maskers (via both auditory and audio-visual modalities). However, the developmental effect of context usage on audio-visual speech perception has not been thoroughly researched.

2.9.5 Non-speech audio-visual illusions

Speech perception evolves throughout life, and grows, and matures from childhood to adulthood (Burnham et al., 1991). In everyday life, multisensory information from an

individual's surroundings assists in forming a coherent perception of the world. Visual and auditory information, in particular, often convey consistent information; for instance, when an individual is playing tennis and simultaneously hears the ball hit a racket. A perceptual illusion can occur when incongruent information from different modalities is presented simultaneously. This was demonstrated in the sound-induced flash illusion, in which viewers perceive a single flash of light as a double flash if it coincides with two auditory beeps (Shams et al., 2000). Similarly, auditory perception can also be hampered by visual information. These findings have sparked debate about the nature of the brain's sensory pathways and how multisensory information combines to produce unified perspectives. Since other modalities can influence sensory pathways, they can be cross-modal (Shams & Kim, 2010; Shimojo, 2001).

2.9.6 Incongruent audio-visual speech in children

Infants are also responsive to the correlation of auditory and visual speech, and this was noted in infants' ability to match heard and seen vowels (Kuhl & Meltzoff, 1984; Patterson & Werker, 2003).

Kuhl and Meltzoff (1996) conducted a novel study which had 4–5-month-old infants as participants. These infants were exposed to side-by-side videos of faces with no audio articulating English vowels / ɑ/ and /i/. At the same time, infants were exposed to a stream, of auditory signs made up of either / ɑ/ or /i/. Results showed that infants looked for a longer time at the face articulating the matching vowel compared to the face displaying the incongruent vowel. These results served as evidence that infants are responsive to the correlation that exists between auditory and visual cues of speech and that this responsiveness may be prevalent much earlier than originally thought. After the results of this study, Patterson and Werker (2002) recreated the above study using two-month-old infants who similarly displayed strong responsiveness to the match of visual and heard speech sounds in their L1.

Based on the results of the above findings, later research studies concluded that infants could not only match heard and seen speech in their native language but also in languages they are unacquainted with. Asserting that infants could match heard and seen speech of sound contrasts not found or are insignificant in their native language. This was seen in the non-canonical bilabial trill /B/consonants that 8-month-old Japanese infants could match even though it is not

present in their speech sound reserve (Mugitani et al., 2008). Similarly, in English speech sounds of the voiced labiodental fricative /v/ and voiced bilabial stop /b/, Spanish-learning infants were able to match the heard and seen signals of this sound even though it does not appear contrastively in Spanish (Skoruppa et al., 2009). These studies reveal that even in infants who lack language experience, audio, and visual modalities aid in the perception of speech.

Through their early months, infants experience particularly informative sensory cues (Lewkowicz, 2006). Infants have to combine both auditory and visual modalities to help shape their language background which allows them to comprehend (Mazzaro et al., 2016; Pons et al., 2015). Based on available research it is difficult to avoid the benefit visual features bring to speech perception. Amongst their benefits, visual signals operate as an amplifier of auditory speech as they aid in the salience of speech and supply necessary audio-visual information (Campbell, 2007). Thereafter, visual information present to infants such as the talker's face allows them to acknowledge their social partner's purpose (Tomasello & Carpenter, 2007). Lastly, exposure to talkers' faces might assist infants to imitate which could lead to assistance in speech production (Messum & Howard, 2015).

2.10 The McGurk Effect

The auditory and visual effect is best defined by the "McGurk Effect" (McGurk & MacDonald, 1976). This well-documented widely known phenomenon was researched by the latter authors (using over 40 years of research) to investigate audio-visual integration in speech perception (Rosenblum, 2019). The phenomenon demonstrated how auditory and visual modalities could combine content to form a unitary percept (Tiippana, 2014). McGurk and MacDonald (1976) observed 'the McGurk effect,' when a face articulating the syllable [ga͡] was presented alongside an auditory /ba/ syllable. Participants in that study revealed that they perceived it as the syllable /da/, resulting in two conflicting cues. The response of /da/ to visual [ga] and auditory /ba/ is known as fused perception, and it has received significant attention (Green, 2013). In addition to the fused perception, the McGurk effect includes an additional perception known as the combined perception (Green & Gerdeman, 1995). This combined perception occurs when the presented stimulus is visual /ba/ with auditory /ga/ and the response is /gba/.

Several research studies have discovered that the McGurk effect is comprehensive in a variety of capacities. The McGurk effect can be detected in both consonants and vowels (Traunmüller & Öhrström, 2007; Valkenier et al, 2012). Traunmüller and Öhrström (2007) discovered that when the auditory syllable /geg/ was combined with the visual syllable /gyg/, the Swedish high unrounded front vowel /e/ was perceived as the high rounded front vowel /ø/. When audio-visual incongruent vowels in Dutch were presented to Dutch speakers. Valkenier et al. (2012) discovered through the McGurk effect that the unrounded high vowel /I/ was perceived when the auditory rounded mid-high vowel /Y/ was synchronized with the visual unrounded mid-high vowel /e/.

Aside from segmental features, studies (Zhang et al., 2019) discovered that the McGurk effect in lexical tones which are suprasegmental however, the effect of visual information is minimal. Zhang et al. (2019) created sensory inputs with incongruent tonal auditory and tonal visual cues. Using both the response time paradigm and event-evoked potential methodology, Zhang et al. (2018) discovered that adding visual information of corresponding articulatory movement improves lexical tone perception and that incongruent visual information changes the duration perception of lexical tones. Another example of the McGurk effect's reliability is that it is an automatic and involuntary process that is unchanged even when participants are told that visual and auditory inputs are incongruent (Zhang et al, 2018). Sekiyama and Burnham (2008) proposed two methods to study the development of audio-visual speech processing, one of which is the "differential experience method" and the other is the "ontogenetic method".

The 'differential experience method' is used to investigate the McGurk effect across languages, while the 'ontogenetic method' is used to analyze it across developmental stages. The differential experience method looks into:

> "Individuals of the same age brought up in functionally different environments on a common task to investigate the effect that the type of exposure or experience may have on development, whereas the ontogenetic method looks into comparing the abilities of individuals of different ages brought up in functionally identical environments on a common task to investigate the amount of experience or maturation on development" (Sekiyama & Burnham, 2008).

Tiippana et al. (2014) discuss the issue of defining the McGurk effect in a review, asserting that it may be challenging to determine the prevalence of the McGurk effect owing to the various definitions used by different studies (Tilppana et al., 2023). The fusion response is thought by some to be the true McGurk effect because it produces a third or different syllable than the auditory or visual syllables. This definition, however, excludes other syllable combinations that produce a blend of auditory and visual information, such as /bga:/. Tiippana (2014) proposes the classical definition, in which an individual reports anything other than the auditory percept because this definition accounts for all McGurk syllable combinations (McGurk & MacDonald, 1976), which means that all incorrect auditory syllable responses are counted as an illusory effect.

The robustness of the McGurk effect has also been demonstrated in different conditions, such as with changes in auditory or visual cues (Campbell, 2007). Although the McGurk effect is robust in the aforementioned aspects, there is no concrete understanding of whether the McGurk effect is (1) robust across developmental stages and (2) across languages. Individuals who were tested in the original study carried out one test followed by the next stage of the study carried out two months later (Strand et al, 2014). A year later it was concluded that the McGurk effect was prevalent at both stages (Basu-Mallick et al., 2015).

There have been minimal studies on the relationship between the McGurk effect and connected speech. The majority of studies have focused on this relationship solely in laboratory settings and have not tapped into its effects in various adverse environments. This study will aim to further explore the effects of speech perception in various conditions in bilingual individuals.

2.10.1 Stimuli used in previous McGurk studies

The stimuli used in different studies with the McGurk effect as the framework utilised different stimuli. The original McGurk (1975) study made use of English /ba/, /ga/, /pa/ and /ka/ syllables. Sekiyama (1997) and Sekiyama and Tohkura (1993) used ten syllables in English and Japanese which consisted of /ba/, /da/, /ga/, /ka/, /ma/, /na/, /pa/, /ra/, /ta/ and /wa/. These syllables were used to create 100 stimuli and were used for both Japanese and English participants. In another study, audio-visual stimuli were used to create audio /apa, aba, ata, ada,

ama, ana/ which was integrated with visual /ata, ada, apa, aba, ana, ama/ (Vroomen & de Gelder, 1995).

These stimuli were used in studies consisting of Dutch and Cantonese participants. Thereafter, Chen and Hazan (2009) produced three incongruent stimuli made up of an auditory-[ba]/visual-[ga] and auditory-[da]/visual-[ba] and an auditory-[ga]/visual-[ba] used in both Chinese and English participants. From these studies, it can be noted that syllables made up of an open vowel /a/ and either a bilabial, alveolar or velar stops are the most popular stimuli sets in studies using the McGurk Effect as its foundation However when there is incongruency in the native language in the stimuli and the L1 of the participant the "non-native speaker" effect occurs.

2.10.2 Differences between McGurk effect studies

Apart from different stimuli being used when studying the McGurk Effect, there have been an array of different research designs regarding to the number of participants used and the age groups who participated. Earlier studies such as Sekiyama (1997) and Sekiyama and Tohkura (1991, 1993) made use of small participants sample sizes ranging with few (10-20) participants. However, this was criticized by Magnotti et al. (2015). The latter author instead used 162 Mandarin speakers native to China and 145 English participants from the United States of America (USA). The majority of studies made use of between 18 and 24 participants from a particular language group (Chen & Hazan, 2009, Vroomen & de Gelder, 1995). Magnotti et al. (2015) made use of groups aged between 14 and 23 years of age for the Chinese group while the English groups aged ranged between 18 and 26 years of age . Pearl et al. (2009) argued that adolescents carried out the tasks more effectively than adults did. This shows that is important to have a large age range across the study (Pearl et al., 2009).

2.10.3 The McGurk effect across various languages

The initial study which used the McGurk effect was carried out among English speakers and has since been well-studied in the English language (Sekiyama, 1994). Studies revealed that English speakers tend to rely more heavily on visual information compared to Japanese speakers (Sekiyama & Burnham,, 2008). Sekiyama also formed the fundamentals of the McGurk effect in Japanese (Sekiyama, 1997; Sekiyama & Burnham, 2008; Sekiyama & Tohkura, 1991; Sekiyama & Tohkura, 1993). The author states that L1 Japanese speakers had

less dependence on lip reading compared to English speakers even in instances when the Japanese participants were introduced to non-native cues. Researchers argue that reliance on visual information a result of differences in linguistic properties as Japanese is not as phonologically complex as English as it contains fewer vowels and consonants and no consonant clusters (Sekiyama & Burnham, 2008; Sekiyama & Tohkura, 1993; Tona et al., 2015). Other research studies assert that the differences are cultural as Japanese speakers are likely to avoid looking at a speaker (Sekiyama, 1997; Sekiyama & Tohkura, 1993; Sekiyama & Tohkura, 1991).

Based on these findings researchers looked into Chinese participants who have common cultural practices as Japan. Chinese participants recorded a weaker susceptibility to the McGurk effect compared to the Japanese participants despite these cultural similarities. The reliance of the Chinese participants on lip-reading was lower than that of L1 Japanese speakers (DeKeyser & Preito Botana, 2014). The latter authors declared that this result is largely owing to lexical tones in Mandarin and Cantonese (Dorado Solarte, 2023).

Regarding tonal differences in various studies, Sekiyama et al. (2003) stated that participants with tonal languages (languages that rely on pitch) will depend more heavily on auditory cues; therefore the McGurk effect may not be presented as effectively. This is because speakers of tone languages do not depend as much on visual cues to categorize lexical meanings (Sekiyama & Burnham, 2008). The study by Han et al. (2020) revealed that there is a slight McGurk effect in tone languages. However, this does not contest that a strong McGurk effect may not be elicited in tone languages that possess a bigger array of vowels and consonants.

2.10.4 Theoretical explanation of the McGurk effect

Earlier theories of speech perception such as Diehl (1987) stressed the auditory modality and downplayed the role of the visual modality in speech perception. The theories listed below will aid in understanding the effects of the visual modality in perception.

2.10.4.1 The fuzzy logic model of perception

The Fuzzy Logical Model of perception places importance on decoding the acoustic signal and can therefore be classified as an auditory model of speech. This model argues that speech perception is not reserved for human begins therefore there is no difference when people

process speech or other sounds (Massaro, 1983). A study was carried out that used audio-visual stimuli on a continuum from [ba] to [da]. Response times were recorded to evaluate the speed of the decision. The study found that response times were slower when the auditory and visual information was incongruent compared to when it was congruent. This showed that when presented with ambiguity the processing time was prolonged. This indicates that each modality is processed separately but side by side and that the combination of the audio and visual modalities takes place later (Massaro, 1983; Massaro & Cohen, 1983).

2.10.4.2. Revised motor theory

Originally known as the "motor theory of speech perception" (Liberman et al., 1957), the "revised motor theory" (Liberman & Mattingly, 1985) argues that the motor commands needed for producing speech are also used for perceiving speech. This theory highlights recognising the particular cues of the talker rather than the auditory cue. This model describes the illusion of the McGurk effect as confirmation speech perception requires decoding gestures. In general, this theory argues that speech is bimodal rather speech perception depends on the coming together of audio and visual information at an early stage which results in a particular gesture.

2.10.5 Limitations of the McGurk Effect

There have been general ideas relating to whether the McGurk effect draws on the processes of integration which are necessary for speech communication. These ideas have also related to whether the McGurk effect was the best way to approach integration (Alsius et al, 2018). Some studies argue that McGurk effect responses could be due to ambiguous input rather than an audio-visual fusion (Gonzales et al., 2021). Other studies have argued that a McGurk response should be linked to an illusion that only exists under particular tasks and conditions (Getz & Toscano, 2020). Hence, individual contrasts in the McGurk Effect among normal-hearing participants could be a result of audio-visual fusion, visual-only encoding or both.

Differences in McGurk's responses in studies could therefore be attributed to the age of a population or lip reading abilities and can be seen in the variation of responses as the benefit from visual speech is dependent per person (Sommers et al., 2005; Van Engen et al., 2022). Similarly, audio-visual perception has been known to vary among various clinical populations ranging from autism spectrum disorders to language development disorders (Irwin & DiBlasi, 2017). However, it is known that the stimuli presented can affect susceptibility to the McGurk

effect, therefore concepts such as the speaker used in the video and the video quality etc could all influence how the McGurk effect is perceived.

2.11 Speech perception in the L2

Second language acquisition is used to describe any language that is learnt after an individual has learnt their home language (L1) (Makina & Lephalala, 2010; Saville-Troike & Barto, 2017). The process of Second Language Acquisition is made up of three stages: the initial state, the intermediate state and the final state (Troike, 2006).

The initial state, L2 learning is distinguished by the use of the L1 knowledge, world knowledge which can serve to either assist or interrupt L2 development. The intermediate state is described by the transfer of L1 knowledge to L2 learning and development. The intermediate state starts when a speaker begins to transfer meaningful communication in the L2. The final state can be described as a continuation of the intermediate state which follows either interference or fossilization. Fossilization can be described as the period in which the L2 speaker cannot achieve native competence in the L2 (Wang, 2022).

Various studies have shown that a person's L1 phonemic inventory shapes how L2 speakers process visual input. This was shown by Werker et al., (1998) where French participants mistook the interdental fricative /ð/ for /d/ or /t/. This indicated that L2 speakers alter their concepts based on the phonology of their L1 (Werker et al., 1998). A study which included L1 Spanish participants learning English introduced participants to stimuli which varied visually such as /b/ and /v/ and those that didn't vary visually such as /p/ and /b/ (Thompson & Hazan, 2010). The study determined that at the beginning stage of language acquisition L2 speakers are not as reactive to visual cues as late language learners. Moreover, the study noted that stimuli such as /b/ and /v/ which have allophonic descriptions in Spanish cannot be differentiated as easily as two sounds with phonetic descriptions (Thompson & Hazan, 2010).

A further study showed that L2 Spanish speakers of English produced limited errors with regard to the identification of manner and place of articulation when visual cues from lip-reading were accessible. In general, it is understood that perceiving unfamiliar sounds in the L2 is more strenuous compared to perceiving L2 sounds that also exist in the L1 (Hommel, 2018; Wang et al., 2008; Wang et al., 2009). In a study carried out by Tayler and colleagues, it was noted that when recognising fricatives Spanish participants utilised similar approaches

as English L1 participants whereas Dutch participants utilised a different approach showing that languages which share similar phoneme inventories allow for individuals to use similar strategies in sound recognition (Tyler & Cutler, 2009).

The extent to which L2 speakers rely on visual cues during speech perception is also based on attributes such as the number of phonemes and the role of suprasegmental features such as lexical tone. This was seen in Sekiyama (1997) study which saw that L1 Japanese speakers had to differentiate a limited number of phonemes in contrast to L1 English speakers meaning that L1 English speakers may rely more on visual cues. Similarly, if a speaker has a tonal L1 they might rely more on an auditory modality as it contains greater information in Sekiyama (1997). Therefore the speaker's L1 and the methods they employ in their L1 perception may affect their learning strategies in the L2.

2.11.1 Speech perception in noise

Due to the reduction processes that occur frequently, casual speech with pronunciation variations and ambiguities present a greater listening challenge to non-native listeners than carefully pronounced speech (Broersma & Scharenborg, 2010). The authors assert that native English speakers adhere to a convenience of articulation principle in which differences between segments (consonants or vowels) are kept to a minimum (Broersma & Scharenborg, 2010). As a result, reduced forms of connected speech are prevalent, wherein sounds are deleted or changed. Assimilation, contraction, and elision are common examples of connected speech phonological processes. Bilingual learners may face difficulties with these reduced forms, particularly when their first and second languages do not hold common characteristics in reduced variants of pronunciation (Mitterer & Tuinman, 2012). This challenge has mostly been evidenced in noise-free laboratory settings (Mitterer & Tuinman, 2012) and has received minimal focus in less ideal listening conditions such as multi-talker noise.

Listeners in everyday situations face unfortunate listening conditions in which auditory, acoustic, or linguistic signals pertinent to speech recognition are degraded or completely lost. Conversations in noisy environments, for example, are difficult as multiple background talkers create competing speech signals. These settings impair speech perception even in native English speakers, but they are even more complex for non-native English speakers (Dietz et al., 2011).

Characterizing the impact of these unfavourable listening conditions on speech perception is critical for improving the validity of second-language-connected speech perception (Vlasenko et al., 2012). The existing literature on non-native perception of English speech under unfavourable listening conditions has primarily focused on additive noise, which includes energy from other sound sources (Liu & Shi, 2019). Listeners in a noisy environment must distinguish between competing noise signals to identify speech targets (Liu & Shi, 2019).

2.11.2 Uncertainty in audio and visual cues

Two experiments studied whether CV syllables to understand which phonemes created the most confusion in audio-only (Miller & Nicely, 1955) and visual-only conditions (Walden et al., 1977). The basis of both studies was formed around hierarchical cluster analysis that shows similarities between consonants through a correlation tool. The main findings of the study were that uncertainty among English consonants varies between auditory and visual perception. For example, the phonemes /b/ and /v/ created confusion in the auditory condition as both are voiced. The same /b/ and /v/sounds are however, easy to tell apart in the visual condition as /b/ is a bilabial phoneme while /v/ is a labiodental phoneme. Therefore, when information in the visual modality is limited it can be clarified by the auditory modality and vice versa relating that the audio-visual modality is the most accurate in the perception of speech (Chen & Rao, 1998, Hazan et al., 2006; Potamianos et al., 2004).

2.11.3 Audio-visual perception across languages

An individual's native language is an important aspect to consider when studying speech perception. Several studies have been conducted to determine whether audio-visual integration is language-specific or universal. Earlier studies using synthetic stimuli to test the McGurk effect found similar processing for native speakers of English, Spanish, Japanese, and Dutch speakers (Massaro et al., 1993) however differences owing to language characteristics were observed. These differences are linked to differences in phonological structures or cultural habits.

Sekiyama and Tohkura (1993) state that one of the most distinct differences occurred among Japanese speakers as they tend to avoid eye contact due to cultural values which results in them missing a significant amount of visual information. According to Sekiyama (1997), Chinese Mandarin native speakers exhibit a considerably weaker magnitude of the McGurk effect than American English participants. Similarly, studies conducted by Chen and Hazan (2009) and Magnotti et al. (2015) revealed that there was no distinction in the extent of the McGurk effect between American English speakers and Mandarin Chinese speakers. Cantonese speakers however displayed a considerably weaker McGurk effect than Australian English speakers, according to Burnham and Lau (1998). When Cantonese speakers were compared to Dutch speakers Vroomen and de Gelder (1995) discovered that Cantonese participants were more influenced by vision (wider McGurk effect) than the Dutch (Vroomen & de Gelder, 1995). It was declared that bilingual school-aged children showed no difference in audio-visual speech perception compared to monolinguals (Reetzke et al., 2016). However, Marian et al. (2018) and Burfin et al. (2014) noted that adult bilinguals of English-Korean and Catalan-Spanish had contrasts in the audio-visual perception of monolingual and bilingual adults where the bilinguals displayed heightened sensitivity to audio-visual cues (Burfin et al., 2014; Marian, et al., 2018). These studies display the arrived results in research of audio-visual among monolinguals and bilinguals sample groups.

The stimuli used in previous studies have also been varied. McGurk and MacDonald (1976) produced stimuli from English syllables /ba/, /ga/, /pa/, and /ka/ in the original McGurk effect study. Studies conducted by Sekiyama (Sekiyama, 1997; Sekiyama & Tohkura, 1993), saw ten syllables being used in English and Japanese to create 100 stimuli (10 audio x 10 visual) in each language, and both 100 English and 100 Japanese stimuli were presented to both Japanese and English participants. Chen and Hazan (2009) used three incongruent audio-visual stimuli in their study: (1) auditory-ba/visual-ga, (2) auditory-da/visual-ba, and (3) auditory-ga/visual-ba, which were created by two Chinese and two English native speakers (one male and one female in each language), and participants who spoke both Chinese and English. Like McGurk and MacDonald (1976), Magnotti et al. (2015), included English /ba/, /ga/, /pa/, and /ka/. In another study by Magnotti et al. (2015), all nine stimuli performers were Americans, and the faces were non-Asian (e.g., Caucasian), with one group of Chinese and one group of Americans as participants. In earlier McGurk studies, syllables containing the open vowel /a/ and one of the bilabials, alveolar, or velar stops were most frequently used as stimuli.

An interesting pheromone brought about through studies of the McGurk effects is the non-native speaker effect. This effect hypothesises that when the stimuli are non-native, people use more visual information (Chen & Hazan, 2009). This implies that when listening to non-native speech, participants face a more challenging time processing the auditory speech than when listening to their native language, as they borrow more information from the visual modality when perceiving a foreign language. Chen and Hazan (2009) discovered that when the stimuli were in Chinese, English perceivers used considerably more visual information, whereas Chinese native perceivers did not reveal such a difference. The authors proposed the possible explanation that while none of the English participants had studied Chinese, all the Chinese participants had studied English for at least six years in school (Chen & Hazan, 2009). Previous studies have concluded that (non-)nativeness matters in audio-visual speech perception. They did not, however, distinguish whether the different degree of visual influence is due to non-native auditory speech information (the non-native speaker effect) or non-native visual cues (the other-race effect).

Various studies concerning the perception of L2 sounds have shown the effect of the L1-accurate L2 perception (Flege, 1993; Iverson & Kuhl, 1995). Similarly, other studies regarding L2 perception have discussed the influence of accurate perception on accurate production (Bohn & Munro, 2009; Escudero et al., 2009; Fledge et al., 1995). Studies have shown that it is likely that at earlier stages if L2 learning, adults are language-specific perceivers and that they are likely perceiving the L2 through their L1 sound system knowledge (Best et al., 2001; Harnsberger, 2001; Schmidt, 2018).

The most evident difference between L1 perception and L2 perception is that L2 speakers perceive the sounds of L2 through the L1 (Flege, 1981). Moreover, Flege (1993) states that the production of L2 sounds may also be affected due to the mental representations held from the L1. According to the Speech Learning Model (SLM) the link between the L1 and L2 categories is essential in perceiving and producing L2 sounds (Flege et al., 1995). The SLM declares that as L2 speakers gain experience they might be able to differentiate between phonetic differences. Therefore this model argues that the perceived link between a person's L1 and L2 categories is essential in correctly perceiving and producing L2 speech sounds.

2.11.4 Properties affecting L2 perception

Age is another factor that can influence L2 acquisition. Younger language learners seem to have greater mastery than adults. The advantages are noted in the brain systems of younger people (*Second Language Acquisition and the Critical Period Hypothesis*, n.d). The age at which the person started learning the L2 saw Flege et al. (1995) introduce the terms "age of arrival" and "length of residence" to highlight the two properties that affect L2 acquisition. However, these individuals have less experience with their L1 and, therefore, less interference in the L2 (Best, 1994 as cited by Strange, 1995; Bohn & Munro, 2009). Another age-related element that can influence L2 acquisition is the L2 speaker's onset age. Relating to this is a study carried out by Massaro et al. (1986) who revealed that 6 to 10-year-old participants are unlikely to be affected by visual details compared to adults in consonant perception (Massaro & Egan, 1996). For the reason that responsiveness to specific acoustic cues rises in the first 10 years of life (Mayo & Turk 2004). It is expected that L2 speakers lose responsiveness to acoustic cues that are not pertinent in their L1. Therefore, even though L2 speakers are aware of articulatory differences they find it difficult to match it to relevant phonetic labels (Hazan et al., 2005).

The above-mentioned concepts- age of arrival and length of residence (Flege, 1995) are affected by the amount a person uses their L1 or L2. To assert this stance, the authors used a cohort of Italian immigrants in Canada and tested their eleven English vowel production (Piske et al, 2002). The group was divided into early learners of English and late learners of English where they were subsequently rated on their intelligibility by Canadian English speakers. The results revealed that their L1 (Italian) did hinder their L2 categories regardless of whether they were early or late learners of English. These results highlight that difficulties in L2 acquisition could be due to the L1 categories.

The type of learning elicited in the task for L2 speakers is another factor to consider that might affect perception. There are various studies used to obtain results- some studies aim to highlight the improvement in perception while others study the effects of language training (Iverson et al., 2005; Nishi & Kewley-Port, 2007; Wang & Munro, 2004). Similarly, some focus on particular cues to study the effects of attention while others look at both perception and production (Bradlow et al., 1999; Hazan et al., 2005; Kondaurova & Francis, 2010; Lambacher

et al., 2005). Therefore, the type of task used could also affect perception or account for variances among participants.

Hazan et al. (2006) pointed out three groups of visual speech classes based on their existence in L1 or L2 speakers. The first is when a visual class is evident in both the L1 and L2, the second is when the visual class occurs in the L2 but not the L1 and the third one is when the visual class is apparent in both the L1 and L2 (Wang et al., 2009). Hazan et al., (2006) hence state that because of L1 experience an L2 speaker could lose responsiveness to visual classes that are not apparent in their L1. Similarly, an L2 speaker may not struggle to perceive L2 sounds if they are present in the L1. Regardless, audio-visual teaching is said to facilitate L2 speakers' link of non-native speech sounds with specific visual classes (Erdener, 2012; Hazan et al., 2005).

The cognitive demand required for a task may also affect L2 perception and account for contrasts in a participant's performance. Using this variable a test was conducted that looked at the English /æ/-/ɛ/ contrast of 55 late Dutch-English bilinguals using categorisation, word identification and a lexical decision task (Díaz et al, 2012). The study concluded that late L2 speakers achieved better in lower-level phonetic recognition compared to lexical decision tasks. This shows that participants may achieve better results in tasks which require less cognitive demand.

In addition to this, other factors such as differences in gender (Alm & Behene, 2015; Gelfer & Bennett, 2013; Reid et al, 2003), motivation (Gardner et al, 1985; Gürbüz & Cabaroğlu, 2021; Lidestam & Beskow, 2006), cognitive ability (Fostick et al., 2013; Heinrich et al., 2016) could also account for contrasts in L2 speech perception. Furthermore, lip reading abilities (Demorest et al., 1996), and responsiveness to visual cues (Sennema et al., 2003) could also affect L2 perception. These specific factors could also account for differences in studies with participants of similar language backgrounds.

2.11.5 Difficulty learning tone in an l2

It is known that the existence of tone in L2 speech presents a challenge for new language learners. Tone poses a particular challenge to L2 speakers particularly when it does not exist in their L1 (Francis et al, 2008; So & Best, 2014). However, the extent of this difficulty has seldom been explored. Wells (2010) states that in tone languages (e.g. isiZulu) pitch operates

like vowels and consonants to create a distinction between a word's meaning. However, non-tonal languages (such as English) usually use pitch to provide intonational cues. These serve as discourse types or intonational cues. The three main difficulties that non-tonal L1 speakers face lie in: (1) challenges with the perception and distinction of tonal pitch patterns at the phonetic-acoustic level; difficulty encoding pitch patterns at the syllable level; and (3) problems in utilising abstract tone categories such as lexical cues (Wong et al., 2006).

Research studies have shown that L1 tonal language speakers are likely to outdo non-tonal L1 speakers in distinguishing novel tones (Hallé et al., 2004; Leather, 1987; Mok & Zuo, 2012; Wang et al., 2001; Wayland & Guion, 2004). However, based on the perceptual assimilation, the success of an individual's tone is subject to the similarities of the tonal systems between the L2 and the speaker's L1 based on the Perceptual Assimilation Model (e.g. Best, 1994 as cited by Strange, 1995; Francis et al., 2008; So & Best, 2010; Wayland & Guion, 2004). Non-tonal L1 speakers are said to approach tone very differently compared to L1 tone speakers and the differences can be found behaviourally and neurologically (Gandour et al., 2003; Hallé et al., 2004; Krishnan et al., 2005; Wang et al, 2004; Wong & Perrachione, 2006; Xu et al, 2006). In L1 tone speakers, short-term training has been shown to be effective in non-native tonal recognition, however, on non-tonal speakers, the results have been varied (Francis et al., 2008; Wang et al., 2004; Wayland & Guion, 2004).

