DEVELOPMENT OF AUDITORY-VISUAL SPEECH PERCEPTION IN YOUNG CHILDREN

VAHİT DOĞU ERDENER
B.Sc. METU, M.A. (Hons.) UWS

MARCS AUDITORY LABORATORIES & SCHOOL OF PSYCHOLOGY
UNIVERSITY OF WESTERN SYDNEY

SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS OF THE DEGREE OF DOCTOR OF PHILOSOPHY IN PSYCHOLOGY

MARCH 2007
To
Nezihe & Vecih Kemikoğlu,
whose linguistic influence and love were beyond their life times
Abstract

Unlike auditory-only speech perception, little is known about the development of auditory-visual speech perception. Recent studies show that pre-linguistic infants perceive auditory-visual speech phonetically in the absence of any phonological experience. In addition, while an increase in visual speech influence over age is observed in English speakers, particularly between six and eight years, this is not the case in Japanese speakers. This thesis aims to investigate the factors that lead to an increase in visual speech influence in English speaking children aged between 3 and 8 years. The general hypothesis of this thesis is that age-related, language-specific factors will be related to auditory-visual speech perception.

Three experiments were conducted here. In Experiment 1 children of four ages, 5, 6, 7, and 8 years, and adults were given tests of auditory-visual speech perception, language specific speech perception (the relative degree of the native speech perception compared with non-native speech perception), articulation, and reading. In Experiment 2, 3- and 4-year-old preschool children were tested for their auditory-visual, language specific speech perception, vocabulary knowledge and executive functions. In Experiment 3, children with and without phonological speech disorder were tested on auditory-visual speech perception, and executive function abilities.

Results show that in linguistically challenging periods, such as school onset and reading acquisition, there is a strong link between auditory-visual and language specific speech perception, and that this link appears to help cope with new linguistic challenges. However this link does not seem to be present in adults or preschool children, for whom auditory-
visual speech perception is predictable from auditory speech perception ability alone. Children with and without speech disorder did not differ from each other on general auditory-visual speech perception ability, but speech-disordered children’s lipreading performance was poorer than that for normal speech children. In addition, unlike children with normal speech, auditory-visual speech perception did not improve over age in speech disordered children suggesting some relationship between speech disorder and relatively poor auditory-visual speech perception.

In general the results of this thesis, combined with earlier findings, suggest that auditory-visual speech perception becomes more aligned with other linguistic skills, specifically language specific speech perception with age. Implications of these results in relation to existing models of auditory-visual speech perception and directions for future studies are discussed.
This work has not previously been submitted for a degree or diploma in any university. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

__________________________

Vahit Doğu Erdener
30 December 2006
TABLE OF CONTENTS

ABSTRACT ............................................................................................................................. IV

TABLE OF CONTENTS ......................................................................................................... VII

LIST OF FIGURES ............................................................................................................... XVIII

LIST OF TABLES .................................................................................................................. XXII

LIST OF FORMULAS ........................................................................................................... XXV

ACKNOWLEDGEMENTS ...................................................................................................... XXVI

CHAPTER 1  INTRODUCTION .............................................................................................. 1

1.1  I COULD SEE YOUR LIPS MOVE .............................................................................. 2

CHAPTER 2  ARTICULATION AND SPEECH PRODUCTION ............................................. 5

2.1  ANATOMY OF ARTICULATION ................................................................................. 6

2.1.1  Larynx & Vocal Folds ............................................................................................. 7

2.1.2  Velum ....................................................................................................................... 8

2.1.3  Tongue .................................................................................................................... 8

2.1.4  Teeth ....................................................................................................................... 9

2.1.5  Mandible ............................................................................................................... 10

2.1.6  Lips ....................................................................................................................... 12

2.2  ELEMENTS OF SPEECH AND LANGUAGE ............................................................... 12

2.3  SPEECH PRODUCTION .......................................................................................... 13

2.3.1  Laryngeal Variation and the Role of Vocal Folds in Speech Production ........ 13

2.3.2  Supralaryngeal Variation ...................................................................................... 15

2.4  VOWELS, CONSONANTS AND TONES AS SPEECH SEGMENTS AND THEIR PRODUCTION .................................................................................................................. 15

2.4.1  Periodic and Aperiodic Speech Sounds ................................................................. 15

2.4.2  Vowels as Speech Segments and Their Production ........................................... 16

2.4.2.1  Diphthong Production ..................................................................................... 18

2.4.3  Aperiodic Speech Sounds .................................................................................. 18

2.4.4  Consonants as Speech Segments and Their Production .................................. 18

2.4.4.1  Voicing ......................................................................................................... 19
2.4.4.2 Place of Articulation ................................................................. 19
2.4.4.3 Manner of Articulation ............................................................... 20
2.4.5 Tones as Speech Segments and their Production ........................... 21
2.4.6 Prosodic Features of Speech Signal .............................................. 22
2.4.7 Stress ............................................................................................ 22
2.4.8 Rhythm ......................................................................................... 22
2.4.9 Intonation ..................................................................................... 23
2.5 ARTICULATION AND ITS CONSEQUENCES ................................... 23

CHAPTER 3 AUDITORY SPEECH PERCEPTION AND DEVELOPMENT