2.12 Theoretical Framework

The two most prominent approaches regarding L2 sound acquisition are the *Speech Learning Model* (SLM) by Flege (1991, 2003; Flege et al., 1995; Flege et al., 1999) and the *Perceptual Assimilation Model* (PAM) (Best et al., 2001; Bohn & Munro, 2009; Best & Strange,1992).

The phonological acquisition model proposed by Fledge (1995) (that is, SLM) demonstrated that an element controlling the degree of accuracy of L2 sound perception and production is the connection between a person's L1 and L2 phonetic system. The model asserts that sounds are categorized and produced depending on how they are perceived- either as being a new sound, a similar sound or a sound identical to an individual's L1 (Kucirkova, 2011). When new sounds are perceived an L2 speaker is more likely to perceived the sound correctly as they do not match any L1 sound (Kucirkova, 2011). Therefore, words that are identical to a person's L1 will be more difficult to perceive. The SLM also considers age difference in L2 learning

specifically in the perception and production of L2 vowels and consonant sounds (Hanulíková et al., 2012).

This model investigated the link between the production and perception of speech sounds and argued that a speaker's L2 production was determined by their L1 phonemes (DeGroot, 2011; So, 2010). The model states that the rate at which an L2 speaker learns the L2 sounds is based on the acoustic space in relation to their L1 sound categories (So & Best, 2010). Thus, SLM argues that the perception of an L2 speaker can be predicted based on the acoustic differences and similarities between two languages. Similarly, the SLM asserts that learning a second language is more strenuous when the phonetic space becomes dedicated to the first language (Walley et al., 2007). In the L1, the formation of perceptual phoneme categories begins during the first year of infancy. Learning new perceptual categories in the L2 gets tougher when individuals become more established in their L1 categories (Akbarian, 2015; Simonet, 2012).

Therefore, the rate at which an L2 speaker achieves native-like perception is based on how well they can tell apart the phonetic differences in the L1 and L2 phonemes (Schouten, 2009). Thus, when there are distinct differences between the L1 and L2 a person is likely to create a new L2 phonetic category particularly when their L1 has fewer sounds then the L2. (Flege, 1995). In contrast, when the L1 and L2 have similar sounds new L2 speakers will find it more challenging to create L2 sounds as distinct phonemes; this makes it more challenging for them to formulate new sound categories (Flege, 1995).

The other phonological acquisition model is known as PAM. This model states that the perception of L2 speakers of non-native contrasts is determined by their L1 (Best, 1994 as cited by Strange, 1995). According to PAM, at the stage of infancy and early childhood stage, perception is suited to the L1. Therefore, sounds similar to the L1 are natural (Bohn & Munro, 2009). The model also holds that through perceptual assimilation L2 phonemes are classified based on related articulatory gestures that occur in a person's L1 and L2 phonemes (Guion et al., 2000). PAM thus, outlines various patterns of perceptual assimilation of L2 segments to L1 phonological categories, which are affected by the perceived phonetic similarity between L1 and L2 segments (Best, 2001). In order to account for L2 acquisition, PAM was expanded into PAM-L2 (Best & Taylor as cited by Best et al., 2016) which stated that language learning at any age is greatly affected by a person's language experience. PAM-2 revealed that a person's L1 does not necessarily affect their L2 perception (Best & Taylor, 2007 as cited by Best et al.,

2016). However, the perception of non-native sounds are connected to their L1 sounds. Therefore, when an L2 speaker learns new sounds in the L2 they assimilate these new sounds based on the acoustic properties of their L1 (Tyler, 2019). There are various patterns of assimilation mentioned in the model. Both PAM and PAM-L2 state that the first pattern of assimilation takes place when a non-native contrast is perceived and has similar properties to the L1 (Tyler, 2019). The second pattern of assimilation occurs when speech sound is unrecognizable and therefore cannot be assimilated to a native category. The third pattern occurs when a sound is heard but has no resemblance to a native cue (Bohn & Munro, 2009). In essence the PAM-L2 relates that the way in which an L2 speaker learns words is based on how the L2 sound is perceived and at which stage of development it takes place at (Tyler, 2019). Therefore, as an L2 speaker develops they acquire more of the L2 vocabulary whereby phonological fossilization will occur (Bundgaard-Nielson et al., 2012)

The one clear contrast between SLM and PAM is centred around the topic of phonetic category. According to SLM, the mechanisms needed to acquire L1 sounds and phonetic categories remain the same over a person's life time and are thus accessible to L2 leaners (Bohn & Best, 2012, Georgiou, 2021). PAM however, states that instead of developing a new category an individual's L1 category can expand over time to include the acoustic properties of the L2 language (Bohn & Best, 2012).

2.13 Chapter summary

The following chapter provided an overview of the research underlying the study. The topic of bilingualism which forms the basis of this study was discussed. Bilingualism mentioned in this study will refer to sequential bilinguals as explained by Castilla et al., 2009 and Kohnert, 2008. The properties that form the basis of language were also discussed. This referred to properties such as phonemes and the phonetic and phonology of isiZulu and English. The models of audio-visual integration were also explained; the three stages included later integration, early integration and multistage integration. The three conditions used in this study including the auditory condition, visual condition and audio-visual condition were also referenced. Thereafter the theoretical framework was discussed. The models that made up the theoretical framework included the SLM (Flege, 1991, 2003; Flege et al., 1995; Flege et al., 1999) and the PAM (Best et al., 2001; Best & Strange,1992; Bohn & Munro, 2009).

CHAPTER 3

Methodology

3.1 Introduction

This chapter discusses the methodology used to collect and analyse data in order to respond to the study's research questions. This study seeks to investigate the extent to which young adult bilinguals benefit from audio-visual speech. The methodology employed in this study employs a similar method from a study titled "*Audio-visual integration in noise: Influence of auditory and visual stimulus degradation on eye movements and perception of the McGurk effect* (Stacey et al., 2020)".

Audio-visual integration in noise: Influence of auditory and visual stimulus degradation on eye movements and perception of the McGurk effect.

For the benefit of this chapter, we will readdress the main components of the study. The study aimed to explore the effects of modality on audio-visual speech. In order to explore this subject, the study set out to understand what the effect of modality is on speech perception. Similarly, we sought to understand the extent to which different conditions affect audio-visual speech.

We aimed to respond to the following investigative questions:

1. What is the effect of modality on audio-visual perception?

2. How do different conditions affect the way L2 bilinguals perceive audio-visual speech?

3.2 Research design

To fully answer the above-mentioned research questions pertaining to this study, a quantitative research design was implemented for this study. Such a research design is directed at constructing general laws and occurrences across diverse environments which allows for the support or rejection of a class or category of beliefs (Nietzel et al., 2021). Similarly, this design allows for a thorough and in-depth investigation of an occurrence or event through the data set and introduces a comprehensive description of the research method (Hignett & McDermott, 2015). Quantitative data involves quantifying and analysing variables in a data set. The

research question is thus answered through numerical data which is obtained through mathematical methods such as statistics (Apuke, 2017). A quantitative research design aims to test a hypothesis, evaluate cause and effect and constructing a prediction (Apuke, 2017). In essence, a quantitative approach aims to report on the results acquired through the methodological and unbiased data collection process through numerical data which is then statistically analysed (Creswell, 2008; Dornyei, 2007).

The data collected in quantitative studies is collected in a methodological and unbiased procedure and is often analyzed through statistical methods as mentioned above. (Almeida et al., 2017; Dornyei, 2007; Nietzel et al., 2021). Taking into account the research question and the nature of the study variables, a quantitative approach was adopted for the present study and a Microsoft Excel spreadsheet was used to analyse the data set. Analysing data in a quantitative method allowed for variables to be reconstructed into a practical form. Quantitative methods enable key comparisons to be made and for important correlations in data sets to be quantified. Therefore, this practical form of data permits prompt analysis of the data set particularly when the data set deals with extensive volumes of data (Nietzel et al., 2021). A large amount of quantitative research designs aid in informing the procedure for analysing new data sets (Creswell, 2008; Rashid & Sipahi, 2021). A key advantage of a quantitative research method, however, is that it is believed to be scientifically objective and rational because of its statistics and mathematical basis (Nietzel et al., 2021).

Besides adopting a quantitative design, this study utilised descriptive statistics. Descriptive statistics can be defined as an approach that analyses, explains and summarises data coherently and systematically (Marshall & Jonker, 2010). Descriptive statistics allows for data to be simplified as percentages or visual summaries such as histograms or bar graphs (Zook & Pearce, 2018). This method allows researchers to display and summarise a data set.

The above-mentioned paragraph described the first noteworthy element of this study which was the application of the quantitative approach in investigating the extent to which adult bilinguals benefit from different modalities in audio-visual speech perception. Using a quantitative method.

3.3 Sampling method

Regarding sampling, this study employed a snowballing method to recruit participants. The snowballing method is a frequently utilised method in quantitative studies and is described as an approach carried out through survey and data registration particularly when the data set is detailed or rare (Dragan & Isaic-Maniu, 2022). Owing to the uniqueness of the data set the participants may be difficult to identify. The most overarching definition for the snowballing method is a technique in which one interviewee provides the researcher with the name of at least one other potential interviewee. That interviewee subsequently provides the researcher with another name, and the sample then grows (Naderifar et al., 2017). The snowballing method is a convenient, efficient and cost-effective technique for gathering participants that fit the study's criteria (Parker et al., 2019).

However, using this sampling method, however, means that individuals in a random sample do not have the same probability of being included in the sample therefore the data would be generalisable to a particular population only. Although the snowball sampling method is time-consuming the researcher has the chance to communicate with the participants to identify other recruits (Levy & Ellis, 2011; Parker et al., 2019). Although the snowballing approach has drawn criticism various scholars refute this criticism stating that the intention of the researcher is not to merely generalize a population but rather to grow a comprehensive investigation of a focal phenomenon (Akcam et al., 2019).

To summarise, compared to various other sampling methods and considering the practical factors of recruiting participants who were eligible to participate in the experiment and satisfy the strict criteria, the most feasible technique for identifying participants for this research was using the snowballing sampling method.

3.4 Participants

Participants in this study were identified through particular characteristics. The details of the research participants are presented as follows:

The first participant was selected through the University of the Witwatersrand (Wits) Language Policy teaching process, thereafter a snowball technique was used to gather participants. First adopted by the university in 2015, the policy requires all first-year students to study one of the

national languages (a selection of English, isiZulu, Sesotho, or South African Sign Language). For this study, only new L2 isiZulu learners were eligible to participate and had to be between 17 and 29 years of age. A background questionnaire stating the participant's home language, age, and number of years of exposure to isiZulu ensured homogeneity. Therefore, the common characteristics among participants were their age group (between 17 and 29 years), their tertiary institution (University of the Witwatersrand) and that they were newly acquiring L2 isiZulu.

The complete data set consisted of 12 participants which was further broken down into 11 females and 1 male. Table 1 represents the linguistic composition of the data sample.

Table 1: Data Sample

# of Participants	Sex	L1	Average Age	#Months learning L2
11	Female	English	19	8-12
1	Male	English	21	36

Thus, the participants were all recruited from the University of the Witwatersrand and the prerequisite was that all participants were early isiZulu learners. The study used early L2 isiZulu speakers as it is well known that early L2 language adult learners of isiZulu find learning the sound of the L2 challenging (Escudero et al., 2009). Therefore, employing this group will allow us to investigate the effect of modality on early L2 speech and how different conditions affect late L2 speakers' perception of audio-visual speech.

3.4.1 Language background questionnaire

Questionnaires recording a participant's bilingual experience are commonly used in research studies on language and cognitive development. These questionnaires are used to identify the bilingual profile of participants. Although a language background questionnaire is common, they vary in the type and number of questions asked. Some questionnaires are detailed and extensive (such as in Liu et al., 1992) while other are short and informal.

In this study participants were asked three questions:

1. Their current age

2. Their native/home language

3. The number in months they had been exposed to isiZulu as an L2

Participants were given the given the language background questionnaire before the experiment began. To comply with the ethical principle's participants were given a pseudonym. Table 2 below sets out the answers received in response to the questionnaire from the questionnaire:

Table 2 Shows the data gathered from the linguistic questionnaire

Pseudonym	**Age**	**L1**	**Months of experience**
P 1	18	English	8
P 2	19	English	8
P 3	19	English	8
P 4	19	English	8
P 5	19	English	8
P 6	20	English	9
P 7	17	English/Sesotho	9
P 8	18	English	9
P 9	18	English/ Sepedi	10
P 10	25	English	12
P 11	22	English	12
P 12	21	English	36

The questionnaire revealed that the L1 spoken by the majority of participants was English. Two participants each had one native language respectively, namely Sesotho and Sepedi. The

duration the participants had been exposed to isiZulu as an L2 ranged between 8 and 36 months.

3.5 Procedure

The study occurred in a soundproof laboratory, with only the researcher and one participant present at a time. The participants were each played pre-recorded video clips .The talker in the clips was a 23-year-old male student who is an L1 isiZulu speaker and who is a Master's degree candidate in Linguistics. The student was video-recorded by the researcher. For the recording, the talker wore neutral clothing and was filmed in front of a neutral background. When it was their respective session in the soundproof laboratory, the participants were each seated approximately 45 cm in front of a desk with a 13-inch MacBook Air laptop and comprised 23 videos. Individuals were also made to wear headphones for optimal listening. The duration of the videos varied between 1-9 seconds.

Before the experiment began participants were asked to observe the videos closely and to repeat what they observed out loud. The experiment was approximately 10-15 minutes and was recorded with a camcorder. The video camera used was a Sony DCR-SR190E Digital Video Camera Recorder. The task was performed and filmed individually.

Before participation, individuals were given a participant information sheet that introduced them to the study thereafter each participant had to fill out a consent form accepting the procedure of the study. Participants could withdraw from the study at any time. Figure 10 is an illustration of the participants placement during the study.



Figure 10: An illustration of the participant's placement during the laboratory session

3.6 Pilot study

Before the formal data collection began a pilot study was carried out. This was done to examine the feasibility of the study's procedure and to examine the research materials. Carrying out the pilot study minimised and eliminated the risk of any flaws coming out during the formal data analysis.

The pilot study consisted of three participants (three females) who met the age criteria and were learning isiZulu as an L2. Of the three participants, two were native English speakers whilst the other was a native English and Afrikaans speaker. The intent of carrying out the pilot study was to establish the feasibility of the study, to check the practicality of the method and to determine the time each participant needed to carry out the study.

The key finding of the pilot study was that participants found the length of the study to be long and time-consuming. In order to rectify this problem the stimuli were reduced in the final study. The pilot consisted of an average of eighteen-word pairs (totalling 15-word pair clips and 3 incongruent clips). This was reduced in the final study to thirteen-word pairs (totalling 10-word pairs and 3 incongruent clips).

The results of the pilot study revealed that participants perceived words well overall in the audio-only condition. For the visual-only conditions, all three participants were unable to perceive the word in the visual-only condition therefore, for the final study we introduced a multiple-choice list of similar-sounding phonemes whereby participants could choose the sounds they believed they observed.

Therefore, carrying out the pilot study minimised and eliminated the risk of any flaws coming out during the formal data analysis. The data from the pilot study was not used in the analysis or conclusions of the study as the changes made in the final study altered the procedure used in the pilot.

3.7 The video stimuli

Participants were introduced to four different conditions. The four conditions were made up of audio-only, visual-only, audio-visual and incongruent conditions. Each condition was further broken down into sub-conditions which consisted of word pairs that were either a tonal pair, similar word pair or minimal pair. For the purpose of this study, similar word pairs are also

referred to as minimal pairs. Although differences in tone between words can also be classified as minimal pairs, for the purpose of this study, two words differentiated by tone are referred to as a tonal pair.

The stimuli consisted of 23 videos made up of various isiZulu tonal pairs and minimal pairs. Made by video recording a 23-year-old L1 isiZulu speaker. The stimuli were retrieved from "A linguistic analysis of isiZulu" (Poulos & Msimang, 1998) and some stimuli were gathered from an isiZulu speaker and will be broken down into its specific components below. The types of words selected for the stimuli were tonal pair words, similar-sounding words and minimal pair words. The study chose to use tone pair words as isiZulu is a tone language where differences at the word level affect lexical meaning (Westphal, 1955). At the syllable level, isiZulu uses a tonal variation to change the meaning between words (Chen et al., 2020). This means that phonologically similar words could have a change of meaning because of tone patterns. There are two general tone levels in isiZulu known as high (H) and low (L). Therefore in the word "índòdà" (man) the tone will be realised as HLL.

According to Putri (2015), minimal pairs are commonly used in listening practice and oral production. Within the field of speech pathology and speech therapy, minimal pairs are used to test tone production and perception. Minimal pair is an approach utilised in teaching pronunciation as the difference in sound can affect the perception, comprehension and fluency in communication (Putri, 2015). Although the sound differences between minimal pairs may seem obvious, it can be challenging for new L2 speakers to contrast the difference in sound. Therefore, using minimal pairs as the stimuli allows us to test the extent to which different modalities affect L2 speakers' perception of non-native cues. Similarly, using tone pairs and minimal pairs to study speech perception can aid in understanding how new L2 speakers perceive words in different conditions.

In the incongruent condition, the McGurk effect was used to investigate how different conditions affect the way L2 speakers perceive audio-visual speech. The McGurk effect is an auditory-visual illusion that describes how individuals combine information from speech sounds (Tona et al., 2015). This illusion exhibits multi-sensory information and describes how visual and auditory information are combined into and integrated percept. Tiippana (2014) states that the McGurk effect is the perfect device to study multisensory information in speech perception. Audio-visual integration is reflected from the McGurk effect .

The section below will introduce the four conditions and will explain the procedure each condition followed. 3.7.1 Audio-only condition

In the audio-only condition (tonal pairs), the stimuli consisted of the following words:

bònà	they	(Poulos & Msimang, 1998)
bónà	To see	(Poulos & Msimang, 1998)

The word “bònà” (they) is made up of a low high tone while the word “bónà” (to see) is made up of a high low tone as indicated above. Figures 11 and 12 depict an image of the video clips participants were introduced to.

Dúmà	be tasteless	(Poulos & Msimang, 1998)
dùmá	roar; be famous	(Poulos & Msimang, 1998)

The first word in the tonal pair “dúmà” (to be tasteless) consists of a high low tone and the second word “dùmá” (roar/ be famous) has a low high tone in the word

The audio-only minimal pair words were made up of the following words:

usi.si	sister	n/a
isi:su	stomach	n/a

In usisi (sister) and isisu (stomach) the distinguishing factor in this pair is the front vowels [i] and [ɛ]. The [i] sound is high, front and unrounded whilst the [ɛ] is mid, front unrounded. Figure 13 and 14 are illustrations of the audio-only condition stimulus participants were introduced to.



Figure 12: An illustration of the word " usisi" (sister) in the audio-only condition



Figure 11: An illustration of the word" isisu" (stomach) in the audio-only condition

[Permission to use this photograph was granted by the research assistant]

ibhola	ball	(Ncobela, 2015)
ibhala	bar/ wheelbarrow	(Ncobela, 2015)

In "ibhola" (ball) and "ibhala" (bar/wheel barrow) the contrast can be found in the back vowels [o] and [a]. The [o] is mid, back unrounded sound whilst the [a] is an open central unrounded.

The procedure for the audio-only condition was as follows:

In this condition, participants were exposed to eight clips. The first four clips were made up of two tonal pair words while the second four clips were made up of two sets of similar sounding word pairs. Participants were introduced to the first word in the tonal pair through a set of headphones with no visual access; they were then asked to recall aloud what they had heard in the clip. Thereafter the participants were introduced to the second word in the tonal pair and were asked to recall what they heard. After being exposed to both words in the tonal pair participants were then asked if the words they heard were the same or not.

The next set of audio-only words consisted of similar-sounding words. The two similar-sounding words were paired with each other. For this stimulus participants were introduced to one word in the form of a clip where they recalled the word they heard thereafter they were introduced to the second word where they were asked to recall what they heard. Figure 15 is an illustration of how participants experienced the audio-only condition.



Figure 13: An illustration of the audio-only condition where partcipants were seated in front of a blank screen as they observed an auditory stimuls

3.7.2 Visual-only condition

The Minimal Pair in the visual-only condition was made up of the following words:

isithelo	fruit	n/a
isitini	brick	n/a

In “isithelo” (fruit) and “isitini” (brick) the area of interest in the above words is the alveolar [t^h] in [isit^h ɛlɔ] and the alveolar [t’] sound in [isit’ini]

ukusala	to remain	(Poulos & Msimang, 1998)
ukuzala	to bear	(Poulos & Msimang, 1998)

In “ukusala” (to remain) and “ukuzala” (to bear) the key difference can be found in the voiceless, alveolar, fricative [s]sound compared to the voiced, alveolar, fricative [z] sound

The visual-only condition procedure is outlined below:

In this condition, participants were exposed to clips through a visual modality with no access to any auditory signal or cues. The condition was made up of two pairs of similar sounding words with each pair consisting of two words. Participants were asked to watch a clip with a word being pronounced. The clip consisted of no auditory cues or signals, thereafter they had a second opportunity to review the same clip again after which they were asked to choose a

word from a list that looked most similar to the word they observed. The Figure 16 is an example of what participants would have encountered whilst Figure 17 is an example of words participants could choose from that they felt best resembled the word they observed



Figure 14: An example of the visual stimulus "isithelo"

[Permission to use this photograph was granted by the research assistant]

.

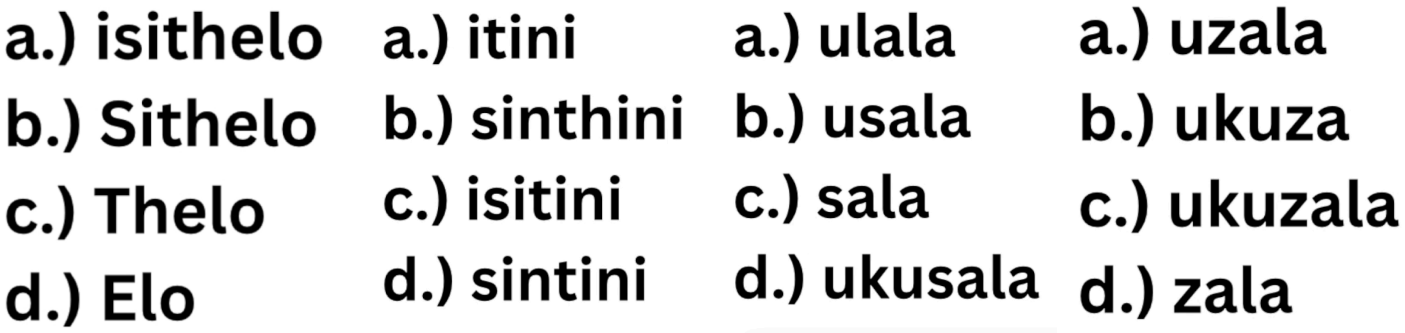


Figure 15: An example of the option participants encountered during the visual-only condition3.7.3 Audio-visual condition

The audio-visual condition in this study was made up of four word pairs the first two sets of word pairs consisted of tonal differences and can be seen below:

ùkúsìndà	to smear	(Poulos & Msimang, 1998)
ùkùsíndà	to be heavy; to recover	(Poulos & Msimang, 1998)

In the minimal pair “ùkúsìndà” (to smear) and “ùkùsíndà” (to be heavy/to recover) the first word “ùkúsìndà” (to smear) has a low, high, low, low tone the second word “ùkùsíndà” (to be

heavy/to recover) however, has a high, low, high, low tone. Figure 18 and 19 are illustrations of the clips participants were exposed to in the audio-visual condition.



Figure 16: An illustration of the word ùkúsìndà (to smear) in the audio-visual condition



Figure 17: An illustration of the word ùkùsíndà (to be heavy; to recover) in the audio-visual condition

[Permission to use this photograph was granted by the research assistant]

ùkwèthukà	to insult	(Poulos & Msimang, 1998)
úkwéthùkà	to get a fright	(Poulos & Msimang, 1998)

In "ùkwèthukà" (to insult) we observe a high, low, high, low tone while a low, high, low, low tone can be seen in úkwéthùkà (to get a fright).

The remaining two-word pairs were made up of minimal pairs and can be seen below:

ukupheka	to cook	(Poulos & Msimang, 1998)
ukubheka	to look	(Poulos & Msimang, 1998)

In this minimal pair, the area of interest is the [p^h] sound in "ukupheka" (to cook) compared to the [b^h] labial, plosive sound in "ukubheka" (to look).

isitho	organ	(Poulos & Msimang, 1998)
isipho	gift	(Poulos & Msimang, 1998)

The area of concern in this minimal pair is the distinction of the [t^h] in “isi<u>th</u>o” (organ) and the [p^h] in “isi<u>ph</u>o” (gift). Figure 20 and 21 are an illustration of the video clips in the audio-visual condition.



Figure 18: An illustration of the word isitho (organ) in the audio-visual condition



Figure 19: An illustration of the word isipho in the audio-visual condition

[Permission to use this photograph was granted by the research assistant]

The procedure for the audio-visual condition was as follows:

The audio-visual condition was made of two tonal congruent pairs and two minimal pair tasks. In this condition, participants were first introduced to the tone words. Each tone set was made up of two words and this condition consisted of two tone sets. For the tone words participants were introduced to the first word in the set thereafter they recalled what they observed in the clip. After this, the corresponding tone pair in the set was played and participants were asked to recall what they had heard. After hearing both words participants were asked if the two words they observed were the same or not. Failure to recognise the difference between the two words meant that the participant did not detect the difference in tonality between the two words.

For the minimal pair participants were introduced to the first word in the minimal pair and they were asked to recall the word. Thereafter, they were asked to recall the second word in the minimal pair. Recalling the correct word highlighted the correct perception of the word. The same method was followed for the second minimal pair word set.

3.7.4 Incongruent stimuli

The incongruent condition utilised the McGurk effect (1976) which saw the audio of one word being dubbed over the visual of another word. This study consisted of three incongruent stimuli broken down below:

isitho (vocal)	organ	(Poulos & Msimang, 1998)
isipho (visual)	gift	(Poulos & Msimang, 1998)

The vocal form of "isitho" (organ) was dubbed over the visual form of "isipho" (gift). In this minimal pair, the difference noted is in the voiceless, alveolar, plosive of [t] compared to the voiceless, bilabial, plosive of [p].

ukulima (vocal)	to plough	(Poulos & Msimang, 1998)
ukulila (visual)	to cry	(Poulos & Msimang, 1998)

The vocal of "ukulima" (to plough) was dubbed over the visual of "ukulila" (to cry). In this pair, the contrast can be seen in the voiced, bilabial, nasal of [m] compared to the voiced, alveolar, lateral approximant of [l].

ukusala (vocal)	to remain	(Poulos & Msimang, 1998)
ukuzala (visual)	to bear	(Poulos & Msimang, 1998)

The vocal of "ukusala" (to remain) was dubbed over the visual "ukuzala" (to bear). Here, the minimal pair differs in the voiceless, alveolar, fricative of [s] and the voiced, alveolar, fricative of [z].

The procedure for the incongruent condition was as follows:

The incongruent effect was based on the McGurk effect (McGurk & McDonald, 1976) which is produced when a video recording of an utterance is combined with an audio recording of another utterance. With regards to incongruency in audio-visual perception, McGurk and McDonald's (1976) study served as the basis for a plethora of studies. The McGurk effect is used to measure the weightings of auditory and visual information. (McGurk & McDonald, 1976). In essence, the tests measure the degree to which visual perception is affected when the auditory and visual components are not integrated. However, the majority of McGurk effect

studies have used syllables as stimuli to study the phenomenon. This study made use of words that would provide insight into audio-visual speech perception and lexical knowledge. The evidence from this study was important as it showed that incongruent visual and auditory information can result in in fused or combined precepts.

The incongruent condition consisted of three sub-tasks and each was made up of three tonal pairs. The first tonal pair consisted of the word “isitho” (organ) which was placed with the visual “isipho” (gift). The second pair was made up of the word “ukulima” (to plough) which was placed over a visual “ukulila” (to cry) and lastly the audio word “ukusala” (to remain) was placed over a visual “ukuzala” (to bear). Participants were introduced to the video and were asked to recall what they observed in the video. Each video was only introduced to participants once. The presentation of the videos, however, was randomized for each participant, meaning every participant was presented with a unique, randomized sequence of the stimuli.

3.8 Coding and transcription

After each participant had undergone the laboratory session the data was coded on an Excel spreadsheet (See Appendices 4, 5, 6 and 7). Each participant’s results were recorded according to their pseudonym. The results from each section were averaged and the average was then used to determine the percentage of correct answers. The spreadsheet was separated according to the condition and will be explained below:

3.8.1 Audio-only condition coding

In the audio-only condition, the coding was separated according to each stimulus (See Appendix 4). The first column consisted of Audio-Tonal Pair 1 (A-TP1) which was sub-divided into three headings namely, Audio 1 Identification of Low High Tone (A1 Id of LHT), Audio 2 Identification of High Low Tone (A2 Id of HLT) and Identification of Difference (Id of DIFF). Table 3 below provides a visual representation of the coding scheme for the audio-only condition:

Table 3: A representation of the audio-only tonal pair coding schema

A-TP 1			A-TP 2		
A1 ID OF LHT	A2 ID OF HLT	ID OF DIFF	A1 ID OF HLT	A2 ID OF LHT	ID OF DIFF

Under each heading, if the word was perceived correctly, a score of 2 was recorded. If the participant produced an incorrect response, a score of 1 was recorded. The sub-heading ID OF DIFF, if the participant was able to identify the tonal difference between the two words they were scored a 2 while failure to recognise the tonal difference was scored a 1.

The audio-only minimal pair consisted of two word pairs A-MP 1 (Audio-only minimal pair 1) and A-MP2 (Audio-only Minimal Pair 2). Each word pair was made up of two words. If the word was perceived correctly the participant was scored a 2 while an incorrect perception of the word was scored a 1. Table 4 is a visual representation of the coding schema for the audio-only minimal pair condition:

Table 4: A representation of the audio-only minimal pair coding schema

A-MP 1		**A-MP 2**	
MP ID WD1	MP ID WD 2	MP ID WD 1	MP ID WD 2

3.8.2 Visual-only condition coding

The coding for the visual-only condition can be found in Appendix 5. In this condition, after watching each stimulus twice participants could choose from a list of four options the word they believed they observed. A correct perception of the word from the list saw participants score a 2 while incorrectly choosing a word from the options resulted in a score of 1. The table below is a visual representation of the coding scheme:

Table 5: A representation of the visual-only condition coding schema

V-WP		V-MP	
ID OF WD 1 "isithelo"	ID OF WD 2 "isithini"	MP ID W1 "ukuzala"	MP ID WD 2 "ukusala"

3.8.3 Audio-visual condition coding

The audio-visual condition was coded in the following manner (See Appendix 6). The first part of the audio-visual condition was the tonal pair. For each pair, participants had to first recall the word they observed. If they perceived the correct word they were scored a 2 and if a word was perceived incorrectly they were scored a 1. After observing both words participants were asked whether there was a tonal difference present between the two words. ID OF DIFF column measured the tone difference between the two words. If the tone was detected between the two words participants were scored a 2 while failure to recognise the tone was scored a 1. Table 6 below represents the coding schema of the audio-visual tonal pair condition:

Table 6: A representation of the audio-visual tonal pair coding schema

AVTP 1			AVTP 2		
AV 1.1: Tonal LHLL	**AV 1.2 Tonal HLHL**	**ID OF DIFF**	**AV 2.1: Tonal HLHL**	**AV 2.2: Tonal LHLL**	**ID OF DIFF**

The audio-visual minimal pair condition (AVMP) consisted of two minimal word pairs (AVMP 3 and AVMP 4). These word pairs each consisted of two words. A correct identification of the word was scored a 2 while an incorrect perception of the word was scored a 1. Table 7 table below illustrates the schema of the audio-visual minimal pair condition:

Table 7: A representation of the audio-visual minimal pair coding schema

AVMP 3		AVMP 4	
AV 3.1: MP	**AV 3.2 MP**	**AV 4.1 MP**	**AV4.2 MP**

3.8.4 Incongruent condition coding

The incongruent condition was coded as follows:
The incongruent stimuli was coded as shown in Table 8, with the incongruent condition consisting of three stimuli (See Appendix 7). As seen in the Table 8, the stimuli were labelled Incongruent 1, Incongruent 2 and Incongruent 3.

Table 8: A representation of the incongruent condition coding schema

INCONGRUENT STIMULI 1	INCONGRUENT STIMULI 2	INCONGRUENT STIMULI 3

After being exposed to each stimulus participants had to recall what word they observed. The data was coded as follows: if the participants recalled the visual of the word it was coded a "1" if the participant recalled the vocal of the word it was coded a "2" and if the word recalled matched neither the visual nor the vocal of the word it was coded a "3". After gathering this information the data was recorded on an Excel spreadsheet where the averages and percentages were calculated for the number of visual, vocal and incongruent responses from participants.

3.9 Reliability and Validity

The pilot study allowed us to determine how reliable the research method was. As mentioned above, after conducting the pilot study relevant changes were made to achieve a precise method (Akram et al, 2019). Similarly, documenting the data collection process according to (Yin, 2003) enhances a study's reliability and was used later for inter-reliability testing. Similarly, peer debriefing took place weekly during the data collection and throughout the transcription and coding of the data in order to enhance the accuracy of the study (Creswell, 2008)

3.10 Ethical considerations

Participation in the study only took place once ethical clearance was granted from the Health Research Ethics Committee (HREC) at the Witwatersrand (See Appendix 3) for the ethical clearance certificate). All research procedures in this study strictly complied with the guidelines set out by the ethics committee. Before the commencement of the laboratory sessions, participants were given a participation information sheet as well as a consent form to complete (see Appendices). This information granted participants a well-rounded view of the study and what was expected from them and allowed them to make an informed decision before participation. Participation in this study was completely voluntary and individuals could withdraw from the study if they wished to. After reading the Participation Information Sheet individuals had the right to refuse participation without giving a reason. They were also made aware that their answers in the research task would be used in current and future analyses only. The identities of participants were kept anonymous and the results of the study divulged no identifying factors of individuals.

3.11 Summary of the chapter

The following chapter outlined the methodological strategy utilised in this research study. The process of data collection and the data analyses used were discussed in detail. Overall, this study made use of the quantitative research methodology with a snowballing sampling technique. By adopting this method the strict participant criteria were adhered to. Participants were students at the University of the Witwatersrand aged between 17 and 29 years who were new L2 learners of isiZulu. In participating in the study, they were introduced to a variety of modalities and stimuli as mentioned above. The data for each condition was coded and recorded as explained above. The next chapter will present the analysis of the data collected.

CHAPTER 4

Analysis

4.1 Introduction

In investigating the effect of modality on audio-visual speech, this following chapter reports on the results and analysis of the extent to which young adult bilinguals benefit from audio-visual speech. This study investigates the effect of modality on audio-visual speech perception. For the purpose of the present study, we will revisit the research questions in order to analyse the results.

> Research question 1: What is the effect of modality on audio-visual perception?
>
> Research question 2: How do different conditions affect the way L2 bilinguals perceive audio-visual speech?

In order to study the effect of modality and the extent to which different conditions affect how L2 individuals perceive audio-visual speech we will first present the results of various conditions undertaken by participants. The cohort of this study consisted of university students between the ages of 17- and 29-years learning isiZulu with more than 5 months of exposure to learning isiZulu.

The different conditions undertaken by participants were as follows:

4.1.1 Audio-only condition

These stimuli consisted of four tasks. The first two sub-tasks consisted of tonal pairs while the second two sub-tasks consisted of a minimal pair set. In the tonal pair sub-task participants were exposed to a word which they had to recall aloud. After hearing and recalling the second word pair in the tonal tasks sub-task participants were asked to identify if a tonal difference was present. The tonal pairs were:

Tonal Pair sub task 1: [bòna] (they) and [bónà] (to see)

Tonal Pair sub task 2: [dúmà] (be tasteless) and [dùmá] (roar)

Minimal Pair: sub task1: [usisi] (sister) and [isisu] (stomach)

Minimal Pair sub task 2: [ibhola] (ball) and [ibhala] (bar)

4.1.2 Visual-only Condition

In this stimulus set, participants viewed a clip of an individual voicing out a word without any audio. Participants were asked to recall the word and were subsequently shown the video again wherein a set of four options were given and they were asked to say the word that sounded most alike to the word they watched. Figures 22 and 23 below are an illustration of the clip participants were asked to watch followed by the options provided to them.



Figure 20: An illustration of the visual-only stimuli participants were exposed to



Figure 21: An illustration of the options participants received after viewing the visual-only stimulus

The stimulus in the visual-only condition consisted of:
“isithelo” “(fruit) and “isitini” (brick)
“ukusala” (to remain) and “ukuzala” (to bear)

4.1.3 Audio-visual only condition

In this task, participants were exposed to four sub-tasks. The first two-sub tasks consisted of tonal pairs while the last two sub-tasks consisted of minimal pairs. Individuals were asked to watch a video of an individual saying a word that consisted of both audio and visual stimuli. Thereafter they were asked to recall the word they observed.

The stimulus in the audio-visual condition tonal condition was:
“ùkúsìndà” (to smear) and “ùkùsíndà” (to be heavy/to recover)
“ùkwèthukà” (to insult) and “úkwéthùkà” [to get a fright]

The minimal pair stimulus in the audio-visual condition were:
“ukupheka” (to cook) and “ukubheka” (to look)
“isitho” (organ) and “isipho” (gift)

4.1.4 Incongruent condition

The incongruent task saw participants being exposed to three words that consisted of an audio and visual stimulus. The incongruent stimuli however saw a mismatch in the audio and visual stimulus as the audio of one word was dubbed over the mismatched visual of another word. After viewing the video participants were asked to say what they observed.

Example: Individuals were exposed to a video of an audio “isitho” [organ] over a visual “isipho” [gift]
The following stimulus made up the audio-visual condition:
Vocal “isitho” (organ) over a visual “isipho” (gift)
Vocal “ukulima” (to plough) over a visual “ukulila” (to cry)
Vocal “ukusala” (to remain) over a visual [ukuzala] (to bear)

All participants in the study successfully performed the tasks. This chapter will present the findings and quantitative analysis of each condition. The following sections will present the findings per sub-task.

4.2 Audio-only condition findings

4.2.1 Audio-only tonal pairs

In the first sub-condition participants were provided with two sets of tonal pairs. The first tonal pair (TP 1) consisted of the words [bòna] (they) and [bónà] (see) with the first word being made up of an LHT and the second word being a HLT. The second tonal pair (TP 2) consisted of the words [dúmà] (be tasteless) and [dùmá] (roar; be famous). The first word in this pair contained a HLT whilst the second word had a LHT. Participants had to first recall the first word they heard and thereafter the second word after hearing both words in the pair participants were asked whether the two words were the same or not. Asking this question allowed us to test whether the participants observed the tonal difference between the two words. Tone detection has been labelled ID (identification of the tone). For a correct answer, the code was a "2", and for an incorrect answer, the coding was labelled a "1" to allow for analysis. After gathering the data, the answers were averaged, and a percentage was then generated. The summary of results for the tonal pair subtask can be seen in Table 9 below:

Table 9: Table showing the tonal pair stimulus in audio-only condition

	TP 1			**TP 2**		
	ID OF LHT	ID OF HLT	ID OF DIFF	ID OF HLT	ID OF LHT	ID OF DIFF
Average $\bar{x}$	2,00	2,00	1,75	1,92	1,92	1,58
Percentage %	100,0	100,0	87,50	95,83	95,83	79,16

Table 9 displays the overall results of the two tonal pairs presented to participants. In the tonal pair stimulus, participants were exposed to a word through an auditory modality and were

asked to recall the word thereafter they were exposed to the second word of the tonal pair where they were asked to recall the word. Similar averages were calculated across the tonal pairs with the identification of TP 1 being 100,0% (for the identification of LHT and HLT) and the identification of TP2 at 95,83% (for the identification of HLT and LHT). Overall, a higher success rate was noted in the perception of the first tonal pair (TP1) compared to the second tonal pair (TP 2) where one participant was incorrect however, both stimuli were answered well overall.

The second part of the task was to ask participants if they perceived a difference in the tone within the tonal pair. The results can be found in Table 9 under the column "Id of DIFF". This column refers to whether participants were able to identify that there was a tonal difference between the word pair. In TP 1 87,50% of participants were able to identify to difference between the LHT tone and the HLT tone compared to TP 2 where only 76,19% of participants were able to identify the difference between the HLT and LHT. The results for this can be seen below (Figure 24).

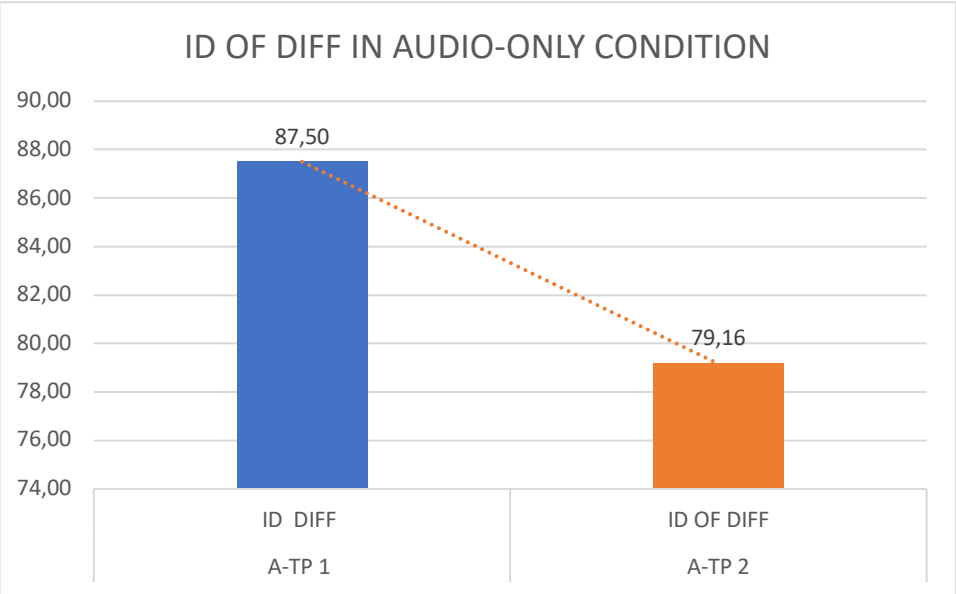


Figure 22: A bar graph displaying the identification of the difference between the audio-only tonal pair words

As seen in the graph above participants had more difficulty in identifying the tonal difference between the second tonal pair "dúmà" (be tasteless) and "dùmá" (roar) compared to the first pair "bòná" (they) and "bónà" (see). Therefore, in terms of tonal perception, they had a better

chance of perceiving tonal differences when they were introduced to the LHT first as can be seen in “bòna” (they) and “bónà” (to see). The results revealed that although participants were able to distinguish tonal pair words individually, they found difficulty in distinguishing the tonal difference between the word pairs.

Interestingly, although the perception of the individual word was detected well, participants experienced more difficulty in identifying the “Id of DIFF” highlighting an increased difficulty in identifying the tonal difference between the word pairs.

4.2.2 Audio-only similar word pairs

The second sub-condition saw participants being introduced to two sets of minimal pair words through an audio-only stimulus. The first minimal pair was made up of the words “usisi” (sister) and “isisu” (stomach). The distinction between these two words is the vowel [i] and the vowel [ɛ]. The second pair consisted of the words “ibhola” (ball) and “ibhala” (roar). The difference between these words is the back vowels [o] and [a]. Participants were introduced to the word pairs and after every word they recalled what they heard. If participants correctly detected the word, they were scored a 2 and if they incorrectly recalled a word, they were scored a 1. Table 10 below is a summary of the participant's performance:

Table 10: Table showing the results of minimal pairs in the audio-only condition

	A-MP 1		**A-MP 2**	
	MP ID W1	MP ID WD 1	MP ID W2	MP ID W2
Average $\bar{x}$	1,83	2,00	1,75	1,83
Percentage%	92,66	100,0	87,50	91,66

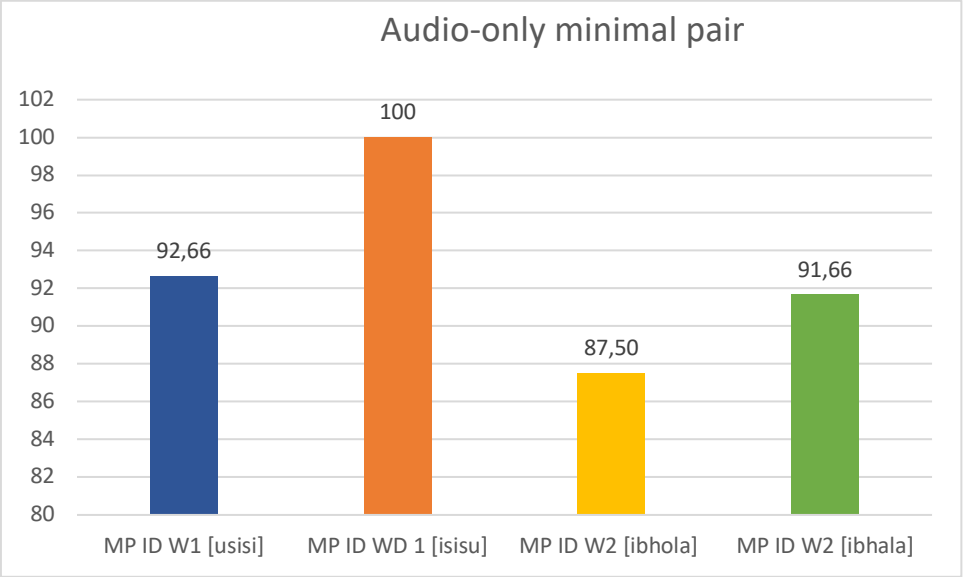


Figure 23: A bar graph displaying the results of minimal pair words in the audio-only condition

Figure 25 displays the results of participants in the audio-only minimal pair sub-task in percentages and averages. In this task, participants were asked to recall the words presented to them after hearing them through an auditory stimulus. In the first minimal pair- "usisi" (sister) and "isisu" (stomach) participants perceived the word "isisu" (stomach) better as participants received a 100% in recalling the word, the corresponding minimal pair "usisi" (sister) performed well too with the word being recalled correctly by a large number of participants with a 92,66% correct recall percentage. Participants had the most difficulty in recalling the word "ibhola" (ball) suggesting difficulty in perceiving the mid, back, rounded [o] in comparison to the open, central unrounded [a] sound found in [ibhala] "bar". These results highlight that participants perceived the [ɛ] sound best compared to the vowel sounds [i], [o] and [a] through an auditory stimulus.

The study had two audio-only sub-conditions: tonal pair and minimal pair. The results showed that participants were able to identify between the individual words very well, however, some participants found it difficult to distinguish the tonal difference between the tonal word pairs. This shows that L2 speakers had success in identifying isiZulu words through an audio stimulus however they experienced some difficulty in identifying tone differences between words through an auditory modality only particularly with [i], [o] and [a] sounds.

4.3 Visual-only condition findings

The second condition that participants were exposed to was visual only. In this condition, individuals were exposed to clips with no audio. The words in these clips were two minimal pair words- the first similar-sounding word pair was made up of the words “isithelo” (fruit) and “isithini” (brick) and the second pair consisted of the words “ukusala” (to remain) and “ukuzala” (to bear). Participants were exposed to the video twice before being given options wherein they selected the word that sounded most like the clip they observed. The results were coded in the following way: participants who correctly identified the word scored a “2” whilst an incorrect identification was scored with a “1”. The table below (Table 11) is the summary of the visual-only condition results in averages and percentages:

Table 11: Table showing the results of tonal pair in the visual-only condition

	ID of [isithelo]	ID of [isitini]	ID of [ukusala]	ID of [ukuzala]
Average $\bar{x}$	1,67	1,42	1,33	1,42
Percentage %	83,33	70,83	66,67	70,83

The above table is a summary of the results produced by participants in the visual-only condition. The table above displays the results of the visual-only condition. Participants best perceived the alveolar [t^h] visually as seen in the word [isitelo] “fruit”. Although the alveolar [t’] in “isi**t**ini” (brick) was perceived well overall participants perceived “isithelo” (fruit) more correctly. The voiced, alveolar, fricative [z] performed well whilst participants found it most challenging to perceive the voiceless [s] sound through a visual stimulus. The voiceless [s] sound was apparent in the word “ukusala” (to remain) and it was where participants had the highest number of errors in perceiving the word. This shows that participants found it easier to visually perceive the voiced [z] sound compared to the voiceless [s] sound whilst the word “isithelo” (fruit) and “isitini” (brick) yielded similar results. It is important to not that in this condition participants had the chance to perceive the word twice and were also given options to decide which word matched the word they perceived. After watching the stimulus once participants had difficulty in recalling what they observed it was only after perceiving the word

for a second time together with a list of options were participants able to decide what they perceived visually.

Similarly, the alveolar [t^h] was perceived correctly more often compared to the alveolar [t'] sound. The illustrations below are a visual representation of the difference in the [s] and [z] sound. From these sounds, participants had the most difficulty in perceiving the [s] sound due to the similarity in the visual pronunciation of the sound. Overall, partcipants performed adequately in the visual-only modality.

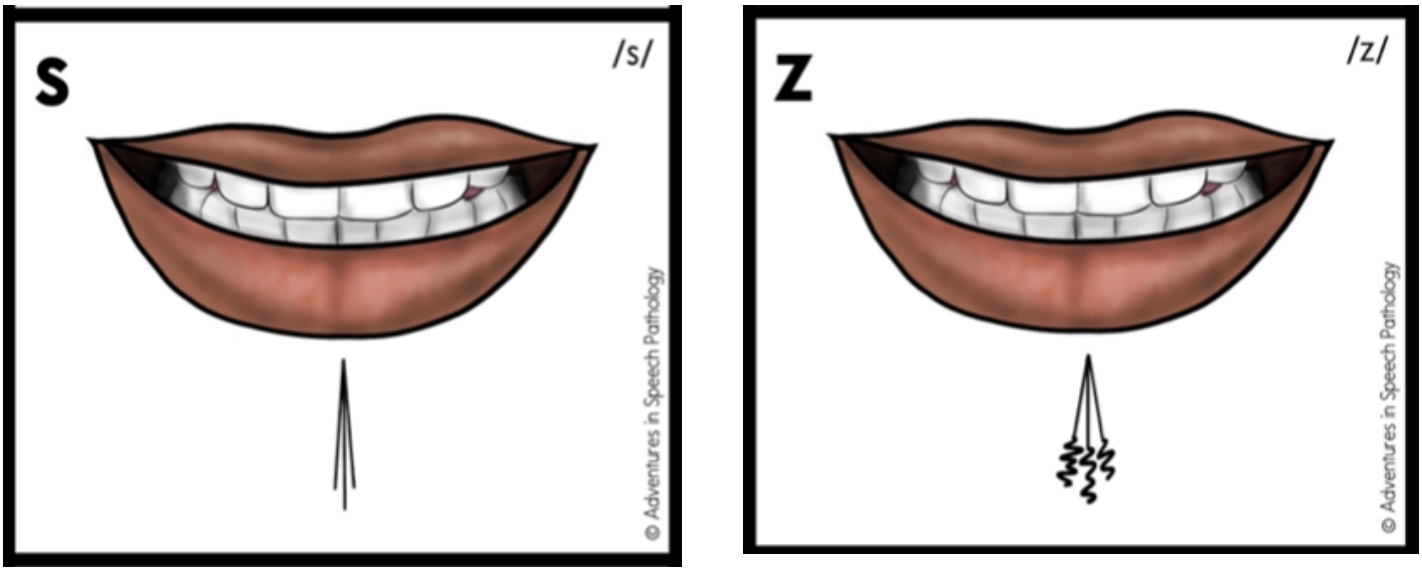


Figure 24: An illustration of the visual cues in the /s/ and /z/ sound

4.4 Audio-visual condition

Overall, in the visual-only condition participants performed equally however some sounds proved to be more difficult to perceive through a visual-only stimulus. The voiceless alveolar fricative of the /s/ and the voice alveolar nasal of the /n/ were overall the most difficult sounds to perceive through a visual-only stimulus.

4.4.1 Audio-visual tonal pair

The third condition in the study was audio-visual. Participants were exposed to a clip that consisted of both audio and visual modalities and were they asked to recall what they observed in the clip. The words in the clip consisted of two tonal pairs and two minimal pairs. For the

first audio-visual tonal pair (AVTP 1) the words participants were exposed to were “ùkúsìndà” (to smear) and “ùkùsíndà(to be heavy/to recover. The word “ùkúsìndà” (to smear) is made up of a low, high, low, low tone (LHLLT) while “ùkùsíndà” (to be heavy/to recover) has a high, low, high, low tone (HLHLT). The second audio-visual tonal pair (AVTP 2) was “ùkwèthukà” (to insult) which has a high, low, high, low tone (HLHLT) and [úkwéthùkà] “to get a fright” consisting of a low, high, low, low tone (LHLLT). Participants were first asked to identify both words in the pair by recalling what they observed in the tonal pair before being asked if there was a tonal difference between the words. Identification of the tonal difference was recorded as “ID OF DIFF” as seen in the table below. Correct identification of the word was labelled as 2 and an incorrect identification was labelled as 1 similarly, if a tonal difference was detected between the word pair it was coded as 2 and no tonal detection was labelled as 1.

Table 12: Table showing the results of tonal pairs in the audio-visual condition

	AVTP 1			**AVTP 2**		
	LHLLT [ùkúsìndà]	**HLHLT [ùkùsíndà]**	**ID OF DIFF**	**HLHLT [ùkwèthukà]**	**LHLLT [úkwéthùkà]**	**ID OF DIFF**
Average $\bar{x}$	1,92	2	1,83	2	1,92	1,67
Percentage %	95,83	100,0	91,66	100,0	95,83	83,33

Table 12 represents the results from the two audio-visual tonal pairs participants encountered. Participants were exposed to both the audio and visual modalities in these clips. Overall, participants were successful in recalling the words they observed from the clips this can be noted in AVTP 1 for the word “ùkùsíndà” (to be heavy/to recover) which received a 100% correct recall/identification of the word . Its pair word in the LHLLT column “ùkúsìndà” (to smear “ùkwèthukà” (to insult) and “úkwéthùkà (to get a fright) showcased good recall of the words with 100% and 95.83% correct perception recorded for each word respectively.

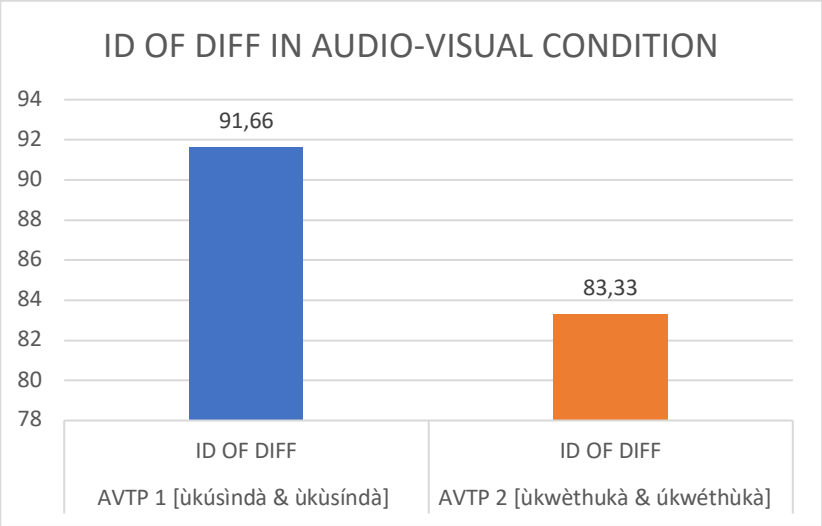


Figure 25: A bar graph displaying the tone detection in the audio-visual condition

Figure 27 illustrates the word pair "ùkúsìndà" (to smear) and "ùkùsíndà" (to be heavy/to recover) had a higher success rate of tone detection compared to "ùkwèthukà" (to insult) and "úkwéthùkà" "to get a fright". This shows that participants detected the tonal difference better when the LHLLT word was introduced first. However, the tone detection between both tone pairs were high.

4.4.2 Audio-only minimal pair findings

In this sub-condition termed audio-visual minimal pair (AVMP) participants observed clips with two sets of minimal pairs (MP) through an audio-visual modality. In this condition, the minimal pair word sets consisted of [ukupheka] "to cook and [ukubheka] "to look" and [isitho] "organ" and [isipho] "gift". After observing each word participants were asked to recall what they observed in the clip. The results were calculated as percentages below:

Table 13: Table showing the results of minimal pairs in the audio-visual condition

	AVMP 3		**AVMP 4**	
	AVMP 3.1	**AVMP 3.2**	**AVMP 4.1**	**AVMP 4.2**
Average $\bar{x}$	2,00	2,00	2,00	2,00
Percentage %	100,0	100,0	100,0	100,0

In the audio-visual condition, participants were exposed to two sets of minimal pair words. The first word pair “ukupheka” (to cook) and “ukubheka” (to look) both had a 100% recall rate amongst participants. This was also seen in the second minimal pair set “isitho” (organ) and “isipho” (gift) were a 100,0% of participants successfully recalled both words. The audio-visual condition displayed the best overall results with participants finding it easier to perceive L2 words when both the audio and visual stimulus was present. In all, 91,66% and 83,33% of participants were also able to identify the tonal difference between the tonal pairs showing that participants did benefit when having both modalities present.

4.5 Incongruent condition findings

To answer the extent to which incongruent speech affects audio-visual speech perception an incongruent condition was created. In this condition, participants were exposed to three minimal pair stimuli that were incongruent. The incongruency based on the McGurk effect (1976) saw the audio of one word of the minimal pair being voiced over the visual of the other word. For example, in the first minimal pair of incongruent stimuli (MPI 1), the word “isitho” (organ) was vocalised over the visual of the word “isipho” (gift). The second stimulus (MPI 2) saw the audio of the word “ukulima” (to plough) merged over the visual of “ukulila” (to cry). The last stimulus (MPI 3) consisted of the visual “ukuzala” (to bear) over the audio “ukusala” (to remain). Participants were exposed to the clip wherein they were asked to repeat what they observed after watching the clip. Participants who recalled the visual of the word were scored a “1” the auditory recall of the words was scored a “2” whilst the recall of neither the visual or vocal of the words was scored a "3" and was labelled incongruent.

Table 14: Table showing the results in the incongruent condition

	Incongruent 1	Incongruent 2	Incongruent 3
% Incongruent	75,0	83,30	0
% Vocal	25,0	0	100,0
% Visual	0	16,70	0

Overall, in the first stimulus (Incongruent 1) 75,0 % of the words produced by participants were incongruent meaning that the words they reported consisted of neither the visual nor the vocal of the word and were made up of syllables that were not present in either word. Of the 75% incongruent perceptions 66,70% of participants perceived the word "isi**p**to" whilst 8,30% perceived the word "isi**m**to". This shows that in the incongruent condition the audio [t^h] and the visual [p^h] was perceived most often as a [p] or a [m] sound which are both bilabial. Similarly, 25,0% of the participants produced the vocal of the word meaning the vocal modality was stronger than the visual modality. Therefore, the audio of the word [isitho] "organ" dubbed over the visual word [isipho] "gift" produced mostly incongruent perception whilst the remaining 25,0% perceived the audio of the word.

In the second stimulus (Incongruent 2) 83,33% per cent of participants produced an incongruent word. The 83,33% incongruency was made up of 58,33% of participants who perceived the word "ukuli**n**a", 16,67% who perceived the word "ukuli**ng**a" and 8,33% who perceived the word "ukulile". The incongruency was created by dubbing the audio of [ukuli**m**a] "to plough" with the visual of [ukuli**l**a] "to cry". The minimal pair contrasted in the [m] and [l] sound. Therefore, the 58,33% of participants that perceived the word "ukulina" merged the acoustic properties of the [m] (voiced, bilabial, nasal) and [l] (voiced, alveolar, lateral approximant) sound and perceived an [n] (voiced, alveolar, nasal) sound. The other 16,67% who perceived the word "ukuli**ng**a" perceived the [m] and [l] sound as [ng]. Another 16.67% produced the visual of the word "ukulila" (to cry) whilst 8, 33% perceived the word "ukulile." This relates that the visual modality took precedence over the vocal modality as no participant perceived the word "ukulima" (to plough). This shows that a majority of participants perceived an incongruent word whilst the remaining 16,67% perceived the alveolar lateral approximant [l] sound which was through the visual stimulus.

Interestingly, in the last stimulus, 100,0% of participants produced the vocal of the word "ukusala" (to remain) relating that the vocal modality was favoured over the visual modality. Due to 100% of participants observing the word "ukusala" (to remain) the voiceless, alveolar, fricative in [s] was perceived better through an auditory stimulus compared to the voiced alveolar fricative of the [z] which was presented visually.

In the first minimal pair when the audio of the word [isi**th**o] (organ) was dubbed over the visual of the word [isi**ph**o] (gift) the difference in the word is in the [t^h] and [p^h] of the sound. Both the [t^h] and [p^h] are plosives with [t^h] being an alveolar and [p^h] being a labial. Therefore, the merging of an audio plosive [t^h] and visual labial [p^h] resulted in 75,0% incongruent perception and 25,0% vocal perception. In the second stimulus when the vocal [ukuli**m**a] (to plough) was dubbed over a visual [ukuli**l**a] "to cry", 83,33 % per cent of participants perceived an incongruent form of the word while 16.70% perceived the visual of the word. This shows that when the properties of the place and manner of articulation of two sounds differ there is a higher chance of an incongruent word being perceived. In the last stimulus, the audio of [uku**s**ala] (to remain) was dubbed over the visual of [uku**z**ala] (to bear) had no incongruent responses as all the participants recalled and perceived the audio of the word which was [ukusala] (to remain). What can be noted here is that when the two sounds [s] and [z] shared the place or articulation, and manner of articulation no incongruent response was perceived.

4.6 Summary of chapter

This chapter provided an analysis of the results obtained. The results displayed that participants could successfully recall words in an auditory condition however, it was more difficult to detect tone differences in an audio-only condition. The visual-only condition highlighted the difficulty L2 speakers had in perceiving lexical cues visually. Overall, the results revealed that in the audio-visual condition, participants performed best in recalling words and detecting time between words. Providing participants with both an auditory stimulus and a visual stimulus yielded the best perception of lexical cues in L2 speakers.

The incongruent condition saw the most variation in responses. The results varied based on the stimuli presented to the participants. In the first incongruent stimulus, the majority of participants perceived an incongruent word while the remaining 25% perceived the vocal form of the word. In the second incongruent stimulus, most of the participants perceived an incongruent word whilst 16,70% perceived the visual form of the word. While in the last stimulus no participants perceived an incongruent word as all of the participants recalled the audio of the word. It is interesting to note that when the phonemes shared a place and manner of articulation the stimulus did not yield an incongruent perception. The following analysis will be discussed in the next chapter.

CHAPTER 5

Discussion and conclusion

5.1 Introduction

The purpose of this study was to examine the extent to which young adult L2 bilinguals benefit from audio-visual speech. In order to achieve this, young adult learners aged between 17 and 29 years who were late isiZulu learners were selected. The participants were university students with various levels of exposure to isiZulu as an L2 ranging from 8 to 36 months of exposure. This study aimed to answer the following question:

Research Question 1: What is the effect of modality on speech perception in young bilingual adults?

Research Question 2: How do different stimuli affect the way L2 bilinguals perceive audio-visual speech?

In order to answer the above questions four conditions with varying perception modalities were introduced to participants. The conditions were:

a.) Audio-only condition

b.) Visual-only condition

c.) Audio-visual condition

d.) Incongruent condition

The findings were conducted to respond directly to the research questions. The main areas of focus identified were the difficulty in detecting tonality in a visual modality, the modality which received the best results was the audio-visual modality, while the most varied results occurred in the incongruent condition.

The following discussion is broken down in line with the study's research questions and the main identified themes from the results are explained below under each headed question.

Research Question 1: What is the effect of modality on speech perception in young bilingual adults?

In order to answer Research Question 1, three conditions were conducted, and the overall results are discussed accordingly.

In Condition 1: Audio-only condition

Using audio clips consisting of various tonal and minimal word pairs participants were asked to recall what they had heard on these clips. Overall, the results revealed that participants performed well in detecting words they had heard from a vocal-only modality only (across both tonal and minimal pair stimuli). However, the participants faced difficulty in detecting the tonal difference between two words of a tonal pair. Similarly, participants performed better in identifying the tonal difference between words when they were first introduced to the LHT word first. This experiment also revealed that individuals recalled and found it easier to make a distinction between the [i] sound compared to the [u] sound and that the [o] sound was not perceived as well as the [a] sound.

In Condition 2: Visual-only condition

In this condition, participants were twice introduced to a visual (without audio) clip before they were asked to choose from among four words the one word they perceived to be in the visual clip. Throughout the study, participants across the study had increased difficulty in making out the words with a visual-only modality. Only after the second exposure to the clip together with the four options did they choose the word they believed to be most similar. The experiment revealed overall that the [z] and [t^h] was perceived well the sound that participants had the most difficulty in perceiving with the visual-only stimulus was the [s] sound.

In Condition 3: Audio-visual condition

In this experiment, participants were asked to recall the words they observed. The clips participants were exposed to consisted of both the audio and visual modalities. In the audio-visual condition, participants performed well in recalling words across both the tonal and minimal pair stimuli. The key finding was that participants found it easier to detect the tonal difference between tonal pairs when they were exposed to the LHLLT word first. For the minimal pair stimuli, participants performed well overall in recalling the word they were exposed to.

Research question 2: How do different stimuli affect how young L2 bilinguals perceive audio-visual speech?

To answer this question, participants were introduced to incongruent stimuli and the results were as follows:

In this experiment, participants were exposed to three incongruent stimuli. The incongruency was created by dubbing the audio of one sound over the visual of another sound (McDonald & McGurk, 1976). The results of each stimulus varied widely and will be explained below:

For the first stimulus the audio of the word “isitho” (organ) was dubbed over the visual of the word “isipho” (gift). After being exposed to the clip once the participant's results revealed that 25% of the participants recalled observing the vocal of the word “isitho” (organ). 8,30% of participants recalled the word “isi**m**tho” here participants added the [m] sound, which is a voiced, bilabial, nasal before the vocal of the word “isitho” (organ) to create “isimtho”. The other 66,70% of participants, however, recalled a newly created word “isiptho”. This newly created word “isipto” was created by participants as they fused the vocal [t] (voiceless, alveolar, plosive) sound in “isi**t**ho” (organ) with the visual of [p] (voiceless, bilabial, plosive) found in the visual of the word “isi**p**ho” (gift) participants, therefore, eliminated the [h] sound (voiceless, glottal, fricative).

In the second stimulus the vocal of the word “ukulima” (to plough) was dubbed over the visual of the word “ukulila” (to cry). In this stimulus, results revealed that 16,7% of participants

recalled the visual of the word “ukulila” (to cry). The other 83,30% of participants either recalled the word “ukuli**ng**a” In this realization participants eliminated the [m] (voiced, bilabial, nasal) and [l] (voiced, alveolar, lateral approximant) sound and added a [ng] sound to create “ukuli**ng**a” or recalled the word “ukulina”- in this word, participants fused acoustic properties from both the [m] (voiced, bilabial, nasal) and [l] (voiced, alveolar, lateral approximant) sounds and replaced it with the [n] sound (voiced, alveolar, nasal).

The third stimulus saw the vocal of the word “ukusala” (to remain) being dubbed over the visual of the word [ukuzala] “to bear”. This stimulus was clear-cut as all participants recalled the vocal of the word [ukusala] “to remain”. This revealed that the vocal of the [s] (voiceless, alveolar, fricative) was realized better than the visual of [z] (voiced, alveolar, fricative).

Based on the above findings, each question will now be broken down and discussed.

5.2 Research question 1:

5.2.1 What is the effect of modality on speech perception in young adult bilinguals?

To investigate the above question, participants were introduced to three different conditions. Each condition contained stimuli from different modalities (i.e. audio-only, visual-only and audio-visual stimuli). This study chose to use a recalling task as recall protocols supply insight into the organization of stored information, retrieval and perception.

In the audio-only condition of tonal pairs, overall participants performed well overall in recalling the tonal and minimal pairs they were exposed to. The difficulty, however, arose after hearing both word pairs. Participants were asked whether or not the words, they heard were the same or not. This answer provided insight into whether participants could distinguish tonal differences between words through an audio-only stimulus. In the first stimulus pair “bònà” (they) and “bónà” (to see) 87,50 % of participants observed a tonal difference between the words while in the stimulus pair “dúmà” (be tasteless) and “dùmá” (to roar) only 79, 19% of participants observed a tonal difference in the word. In this study the participants distinguished the tonal difference more often when exposed to the LHT word first and thereafter the HLT word. In the audio-only minimal pair condition, participants performed well in recalling what they heard. The word with the most incorrect responses was [ibhola] “ball” with the distinguishing [o] sound.

Pisoni (1993) related that introducing L2 listeners to various instances of L2 sounds presented through different modalities and in different conditions could lead to successful perceptual training. The perception of speech does not follow a specific pathway for all listeners; thus speakers construct perception in a way that will best adjust the acoustic-phonetic properties of that particular language's conditions. The accuracy of speech perception is best represented through the contrasts that exist in the acoustic reality and the actual understanding of one's native language. For example, an experiment revealed that American English and Japanese listeners differed considerably in their perception of /ra/ and /la/ when these syllables were introduced in the context of speech but not when they were introduced in non-speech contexts (Goto, 1971; Miyawaki et al., 1975). This revealed that Japanese and American English listeners executed uniformly when perceiving the central acoustic properties that differentiate the English consonants of /r/ and /l/ when performed in a non-speech context. This shows that when participants heard the acoustic properties that contrasted the two consonants in a speech context, their language-specific knowledge directed their prejudice between such contrasts. However, when the auditory contrast was found in a non-speech context an individual's general auditory processing guided their prejudice. Similarly, in another study on the perception of phonotactics in French and Japanese listeners, it was revealed that participants phonologically recognised the contrasts permitted in their linguistic repertoire while they auditorily distinguished illegitimate differences (Morison et al., 2023). These studies suggest that the variation between speech perception and auditory perception exists based on how the individual observes the phonological changes differently from the auditory changes. This is largely based on their L1 which in this study was English.

How individuals utilise articulatory cues is based on the participant's L1, the number of visemes in the L1 phonetic inventory and whether or not it is a tone language. These are all factors that could affect how L2 speaker's perceive non-native cues (Li, 2015). In this study, participants had more difficulty in detecting tone through an auditory condition compared to an audio-visual condition. However, L1 English speakers of a non-tone language most likely do not have effective sensitivity to tone as tone does not exist in their L1. Thus, as new adult bilinguals, they may have difficulty in perceiving non-native contrasts especially when they do not exist in their native L1 (Pereira Reyes & Hazan, 2021). The result of various research studies have been in agreement with this, concluding that learning lexical tone as an L2 speaker

is challenging (Schaefer & Darcey, 2014; Wang, 2013; Wayland & Guion, 2004). Non-tonal L1 speakers do not perceive tones categorically (Hallé et al., 2004; Peng et al., 2010; Wang et al., 2020) and their perception of tone varies across languages and conditions (Wang et al., 2020).

In the visual-only condition, participants performed well overall in recalling non-native words. The detection of tonality was marginally higher in the audio-visual condition. However, the visual-only condition yielded satisfactory results. The satisfactory rate of tone detection in the audio-only condition could also be the result of the stimuli not being as challenging for the L2 speaker. It is however, important to note that in the visual condition, additional cues were provided to participants (such as being exposed to the stimuli twice and having a list of words to choose from where participants could pick the word that they believed best resembled what they observed in the clip).

In the visual-only condition, participants had difficulty recalling the word after being exposed to it for the first time. Only after the second exposure and through providing participants with a set of options were they able to select a word they found best resembled what they observed in the video clip. The visual-only word that was perceived best visually only was “isithelo” (fruit); its word pair “isitini” (brick) also performed well. In the second stimulus, participants perceived the [z] sound in “uku**z**ala” (to bear) which is a voiced, alveolar, fricative better than they perceived the [s] voiceless, alveolar, fricative in the word “ukus**a**la” (to remain) through a visual modality. In the visual-only modality, participants were not able to grasp the word at first and needed additional exposure to the word together with options in order to grasp the word they had observed visually.

In the visual modality of this study, participants were adequate in perceiving words from a visual modality. However, participants were exposed to the stimulus twice and were provided options for the word they believed best resembled the word they had observed. The good perception rate seen in the results may be because participants were provided with options of words they thought best suited what they were exposed to. Therefore, it could be said that participants were proficient enough to relatively easily comprehend the phoneme contrasts without great difficulty. This reveals that non-native L2 speakers do in fact utilise visual cues

to comprehend what they have seen. This is in agreement with studies such as Montero Perez (2022) who revealed that audio-visual exposure to L2 speech aids in L2 development.

The extent to which L2 speakers make use of visual cues in speech perception is dependent on the informational value of the cues. The use of visual cues differs based on how much information they relay. The degree of salience of the visual cue is another prominent factor this can be seen in bilabial and labiodental contrasts which are particularly salient and influence perception more often than alveolar sounds (Hazan et al., 2006). This is relevant as it provides information on the extent to which visual cues can aid non-native L2 perception. Therefore, audio-visual information can aid in the perception of L2 contrasts particularly when they do not exist in the L1 (Hazan et al., 2006). This evidence correlates with the study as it revealed that visual information can only aid perception based on the information value they carry.

Hazan et al., (2006) studied how visual cues affect perception in non-native consonant contrasts. The study investigated the visual salience of English contrasts and used (/b/-/p/) and the (v) consonants together with the /r/ and /l/ contrasts. These visual contrasts were tested in Spanish, Japanese, and Korean speakers of British English. As in the present study, the participants were introduced to three different conditions- Video-only, Audio-only and Audio-visual. For the (/b/-/p/) and the (v) consonants, participants achieved well in the audio-visual and audio conditions highlighting the importance of the audio-visual cue. With regards to the /r/ and /l/ contrasts the study found no audio-visual benefit. The results from Hazan et al. (2006) were similar to this study as both highlight the audio-visual benefit when providing in non-native cues.

Kluge et al. (2009) used the basis of Hazan et al. (2006) as the basis for research into the use of visual cues in the perception of English cues in Brazilian English as a Foreign Language (EFL). Kluge et al. (2009) used word-final nasals /m/ (bilabial) and /n. (alveolar) to investigate any trouble Brazilian English learners could have when they encounter nasal consonants in final-word positions. Although Kluge et al. (2009) study revealed that the audio-visual condition resulted in the correct identification of both bilabial and alveolar nasal consonants the study also revealed that the visual-only condition was not the most strenuous. Therefore, the studies of both Hazan et al. (2006) and Kluge et al. (2009) align with the results in this study highlighting the importance of the audio-visual condition in the perception and identification of non-native contrasts.

The difficulties L2 speakers had in detecting isiZulu words through a visual-only condition may be due to L2 speakers having to learn how to make use of L2 visual cues. Hazan et al. (2006) and Wang et al. (2009) also argue that the perception of L2 cues is based on the experience participants had with visual cues in their L1. Therefore, making non-native L2 speakers aware of visual cues they could benefit from could allow them to benefit from visual cues.

In the audio-visual modality participants performed well in recalling the word they had been exposed to through the different stimuli. There was a high success rate in the identification and recall of L2 isiZulu tonal and minimal pairs. Despite some errors in the tonal pair stimulus, participants performed well in recalling what they observed from an audio-visual modality. Compared to the audio-only modality, participants had a higher success rate in detecting the existence of a tonal difference through an audio-visual modality compared to an audio-only modality.

Rosenblum (2008) reported that during L1 perception of auditory and visual information is assessable and affects speech perception. The weighting of these two modalities however could be influenced by an individual's visual speech-reading capacity, phonemic context of sounds and the particular environment in which they perceive the sound or language (Massaro et al., 1993; McGurk & MacDonald, 1976; Owens & Blazek, 1985). The bimodality of speech in L1 has been well established (Rosenblum, 2008). The focal point of L2 speech perception has been centred around the auditory channels, while the influence of visual cues in L2 perception has not been studied as intensively although some studies have found benefit (Hardison, 2009, 2003; Hazan et al., 2005; Hazan et al., 2006; Wang et al., 2008). This is in line with the current study that found that L2 speakers had performed well in recalling L2 words and detecting tonality when the stimulus presented to them consisted of audio and visual modalities.

Studies have also suggested that adults pay more attention to the mouth of a speaker when perceiving fluent L2 audio-visual speech (Birulés et al., 2020). The difficulty of the task presented to participants could also influence the extent to which L2 speakers rely on audio-visual cues accessible to them. Visual cues support auditory speech in multiple ways. For example, lip movements supply high-level phonemic representants (Bourguignonet et al., 2020), and visual motion information and provide insight into time (McGrath & Summerfield, 1985). If adult speakers assign their attention to particular features of the talker such as their

eyes or mouth it is plausible that their proficiency in an L2 could affect their distribution of attention whether it be to the talker's eyes or mouth. However, it is rare for a late L2 speaker to achieve native-like fluency (Lecumberri et al., 2010) hence, compared to L1 speakers, L2 speakers may always be reliant on visual cues such as the talker's mouth compared to L1 speakers.

In contrast, other studies have revealed that adults rely more on visual cues such as the talker's mouth when they have to perceive L2 speech in noisy conditions in contrast to their L1 (Barenholtz, 2016; Buchan et al., 2008, Lansing & McConkie, 2003). This is because perceiving L2 speech is cognitively more demanding than perceiving native L1 speech (Borghini & Hazan, 2018). Studies have also shown that non-native speakers focus on the talker's mouth regardless of their L2 proficiency (Birulés et al, 2020). This relates to audio modality together with a visual modality offer supplementary cues which likely allow L2 speakers to conquer the processing of uncommon linguistic input.

The overall satisfactory results of the participants in the audio-visual condition highlighted that both auditory and visual modalities enhance the perception of non-native cues. Participants performed well in detecting tonality in the audio-visual condition compared to the audio-only condition as the visual cues aided their perception. This is in line with the main findings of Barenholtz and colleagues (2016) whose study highlighted that participants' fixation on the mouth increased when they were presented with non-native language stimuli. L2 development is also stimulated by audio-visual cues (Long, 2019, Murphy et al, 2020). This shows that multimodal input of both the auditory and visual stimulus increased perception.

When learning an L2 individuals have to perceive new phonemic categories that are weighted differently from the individual's L1 (Cutler & Broesma, 2011, Strange & Shafer, 2008; Ylinen et al, 2010). This means that cues that are important in the person's L1 may not be as crucial in the L2 (Holt & Lotto, 2006); therefore it is likely that new L2 speakers depend on secondary cues in perception tasks rather than the important cues used by L1 speakers of that language. Moreover, it can be said that new L2 learners may weigh language cues differently depending on their L1. This means it is necessary for L2 learning and teaching for new speakers to be guided through the correct modality (that is audio-visual) to focus their attention on particular cues for effective perception and identification of L2 speech sounds. Therefore, the modality in which new L2 speakers perceive L2 speech sounds has consequences for their language

learning particularly in instances when: (1) their L1 has transparent orthography and their L2 has an opaque orthography and (2) between tonal and non-tonal languages.

Results in the study also noted that the best modality for L2 speakers to perceive L2 words and detect tonality between words is in the audio-visual modality. The combination of visual and verbal cues associated with the audio-visual modality provides a positive impact on L2 learning. Visual imaging found in visual learning strengthens new L2 speakers meaning inference and the recognition of vocabulary (Chung, 1996). In this study in the audio-visual condition, the stimulus consisted of isiZulu tone word pairs and minimal pairs; here the tone words together with the minimal pair words saw participants being introduced to relatively unfamiliar sound contrasts (such as tone they may not have encountered often in their L1). Participants performed well in recalling minimal pair and tone pair words and had even greater success in detecting tonality between word pairs in the audio-visual condition. This is in line with the findings of other researchers who revealed, particularly when they deal with non-native speech sounds, L2 listeners benefit from lip reading in audio-visual conditions particularly when they deal with non-native speech sounds (Cutler et al., 2008; Wang et al., 2008; Wang et al., 2009). The lip reading benefit in the audio-visual condition extends to affecting the auditory speech process. This can be seen when visual information provides high-level phonemic representations (Bourguignon et al., 2019) and visual motion information (McGrath & Summerfield, 1985; Vroomen & Keetels, 2010). Moreover, studies on the effects of lip-reading on speech perception have highlighted the facilitating role of visual articulatory information in the perception of speech (Aller et al., 2022; Piquard-Kipffer et al., 2021; Waldon et al., 1990). Therefore, it can be concluded that together with acoustic cues, visual cues aid in L2 word perception.

A property of speech with a different function in different languages is pitch variation. Based on how pitch is used, certain elements in tonal languages are unique in Indo-European languages. Specifically, in English intonational variation contributes to pragmatic and emotional information (Wichmann, 2004, 2005). However, pitch in the English language does not change word meaning. Therefore, English speakers do not treat tone in a distinct way; nor do they adjust tone in lexical representation (Chen et al., 2016; Hay et al., 2014; Liu & Kager, 2016; Quam & Swingley, 2010). Similarly, even though non-tonal speakers are cognisant of the acoustic properties of lexical tone these tones remain a strange suprasegmental feature.

Results in this study revealed that new adult bilinguals do face difficulty in detecting tonal differences between words. This is in line with results from various other studies which explored adults' learning tones and found that lexical contrasts are challenging for L2 speakers (Cooper & Wang, 2013; Hao, 2012; Schaefer & Darcey, 2014; Showalter & Hayes-Harb, 2013; Wayland & Guion, 2004). This difficulty is more noticeable for speakers whose L1 does not make use of pitch height or contour in a contrastive technique (Antoniou & Chin, 2018). Various studies revealed that new non-tonal L2 speakers utilise very different techniques to understand lexical tone compared to native tonal speakers (Francis et al, 2008; Peng et al., 2010, Wu & Lin, 2008). Similarly, Mattock et al. (2007) reveal that different tone categories exist across languages. In native Mandarin for example native speakers deal with frequency contour when perceiving tonal contrasts wherein they begin to zone in on the direction of change (e.g., the rising versus falling) of pitch height (Liu & Samuel, 2004). Wang et al. (2003) revealed that native English speakers make us of the height of tones for Mandarin tonal contrasts only which results in inaccuracy during tonal processing (Wang et al., 2003). Concerning this study, isiZulu is a tonal language which means that a slight difference in tone to various words could lead to a change in the meaning in words. Differences in tone are also crucial in distinguishing one construction or tense from another. The lexical tone is thus a demanding facet of L2 learning. In studies by Lin (1985) and Qin et al. (2021), L2 speakers exhibited a poor performance rate even after some exposure to tone. Adults have particular difficulty perceiving non-native contrasts. Studies researching the learning of tones of by adults show that lexical tone contrasts are challenging for new L2 speakers (Showalter & Hayes-Harb, 2013). L2 speakers with a non-tonal L1 struggle to recognize tones categorically (Peng et al., 2010; Wu & Lin, 2008) and the perception of these tone categories differ across languages and various conditions (Qin & Jongman, 2016). Therefore, both native and non-native tonal speakers depend on the feature of pitch which aids in their native language.

In the audio-only and audio-visual conditions concerning the tonal pair words, the first word in the tonal pair was presented to participants wherein they recalled the word they observed; thereafter, they had to recall the word they perceived and were asked whether the two words were the same or not. The method in which participants were introduced to the stimuli could have affected whether they perceived the tonal difference in the word pair. Interestingly, additional factors such as time could affect (phonological to basic auditory) sound discrimination of speech. The study by Werker and Logan (1985) showed that English

participants were able to perceive the contrasts between the dental and retroflex Hindi stops whenever the time between the stimuli where the distinction occurred was narrowed. This means that when the stimuli are introduced together the auditory properties aid in differentiating the words compared to when there is a prolonged time between them, and listeners only depend on abstract phonological representations.

Another component that could have affected the perception of stimuli is the L1 of participants. All the participants in this study had English as their home language. Numerous components have been shown to affect the perception of L2 sounds. The link between an individual's native language and L2 sounds can influence how new L2 phonemes are perceived (Iverson & Evans, 2009). This can be seen in Spanish participants who at first perceive English /i:/-/ɪ/ contrasts as two separate entities of their Spanish L1 /i/. This is because the English /i:/ is closely perceived with the Spanish /i/category in contrast to the English/ɪ/ which is perceived as a substandard of Spanish /i/ (Flege et al., 1995; Flege & MacKay, 2004; Morrison, 2008). In addition to the above information, the size of a particular language inventory could also affect the perception of L2 speech. Individuals whose L2 have a smaller vowel inventory display substandard recognition and learning performance compared to individuals with a large vowel inventory (Iverson & Evans, 2009).

Regarding the theoretical frameworks the study aligns with the PAM which holds that an a speaker's L2 perception of non-native contrasts is determined by their L1 (Bohn & Munro, 2009). This can be seen as participants had difficulty with aspects such as detecting tone. Tonal detection is a new concept for the participants as it is not present in their L1 English, hence they would have more difficulty with acquiring this concept.

Therefore the results of this study highlight and are in support of the notion that audio-visual stimuli are helpful in non-native speech perception. Audio-visual cues aid in perceiving speech that is more demanding (Ross et al., 2006). Although L2 speakers can effectively recall stimuli from an auditory stimulus, through an auditory-only or visual-only stimulus they may struggle perceiving non-native cues such as tone. The integration of both auditory and visual stimuli aids L2 speakers in improving and developing their perception skills which is necessary in non-native language development.

5.3 Research Question 2:

5.3.1 How do different stimuli affect the way L2 bilinguals perceive audio-visual speech?

In exploring incongruent stimuli as a measure of audio-visual integration an important finding in Experiment 4 was the variation in responses across the incongruent stimuli. Individual differences in the perception of incongruent stimuli within the participant's group were noted and will be explained below:

This research study aimed to investigate the use of incongruent stimuli as a measure of audio-visual integration. This was accomplished by introducing participants to incongruent stimuli where the audio of one word was dubbed over the visual of another word. Three occurrences could have taken place in this condition after being exposed to the incongruent stimuli: first, participants could have perceived the audio form of the word, secondly, the participants could have perceived the visual form of the word or lastly, the participants could perceive an incongruent word meaning it was not the visual or vocal of the word, rather a new form of the word.

To answer the question above, the auditory-visual speech was explored with the partcipants using the McGurk effect. The stimuli used were in the participant's non-native language of isiZulu. The McGurk effect used incongruent syllables as a measure in this study we used minimal pairs of isiZulu wherein the audio of one minimal pair word was dubbed over the visual of the corresponding minimal pair word. Previous studies have revealed that the processes of speech perception tend to depend on the projection value between the audio and visual mental representations within the individual's native language (Hazan et al., 2006; Jesse & Massaro, 2010). Based of the two modalities and centered on these projection values based on the two modalities a choice is made on the perceived speech. The native language mental representation which best matches the audio-visual speech input is likely to be the one that is perceived.

In testing for the McGurk effect, the results revealed consistent variability in responses across all three incongruent stimuli. It was observed in the first incongruent stimuli when the vocal of the word "isitho" was dubbed over the visual of the word "isipho" the McGurk effect was elicited in 75,0% of participants. Within this 75,0%, 8,30% of participants recalled the word

"isimto" whilst 66,70% of participants recalled the word "isipto". The 8,30% accounts for one participant that deleted the aspiration in both the vocal and visual of the word and replaced it with the consonant [m] (voiced, bilabial, nasal). In both instances, participants removed the aspirated sounds of [t^h] (aspirated, alveolar, stop) and [p^h] (aspirated, bilabial, stop) and either replaced them with another consonant or fused the two aspirated sounds. In the elicited word "isipto" 66,7% of participants fused the [p] sound found in the audio of the word [isi**p**ho] with the [t] sound found in the visual representation of the word [isi**t**ho]. The remaining 25% of participants recalled the vocal (isitho) of the word revealing the vocal dominated the visual of the word. These results show that when the McGurk effect was not elicited participants favored the vocal modality of the word.

In the second stimulus where the vocal "ukulima" was dubbed over the visual of the word [ukulila], 16, 70% of participants recalled the word "ukulinga". The [ng] sound was created by deleting the [ma] sound in "ukuli**ma**" and the [la] sound in "ukuli**la**". The majority of participants (58,0%) in this stimulus however produced the word [ukulina]. Essentially participants perceived the vocal [m] sound with the visual [l] as an [n] sound. The [m] acoustically is a voiced, bilabial, nasal while the [l] is a voiced, alveolar, lateral approximant, participants therefore borrowed the acoustic properties from each of these sounds to and perceived the [n] (a voiced, alveolar, nasal). The remaining 25% of participants, however, perceived the visual of the word which was [ukulila]. Although the responses to the McGurk effect varied the average McGurk perception was consistent throughout the study revealing that the stimuli elicited the illusion in the majority of the participants. These results reveal that when the McGurk Effect was not elicited participants favored the visual perception of the word.

Interestingly in the last stimulus where the vocal of the word [ukusala] was dubbed over the visual of the word [ukuzala], all participants perceived the vocal modality of the word. This shows that the [s] voiceless, alveolar fricative is perceived better vocally than the [z] voiced, alveolar, fricative. The key element in this stimulus is that the dominant difference between the two words the [s] and [z] shared the same acoustic properties in the place of articulation and manner of articulation, respectively; therefore, participants did not perceive a new sound rather the vocal modality seemed to be stronger than the visual modality. This is new evidence of the effect of modality on audio-visual perception. This stimulus also revealed that L2 isiZulu

listeners rely less on visual cues during the perception of speech when the incongruency of the sound shares acoustic properties of manner and place of articulation.

The results of the three stimuli presented to participants results varied considerably across each stimulus. These results are congruent with studies investigating the use of visual cues in native language; in the latter studies, contrasts were found in the percentage of McGurk elicitations between Chinese (Cantonese), Japanese and English participants (Sekiyama, 1997, Sekiyama & Burnham, 2008). Distinctions were also present in the McGurk effect between Japanese, Spanish and English individuals (Massaro et al., 1993). Interestingly, it was revealed that Japanese participants depended less than Spanish listeners on visual cues (Hazan et al., 2006). The reasoning behind this finding was provided by Sekiyama (1997) who revealed that decreased dependence on visual cues found in Chinese. The latter language uses tones while Japanese uses pitch accents. however, this reasoning cannot broadly be applied to this study as the majority of participants spoke English as an L1 and this does not rely on pitch accents while the L2 in this study, isiZulu, however, is a tonal language.

This data can be linked to the Fuzzy Logical Model of Perception proposed by Massaro (1989) regarding perceptual bias which is based on the visual mental representations of the native language. The process of speech perception is thus malleable in situations where there is a change in weighting from an auditory or visual input depending on the necessary information acquired from each modality. Luce and Pisoni (1998) proposed the Neighborhood Activation Model which suggests that the perception of audio-visual speech is dependent on the declared power of an individual's native language visual cues and the weight of phonetic visual neighbourhood**.** Therefore, it can be said that audio-visual perception of incongruent stimuli in isiZulu is an intricate process that is often based on the interplay between the relevant audio and visual cues present. When there is a clear match between the auditory and visual cues the matching native language representation is then said to be strong (Peelle & Sommers, 2015). The place of articulation can also affect the visual perception of the phoneme (Alsalmi, 2016). Alternatively, when the auditory and visual cues are incongruent, then the speech input which has greater predictive power will largely control the perception process.

To account for contrasts in mental perception, previous research studies have distinguished participants as being strong or weak perceivers of the McGurk effect (Berger & Ehrsson, 2013)

in children (Nath et al., 2011) and adults (Strand et al., 2014). Various factors such as phoneme space, perceptual strain in labial versus non-labial discrimination and cultural traits can influence the McGurk effect. In this study, it was noted that when the merge incongruent stimuli consisted of the same place of articulation and manner of articulation participants were more likely to observe either the visual or vocal modality instead of perceiving a new sound. Previous studies have also noted variation in incongruent stimuli, Getz and Toscano (2020), for example, illustrated a notable top-down effect on the illusion and asserted that the McGurk effect should not be considered a powerful illusion rather it should be seen as a consequence that exists for some individuals under certain task and stimulus conditions. Another study similarly revealed that some of the variation of the McGurk effect was due to talker distinctions (such as facial features or the articulation of syllables) (Jiang & Bernstein, 2011). Yet, other researchers found that individuals who are more inclined to the perception of auditory stimuli are less likely to perceive the McGurk effect (Proverbio et al., 2016). Therefore, it can be revealed that the perception of incongruent stimuli is based on the exposure to a particular stimulus and the cultural conditions of a particular individual. However, in this study, we found that it was found that when incongruent isiZulu stimuli share the same place and manner of articulation participants are less likely to perceive a new sound.

Using the McGurk effect as a tool for multisensory integration, however, means that it is not possible to ascertain whether the stimuli the observer is exposed to will elicit a response matching the actual percept (Tona et al.,2015). When integration does not take place, it is likely the participant will observe the components separately and perceive either what they heard, or they saw (Tiippana, 2014). However, when a participant perceives a word that differs from the stimulus percept, this means that they may not rely on one modality alone rather they fuse the information from both (Tippana, 2014). The variation of responses amongst participants in this study could mean that participants relied on both auditory and visual stimuli during perception.

Therefore, the results yielded in this study reveal that although there is a variation in the incongruent condition responses the McGurk effect illusion was elicited in the majority of the stimuli. Similarly, when the audio and visual stimuli are incongruent, the speech input with greater power is likely that the speech input with greater power will controls the individual's perception.

5.3 Conclusion

This study examined the extent the which young adult bilinguals benefit from audio-visual speech. In examining this the study focused on the effects of modality on speech perception and how different stimuli affect how L2 bilinguals perceive speech. It was revealed that participants were able to accurately recall words well through an audio-only modality and an audio-visual modality. However, participants had more difficulty in picking up the difference in tone between tonal pair words in the audio-only condition compared to the audio-visual condition. Therefore, it was found that L2 learners perceived and recalled speech best through an audio-visual modality where they had greater success in identifying and detecting tonal differences between tone word pairs. Moreover, this study revealed how incongruent speech is perceived differently depending on the stimulus presented to participants. The results yielded in this study differed depending on the stimulus being presented. Interestingly, it was found that incongruent speech does not result in the creation of a new sound when the incongruent stimulus made up of the vocal and visual of the word share the same place and manner of articulation. In this instance, the study showed that all participants perceived the vocal modality of the word.

Visual information from audio-visual speech is not essential for speech perception however, when the L2 speakers are confronted with unfamiliar speech sounds not found in their L1 (i.e tone, the visual modality together auditory modality can ease speech perception of non-native cues and can lessen the cognitive load. This provides insight as to why L2 speakers may face comprehension difficulties in communication when the visual modality is missing. This gives rise to the importance of visual input in educational settings for new L2 speakers as the fusion of audio and visual modalities of speech will aid in their comprehension and perception of speech non-native speech sounds. In context to the world, we live in today, individuals are utilising multimedia to aid L2 learning.

Multimedia comes in the form of visual stimuli that are used in L2 learning and testing. As seen in this study new L2 speakers may benefit from audio-visual learning meaning that new L2 speakers may benefit from audio-visual teaching instead of visual teaching alone. Similarly, L2 teachers have to be aware of the benefit of audio-visual modalities on L2 perception

(Hardison, 2018). Therefore, results from this study provide evidence of the importance audio-visual modalities play in the perception of L2 speech.

The following study contributes to filling a research gap because limited research regarding L2 speech perception of Bantu languages has been carried out. Similarly, there is very little evidence regarding which modality can best aid L2 learning of late Bantu languages. Exploring the effects of modality in L2 perception can aid in formulating appropriate learning material for L2 speakers. Similarly, developing appropriate multimodal learning materials can aid various groups including hard-of hearing individuals, children with phonological disorders and speech and language therapists in their clinical intervention.

5.4 Limitations and future recommendations of the current study

5.4.1 Limitations

The first limitation of this study is the relatively small sample size. The strict criteria participants had to meet limited the number of participants, which meant it was challenging to fully comprehend the nature of individual differences found in the perception of isiZulu as an L2 speaker. Similarly, the sample size was not representative of genders in the general population, and this meant that the effect of gender on L2 speech perception could not be explored. While the findings provide insight into the different modalities that will aid in an L2 speaker's grasp of non-native stimuli, the small sample size makes it difficult to fully grasp the individual contrasts in L2 auditory and visual processing in speech perception. Understanding the perception of a language by native and non-native speakers of isiZulu may provide insight and a more nuanced view into the role of modality in second-language perception.

Another limitation could be the choice of the recall task used in the audio-only, visual-only and incongruent conditions. A recall task may be difficult for early L2 speakers as they have to rely on their listening comprehension and working memory capacity. Therefore, although a participant may comprehend the stimuli, they might have difficulty remembering the word.

Despite the above-mentioned limitations, this study provided valuable information towards learning isiZulu as second language. The findings can aid L2 teachers and curriculum planners

in formulating effective L2 learning strategies that accommodate for the use of audio-visual learning.

5.4.2 Future recommendations

Knowing that audio-visual integration changes throughout a person's lifetime, it will be interesting to note the differences between young children's and young adults' perceptions of their second language and how age might be a factor in a person's L2 susceptibility. Using integrated eye-tracking software this study will prove beneficial in understanding the nuances related to L2 speech perception.

It will also be interesting to determine whether the perception of non-native lexical cues will differ when the stimuli are marred by obstacles such as noise. This factor was not included in the study therefore it was easy to determine the perception across audio, visual and audio-visual conditions. However, adding an element such as noisy listening conditions could yield different results or enhance the current findings.

5.5 Summary of chapter

This thesis provided new insights into how young adult bilinguals perceive speech. Regard the first research question, the study concluded that the integration of both auditory and visual modalities saw the best results in the perception of non-native cues. The results of this study support the need for of audio-visual cues in L2 perception. The integration of audio and visual stimuli can aid new speakers and provide them with necessary cues to improve their L2 perception, particularly in contrasts not found in their L1 language inventory. Concerning the second research question, the study concluded that incongruent stimuli of non-native cues resulted in varied responses among participants depending on the non-native stimuli provided to them. However, it was noted that when incongruent stimuli sounds shared the same manner and place of articulation, they are less likely to perceive a new word.

In conclusion, this study sought to find the extent to which young adults benefit from audio-visual speech. The difficulty in detecting tone in audio-only conditions is in line with research carried out by Showalter & Hayes-Harb, 2013; Wang et al., 2003, Wayland & Guion, 2004 and

Wang et al., 2013 whose studies showed that lexical tones can be difficult for L2 speakers to detect. Similarly, the results from the present study showed that the audio-visual modality yielded the best perception. The latter results are in line with studies carried out by Hazan et al. (2006). Overall, modality in second language learning is integral as it aids new speakers in sound perception and in achieving second language fluency. Using different modalities provides L2 speakers with nuanced and comprehensive skills in acquiring a second language. Curriculum planners need to utilise such modalities in the teaching of a second language, particularly when there are tone differences between the L1 and L2.

CHAPTER 6

References

Akbarian, I. (2015). Jiménez Catalán, Rosa María (Ed.). Lexical Availability in English and Spanish as a Second Language. New York: Springer. xiv + 205 pages. ISBN: 978-94-007-7157-4. *International Journal of English Studies*, *15*(2), 97. https://doi.org/10.6018/ijes/2015/2/218611

Akcam, B. K., Guney, S., & Cresswell, A. M. (2019). Research Design and Major Issues in Developing Dynamic Theories by Secondary Analysis of Qualitative Data. *Systems*, *7*(3), 40. https://doi.org/10.3390/systems7030040

Aldridge, M. A., Braga, E. S., Walton, G. E., & Bower, T. G. R. (1999). The intermodal representation of speech in newborns. *Developmental Science*, *2*(1), 42–46. https://doi.org/10.1111/1467-7687.00052

Aller, M., Økland, H. S., MacGregor, L., Blank, H., & Davis, M. H. (2022). *Differential Auditory and Visual Phase-Locking Are Observed during Audio-Visual Benefit and Silent Lip-Reading for Speech Perception*. The Journal of Neuroscience. https://doi.org/10.1523/jneurosci.2476-21.2022

Alm, M., & Behne, D. M. (2015). *Do gender differences in audio-visual benefit and visual influence in audio-visual speech perception emerge with age?* Frontiers in Psychology. https://doi.org/10.3389/fpsyg.2015.01014

Almeida, F., Faria, D., & Queirós, A. (2017). *Strengths and Limitations of Qualitative and Quantitative Research Methods*. ResearchGate. https://doi.org/10.5281/zenodo.887089

Alsalmi, J. (2016). *Auditory-Visual Integration during the Perception of Spoken Arabic*. Etheses.whiterose.ac.uk. https://etheses.whiterose.ac.uk/13320/

Alsius, A., Paré, M., & Munhall, K. (2018). Forty Years After Hearing Lips and Seeing Voices: the McGurk Effect Revisited. *Multisensory Research*, *31*(1–2), 111–144. https://doi.org/10.1163/22134808-00002565

Altieri, N. (2014). Multisensory integration, learning, and the predictive coding hypothesis. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00257

Altieri, N. A., Pisoni, D. B., & Townsend, J. T. (2011). Some normative data on lip-reading skills (L). *The Journal of the Acoustical Society of America*, *130*(1), 1–4. https://doi.org/10.1121/1.3593376

Ambrose, S. E., Fey, M. E., & Eisenberg, L. S. (2012). Phonological Awareness and Print Knowledge of Preschool Children With Cochlear Implants. *Journal of Speech, Language, and Hearing Research*, *55*(3), 811–823. https://doi.org/10.1044/1092-4388(2011/11-0086)

Amengual, M. (2017). Type of early bilingualism and its effect on the acoustic realization of allophonic variants: Early sequential and simultaneous bilinguals. *International Journal of Bilingualism*, *23*(5), 954–970. https://doi.org/10.1177/1367006917741364

Anthony, J. L., & Francis, D. J. (2005). Development of Phonological Awareness. *Current Directions in Psychological Science*, *14*(5), 255–259. https://doi.org/10.1111/j.0963-7214.2005.00376.x

Antoniou, M., & Chin, J. L. L. (2018). What Can Lexical Tone Training Studies in Adults Tell Us about Tone Processing in Children? *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00001

Apuke, O. D. (2017). Quantitative Research Methods : A Synopsis Approach. *Kuwait Chapter of Arabian Journal of Business and Management Review*, *6*(11), 40–47. https://doi.org/10.12816/0040336

Arbona, E., Seeber, K., & Gullberg, M. (2023). *The role of manual gestures in second language comprehension: a simultaneous interpreting experiment*. Frontiers in Psychology. https://doi.org/10.3389/fpsyg.2023.1188628

Arnold, P., & Hill, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology*, *92*(2), 339–355. https://doi.org/10.1348/000712601162220

Aubanel, V., Davis, C., & Kim, J. (2016). Exploring the Role of Brain Oscillations in Speech Perception in Noise: Intelligibility of Isochronously Retimed Speech. *Frontiers in Human Neuroscience*, *10*. https://doi.org/10.3389/fnhum.2016.00430

Auer, E. T., Bernstein, L. E., Sungkarat, W., & Singh, M. (2007). Vibrotactile activation of the auditory cortices in deaf versus hearing adults. *NeuroReport*, *18*(7), 645–648. https://doi.org/10.1097/wnr.0b013e3280d943b9

Bahal, A. (2012). Advances in Automatic Speech Recognition: From Audio-Only To Audio-Visual Speech Recognition. *IOSR Journal of Computer Engineering*, *5*(1), 31–36. https://doi.org/10.9790/0661-0513136

Bahrick, L. E., Hernandez-Reif, M., & Flom, R. (2005). The Development of Infant Learning About Specific Face-Voice Relations. *Developmental Psychology*, *41*(3), 541–552. https://doi.org/10.1037/0012-1649.41.3.541

Baptista, R. (2001). Geographical Clusters and Innovation Diffusion. *Technological Forecasting and Social Change*, *66*(1), 31–46. https://doi.org/10.1016/s0040-1625(99)00057-8

Barenholtz, E., Mavica, L., & Lewkowicz, D. J. (2016). Language familiarity modulates relative attention to the eyes and mouth of a talker. *Cognition*, *147*, 100–105. https://doi.org/10.1016/j.cognition.2015.11.013

Basu Mallick, D., F. Magnotti, J., & S. Beauchamp, M. (2015). Variability and stability in the McGurk effect: contributions of participants, stimuli, time, and response type. *Psychonomic Bulletin & Review*, *22*(5), 1299–1307. https://doi.org/10.3758/s13423-015-0817-4

Beattie, R. L., & Manis, F. R. (2012). Rise Time Perception in Children With Reading and Combined Reading and Language Difficulties. *Journal of Learning Disabilities*, *46*(3), 200–209. https://doi.org/10.1177/0022219412449421

Bejjanki, V. R., Clayards, M., Knill, D. C., & Aslin, R. N. (2011). Cue Integration in Categorical Tasks: Insights from Audio-Visual Speech Perception. *PLoS ONE*, *6*(5), e19812. https://doi.org/10.1371/journal.pone.0019812

Benguerel, A. P., & Pichora-Fuller, M. K. (1982). Coarticulation Effects in Lipreading. *Journal of Speech, Language, and Hearing Research*, *25*(4), 600–607. https://doi.org/10.1044/jshr.2504.600

Berger, C., & Ehrsson, H. (2013). Mental Imagery Changes Multisensory Perception. *Current Biology*, *23*(14), 1367–1372. https://doi.org/10.1016/j.cub.2013.06.012

Bergeson, T. R., Houston, D. M., & Miyamoto, R. T. (2010). Effects of congenital hearing loss and cochlear implantation on audio-visual speech perception in infants and children. *Restorative Neurology and Neuroscience*, *28*(2), 157–165. https://doi.org/10.3233/rnn-2010-0522

Berken, J. A., Chai, X., Chen, J. K., Gracco, V. L., & Klein, D. (2016). Effects of Early and Late Bilingualism on Resting-State Functional Connectivity. *The Journal of Neuroscience*, *36*(4), 1165–1172. https://doi.org/10.1523/jneurosci.1960-15.2016

Bernstein, L. E., & Liebenthal, E. (2014). Neural pathways for visual speech perception. *Frontiers in Neuroscience*, *8*. https://doi.org/10.3389/fnins.2014.00386

Bernstein, L. E., Jiang, J., Pantazis, D., Lu, Z., & Joshi, A. (2010). Visual phonetic processing localized using speech and nonspeech face gestures in video and point-light displays. *Human Brain Mapping*, *32*(10), 1660–1676. https://doi.org/10.1002/hbm.21139

Bernstein, L. E., Tucker, P. E., & Demorest, M. E. (2000). Speech perception without hearing. *Perception & Psychophysics*, *62*(2), 233–252. https://doi.org/10.3758/bf03205546

Best, C. T., Goldstein, L. M., Nam, H., & Tyler, M. D. (2016). Articulating What Infants Attune to in Native Speech. *Ecological Psychology*, *28*(4), 216–261. https://doi.org/10.1080/10407413.2016.1230372

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. *The Journal of the Acoustical Society of America*, *109*(2), 775–794. https://doi.org/10.1121/1.1332378

Best, C. T., & Strange, W. (1992). Effects of phonological and phonetic factors on cross-language perception of approximants. *Journal of Phonetics*, *20*(3), 305–330. https://doi.org/10.1016/s0095-4470(19)30637-0

Bhattacharjee, B. (2014). *An Investigation through Different Types of Bilinguals and Bilingualism*. Assamsilchar. https://www.academia.edu/8436489/An_Investigation_through_Different_Types_of_Bilinguals_and_Bilingualism

Bialystok, e., & Herman, J. (1999). Does bilingualism matter for early literacy? *Bilingualism: Language and Cognition*, *2*(1), 35–44. https://doi.org/10.1017/s1366728999000139

Birmingham, E., Bischof, W. F., & Kingstone, A. (2008). Social Attention and Real-World Scenes: The Roles of Action, Competition and Social Content. *Quarterly Journal of Experimental Psychology*, *61*(7), 986–998. https://doi.org/10.1080/17470210701410375

Birulés, J., Bosch, L., Pons, F., & Lewkowicz, D. J. (2020). Highly proficient L2 speakers still need to attend to a talker's mouth when processing L2 speech. *Language, Cognition and Neuroscience*, *35*(10), 1314–1325. https://doi.org/10.1080/23273798.2020.1762905

Bohn, O. S., & Best, C. T. (2012). Native-language phonetic and phonological influences on perception of American English approximants by Danish and German listeners. *Journal of Phonetics*, *40*(1), 109–128. https://doi.org/10.1016/j.wocn.2011.08.002

Bohn, O. S., & Munro, M. J. (2009). Language Experience in Second Language Speech Learning: In Honor of James Emil Flege. *Phonetica*, *65*(4), 272–276. https://doi.org/10.1159/000192797

Borghini, G., & Hazan, V. (2018). Listening Effort During Sentence Processing Is Increased for Non-native Listeners: A Pupillometry Study. *Frontiers in Neuroscience*, *12*. https://doi.org/10.3389/fnins.2018.00152

Borowsky, T. (1989). Structure preservation and the syllable coda in English. *Natural Language and Linguistic Theory*, *7*(2), 145–166. https://doi.org/10.1007/bf00138075

Bourguignon, M., Baart, M., Kapnoula, E. C., & Molinaro, N. (2019). Lip-Reading Enables the Brain to Synthesize Auditory Features of Unknown Silent Speech. *The Journal of Neuroscience*, *40*(5), 1053–1065. https://doi.org/10.1523/jneurosci.1101-19.2019

Bourguignon, M., Molinaro, N., Lizarazu, M., Taulu, S., Jousmäki, V., Lallier, M., Carreiras, M., & De Tiège, X. (2020). Neocortical activity tracks the hierarchical linguistic structures of self-produced speech during reading aloud. *NeuroImage*, *216*, 116788. https://doi.org/10.1016/j.neuroimage.2020.116788

Bozkurt, E., Erzin, E., Erdem, I. E., & Erdem, A. T. (2011). Formant position based weighted spectral features for emotion recognition. *Speech Communication*, *53*(9–10), 1186–1197. https://doi.org/10.1016/j.specom.2011.04.003

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

Brahme, A., & Bhadade, U. (2016). Lip Detection and Lip Geometric Feature Extraction using Constrained Local Model for Spoken Language Identification using Visual Speech Recognition. *Indian Journal of Science and Technology*, *9*(32). https://doi.org/10.17485/ijst/2016/v9i32/98737

Brancazio, L., Miller, J. L., & Paré, M. A. (2003). Visual influences on the internal structure of phonetic categories. *Perception & Psychophysics*, *65*(4), 591–601. https://doi.org/10.3758/bf03194585

Broersma, M., & Cutler, A. (2011). *Competition dynamics of second-language listening*. Quarterly Journal of Experimental Psychology. https://doi.org/10.1080/17470218.2010.499174

Broersma, M., & Scharenborg, O. (2010). Native and non-native listeners' perception of English consonants in different types of noise. *Speech Communication*, *52*(11–12), 980–995. https://doi.org/10.1016/j.specom.2010.08.010

Buchan, J. N., Paré, M., & Munhall, K. G. (2008). The effect of varying talker identity and listening conditions on gaze behavior during audio-visual speech perception. *Brain Research*, *1242*, 162–171. https://doi.org/10.1016/j.brainres.2008.06.083

Bundgaard-Nielsen, R. L., Best, C. T., Kroos, C., & Tyler, M. D. (2011). Second language learners' vocabulary expansion is associated with improved second language vowel intelligibility. *Applied Psycholinguistics*, *33*(3), 643–664. https://doi.org/10.1017/s0142716411000518

Burfin, S., Pascalis, O., Tada, E. R., Costa, A., Savariaux, C., & Kandel, S. (2014). *Bilingualism affects audio-visual phoneme identification*. Frontiers in Psychology. https://doi.org/10.3389/fpsyg.2014.01179

Burnham, D., & Dodd, B. (2004). Auditory–visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology*, *45*(4), 204–220. https://doi.org/10.1002/dev.20032

Burnham, D. K., Earnshaw, L. J., & Clark, J. E. (1991). Development of categorical identification of native and non-native bilabial stops: infants, children and adults. *Journal of Child Language*, *18*(2), 231–260. https://doi.org/10.1017/s0305000900011041

Burnham, D., & Lau, S. (1998). *The Effect of Tonal Information on Auditory Reliance in the McGurk Effect*. https://www.semanticscholar.org/paper/The-Effect-of-Tonal-Information-on-Auditory-in-the-Burnham-Lau/2c852e88bc62c277adb89f78ae00159dc026920b

Cabrera, L., & Werner, L. (2017). Infants' and Adults' Use of Temporal Cues in Consonant Discrimination. *Ear & Hearing*, *38*(4), 497–506. https://doi.org/10.1097/aud.0000000000000422

Campbell, R. (2007). The processing of audio-visual speech: empirical and neural bases. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 1001–1010. https://doi.org/10.1098/rstb.2007.2155

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: Human Perception and Performance*, *27*(1), 115–127. https://doi.org/10.1037//0096-1523.27.1.115

Cantiani, C., Dondena, C., Molteni, M., Riva, V., & Lorusso, M. L. (2023). Intergenerational longitudinal associations between parental reading/musical traits, infants' auditory processing, and later phonological awareness skills. *Frontiers in Neuroscience*, *17*. https://doi.org/10.3389/fnins.2023.1201997

Castilla, A. P., Restrepo, M. A., & Perez-Leroux, A. T. (2009). Individual differences and language interdependence: a study of sequential bilingual development in Spanish–English preschool children. *International Journal of Bilingual Education and Bilingualism*, *12*(5), 565–580. https://doi.org/10.1080/13670050802357795

Chen, A., Liu, L., & Kager, R. (2016). Cross-domain correlation in pitch perception, the influence of native language. *Language, Cognition and Neuroscience*, *31*(6), 751–760. https://doi.org/10.1080/23273798.2016.1156715

Chen, J., Best, C. T., & Antoniou, M. (2020). Native phonological and phonetic influences in perceptual assimilation of monosyllabic Thai lexical tones by Mandarin and Vietnamese listeners. *Journal of Phonetics*, *83*, 101013. https://doi.org/10.1016/j.wocn.2020.101013

Chen, T. H., & Massaro, D. W. (2004). Mandarin speech perception by ear and eye follows a universal principle. *Perception & Psychophysics*, *66*(5), 820–836. https://doi.org/10.3758/bf03194976

Chen, T. H & Rao, R. (1998). Audio-visual integration in multimodal communication. *Proceedings of the IEEE*, *86*(5), 837–852. https://doi.org/10.1109/5.664274

Chen, Y., & Hazan, V. (2009). Developmental factors and the non-native speaker effect in auditory-visual speech perception. *The Journal of the Acoustical Society of America*, *126*(2), 858–865. https://doi.org/10.1121/1.3158823

Chen, Y., Lin, B., Lu, L., & Zhou, G. (2020). Can internal audit functions improve firm operational efficiency? Evidence from China. *Managerial Auditing Journal*, *35*(8), 1167–1188. https://doi.org/10.1108/maj-01-2019-2136

Chung, J. M. (1996). The Effects of Using Advance Organizers and Captions to Introduce Video in the Foreign Language Classroom. *TESL Canada Journal*, *14*(1), 61. https://doi.org/10.18806/tesl.v14i1.678

Clark, R. C., Mayer, R. E., & Thalheimer, W. (2003). E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning. *Performance Improvement*, *42*(5), 41–43. https://doi.org/10.1002/pfi.4930420510

Constitution of the Republic of South Africa, 1996 | South African Government. (1996). Www.gov.za. https://www.gov.za/documents/constitution/constitution-republic-south-africa-1996-04-feb-1997

Cooper, A., & Wang, Y. (2013). Effects of tone training on Cantonese tone-word learning. *The Journal of the Acoustical Society of America*, *134*(2), EL133–EL139. https://doi.org/10.1121/1.4812435

Cornut, C., Mahé, G., & Casalis, S. (2021). L2 word recognition in French–English late bilinguals: Does modality matter? *Bilingualism: Language and Cognition*, *25*(1), 121–136. https://doi.org/10.1017/s1366728921000511

Cosi, P., Dugatto, M., Ferrero, F., Caldognetto, E., & Vagges, K. (1996). Phonetic recognition by recurrent neural networks working on audio and visual information. *Speech Communication*, *19*(3), 245–252. https://doi.org/10.1016/0167-6393(96)00034-9

Crandell, C. C., & Smaldino, J. J. (2000). Classroom Acoustics for Children With Normal Hearing and With Hearing Impairment. *Language, Speech, and Hearing Services in Schools*, *31*(4), 362–370. https://doi.org/10.1044/0161-1461.3104.362

Creswell, J. W. (2008). Editorial: Mapping the Field of Mixed Methods Research. *Journal of Mixed Methods Research*, *3*(2), 95–108. https://doi.org/10.1177/1558689808330883

Cummins, J. (1978). Bilingualism and the Development of Metalinguistic Awareness. *Journal of Cross-Cultural Psychology*, *9*(2), 131–149. https://doi.org/10.1177/002202217892001

Cunillera, T., Càmara, E., Laine, M., & Rodríguez-Fornells, A. (2010). Speech segmentation is facilitated by visual cues. *Quarterly Journal of Experimental Psychology*, *63*(2), 260–274. https://doi.org/10.1080/17470210902888809

Cusack, R., Decks, J., Aikman, G., & Carlyon, R. P. (2004). Effects of Location, Frequency Region, and Time Course of Selective Attention on Auditory Scene Analysis. *Journal of*

Experimental Psychology: Human Perception and Performance, *30*(4), 643–656. https://doi.org/10.1037/0096-1523.30.4.643

Cutler, A., Garcia Lecumberri, M. L., & Cooke, M. (2008). Consonant identification in noise by native and non-native listeners: Effects of local context. *The Journal of the Acoustical Society of America*, *124*(2), 1264–1268. https://doi.org/10.1121/1.2946707

DeGroot, A. M. (2011). *Language and Cognition in Bilinguals and Multilinguals*. Google Books. Retrieved March 7, 2024, from https://books.google.co.za/books?hl=en&lr=&id=7sp5AgAAQBAJ&oi=fnd&pg=PT15&dq=DeGroot,+2011+l2+production&ots=sJZQlMcxgN&sig=3Se7X30uaRHHXrBfPy17pHkXpPk&redir_esc=y#v=onepage&q=DeGroot%2C%202011%20l2%20production&f=false

Dekeyser, R., & Prieto Botana, G. (2014). The Effectiveness of Processing Instruction in L2 Grammar Acquisition: A Narrative Review. *Applied Linguistics*, *36*(3), 290–305. https://doi.org/10.1093/applin/amu071

Demorest, M. E., Bernstein, L. E., & DeHaven, G. P. (1996). Generalizability of Speechreading Performance on Nonsense Syllables, Words, and Sentences. *Journal of Speech, Language, and Hearing Research*, *39*(4), 697–713. https://doi.org/10.1044/jshr.3904.697

Desantis, A., & Haggard, P. (2016). How actions shape perception: learning action-outcome relations and predicting sensory outcomes promote audio-visual temporal binding. *Scientific Reports*, *6*(1). https://doi.org/10.1038/srep39086

Díaz, B., Mitterer, H., Broersma, M., & Sebastián-Gallés, N. (2012). Individual differences in late bilinguals' L2 phonological processes: From acoustic-phonetic analysis to lexical access. *Learning and Individual Differences*, *22*(6), 680–689. https://doi.org/10.1016/j.lindif.2012.05.005

Diehl, R. L. (1987). *Auditory Constraints on Speech Perception*. Springer eBooks. https://doi.org/10.1007/978-94-009-3629-4_16

Dietz, M., Ewert, S. D., & Hohmann, V. (2011). Auditory model based direction estimation of concurrent speakers from binaural signals. *Speech Communication*, *53*(5), 592–605. https://doi.org/10.1016/j.specom.2010.05.006

Dijkstra, T., & Kroll, J. F. (2005). *Handbook of Bilingualism*. Google Books. Retrieved March 6, 2024, from https://books.google.co.za/books?hl=en&lr=&id=2fzBDptA5NMC&oi=fnd&pg=PA179&dq=Dijkstra+and+Kroll+(2005&ots=Pr_yHgXHhk&sig=XhIpwLT9MN_4sm5U2kAoezpK4eI&redir_esc=y#v=onepage&q=Dijkstra%20and%20Kroll%20(2005&f=false

Dodd, B. (1979). Lip reading in infants: Attention to speech presented in- and out-of-synchrony. *Cognitive Psychology*, *11*(4), 478–484. https://doi.org/10.1016/0010-0285(79)90021-5

Dorado Solarte, A. (2023). The McGurk Effect Across Languages. *Eureka*. https://doi.org/10.29173/eureka28785

Zoltan Dörnyei. (2007). *Research Methods in Applied Linguistics*. Oxford University Press, USA.

Dragan, I. M., & Isaic-Maniu, A. (2022). An Original Solution for Completing Research through Snowball Sampling—Handicapping Method. *Advances in Applied Sociology*, *12*(11), 729–746. https://doi.org/10.4236/aasoci.2022.1211052

Drijvers, L., & Özyürek, A. (2017). Visual Context Enhanced: The Joint Contribution of Iconic Gestures and Visible Speech to Degraded Speech Comprehension. *Journal of Speech, Language, and Hearing Research*, *60*(1), 212–222. https://doi.org/10.1044/2016_jslhr-h-16-0101

Drijvers, L., & Özyürek, A. (2019). Non-native Listeners Benefit Less from Gestures and Visible Speech than Native Listeners During Degraded Speech Comprehension. *Language and Speech*, *63*(2), 209–220. https://doi.org/10.1177/0023830919831311

du Plessis, E. E. (2020). Student teachers' perceptions, experiences, and challenges regarding learner-centred teaching. *South African Journal of Education*, *40*(1), 1–10. https://doi.org/10.15700/saje.v40n1a1631

Dupont, S., Aubin, J., & Ménard, L. (2005). study of the McGurk effect in 4 and 5-year-old French Canadian children. *ZAS Papers in Linguistics*, *40*, 1–17. https://doi.org/10.21248/zaspil.40.2005.254

Erdener, D. (2012). Basic to applied research: the benefits of audio-visual speech perception research in teaching foreign languages. *The Language Learning Journal*, *44*(1), 124–132. https://doi.org/10.1080/09571736.2012.724080

Erdener, V. D., & Burnham, D. K. (2005). The Role of Audio-visual Speech and Orthographic Information in Nonnative Speech Production. *Language Learning*, *55*(2), 191–228. https://doi.org/10.1111/j.0023-8333.2005.00303.x

Escudero, P., Boersma, P., & Hamann, S. (2009). The linguistic perception of Similar L2 sounds. *De Gruyter EBooks*, 151–190. https://doi.org/10.1515/9783110219234.151

Fastl, H., & Zwicker, E. (2007). *Psychoacoustics*. Google Books. Retrieved March 4, 2024, from https://books.google.co.za/books?hl=en&lr=&id=0zg9hI586kcC&oi=fnd&pg=PA1&dq=Fastl,+H.,+%26+Zwicker,+E.+(2006).+Psychoacoustics:+facts+and+models+(Vol.+22).+Springer+Science+%26+Business+Media.&ots=5-AOicZKs7&sig=WdoonMKd11y38pNgwWMB9L0PvJQ&redir_esc=y#v=onepage&q=Fastl%2C%20H.%2C%20%26%20Zwicker%2C%20E.%20(2006).%20Psychoacoustics%3A%20facts%20and%20models%20(Vol.%2022).%20Springer%20Science%20%26%20Business%20Media.&f=false

Feld, J. E., & Sommers, M. S. (2009). Lipreading, Processing Speed, and Working Memory in Younger and Older Adults. *Journal of Speech, Language, and Hearing Research*, *52*(6), 1555–1565. https://doi.org/10.1044/1092-4388(2009/08-0137)

Fisher, C. G. (1968). Confusions Among Visually Perceived Consonants. *Journal of Speech and Hearing Research*, *11*(4), 796–804. https://doi.org/10.1044/jshr.1104.796

Flege, J. E. (1981). The Phonological Basis of Foreign Accent: A Hypothesis. *TESOL Quarterly*, *15*(4), 443. https://doi.org/10.2307/3586485

Flege, J. E. (1991). Age of learning affects the authenticity of voice-onset time (VOT) in stop consonants produced in a second language. *The Journal of the Acoustical Society of America*, *89*(1), 395–411. https://doi.org/10.1121/1.400473

Flege, J. E. (1993). Production and perception of a novel, second-language phonetic contrast. *The Journal of the Acoustical Society of America*, *93*(3), 1589–1608. https://doi.org/10.1121/1.406818

Flege, J. E., (2003). *Phonetics and Phonology in Language Comprehension and Production*. Google Books. Retrieved March 4, 2024, from https://books.google.co.za/books?hl=en&lr=&id=p7jUYjCndK4C&oi=fnd&pg=PA319&dq=Flege,+J.+E.+(2003).+Assessing+constraints+on+second-language+segmental+production+and+perception.+Phonetics+and+phonology+in+language+comprehension+and+production:+Differences+and+similarities,+6,+319-355.&ots=H6pOwzsB0B&sig=8CrgGSNYY-HWzkTduc8PI2h_1OI&redir_esc=y#v=onepage&q&f=false

Flege, J. E., & MacKay, I. R. A. (2004). Perceiving vowels in a second language. *Studies in Second Language Acquisition*, *26*(01). https://doi.org/10.1017/s0272263104026117

Flege, J. E., Munro, M. J., & Mackay, I. R. (1995). Effects of age of second-language learning on the production of English consonants. *Speech Communication*, *16*(1), 1–26. https://doi.org/10.1016/0167-6393(94)00044-b

Flege, J. E., Yeni-Komshian, G. H., & Liu, S. (1999). Age Constraints on Second-Language Acquisition. *Journal of Memory and Language*, *41*(1), 78–104. https://doi.org/10.1006/jmla.1999.2638

Fort, M., Kandel, S., Chipot, J., Savariaux, C., Granjon, L., & Spinelli, E. (2013). Seeing the initial articulatory gestures of a word triggers lexical access. *Language and Cognitive Processes*, *28*(8), 1207–1223. https://doi.org/10.1080/01690965.2012.701758

Fostick, L., Ben-Artzi, E., & Babkoff, H. (2013). Ageing and speech perception: beyond hearing threshold and cognitive ability. *Jbcpp*, *24*(3), 175–183. https://doi.org/10.1515/jbcpp-2013-0048

Francis, A. L., Ciocca, V., Ma, L., & Fenn, K. (2008). Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers. *Journal of Phonetics*, *36*(2), 268–294. https://doi.org/10.1016/j.wocn.2007.06.005

Fricke, M., Kroll, J. F., & Dussias, P. E. (2016). Phonetic variation in bilingual speech: A lens for studying the production–comprehension link. *Journal of Memory and Language*, *89*, 110–137. https://doi.org/10.1016/j.jml.2015.10.001

Fuchs, A. K., Hagmuller, M., & Kubin, G. (2016). The New Bionic Electro-Larynx Speech System. *IEEE Journal of Selected Topics in Signal Processing*, *10*(5), 952–961. https://doi.org/10.1109/jstsp.2016.2535970

Gandour, J., Dzemidzic, M., Wong, D., Lowe, M., Tong, Y., Hsieh, L., Satthamnuwong, N., & Lurito, J. (2003). Temporal integration of speech prosody is shaped by language experience: An fMRI study. *Brain and Language*, *84*(3), 318–336. https://doi.org/10.1016/s0093-934x(02)00505-9

Gardner, R., Lalonde, R., & Moorcroft, R. (1985). The role of attitudes and motivation in second language learning: correlational and experimental considerations. *Language Learning*, *35*(2), 207–227. https://doi.org/10.1111/j.1467-1770.1985.tb01025.x

Gaskell, M. G., & Snoeren, N. D. (2008). The impact of strong assimilation on the perception of connected speech. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(6), 1632–1647. https://doi.org/10.1037/a0011977

Gatehouse, S., & Gordon, J. (1990). Response times to speech stimuli as measures of benefit from amplification. *British Journal of Audiology*, *24*(1), 63–68. https://doi.org/10.3109/03005369009077843

Gelfer, M. P., & Bennett, Q. E. (2013). Speaking Fundamental Frequency and Vowel Formant Frequencies: Effects on Perception of Gender. *Journal of Voice*, *27*(5), 556–566. https://doi.org/10.1016/j.jvoice.2012.11.008

Genesee, F. (2022). The monolingual bias. *Journal of Immersion and Content-Based Language Education*, *10*(2). https://doi.org/10.1075/jicb.21016.gen

Gentilucci, M., & Cattaneo, L. (2005). Automatic audio-visual integration in speech perception. *Experimental Brain Research*, *167*(1), 66–75. https://doi.org/10.1007/s00221-005-0008-z

Georgiou, G. P. (2021). Toward a new model for speech perception: the Universal Perceptual Model (UPM) of second language. *Cognitive Processing*, *22*(2), 277–289. https://doi.org/10.1007/s10339-021-01017-6

Getz, L. M., & Toscano, J. C. (2020). The time-course of speech perception revealed by temporally-sensitive neural measures. *WIREs Cognitive Science*, *12*(2). https://doi.org/10.1002/wcs.1541

Ginther, A. (2002). Context and content visuals and performance on listening comprehension stimuli. *Language Testing*, *19*(2), 133–167. https://doi.org/10.1191/0265532202lt225oa

Girin, L., Schwartz, J. L., & Feng, G. (2001). Audio-visual enhancement of speech in noise. *The Journal of the Acoustical Society of America*, *109*(6), 3007–3020. https://doi.org/10.1121/1.1358887

Goetz, P. J. (2003). The effects of bilingualism on theory of mind development. *Bilingualism: Language and Cognition*, *6*(1), 1–15. https://doi.org/10.1017/s1366728903001007

González, G. (1989). BILINGUALISM: BASIC PRINCIPLES, 2nd ed.Hugo Baetens Beardsmore. Clevedon, England: Multilingual matters, 1986. Pp. ix + 205. *Studies in Second Language Acquisition*, *11*(1), 92–94. https://doi.org/10.1017/s0272263100007853

Gonzales, M. G., Backer, K. C., Mandujano, B., & Shahin, A. J. (2021). Rethinking the Mechanisms Underlying the McGurk Illusion. *Frontiers in Human Neuroscience*, *15*. https://doi.org/10.3389/fnhum.2021.616049

Goswami, U. (2002). Phonology, reading development, and dyslexia: A cross-linguistic perspective. *Annals of Dyslexia*, *52*(1), 139–163. https://doi.org/10.1007/s11881-002-0010-0

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

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, *108*(3), 1197–1208. https://doi.org/10.1121/1.1288668

Green , K. P. (2013). *The use of auditory and visual information during phonetic processing: implications for theories of speech perception*. Taylor & Francis. https://doi.org/10.4324/9780203098752-3

Green, K. P., & Gerdman, A. (1995). Cross-modal discrepancies in coarticulation and the integration of speech information: The McGurk effect with mismatched vowels. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(6), 1409–1426. https://doi.org/10.1037//0096-1523.21.6.1409

Grosjean, F. (2008). *The Lévy and Grosjean BIMOLA Model*∗. Oxford University Press eBooks. https://doi.org/10.1093/oso/9780199281282.003.0011

Grosjean, F. (2012). Bilingual and monolingual language modes. *The Encyclopedia of Applied Linguistics*. https://doi.org/10.1002/9781405198431.wbeal0090

Grosjean, F., & Byers-Heinlein, K. (2018). *The Listening Bilingual*. Google Books. Retrieved March 4, 2024, from https://books.google.co.za/books?hl=en&lr=&id=Yv9lDwAAQBAJ&oi=fnd&pg=PP8&dq=Grosjean,+F.,+%26+Byers-Heinlein,+K.+(2018).+The+listening+bilingual:+Speech+perception,+comprehension,+and+bilingualism.+John+Wiley+%26+Sons.&ots=NqZ0w5Faan&sig=SNuSMiOadS2B8EKlhEgzLrvvDs4&redir_esc=y#v=onepage&q=Grosjean%2C%20F.%2C%20%26%20Byers-Heinlein%2C%20K.%20(2018).%20The%20listening%20bilingual%3A%20Speech%20perception%2C%20comprehension%2C%20and%20bilingualism.%20John%20Wiley%20%26%20Sons.&f=false

Guion, S. G., Flege, J. E., Akahane-Yamada, R., & Pruitt, J. C. (2000). An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants. *The Journal of the Acoustical Society of America*, *107*(5), 2711–2724. https://doi.org/10.1121/1.428657

Gürbüz, C., & Cabaroğlu, N. (2021). EFL students' perceptions of oral presentations: Implications for motivation, language ability and speech anxiety. *Journal of Language and Linguistic Studies*, *17*(1), 600–614. https://doi.org/10.52462/jlls.41

Halle, M., & Stevens, K. (1962). Speech recognition: A model and a program for research. *IEEE Transactions on Information Theory*, *8*(2), 155–159. https://doi.org/10.1109/tit.1962.1057686

Hallé, P. A., Chang, Y. C., & Best, C. T. (2004). Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners. *Journal of Phonetics*, *32*(3), 395–421. https://doi.org/10.1016/s0095-4470(03)00016-0

Han, Y., Goudbeek, M., Mos, M., & Swerts, M. (2019). Relative Contribution of Auditory and Visual Information to Mandarin Chinese Tone Identification by Native and Tone-naïve Listeners. *Language and Speech*, *63*(4), 856–876. https://doi.org/10.1177/0023830919889995

Hanulíková, A., Dediu, D., Fang, Z., Bašnaková, J., & Huettig, F. (2012). Individual Differences in the Acquisition of a Complex L2 Phonology: A Training Study. *Language Learning*, *62*(s2), 79–109. https://doi.org/10.1111/j.1467-9922.2012.00707.x

Hao, Y. C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of Phonetics*, *40*(2), 269–279. https://doi.org/10.1016/j.wocn.2011.11.001

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

Hardison, D. M. (2009). Visual and auditory input in second-language speech processing. *Language Teaching*, *43*(1), 84–95. https://doi.org/10.1017/s0261444809990176

Hardison, D. M. (2018). Visualizing the acoustic and gestural beats of emphasis in multimodal discourse. *Journal of Second Language Pronunciation*, *4*(2), 232–259. https://doi.org/10.1075/jslp.17006.har

Harnsberger, J. D. (2001). On the relationship between identification and discrimination of non-native nasal consonants. *The Journal of the Acoustical Society of America*, *110*(1), 489–503. https://doi.org/10.1121/1.1371758

Hay, J. F., Graf Estes, K., Wang, T., & Saffran, J. R. (2014). From Flexibility to Constraint: The Contrastive Use of Lexical Tone in Early Word Learning. *Child Development*, *86*(1), 10–22. https://doi.org/10.1111/cdev.12269

Hazan, V., Sennema, A., Iba, M., & Faulkner, A. (2005). Effect of audio-visual perceptual training on the perception and production of consonants by Japanese learners of English. *Speech Communication*, *47*(3), 360–378. https://doi.org/10.1016/j.specom.2005.04.007

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

Heinrich, A., Henshaw, H., & Ferguson, M. A. (2016). Only Behavioral But Not Self-Report Measures of Speech Perception Correlate with Cognitive Abilities. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00576

Hempenstall, K. (1997). The Role of Phonemic Awareness in Beginning Reading: A Review. *Behaviour Change*, *14*(4), 201–214. https://doi.org/10.1017/s0813483900003314

Herrera, S. G., Holmes, M. A., & Kavimandan, S. K. (2012). Bringing Theory to Life: Strategies that Make Culturally Responsive Pedagogy a Reality in Diverse Secondary Classrooms. *International Journal of Multicultural Education*, *14*(3). https://doi.org/10.18251/ijme.v14i3.608

Hignett, S., & McDermott, H. (2015). *Qualitative methodology for Ergonomics. Chapter X. In Wilson, J.R., Sharples, S. (Eds.) Evaluation of Human Work. A practical ergonomics methodology.* (4th ed.).

Hirata, Y., & Kelly, S. D. (2010). Effects of Lips and Hands on Auditory Learning of Second-Language Speech Sounds. *Journal of Speech, Language, and Hearing Research*, *53*(2), 298–310. https://doi.org/10.1044/1092-4388(2009/08-0243)

Holle, H., Obermeier, C., Schmidt-Kassow, M., Friederici, A. D., Ward, J., & Gunter, T. C. (2012). Gesture Facilitates the Syntactic Analysis of Speech. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00074

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *The Journal of the Acoustical Society of America*, *119*(5), 3059–3071. https://doi.org/10.1121/1.2188377

Hommel, M. (2018). The Role of Orthography and Phoneme Inventory in Dutch Students' Speech Perception in the EFL Classroom. *Philologia*, *16*(16), 65–75. https://doi.org/10.18485/philologia.2018.16.16.4

Howson, P. J., Kallay, J. E., & Redford, M. A. (2020). A psycholinguistic method for measuring coarticulation in child and adult speech. *Behavior Research Methods*, *53*(2), 846–863. https://doi.org/10.3758/s13428-020-01464-7

Irwin, J., & DiBlasi, L. (2017). Audio-visual speech perception: A new approach and implications for clinical populations. *Language and Linguistics Compass*, *11*(3), 77–91. https://doi.org/10.1111/lnc3.12237

Iverson, P., & Evans, B. G. (2009). Learning English vowels with different first-language vowel systems II: Auditory training for native Spanish and German speakers. *The Journal of the Acoustical Society of America*, *126*(2), 866–877. https://doi.org/10.1121/1.3148196

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

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *The Journal of the Acoustical Society of America*, *97*(1), 553–562. https://doi.org/10.1121/1.412280

Jansen, S., Chaparro, A., Downs, D., Palmer, E., & Keebler, J. (2013). Visual and Cognitive Predictors of Visual Enhancement in Noisy Listening Conditions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*(1), 1199–1203. https://doi.org/10.1177/1541931213571267

Jerger, S., Tye-Murray, N., & Abdi, H. (2009). Role of Visual Speech in Phonological Processing by Children With Hearing Loss. *Journal of Speech, Language, and Hearing Research*, *52*(2), 412–434. https://doi.org/10.1044/1092-4388(2009/08-0021)

Jesse, A., & Janse, E. (2012). Audio-visual benefit for recognition of speech presented with single-talker noise in older listeners. *Language and Cognitive Processes*, *27*(7–8), 1167–1191. https://doi.org/10.1080/01690965.2011.620335

Jesse, A., & Massaro, D. W. (2010). The temporal distribution of information in audio-visual spoken-word identification. *Attention, Perception, & Psychophysics*, *72*(1), 209–225. https://doi.org/10.3758/app.72.1.209

Jesse, A., Vrignaud, N., Cohen, M. M., & Massaro, D. W. (2000). The processing of information from multiple sources in simultaneous interpreting. *Interpreting. International Journal of Research and Practice in Interpreting*, *5*(2), 95–115. https://doi.org/10.1075/intp.5.2.04jes

Jiang, J., & Bernstein, L. E. (2011). Psychophysics of the McGurk and other audio-visual speech integration effects. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(4), 1193–1209. https://doi.org/10.1037/a0023100

Khoza-Shangase, K., & Masondo, N. (2020). What Are the Current Audiological Practices for Ototoxicity Assessment and Management in the South African Healthcare Context? *International Journal of Environmental Research and Public Health*, *17*(7), 2613. https://doi.org/10.3390/ijerph17072613

Klapp, S. T., & Jagacinski, R. J. (2011). Gestalt principles in the control of motor action. *Psychological Bulletin*, *137*(3), 443–462. https://doi.org/10.1037/a0022361

Kluge, D. C., Reis, M. S., Nobre-Oliveira, D., & Bettoni, M. (2009). *The use of visual cues in the perception of English syllable-final nasals by Brazilian EFL learners*. ResearchGate. https://www.researchgate.net/publication/228485559_The_use_of_visual_cues_in_the_perception_of_English_syllable-final_nasals_by_Brazilian_EFL_learners

Kohnert, K. (2008). Second Language Acquisition: Success Factors in Sequential Bilingualism. *The ASHA Leader*, *13*(2), 10–13. https://doi.org/10.1044/leader.ftr1.13022008.10

Kondaurova, M. V., & Francis, A. L. (2010). The role of selective attention in the acquisition of English tense and lax vowels by native Spanish listeners: Comparison of three training methods. *Journal of Phonetics*, *38*(4), 569–587. https://doi.org/10.1016/j.wocn.2010.08.003

Krishnan, A., Xu, Y., Gandour, J., & Cariani, P. (2005). Encoding of pitch in the human brainstem is sensitive to language experience. *Cognitive Brain Research*, *25*(1), 161–168. https://doi.org/10.1016/j.cogbrainres.2005.05.004

Kroll, J. F., & Bialystok, E. (2013). Understanding the consequences of bilingualism for language processing and cognition. *Journal of Cognitive Psychology*, *25*(5), 497–514. https://doi.org/10.1080/20445911.2013.799170

Kucirkova, N. (2011). *Language and Cognition in Bilinguals and Multilinguals*. Europe's Journal of Psychology. https://doi.org/10.5964/ejop.v7i2.137

Kuhl, P. K., & Meltzoff, A. N. (1996). Infant vocalizations in response to speech: Vocal imitation and developmental change. *The Journal of the Acoustical Society of America*, *100*(4), 2425–2438. https://doi.org/10.1121/1.417951

Kumaravadivelu, B. (2006). TESOL Methods: Changing Tracks, Challenging Trends. *TESOL Quarterly*, *40*(1), 59. https://doi.org/10.2307/40264511

Kunene-Nicolas, R. (2015). Zulu oral narrative development from a speech and gesture perspective. *Per Linguam*, *31*(3). https://doi.org/10.5785/31-3-560

Kyle, F. E., Campbell, R., Mohammed, T., Coleman, M., & MacSweeney, M. (2013). Speechreading Development in Deaf and Hearing Children: Introducing the Test of Child Speechreading. *Journal of Speech, Language, and Hearing Research*, *56*(2), 416–426. https://doi.org/10.1044/1092-4388(2012/12-0039)

Lalonde, K., & McCreery, R. W. (2020). Audio-visual Enhancement of Speech Perception in Noise by School-Age Children Who Are Hard of Hearing. *Ear & Hearing*, *41*(4), 705–719. https://doi.org/10.1097/aud.0000000000000830

Lalonde, K., & Werner, L. A. (2019). Infants and Adults Use Visual Cues to Improve Detection and Discrimination of Speech in Noise. *Journal of Speech, Language, and Hearing Research*, *62*(10), 3860–3875. https://doi.org/10.1044/2019_jslhr-h-19-0106

Lalonde, K., & Werner, L. A. (2021). Development of the Mechanisms Underlying Audio-visual Speech Perception Benefit. *Brain Sciences*, *11*(1), 49. https://doi.org/10.3390/brainsci11010049

Lambacher, S. G., Martens, W. L., Kakehi, K., Marasinghe, C. A., & Molholt, G. (2005). The effects of identification training on the identification and production of American English vowels by native speakers of Japanese. *Applied Psycholinguistics*, *26*(2), 227–247. https://doi.org/10.1017/s0142716405050150

Land, S. (2015). Skilled reading in isiZulu: what can we learn from it? *Journal of Education*, *63*. https://doi.org/10.17159/i63a03

Language Policy - Wits University. (2024). Www.wits.ac.za. https://www.wits.ac.za/about-wits/governance/language-policy/

Lansing, C. R., & McConkie, G. W. (2003). Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. *Perception & Psychophysics*, *65*(4), 536–552. https://doi.org/10.3758/bf03194581

Leather, J. (1987). *Sound Patterns in Second Language Acquisition*. Google Books. Retrieved March 5, 2024, from https://books.google.co.za/books?hl=en&lr=&id=kFG9DwAAQBAJ&oi=fnd&pg=PA59&dq=Leather,+J.+(1987).+F0+pattern+inference+in+the+perceptual+acquisition+of+second+language+tone.+Sound+patterns+in+second+language+acquisition,+59-81.&ots=LVLtTGjdJV&sig=Ciqywa5w6aMF5CNbz6vy1CFexAU&redir_esc=y#v=onepage&q&f=false

Lecumberri, M. L. G., Cooke, M., & Cutler, A. (2010). Non-native speech perception in adverse conditions: A review. *Speech Communication*, *52*(11–12), 864–886. https://doi.org/10.1016/j.specom.2010.08.014

Lei, S. A., & Gupta, R. K. (2010). *College distance education courses: evaluating benefits and costs from institutional, faculty and students' perspectives. | Education | EBSCOhost*. https://openurl.ebsco.com/EPDB%3Agcd%3A15%3A16966831/detailv2?sid=ebsco%3Aplink%3Ascholar&id=ebsco%3Agcd%3A50839014&crl=c

Levy, Y., & Ellis, T. J. (2011). A Guide for Novice Researchers on Experimental and Quasi-Experimental Studies in Information Systems Research. *Interdisciplinary Journal of Information, Knowledge, and Management*, *6*, 151–161. https://doi.org/10.28945/1373

Lewis, G., Jones, B., & Baker, C. (2012). Translanguaging: developing its conceptualisation and contextualisation. *Educational Research and Evaluation*, *18*(7), 655–670. https://doi.org/10.1080/13803611.2012.718490

Lewkowicz, D. J., & Ghazanfar, A. A. (2009). The emergence of multisensory systems through perceptual narrowing. *Trends in Cognitive Sciences*, *13*(11), 470–478. https://doi.org/10.1016/j.tics.2009.08.004

Lewkowicz, M. A. (2006). The Effectiveness of Elite Cues as Heuristics in Proposition Elections. *American Politics Research*, *34*(1), 51–68. https://doi.org/10.1177/1532673x05276510

Li, Y. (2015). English and Thai Speakers' Perception of Mandarin Tones. *English Language Teaching*, *9*(1), 122. https://doi.org/10.5539/elt.v9n1p122

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358–368. https://doi.org/10.1037/h0044417

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1–36. https://doi.org/10.1016/0010-0277(85)90021-6

Lidestam, B., & Beskow, J. (2006). Motivation and appraisal in perception of poorly specified speech. *Scandinavian Journal of Psychology*, *47*(2), 93–101. https://doi.org/10.1111/j.1467-9450.2006.00496.x

Lin, W. C. (1985). *Teaching Mandarin Tones to Adult English Speakers: Analysis of Difficulties with Suggested Remedies*. RELC Journal. https://doi.org/10.1177/003368828501600207

Lindemann, D. (2013). Bilingual Lexicography and Corpus Methods. The Example of German-Basque as Language Pair. *Procedia - Social and Behavioral Sciences*, *95*, 249–257. https://doi.org/10.1016/j.sbspro.2013.10.645

Litovsky, R. Y., McAlpine , D., Rees, A., & Palmer, A. (2010). *The Oxford Handbook of Auditory Science: The Auditory Brain*. Google Books. Retrieved March 5, 2024, from https://books.google.co.za/books?hl=en&lr=&id=Kyq65C3ygGAC&oi=fnd&pg=PA333&dq=Litovsky,+R.+Y.,+McAlpine,+D.,+Rees,+A.,+%26+Palmer,+A.+(2010).+Physiological+correlates+of+the+precedence+effect+and+binaural+masking+level+differences.+The+Oxford+handbook+of+auditory+science:+The+auditory+brain,+2,+333-357.&ots=wKK2MQJnPT&sig=l9KaCUKC9i4dJzXHwz59ToKDk1Q&redir_esc=y#v=onepage&q&f=false

Liu, H., Bates, E., & Li, P. (1992). Sentence interpretation in bilingual speakers of English and Chinese. *Applied Psycholinguistics*, *13*(4), 451–484. https://doi.org/10.1017/s0142716400005762

Liu, L., & Kager, R. (2016). Perception of tones by bilingual infants learning non-tone languages. *Bilingualism: Language and Cognition*, *20*(3), 561–575. https://doi.org/10.1017/s1366728916000183

Liu, S., & Samuel, A. G. (2004). Perception of Mandarin Lexical Tones when F0 Information is Neutralized. *Language and Speech*, *47*(2), 109–138. https://doi.org/10.1177/00238309040470020101

Liu, W., & Shi, P. (2019). Convergence of Optimal Linear Estimator With Multiplicative and Time-Correlated Additive Measurement Noises. *IEEE Transactions on Automatic Control*, *64*(5), 2190–2197. https://doi.org/10.1109/tac.2018.2869467

Liu, S., & Zeng, F. G. (2006). Temporal properties in clear speech perception. *The Journal of the Acoustical Society of America*, *120*(1), 424–432. https://doi.org/10.1121/1.2208427

Long, M. H. (2019). Optimal input for language learning: Genuine, simplified, elaborated, or modified elaborated? *Language Teaching*, *53*(2), 169–182. https://doi.org/10.1017/s0261444819000466

Luce, P. A., & Pisoni, D. B. (1998). Recognizing Spoken Words: The Neighborhood Activation Model. *Ear And Hearing*, *19*(1), 1–36. https://doi.org/10.1097/00003446-199802000-00001

Lusk, L. G., & Mitchel, A. D. (2016). Differential Gaze Patterns on Eyes and Mouth During Audio-visual Speech Segmentation. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00052

Macaluso, E., George, N., Dolan, R., Spence, C., & Driver, J. (2004). Spatial and temporal factors during processing of audio-visual speech: a PET study. *NeuroImage*, *21*(2), 725–732. https://doi.org/10.1016/j.neuroimage.2003.09.049

Magnotti, J. F., Basu Mallick, D., Feng, G., Zhou, B., Zhou, W., & Beauchamp, M. S. (2015). Similar frequency of the McGurk effect in large samples of native Mandarin Chinese and American English speakers. *Experimental Brain Research*, *233*(9), 2581–2586. https://doi.org/10.1007/s00221-015-4324-7

Mahdi, W., Werda, S., & Hamadou, A. B. (2008). A hybrid approach for automatic lip localization and viseme classification to enhance visual speech recognition. *Integrated Computer-Aided Engineering*, *15*(3), 253–266. https://doi.org/10.3233/ica-2008-15305

Makina, B. T., & Lephalala, M. M. K. (2010). *Error analysis : investigating the writing of ESL Namibian learners*. https://uir.unisa.ac.za/handle/10500/4893

Marian, V., Hayakawa, S., Lam, T. Q., & Schroeder, S. R. (2018). *Language Experience Changes Audio-visual Perception*. Brain Sciences. https://doi.org/10.3390/brainsci8050085

Marshall, G., & Jonker, L. (2010). An introduction to descriptive statistics: A review and practical guide. *Radiography*, *16*(4), e1–e7. https://doi.org/10.1016/j.radi.2010.01.001

Mártony, J. (1974). On speechreading of Swedish consonants and vowels. *KTH STL-Qarterly Progress and Status Report*, 11-33.

Massaro, D. W. (1989). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology*, *21*(3), 398–421. https://doi.org/10.1016/0010-0285(89)90014-5

Massaro, D. W., & Cohen, M. M. (1983). Phonological context in speech perception. *Perception & Psychophysics*, *34*(4), 338–348. https://doi.org/10.3758/bf03203046

Massaro, D. W., Cohen, M. M., Gesi, A., Heredia, R., & Tsuzaki, M. (1993). Bimodal speech perception: an examination across languages. *Journal of Phonetics*, *21*(4), 445–478. https://doi.org/10.1016/s0095-4470(19)30230-x

Massaro, D. W., & Egan, P. B. (1996). *Perceiving affect from the voice and the face*. Psychonomic Bulletin & Review. https://doi.org/10.3758/bf03212421

Massaro, D. W., Thompson, L. A., Barron, B., & Laren, E. (1986). Developmental changes in visual and auditory contributions to speech perception. *Journal of Experimental Child Psychology*, *41*(1), 93–113. https://doi.org/10.1016/0022-0965(86)90053-6

Matthews, I., Cootes, T., Bangham, J., Cox, S., & Harvey, R. (2002). Extraction of visual features for lipreading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(2), 198–213. https://doi.org/10.1109/34.982900

Mattheyses, W., Latacz, L., & Verhelst, W. (2013). Comprehensive many-to-many phoneme-to-viseme mapping and its application for concatenative visual speech synthesis. *Speech Communication*, *55*(7–8), 857–876. https://doi.org/10.1016/j.specom.2013.02.005

Mattock, K., Molnar, M., Polka, L., & Burnham, D. (2008). The developmental course of lexical tone perception in the first year of life. *Cognition*, *106*(3), 1367–1381. https://doi.org/10.1016/j.cognition.2007.07.002

Mayo, C., & Turk, A. (2004). Adult–child differences in acoustic cue weighting are influenced by segmental context: Children are not always perceptually biased toward transitions. *The Journal of the Acoustical Society of America*, *115*(6), 3184–3194. https://doi.org/10.1121/1.1738838

Mayyas, H. N., & Farhan, A. (2023). *Students' and Teachers' Attitudes toward Using Audio-visual Chat in Teaching English Language Skills and Areas to the Seventh Grade Students*. https://www.semanticscholar.org/paper/Students%E2%80%99-and-Teachers%E2%80%99-Attitudes-toward-Using-Chat-Mayyas-Farhan/f984535e1bdb6865146bfa6baa0bd69efcf29184

Mazzaro, N., Colantoni, L., & Cuza, A. (2016). Age effects and the discrimination of consonantal and vocalic contrasts in heritage and native Spanish. In *Romance linguistics 2013: selected proceedings of the 43th linguistic symposium on romance languages* (pp. 277-300). Amsterdam, Netherlands: John Benjamins.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review*, *88*(5), 375–407. https://doi.org/10.1037//0033-295x.88.5.375

McGrath, M., & Summerfield, Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *The Journal of the Acoustical Society of America*, *77*(2), 678–685. https://doi.org/10.1121/1.392336

McGuire, G., & Babel, M. (2012). A cross-modal account for synchronic and diachronic patterns of /f/ and /θ/ in English. *Laboratory Phonology*, *3*(2). https://doi.org/10.1515/lp-2012-0014

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. https://doi.org/10.1038/264746a0

Messum, P., & Howard, I. S. (2015). *Creating the cognitive form of phonological units: The speech sound correspondence problem in infancy could be solved by mirrored vocal interactions rather than by imitation*. Journal of Phonetics. https://doi.org/10.1016/j.wocn.2015.08.005

Miller, G. A., & Nicely, P. E. (1955). An Analysis of Perceptual Confusions Among Some English Consonants. *The Journal of the Acoustical Society of America*, *27*(2), 338–352. https://doi.org/10.1121/1.1907526

Mishra, P., & Koehler, M. J. (2006). Technological Pedagogical Content Knowledge: A Framework for Teacher Knowledge. *Teachers College Record: The Voice of Scholarship in Education*, *108*(6), 1017–1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x

Mitterer, H., & Tuinman, A. (2012). The Role of Native-Language Knowledge in the Perception of Casual Speech in a Second Language. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00249

Miyawaki, K., Jenkins, J. J., Strange, W., Liberman, A. M., Verbrugge, R., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception & Psychophysics*, *18*(5), 331–340. https://doi.org/10.3758/bf03211209

Mok, P. P. K., Zuo, D., & Wong, P. W. Y. (2013). Production and perception of a sound change in progress: Tone merging in Hong Kong Cantonese. *Language Variation and Change*, *25*(3), 341–370. https://doi.org/10.1017/s0954394513000161

Montero Perez, M. (2022). Second or foreign language learning through watching audio-visual input and the role of on-screen text. *Language Teaching*, *55*(2), 163–192. https://doi.org/10.1017/s0261444821000501

Moon, C., Lagercrantz, H., & Kuhl, P. K. (2013). Language experienced in utero affects vowel perception after birth: a two-country study. *Acta Paediatrica*, *102*(2), 156–160. https://doi.org/10.1111/apa.12098

Moore, B. C. (2012). Contributions of von Békésy to psychoacoustics. *Hearing Research*, *293*(1–2), 51–57. https://doi.org/10.1016/j.heares.2012.04.009

Morison, L. D., van Reyk, O., Forbes, E., Rouxel, F., Faivre, L., Bruinsma, F., Vincent, M., Jacquemont, M. L., Dykzeul, N. L., Geneviève, D., Amor, D. J., & Morgan, A. T. (2023). Correction: CDK13-related disorder: a deep characterization of speech and language abilities and addition of 33 novel cases. *European Journal of Human Genetics*. https://doi.org/10.1038/s41431-023-01515-5

Morrison, G. S. (2008). L1-Spanish Speakers' Acquisition of the English /i /—/I/ Contrast: Duration-based Perception is Not the Initial Developmental Stage. *Language and Speech*, *51*(4), 285–315. https://doi.org/10.1177/0023830908099067

Mosaka, N. M. (2000). Stress assignment in syllabic structures in Xhosa and Tswana. *South African Journal of African Languages*, *20*(2), 177–185. https://doi.org/10.1080/02572117.2000.10587424

Mugitani, R., Kobayashi, T., & Hiraki, K. (2008). Audio-visual matching of lips and non-canonical sounds in 8-month-old infants. *Infant Behavior and Development*, *31*(2), 307–310. https://doi.org/10.1016/j.infbeh.2007.12.002

Munhall, K., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual Prosody and Speech Intelligibility. *Psychological Science*, *15*(2), 133–137. https://doi.org/10.1111/j.0963-7214.2004.01502010.x

Murphy, V. A., Arndt, H., Baffoe-Djan, J. B., Chalmers, H., Macaro, E., Rose, H., Vanderplank, R., & Woore, R. (2020). *Foreign Language Learning and Its Impact on Wider Academic Outcomes: A Rapid Evidence Assessment*. Retrieved March 5, 2024, from https://eric.ed.gov/?id=ED612981

Naderifar, M., Goli, H., & Ghaljaie, F. (2017). Snowball Sampling: A Purposeful Method of Sampling in Qualitative Research. *Strides in Development of Medical Education*, *14*(3). https://doi.org/10.5812/sdme.67670

Nagels, A., Kircher, T., Steines, M., Grosvald, M., & Straube, B. (2015). A brief self-rating scale for the assessment of individual differences in gesture perception and production. *Learning and Individual Differences*, *39*, 73–80. https://doi.org/10.1016/j.lindif.2015.03.008

Naidoo, S. (2002). The palatalisation process in isiZulu revisited. *South African Journal of African Languages*, *22*(1), 59–69. https://doi.org/10.1080/02572117.2002.10587498

Naidoo, Y., van der Merwe, A., Groenewald, E., & Naudé, E. (2005). Development of speech sounds and syllable structure of words in Zulu-speaking children. *Southern African Linguistics and Applied Language Studies*, *23*(1), 59–79. https://doi.org/10.2989/16073610509486374

Namaziandost, E., & Nasri, M. (2019). The impact of using audio-visual aids on teaching listening among Iranian pre-intermediate efl learners. *LLT Journal: A Journal on Language and Language Teaching*, *22*(2), 246–259. https://doi.org/10.24071/llt.v22i2.2028

Nasir, S. H., Popat, H., & Richmond, S. (2020). The effect of resting morphological lip shape during lip movement: A three-dimensional motion analysis study. *Heliyon*, *6*(6), e04093. https://doi.org/10.1016/j.heliyon.2020.e04093

Nath, A. R., Fava, E. E., & Beauchamp, M. S. (2011). Neural Correlates of Interindividual Differences in Children's Audio-visual Speech Perception. *The Journal of Neuroscience*, *31*(39), 13963–13971. https://doi.org/10.1523/jneurosci.2605-11.2011

Navarra, J., & Soto-Faraco, S. (2005). Hearing lips in a second language: visual articulatory information enables the perception of second language sounds. *Psychological Research*, *71*(1), 4–12 https://doi.org/10.1007/s00426-005-0031-5

Navarra, J., & Soto-Faraco, S. (2007). Hearing lips in a second language: visual articulatory information enables the perception of second language sounds. *Psychological Research*, *71*(1), 4–12. https://doi.org/10.1007/s00426-005-0031-5

Navarra, J., Vatakis, A., Zampini, M., Soto-Faraco, S., Humphreys, W., & Spence, C. (2005). Exposure to asynchronous audio-visual speech extends the temporal window for audio-visual integration. *Cognitive Brain Research*, *25*(2), 499–507. https://doi.org/10.1016/j.cogbrainres.2005.07.009

Ncobela, N. R. (2015). *Research Portal*. Retrieved March 5, 2024, from https://ujcontent.uj.ac.za/esploro/outputs/9910462107691

Neitzel, A. J., Lake, C., Pellegrini, M., & Slavin, R. E. (2021). A Synthesis of Quantitative Research on Programs for Struggling Readers in Elementary Schools. *Reading Research Quarterly*, *57*(1), 149–179. https://doi.org/10.1002/rrq.379

Nicolas, R. K., & Ahmed, S. (2016). Lexical development of noun and predicate comprehension and production in isiZulu. *South African Journal of Communication Disorders*, *63*(2). https://doi.org/10.4102/sajcd.v63i2.169

Nishi, K., & Kewley-Port, D. (2007). Training Japanese Listeners to Perceive American English Vowels: Influence of Training Sets. *Journal of Speech, Language, and Hearing Research*, *50*(6), 1496–1509. https://doi.org/10.1044/1092-4388(2007/103)

Nkomo, D., & Madiba, M. (2012). The Compilation of Multilingual Concept Literacy Glossaries at the University of Cape Town: A Lexicographical Function Theoretical Approach. *Lexikos*, *21*(1). https://doi.org/10.5788/21-1-41

Noesgaard, S. S., & Ørngreen, R. (2015). *The Effectiveness of E-Learning: An Explorative and Integrative Review of the Definitions, Methodologies and Factors That Promote e-Learning Effectiveness*. https://eric.ed.gov/?id=EJ1062121

Nunes, A. (2015). European Portuguese Phonetics: Difficulties for Chinese Speakers – Considerations. *Procedia - Social and Behavioral Sciences*, *192*, 92–100. https://doi.org/10.1016/j.sbspro.2015.06.014

Ong, J. H., Escudero, P., & Stevens , C. J. (2017). *APA PsycNet*. Retrieved March 5, 2024, from https://psycnet.apa.org/record/2018-07730-001

Ouellette, G. P., & Haley, A. (2011). One complicated extended family: the influence of alphabetic knowledge and vocabulary on phonemic awareness. *Journal of Research in Reading*, *36*(1), 29–41. https://doi.org/10.1111/j.1467-9817.2010.01486.x

Owens, E., & Blazek, B. (1985). Visemes Observed by Hearing-Impaired and Normal-Hearing Adult Viewers. *Journal of Speech, Language, and Hearing Research*, *28*(3), 381–393. https://doi.org/10.1044/jshr.2803.381

Oxford, R. L. (2003). Language learning styles and strategies: Concepts and relationships. *IRAL - International Review of Applied Linguistics in Language Teaching*, *41*(4). https://doi.org/10.1515/iral.2003.012

Parker, C., Scott, S. and Geddes, A. (2019) *Snowball Sampling. SAGE Research Methods Foundations. - References - Scientific Research Publishing*. (n.d.). https://www.scirp.org/reference/referencespapers?referenceid=3131783

Pascoe, M., & Jeggo, Z. (2019). Speech Acquisition in Monolingual Children Acquiring isiZulu in Rural KwaZulu-Natal, South Africa. *Journal of Monolingual and Bilingual Speech*, *1*(1). https://doi.org/10.1558/jmbs.11082

Pattamadilok, C., & Sato, M. (2022). How are visemes and graphemes integrated with speech sounds during spoken word recognition? ERP evidence for supra-additive responses during audio-visual compared to auditory speech processing. *Brain and Language*, *225*, 105058. https://doi.org/10.1016/j.bandl.2021.105058

Patterson, M. L., & Werker, J. F. (2002). Infants' Ability to Match Dynamic Phonetic and Gender Information in the Face and Voice. *Journal of Experimental Child Psychology*, *81*(1), 93–115. https://doi.org/10.1006/jecp.2001.2644

Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science*, *6*(2), 191–196. https://doi.org/10.1111/1467-7687.00271

Pearl, D., Yodashkin-Porat, D., Katz, N., Valevski, A., Aizenberg, D., Sigler, M., Weizman, A., & Kikinzon, L. (2009). Differences in audio-visual integration, as measured by McGurk phenomenon, among adult and adolescent patients with schizophrenia and age-matched healthy control groups. *Comprehensive Psychiatry*, *50*(2), 186–192. https://doi.org/10.1016/j.comppsych.2008.06.004

Peelle, J. E., & Davis, M. H. (2012). Neural Oscillations Carry Speech Rhythm through to Comprehension. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00320

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audio-visual speech perception. *Cortex*, *68*, 169–181. https://doi.org/10.1016/j.cortex.2015.03.006

Peng, G., Zheng, H. Y., Gong, T., Yang, R. X., Kong, J. P., & Wang, W. S. Y. (2010). The influence of language experience on categorical perception of pitch contours. *Journal of Phonetics*, *38*(4), 616–624. https://doi.org/10.1016/j.wocn.2010.09.003

Pennington, M. C. (1999). Computer-Aided Pronunciation Pedagogy: Promise, Limitations, Directions. *Computer Assisted Language Learning*, *12*(5), 427–440. https://doi.org/10.1076/call.12.5.427.5693

Pereira Reyes, Y., & Hazan, V. (2021). English vowel perception by non-native speakers: impact of audio and visual training modalities. *Onomázein Revista De Lingüística Filología Y Traducción*, *51*, 112–136. https://doi.org/10.7764/onomazein.51.04

Perez, M. M., & Rodgers, M. P. H. (2019). Video and language learning. *The Language Learning Journal*, *47*(4), 403–406. https://doi.org/10.1080/09571736.2019.1629099

Peters, E., & Muñoz, C. (2020). Introduction to the special issue. *Studies in Second Language Acquisition*, *42*(3), 489–497. https://doi.org/10.1017/s0272263120000212

Peters, M., Weinberg, A., & Sarma, N. (2009). To Like or Not to Like! Student Perceptions of Technological Activities for Learning French as a Second Language at Five Canadian Universities. *The Canadian Modern Language Review*, *65*(5), 869–896. https://doi.org/10.3138/cmlr.65.5.869

Piquard-Kipffer, A., Cavadini, T., Sprenger-Charolles, L., & Gentaz, D. (2021). Impact of lip-reading on speech perception in French-speaking children at risk for reading failure assessed from age 5 to 7. *L'Année Psychologique*, *Vol. 121*(2), 3–18. https://doi.org/10.3917/anpsy1.212.0003

Piske, T., Flege, J. E., MacKay, I. R., & Meador, D. (2002). The Production of English Vowels by Fluent Early and Late Italian-English Bilinguals. *Phonetica*, *59*(1), 49–71. https://doi.org/10.1159/000056205

Pisoni, D. B. (1993). Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. *Speech Communication*, *13*(1–2), 109–125. https://doi.org/10.1016/0167-6393(93)90063-q

Pitt, M. A., & McQueen, J. M. (1998). Is Compensation for Coarticulation Mediated by the Lexicon? *Journal of Memory and Language*, *39*(3), 347–370. https://doi.org/10.1006/jmla.1998.2571

Pons, F., Bosch, L., & Lewkowicz, D. J. (2015). Bilingualism Modulates Infants' Selective Attention to the Mouth of a Talking Face. *Psychological Science*, *26*(4), 490–498. https://doi.org/10.1177/0956797614568320

Posthumus, L. (2016). A critique of 'A re-evaluation of tense in isiZulu.' *Literator*, *37*(1). https://doi.org/10.4102/lit.v37i1.1349

Poulos, G., & Msimang, C. T. (1998). *A Linguistic Analysis of Zulu*. Via Africa. http://books.google.ie/books?id=O_zzAAAACAAJ&dq=A+linguistic+analysis+of+Zulu+Authors:G.+Poulos,+C.+T.+Msimang&hl=&cd=1&source=gbs_api

Poulos, G., & Msimang, C. T. (2001). The ideophone in Zulu. *Typological Studies in Language*, 235–249. https://doi.org/10.1075/tsl.44.19msi

Potamianos, G., Neti, C., Luettin, J., & Matthews, I. (2004). *Audio-Visual Automatic Speech Recognition: An Overview*.

Preston, J., & Edwards, M. L. (2010). Phonological Awareness and Types of Sound Errors in Preschoolers With Speech Sound Disorders. *Journal of Speech, Language, and Hearing Research*, *53*(1), 44–60. https://doi.org/10.1044/1092-4388(2009/09-0021)

Prince, C. G., & Hollich, G. J. (2005). Synching models with infants: a perceptual-level model of infant audio-visual synchrony detection. *Cognitive Systems Research*, *6*(3), 205–228. https://doi.org/10.1016/j.cogsys.2004.11.006

Proverbio, A. M., Massetti, G., Rizzi, E., & Zani, A. (2016). Skilled musicians are not subject to the McGurk effect. *Scientific Reports*, *6*(1). https://doi.org/10.1038/srep30423

Putri, F. Z. (2015). The Effectiveness of Minimal Pairs Drill Towards Students' Ability in Pronouncing Similar Sounds of Words. *Semantic Scholar*. https://api.semanticscholar.org/CorpusID:63242442

Qin, Z., & Jongman, A. (2016). Does Second Language Experience Modulate Perception of Tones in a Third Language? *Language and Speech*, *59*(3), 318–338. https://doi.org/10.1177/0023830915590191

Qin, Z., Zhang, C., & Wang, W. Y. (2021). *The effect of Mandarin listeners' musical and pitch aptitude on perceptual learning of Cantonese level-tones*. Journal of the Acoustical Society of America. https://doi.org/10.1121/10.0003330

Quam, C., & Swingley, D. (2010). Phonological knowledge guides 2-year-olds' and adults' interpretation of salient pitch contours in word learning. *Journal of Memory and Language*, *62*(2), 135–150. https://doi.org/10.1016/j.jml.2009.09.003

Rashid, M. H., & Sipahi, E. (2021). importance of quantitative research in language testing and assessment: in the context of social works. *Linguistics and Culture Review*, *5*(S1), 317–330. https://doi.org/10.21744/lingcure.v5ns1.1413

Reetzke, R., Maddox, W. T., & Chandrasekaran, B. (2016). The role of age and executive function in auditory category learning. *Journal of Experimental Child Psychology*, *142*, 48–65. https://doi.org/10.1016/j.jecp.2015.09.018

Reid, S. A., Keerie, N., & Palomares, N. A. (2003). Language, Gender Salience, and Social Influence. *Journal of Language and Social Psychology*, *22*(2), 210–233. https://doi.org/10.1177/0261927x03022002004

Reisberg, D., Goldfield, A., Dodd, B., & Campbell, R. (1987). *Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli*. Retrieved March 5, 2024, from https://www.semanticscholar.org/paper/Easy-to-hear-but-hard-to-understand%3A-A-lip-reading-Reisberg-Mclean/9b9f3cca398a35d42b5d5c56d87813cc92da8ec6

Richter, K. G., & Lara Herrera, R. (2016). Characteristics and Pedagogical Behaviours of Good EFL Instructors: The Views of Selected Southeast Asian and Mexican SLTE Students. *RELC Journal*, *48*(2), 180–196. https://doi.org/10.1177/0033688216645473

Robinson, C. W., Chandra, M., & Sinnett, S. (2016). Existence of competing modality dominances. *Attention, Perception, & Psychophysics*, *78*(4), 1104–1114. https://doi.org/10.3758/s13414-016-1061-3

Roever, C. (2006). Book Review: New perspectives on CALL for second language classrooms. *Language Teaching Research*, *10*(2), 231–233. https://doi.org/10.1177/136216880601000207

Rosenblum, L. D. (2008). Speech Perception as a Multimodal Phenomenon. *Current Directions in Psychological Science*, *17*(6), 405–409. https://doi.org/10.1111/j.1467-8721.2008.00615.x

Rosenblum, L. D. (2019). Does Sensory Modality Matter? Not for Speech Perception. *Acoustics Today*, *15*(3), 47. https://doi.org/10.1121/at.2019.15.3.47

Rosenblum, L. D., Miller, R. M., & Sanchez, K. (2007). Lip-Read Me Now, Hear Me Better Later. *Psychological Science*, *18*(5), 392–396. https://doi.org/10.1111/j.1467-9280.2007.01911.x

Rosenblum, L. D., & Saldaña, H. M. (1996). An audio-visual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception Performance*, *22*(2), 318–331. https://doi.org/10.1037//0096-1523.22.2.318

Ross, L. A., Molholm, S., Blanco, D., Gomez-Ramirez, M., Saint-Amour, D., & Foxe, J. J. (2011). The development of multisensory speech perception continues into the late childhood years. *European Journal of Neuroscience*, *33*(12), 2329–2337. https://doi.org/10.1111/j.1460-9568.2011.07685.x

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2006). Do You See What I Am Saying? Exploring Visual Enhancement of Speech Comprehension in Noisy Environments. *Cerebral Cortex*, *17*(5), 1147–1153. https://doi.org/10.1093/cercor/bhl024

Ross, B., & Tremblay, K. (2009). Stimulus experience modifies auditory neuromagnetic responses in young and older listeners. *Hearing Research*, *248*(1–2), 48–59. https://doi.org/10.1016/j.heares.2008.11.012

Saville-Troike, M., & Barto, K. (2017). *Introducing second language acquisition*. Cambridge University Press.

Schaefer, V., & Darcy, I. (2014). Lexical function of pitch in the first language shapes cross-linguistic perception of Thai tones. *Laboratory Phonology*, *5*(4). https://doi.org/10.1515/lp-2014-0016

Schmidt, L. B. (2018). L2 development of perceptual categorization of dialectal sounds: A study in Spanish. *Studies in Second Language Acquisition*, *40*(4), 857-882.

Schouten, A. (2009). *The Critical Period Hypothesis: Support, Challenge, and Reconceptualization*. Academic Commons. https://doi.org/10.7916/D82R3R7M

Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., & Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends in Cognitive Sciences*, *12*(3), 106–113. https://doi.org/10.1016/j.tics.2008.01.002

Schwartz, J. L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition*, *93*(2), B69–B78. https://doi.org/10.1016/j.cognition.2004.01.006

Second Language Acquisition and the Critical Period Hypothesis. (n.d.). Google Books. https://books.google.co.za/books?hl=en&lr=&id=d7OPAgAAQBAJ&oi=fnd&pg=PA161&dq=+Bialystok,+E.,+%26+Hakuta,+K.+(1999).+Confounded+age:+Linguistic+and+cognitive+factors+in+age+differences+for+second+language+acquisition.+Second+language+acquisition+and+the+critical+period+hypothesis,+161-181.&ots=5xsbxIxqdn&sig=cHuhUhkIJzPBhnoZR4kwAKg84WU&redir_esc=y#v=onepage&q&f=false

Sekiyama, K. (1997). Cultural and linguistic factors in audio-visual speech processing: The McGurk effect in Chinese subjects. *Perception & Psychophysics*, *59*(1), 73–80. https://doi.org/10.3758/bf03206849

Sekiyama, K., & Burnham, D. (2008). Impact of language on development of auditory-visual speech perception. *Developmental Science*, *11*(2), 306–320. https://doi.org/10.1111/j.1467-7687.2008.00677.x

Sekiyama, K., Kanno, I., Miura, S., & Sugita, Y. (2003). Auditory-visual speech perception examined by fMRI and PET. *Neuroscience Research*, *47*(3), 277–287. https://doi.org/10.1016/s0168-0102(03)00214-1

Sekiyama, K., & Tohkura, Y. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *The Journal of the Acoustical Society of America*, *90*(4), 1797–1805. https://doi.org/10.1121/1.401660

Sekiyama, K., & Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics*, *21*(4), 427–444. https://doi.org/10.1016/s0095-4470(19)30229-3

Sennema, A., Hazan, V., & Faulkner, A. (2003). *The role of visual cues in L2 consonant perception*.

Shams, L., Kamitani, Y., & Shimojo, S. (2000). What you see is what you hear. *Nature*, *408*(6814), 788–788. https://doi.org/10.1038/35048669

Shams, L., & Kim, R. (2010). Crossmodal influences on visual perception. *Physics of Life Reviews*, *7*(3), 269–284. https://doi.org/10.1016/j.plrev.2010.04.006

Shimojo, S. (2001). Sensory modalities are not separate modalities: plasticity and interactions. *Current Opinion in Neurobiology*, *11*(4), 505–509. https://doi.org/10.1016/s0959-4388(00)00241-5

Shook, A., & Marian, V. (2012). The Bilingual Language Interaction Network for Comprehension of Speech. *Bilingualism: Language and Cognition*, *16*(2), 304–324. https://doi.org/10.1017/s1366728912000466

Showalter, C. E., & Hayes-Harb, R. (2013). Unfamiliar orthographic information and second language word learning: A novel lexicon study. *Second Language Research*, *29*(2), 185-200.

Showalter, C. E., & Hayes-Harb, R. (2013). Unfamiliar orthographic information and second language word learning: A novel lexicon study. *Second Language Research*, *29*(2), 185–200. https://doi.org/10.1177/0267658313480154

Silva-Corvalán, C. (2014). *Bilingual language acquisition: Spanish and English in the first six years*. *New York State, New York: Cambridge University Press*.

Simonet, M. (2012). *Dialectal Differences in Spanish Voiced Obstruent Allophony: Costa Rican versus Iberian Spanish*. Retrieved March 5, 2024, from https://scholar.google.com/citations?view_op=view_citation&hl=de&user=GnYMTI8AAAAJ&citation_for_view=GnYMTI8AAAAJ:_FxGoFyzp5QC

Skoruppa, K., Pons, F., Christophe, A., Bosch, L., Dupoux, E., Sebastián-Gallés, N., Limissuri, R. A., & Peperkamp, S. (2009). Language-specific stress perception by 9-month-old French and Spanish infants. *Developmental Science*, *12*(6), 914–919. https://doi.org/10.1111/j.1467-7687.2009.00835.x

Smiljanic, R., & Gilbert, R. C. (2017). Acoustics of Clear and Noise-Adapted Speech in Children, Young, and Older Adults. *Journal of Speech, Language, and Hearing Research*, *60*(11), 3081–3096. https://doi.org/10.1044/2017_jslhr-s-16-0130

Smith, D., & Burnham, D. (2012). Faciliation of Mandarin tone perception by visual speech in clear and degraded audio: Implications for cochlear implants. *The Journal of the Acoustical Society of America*, *131*(2), 1480–1489. https://doi.org/10.1121/1.3672703

So, C. K., & Best, C. T. (2010). Cross-language Perception of Non-native Tonal Contrasts: Effects of Native Phonological and Phonetic Influences. *Language and Speech*, *53*(2), 273–293. https://doi.org/10.1177/0023830909357156

So, C. K., & Best, C. T. (2010). Cross-language Perception of Non-native Tonal Contrasts: Effects of Native Phonological and Phonetic Influences. *Language and Speech*, *53*(2), 273–293. https://doi.org/10.1177/0023830909357156

So, C. K., & Best, C. T. (2014). Phonetic influences on English and French listeners'assimilation of Mandarin tones to native prosodic categories. *Studies in Second Language Acquisition*, *36*(2), 195-221. https://doi.org/10.1017/s0272263114000047

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 And Hearing*, *26*(3), 263–275. https://doi.org/10.1097/00003446-200506000-00003

Sowell, E. R., Thompson, P. M., Leonard, C. M., Welcome, S. E., Kan, E., & Toga, A. W. (2004). Longitudinal Mapping of Cortical Thickness and Brain Growth in Normal Children. *The Journal of Neuroscience*, *24*(38), 8223–8231. https://doi.org/10.1523/jneurosci.1798-04.2004

Stacey, J. E., Howard, C. J., Mitra, S., & Stacey, P. C. (2020). Audio-visual integration in noise: Influence of auditory and visual stimulus degradation on eye movements and perception of the McGurk effect. *Attention, Perception, & Psychophysics*, *82*(7), 3544–3557. https://doi.org/10.3758/s13414-020-02042-x

Stacey, P. C., Kitterick, P. T., Morris, S. D., & Sumner, C. J. (2016). The contribution of visual information to the perception of speech in noise with and without informative temporal fine structure. *Hearing Research*, *336*, 17–28. https://doi.org/10.1016/j.heares.2016.04.002

Statistics South Africa | Census Dissemination. (2022). Census.statssa.gov.za. https://census.statssa.gov.za/#/

Stockman, I. J. (2006). Evidence for a minimal competence core of consonant sounds in the speech of African American children: A preliminary study. *Clinical Linguistics & Phonetics*, *20*(10), 723–749. https://doi.org/10.1080/02699200500322803

Strand, J., Cooperman, A., Rowe, J., & Simenstad, A. (2014). Individual Differences in Susceptibility to the McGurk Effect: Links With Lipreading and Detecting Audio-visual Incongruity. *Journal of Speech, Language, and Hearing Research*, *57*(6), 2322–2331. https://doi.org/10.1044/2014_jslhr-h-14-0059

Strange, W. (1995). Learnability Constraints on Theories of Speech Perception: From Spoken Words to Speech Sounds Review of "The Development of Speech Perception: The Transition from Speech Sounds to Spoken Words." *Language and Speech*, *38*(2), 217–222. https://doi.org/10.1177/002383099503800207

Strange, W., & Shafer, V. L. (2008). Speech perception in second language learners. *Phonology and second language acquisition*, 153-191.

Sueyoshi, A., & Hardison, D. M. (2005). The Role of Gestures and Facial Cues in Second Language Listening Comprehension. *Language Learning*, *55*(4), 661–699. https://doi.org/10.1111/j.0023-8333.2005.00320.x

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*, *26*(2), 212–215. https://doi.org/10.1121/1.1907309

Summerfield, Q. (1992). Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *335*(1273), 71–78. https://doi.org/10.1098/rstb.1992.0009

Surguladze, S. A., Calvert, G. A., Brammer, M. J., Campbell, R., Bullmore, E. T., Giampietro, V., & David, A. S. (2001). Audio–visual speech perception in schizophrenia: an fMRI study. *Psychiatry Research: Neuroimaging*, *106*(1), 1–14. https://doi.org/10.1016/s0925-4927(00)00081-0

Taljaard, P. C., & Snyman, J. W. (1991). *An Introduction to Siswati Phonetics*. Google Books. Retrieved March 5, 2024, from https://books.google.co.za/books/about/An_Introduction_to_Siswati_Phonetics.html?id=dd33cQAACAAJ&redir_esc=y

Tanner, D., McLaughlin, J., Herschensohn, J., & Osterhout, L. (2012). Individual differences reveal stages of L2 grammatical acquisition: ERP evidence. *Bilingualism: Language and Cognition*, *16*(2), 367–382. https://doi.org/10.1017/s1366728912000302

Taylor, R. C., & Gitsaki, C. (2004). *Teaching WELL and loving IT*. https://doi.org/10.4324/9781410610775-14

Thompson, L. A., & Malloy, D. (2004). Attention Resources and Visible Speech Encoding in Older and Younger Adults. *Experimental Aging Research*, *30*(3), 241–252. https://doi.org/10.1080/03610730490447877

Thompson, M., & Hazan, V. (2010). *The impact of visual cues and lexical knowledge on the perception of a non-native consonant contrast for Colombian adults- UCL Discovery*. Retrieved March 5, 2024, from https://discovery.ucl.ac.uk/id/eprint/19505/

Thordardottir, E. (2017). Amount trumps timing in bilingual vocabulary acquisition: Effects of input in simultaneous and sequential school-age bilinguals. *International Journal of Bilingualism*, *23*(1), 236–255. https://doi.org/10.1177/1367006917722418

Tiippana, K. (2014). What is the McGurk effect? *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00725

Tiippana, K., Ujiie, Y., Peromaa, T., & Takahashi, K. (2023). Investigation of Cross-Language and Stimulus-Dependent Effects on the McGurk Effect with Finnish and Japanese Speakers and Listeners. *Brain Sciences*, *13*(8), 1198. https://doi.org/10.3390/brainsci13081198

Tomasello, M., & Carpenter, M. (2007). Shared intentionality. *Developmental science*, *10*(1), 121-125.

Tona, R., Naito, Y., Moroto, S., Yamamoto, R., Fujiwara, K., Yamazaki, H., Shinohara, S., & Kikuchi, M. (2015). Audio–visual integration during speech perception in prelingually deafened Japanese children revealed by the McGurk effect. *International Journal of Pediatric Otorhinolaryngology*, *79*(12), 2072–2078. https://doi.org/10.1016/j.ijporl.2015.09.016

Traunmüller, H., & Öhrström, N. (2007). Audio-visual perception of openness and lip rounding in front vowels. *Journal of Phonetics*, *35*(2), 244–258. https://doi.org/10.1016/j.wocn.2006.03.002

Tremblay, C., Champoux, F., Voss, P., Bacon, B. A., Lepore, F., & Théoret, H. (2007). Speech and Non-Speech Audio-Visual Illusions: A Developmental Study. *PLoS ONE*, *2*(8), e742. https://doi.org/10.1371/journal.pone.0000742

Troike, M. S. (2006). *Introducing Second Language Acquisition*. Google Books. Retrieved March 5, 2024, from https://books.google.co.za/books/about/Introducing_Second_Language_Acquisition.html?id=-Bnu98KHv74C&redir_esc=y

Tsitsipis, L. D. (2006). Multilingualism in the English-speaking World. *Journal of Sociolinguistics*, *10*(2), 255–258. https://doi.org/10.1111/j.1360-6441.2006.0327c.x

Tye-Murray, N., Hale, S., Spehar, B., Myerson, J., & Sommers, M. S. (2014). Lipreading in School-Age Children: The Roles of Age, Hearing Status, and Cognitive Ability. *Journal of Speech, Language, and Hearing Research*, *57*(2), 556–565. https://doi.org/10.1044/2013_jslhr-h-12-0273

Tye-Murray, N., Sommers, M. S., & Spehar, B. (2007). Audio-visual Integration and Lipreading Abilities of Older Adults with Normal and Impaired Hearing. *Ear & Hearing*, *28*(5), 656–668. https://doi.org/10.1097/aud.0b013e31812f7185

Tyler, M. D. (2019). *PAM-L2 and phonological category acquisition in the foreign language classroom*. Retrieved March 5, 2024, from https://www.researchgate.net/publication/333124848_PAM-L2_and_phonological_category_acquisition_in_the_foreign_language_classroom

Tyler, M. D., & Cutler, A. (2009). Cross-language differences in cue use for speech segmentation. *The Journal of the Acoustical Society of America*, *126*(1), 367–376. https://doi.org/10.1121/1.3129127

Valkenier, B., Duyne, J. Y., Andringa, T. C., & Başkent, D. (2012). Audio-visual Perception of Congruent and Incongruent Dutch Front Vowels. *Journal of Speech, Language, and Hearing Research*, *55*(6), 1788–1801. https://doi.org/10.1044/1092-4388(2012/11-0227)

Van Der Merwe, A., & Le Roux, M. (2014). *Idiosyncratic sound systems of the South African Bantu languages : research and clinical implications for speech-language pathologists and audiologists*. Retrieved March 6, 2024, from https://repository.up.ac.za/handle/2263/45212

Van Engen, K. J., Dey, A., Sommers, M. S., & Peelle, J. E. (2022). Audio-visual speech perception: Moving beyond McGurk. *The Journal of the Acoustical Society of America*, *152*(6), 3216–3225. https://doi.org/10.1121/10.0015262

Van Engen, K. J., Phelps, J. E. B., Smiljanic, R., & Chandrasekaran, B. (2014). Enhancing Speech Intelligibility: Interactions Among Context, Modality, Speech Style, and Masker. *Journal of Speech, Language, and Hearing Research*, *57*(5), 1908–1918. https://doi.org/10.1044/jslhr-h-13-0076

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

van Wassenhove, V. (2013). Speech through ears and eyes: interfacing the senses with the supramodal brain. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00388

Vatakis, A., & Spence, C. (2006). Audio-visual synchrony perception for music, speech, and object actions. *Brain Research*, *1111*(1), 134–142. https://doi.org/10.1016/j.brainres.2006.05.078

Venezia, J. H., Fillmore, P., Matchin, W., Lisette Isenberg, A., Hickok, G., & Fridriksson, J. (2016). Perception drives production across sensory modalities: A network for sensorimotor integration of visual speech. *NeuroImage*, *126*, 196–207. https://doi.org/10.1016/j.neuroimage.2015.11.038

Vlasenko, B., Prylipko, D., & Wendemuth, A. (2012). *Towards Robust Spontaneous Speech Recognition with Emotional Speech Adapted Acoustic Models*. Retrieved March 5, 2024, from https://www.researchgate.net/publication/259065425_Towards_Robust_Spontaneous_Speech_Recognition_with_Emotional_Speech_Adapted_Acoustic_Models

Vroomen, J., & de Gelder, B. (1995). Metrical segmentation and lexical inhibition in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(1), 98–108. https://doi.org/10.1037//0096-1523.21.1.98

Vroomen, J., & Keetels, M. (2010). Perception of intersensory synchrony: A tutorial review. *Attention, Perception, & Psychophysics*, *72*(4), 871–884. https://doi.org/10.3758/app.72.4.871

Walden, B. E., Montgomery, A. A., Prosek, R. A., & Hawkins, D. B. (1990). Visual Biasing of Normal and Impaired Auditory Speech Perception. *Journal of Speech, Language, and Hearing Research*, *33*(1), 163–173. https://doi.org/10.1044/jshr.3301.163

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*(1), 130–145. https://doi.org/10.1044/jshr.2001.130

Walker, J. T., & Scott, K. J. (1981). Auditory-visual conflicts in the perceived duration of lights, tones, and gaps. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(6), 1327–1339. https://doi.org/10.1037//0096-1523.7.6.1327

Walley, A. C., Bohn, O., & Munro, M. (2007). Speech learning, lexical reorganization and the development of word recognition by native and non-native English speakers. *Festshrift for James E. Flege*.

Wang, T., Potter, C. E., & Saffran, J. R. (2020). Plasticity in Second Language Learning: The Case of Mandarin Tones. *Language Learning and Development*, *16*(3), 231–243. https://doi.org/10.1080/15475441.2020.1737072

Wang, X. (2013). Perception of Mandarin Tones: The Effect of L1 Background and Training. *The Modern Language Journal*, *97*(1), 144–160. https://doi.org/10.1111/j.1540-4781.2013.01386.x

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

Wang, Y. (2022). The Causation of Lexical Fossilization and Its Pedagogical Implications. *Proceedings of the 2022 5th International Conference on Humanities Education and Social Sciences (ICHESS 2022)*, 96–104. https://doi.org/10.2991/978-2-494069-89-3_13

Wang, Y., Behne, D. M., Jongman, A., & Sereno, J. A. (2004). The role of linguistic experience in the hemispheric processing of lexical tone. *Applied Psycholinguistics*, *25*(3), 449–466. https://doi.org/10.1017/s0142716404001213

Wang, Y., Behne, D. M., & Jiang, H. (2008). Linguistic experience and audio-visual perception of non-native fricatives. *The Journal of the Acoustical Society of America*, *124*(3), 1716–1726. https://doi.org/10.1121/1.2956483

Wang, Y., Behne, D. M., & Jiang, H. (2009). Influence of native language phonetic system on audio-visual speech perception. *Journal of Phonetics*, *37*(3), 344–356. https://doi.org/10.1016/j.wocn.2009.04.002

Wang, Y., Jongman, A., & Sereno, J. A. (2001). Dichotic Perception of Mandarin Tones by Chinese and American Listeners. *Brain and Language*, *78*(3), 332–348. https://doi.org/10.1006/brln.2001.2474

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

Watson, T. L., Robbins, R. A., & Best, C. T. (2014). Infant perceptual development for faces and spoken words: An integrated approach. *Developmental Psychobiology*, *56*(7), 1454–1481. https://doi.org/10.1002/dev.21243

Wayland, R. P., & Guion, S. G. (2004). *Training English and Chinese Listeners to Perceive Thai Tones: A Preliminary Report*. Retrieved March 5, 2024, from https://www.researchgate.net/publication/227960170_Training_English_and_Chinese_Listeners_to_Perceive_Thai_Tones_A_Preliminary_Report

Webb, M. Y., Lederberg, A. R., Branum-Martin, L., & McDonald Connor, C. (2015). Evaluating the Structure of Early English Literacy Skills in Deaf and Hard-of-Hearing Children. *Journal of Deaf Studies and Deaf Education*, *20*(4), 343–355. https://doi.org/10.1093/deafed/env024

Wells, B. (2010). *Tonal repetition and tonal contrast in English carer-child interaction*. Studies in Discourse and Grammar. https://doi.org/10.1075/sidag.23.20wel

Werker, J. F., Cohen, L. B., Lloyd, V. L., Casasola, M., & Stager, C. L. (1998). Acquisition of word-object associations by 14-month-old infants. *Developmental Psychology*, *34*(6), 1289–1309. https://doi.org/10.1037//0012-1649.34.6.1289

Werker, J. F., & Logan, J. S. (1985). Cross-language evidence for three factors in speech perception. *Perception & Psychophysics*, *37*(1), 35–44. https://doi.org/10.3758/bf03207136

Westphal, E. (1955). An Introduction to Tswana Grammar. By Desmond T. Cole. Cape Town: Longmans, Green and Co.1955. Pp. xxxv + 473. 35s. *Africa*, *25*(4), 442–444. https://doi.org/10.2307/1156884

Wichmann, A. (2004). *The intonation of Please-requests: a corpus-based study*. Journal of Pragmatics. https://doi.org/10.1016/j.pragma.2004.03.003

Wichmann, A. (2005). Please – from courtesy to appeal: the role of intonation in the expression of attitudinal meaning. *English Language and Linguistics*, *9*(2), 229–253. https://doi.org/10.1017/s1360674305001632

Wong, P. C., Perrachione, T. K., & Parrish, T. B. (2006). Neural characteristics of successful and less successful speech and word learning in adults. *Human Brain Mapping*, *28*(10), 995–1006. https://doi.org/10.1002/hbm.20330

Wu, X., & Lin, H. (2008). *Perception of Mandarin Tones by Mandarin and English Listeners*. ResearchGate. https://www.researchgate.net/publication/253778622_Perception_of_Mandarin_Tones_by_Mandarin_and_English_Listeners

Wu, Y. C., & Coulson, S. (2015). Iconic Gestures Facilitate Discourse Comprehension in Individuals With Superior Immediate Memory for Body Configurations. *Psychological Science*, *26*(11), 1717–1727. https://doi.org/10.1177/0956797615597671

Xu, H. D., Li, Z., Zhou, Q., Li, C., Wang, Z., Cao, Y., ... & Mao, X. L. (2021). Read, listen, and see: Leveraging multimodal information helps Chinese spell checking. *arXiv preprint arXiv:2105.12306*.

Xu, Y., Gandour, J. T., & Francis, A. L. (2006). Effects of language experience and stimulus complexity on the categorical perception of pitch direction. *The Journal of the Acoustical Society of America*, *120*(2), 1063–1074. https://doi.org/10.1121/1.2213572

Yehia, H. C., Kuratate, T., & Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. *Journal of Phonetics*, *30*(3), 555–568. https://doi.org/10.1006/jpho.2002.0165

Yew, B., Alladi, S., Shailaja, M., Hodges, J. R., & Hornberger, M. (2012). Lost and Forgotten? Orientation Versus Memory in Alzheimer's Disease and Frontotemporal Dementia. *Journal of Alzheimer's Disease*, *33*(2), 473–481. https://doi.org/10.3233/jad-2012-120769

Ylinen, S., Uther, M., Latvala, A., Vepsäläinen, S., Iverson, P., Akahane-Yamada, R., & Näätänen, R. (2010). Training the Brain to Weight Speech Cues Differently: A Study of Finnish Second-language Users of English. *Journal of Cognitive Neuroscience*, *22*(6), 1319–1332. https://doi.org/10.1162/jocn.2009.21272

Zhang, C., Yang, Z., He, X., & Deng, L. (2020). Multimodal Intelligence: Representation Learning, Information Fusion, and Applications. *IEEE Journal of Selected Topics in Signal Processing*, *14*(3), 478–493. https://doi.org/10.1109/jstsp.2020.2987728

Zhang, F., Chen, Y., Li, Z., Hong, Z., Liu, J., Ma, F., ... & Ding, E. (2019). Acfnet: Attentional class feature network for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6798-6807).

Zhang, S., Zhang, S., Huang, T., Gao, W., & Tian, Q. (2018). Learning Affective Features With a Hybrid Deep Model for Audio–Visual Emotion Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *28*(10), 3030–3043. https://doi.org/10.1109/tcsvt.2017.2719043

Zhao, H. (2010). Investigating learners' use and understanding of peer and teacher feedback on writing: A comparative study in a Chinese English writing classroom. *Assessing Writing*, *15*(1), 3–17. https://doi.org/10.1016/j.asw.2010.01.002

Zhyrun, I. (2016). Culture through comparison: creating audio-visual listening materials for a CLIL course. *Latin American Journal of Content & Language Integrated Learning*, *9*(2), 345–373. https://doi.org/10.5294/laclil.2016.9.2.5

Zhyrun, I. (2017). Videos, culture and significant learning: Colombian students speak. In *Conference proceedings of» eLearning and Software for Education «(eLSE)* (Vol. 13, No. 02, pp. 632-639). Carol I National Defence University Publishing House.

Zook, K., & Pearce, J. H. (2018). *Quantitative Descriptive Analysis*. Routledge eBooks. https://doi.org/10.1201/9781315137681-3

Zwicky, A. M., (2017). *Phonemes and features*. https://scholarworks.iu.edu/journals/index.php/iulcwp/article/view/26006

Zwicky, A. M., & Pullum, G. K. (1983). Deleting named morphemes. *Lingua*, *59*(2–3), 155–175. https://doi.org/10.1016/0024-3841(83)90061-x

Appendices

Appendix 1- Participation information sheet

University of the Witwatersrand

School of Literature, Language and Media

Good Day,

My name is Ammaara Seedat and I am a Masters student in Linguistics at the University of the Witwatersrand, in Johannesburg. My supervisor for this study is Dr Ramona Kunene Nicolas. I am conducting a research study about audio-visual speech perception. The study is titled " **Audio-visual Speech Perception amongst bilingual speakers".**

I invite you to participate in a session where you will be asked to view short video clips followed by a question-and-answer session. If you decide to take part, your participation in this research study will last about 30 minutes. This activity will occur at the Linguistics Postgraduate lab at SH 3101 in the presence of myself and a translator.

With your permission, I would like to audio and video record this session. An eye-tracking software will also be used to track eye-gaze movements during the study whereby pictures of participants will also be necessary. The data will be stored on a password-protected drive and computer and may however be used as secondary data for future studies. Only the researcher and translator will have access to the video and audio recordings. Photos acquired during the session may however be used in the research and it cannot be guaranteed that the individuals face will be blurred during analysis of the video.

During this research study, I will need some personal information about you, including what languages you speak and the timeline involved in the acquisition of that particular language.

The results of this study will not include your personal information. Your confidentiality will be upheld as your personal response during the interview will not be disclosed to anyone outside the research team. Please be aware that your face may be disclosed during analyses of eye-tracking movements however your name will remain confidential. With your permission, other researchers may use the data collected from this study, but your name and other personal information will not be passed on.

A decision to participate in this study should be because you want to volunteer. You are not obligated to answer any questions if you do not wish to do so. There will be no direct benefit

if you wish to participate in this study. Participating in this research will not cost anything. You will not be paid for participating in this study.

The study can be stopped at any time if the participant does not feel comfortable.

This study will be written as a research report and will be available on the university library website.

If you have any questions during or before the study please feel free to contact me or my supervisor on the details captured below. If you have any concerns or complaints about the ethical procedures of this research study, you are welcome to contact the University Human Research Ethics Committee (Non-Medical), by telephone +27(0) 11 717 1408, email hrecnon- medical@wits.ac.za.

If at any time you may experience some distress due to the research activity please contact CCDU BRAAMFONTEIN CAMPUS WEST [Main Office] located at the CCDU Building, Braamfontein Campus West, Wits University. Email: infor.ccdu@wits.ac.za
Tel: 011719 9140/32

Yours sincerely, Ammaara Seedat

Researcher:
Ammaara Seedat
Email: 1848552@students.wits.ac.za Contact number: 073 594 0646

Supervisor:
Dr Ramona Kunene Nicolas
Email: Ramona.KuneneNicolas@wits.ac.za Contact number: 011 717 4183

Appendix 2-Consent Form



Audio-visual speech perception amongst bilinguals

Name of researcher: Ammaara Seedat

I, _____ agree to participate in this research project.

I agree to the following:

(Please circle the relevant options below)

The research study was explained to me. I understand what this study is about.	YES	NO
I understand that I can volunteer to take part in the study.	YES	NO
I agree that the interview will be audio recorded.	YES	NO
I agree that the interview will be video recorded	YES	NO
I agree to the use of eye-tracking software whereby photos will be taken to study eye-gaze patterns.	YES	NO
I agree and am aware that my picture may be used for eye-tracking analyses wherein the picture may be unblurred.	YES	NO

I agree that direct quotations and pictures of eye-gaze patterns from my interview may be used by the researcher in their research report.	YES	NO
I agree that my participation will remain anonymous (my name or other identifying data will not be used by the researcher in their research report)	YES	NO
I agree that other researchers may use the information I provide in my interview/focus group/other activity (depending on their own ethics clearance being obtained) but my name and any personal information will not be used or passed on.	YES	NO
I understand that I will not be paid for participating in this research and that I have the right to withdraw from this project at any stage	YES	NO
I agree to answering a questionnaire which will be used for analysis in the study	YES	NO

…… (signature)
…… (name of participant)
…… (date)

…… (signature)
…… (name of researcher/person seeking consent)
…… (date)

Appendix 3- Ethics approval

UNIVERSITY OF THE WITWATERSRAND JOHANNESBURG

Research Office

HUMAN RESEARCH ETHICS COMMITTEE (NON-MEDICAL)
R14/49 Seedat

CLEARANCE CERTIFICATE	**PROTOCOL NUMBER: H23/06/40**
PROJECT TITLE	Audio-visual speech perception amongst bilingual speakers
INVESTIGATOR(S)	Ms A Seedat
SCHOOL/DEPARTMENT	School of Literature, Language and Media/
DATE CONSIDERED	23 June 2023
DECISION OF THE COMMITTEE	Approved Risk Level: Minimal
EXPIRY DATE	04 September 2026

DATE 05 September 2023 **CHAIRPERSON** _____
(Professor J Watermeyer)

cc: Supervisor : Dr R Kunene Nicolas

DECLARATION OF INVESTIGATOR(S)

To be completed in duplicate and **A SIGNED COPY** returned to the Secretary electronically. Unreported changes to the application may invalidate the clearance given by the HREC (Non-Medical)

I/We fully understand the conditions under which I am/we are authorized to carry out the abovementioned research and I/we guarantee to ensure compliance with these conditions. Should any departure be contemplated from the research procedure as approved I/we undertake to submit an amendment of the protocol to the Committee. **I/we agree to completion of a regular progress report. For Minimal and Low Risk studies, this is due annually on 31 December. For Medium and High Risk studies, this is due twice annually on 30 June and 31 December.**

_____ 06 / 09 / 2023
Signature **Date**

PLEASE QUOTE THE PROTOCOL NUMBER ON ALL ENQUIRIES

Appendix 4- Audio-only condition coding

Pseudonym	Age	L1	MOE		A-TP 1			A-TP 2			A-MP 1		A-MP 2	
					ID OF LHT	ID OF HLT	ID OF DIFF	ID OF HLT	ID OF LHT	ID OF DIFF	MP 1 ID WD 1	MP 1 ID WD 2	MP 2 ID WD 1	MP 2 ID WD 2
P 1	18	English	8		2	2	1	2	2	1	2	2	2	2
P 2	19	English	8		2	2	1	2	2	1	2	2	1	2
P 3	19	English	8		2	2	1	2	2	1	2	2	1	1
P 4	19	English	8		2	2	2	2	2	2	2	2	2	2
P 5	19	English	9		2	2	2	2	2	2	2	2	2	2
P 6	20	English	9		2	2	2	2	2	2	2	2	2	2
P 7	17	English/ Sesotho	9		2	2	2	2	2	2	2	2	2	1
P 8	18	English	9		2	2	2	2	2	1	1	2	2	2
P 9	18	English/ Sepedi	10		2	2	2	2	2	2	2	2	2	2
P 10	25	English	12		2	2	2	2	2	1	2	2	2	2
P 11	22	English	12		2	2	2	1	1	2	2	2	2	2
P 12	21	English	36		2	2	2	2	2	2	1	2	1	2
				n= total no. correct	12	12	9	11	11	7	10	12	9	10
				$\bar{x}$	2,00	2,00	1,75	1,92	1,92	1,58	1,83	2,00	1,75	1,83
				%	100,0	100,0	87,50	95.83	95.83	79.16	91,67	100,0	87,50	91.67

Appendix 5-Visual-only condition coding

Pseudonym	Age	L1	MOE		V-WP		V-MP	
					ID OF WD 1 [isithelo]	ID OF WD 2 [isithini]	MP ID W1 [ukuzala]	MP ID WD 2 [ukusala]
P 1	18	English	8		2	1	2	1
P 2	19	English	8		2	1	2	1
P 3	19	English	8		2	2	1	2
P 4	19	English	8		2	2	1	1
P 5	19	English	8		1	2	1	1
P 6	20	English	9		2	1	2	2
P 7	17	English/ Sesotho	9		1	2	1	2
P 8	18	English	9		1	1	1	1
P 9	18	English/Sepedi	10		2	1	2	2
P 10	25	English	12		2	1	1	1
P 11	22	English	12		1	1	1	2
P 12	21	English	36		2	2	1	1
				n= total no. correct	8	5	4	5
				$\bar{x}$	1,67	1,42	1,33	1,42
				%	83,33%	70,83	66,67	70,83

Appendix 6- Audio-visual only condition coding

Pseudonym	Age	L1	MOE		AVTP 1			AVTP 2			AVMP 3		AVMP 4	
					AV 1.1: Tonal LHLL	AV 1.2 Tonal HLHL	ID OF DIFF	AV 2.1: Tonal HLHL	AV 2.2: Tonal LHLL	ID OF DIFF	AV 3.1: MP	AV 3.2 MP	AV 4.1 MP	AV4.2 MP
P 1	18	English	8		2	2	1	2	2	1	2	2	2	2
P 2	19	English	8		1	2	2	2	2	2	2	2	2	2
P 3	19	English	8		2	2	2	2	2	1	2	2	2	2
P 4	19	English	8		2	2	2	2	2	2	2	2	2	2
P 5	19	English	8		2	2	2	2	2	2	2	2	2	2
P 6	20	English	9		2	2	2	2	2	2	2	2	2	2
P 7	17	Sesotho	9		2	2	2	2	2	2	2	2	2	2
P 8	18	English	9		2	2	2	2	2	2	2	2	2	2
P 9	18	Sepedi/ English	10		2	2	2	2	1	1	2	2	2	2
P 10	25	English	12		2	2	2	2	2	2	2	2	2	2
P 11	22	English	12		2	2	1	2	2	1	2	2	2	2
P 12	21	English	36		2	2	2	2	2	2	2	2	2	2
				n= total no. correct	11	12	10	12	11	8	12	12	12	12
				$\bar{x}$	1,92	2,00	1,83	2,00	1,92	1,67	2,00	2,00	2,00	2,00
				%	95,83	100	91,66	100	95,83	83,33	100	100	100	100

Appendix 7- Incongruent condition coding

Pseudonym	Age	L1	MOE	*Code: [1-Visual][2-Vocal] [3-Incongruent]*		INCONGRUENT STIMULI 1	INCONGRUENT STIMULI 2	INCONGRUENT STIMULI 3
P 1	18	English	8			3	1	2
P 2	19	English	8			3	3	2
P 3	19	English	8			3	3	2
P 4	19	English	8			2	3	2
P 5	19	English	8			2	3	2
P 6	20	English	9			3	3	2
P 7	17	Sesotho	9			3	1	2
P 8	18	English	9			3	3	2
P 9	18	Sepedi/English	10			2	3	2
P 10	25	English	12			3	3	2
P 11	22	English	12			3	3	2
P 12	21	English	36			3	3	2
						INCONGRUENT STIMULI 1	INCONGRUENT STIMULI 2	INCONGRUENT STIMULI 3
					% Incongruent	75,0	83,30	0
					% Vocal	25,0	0	100,0
					% Visual	0	16,70	0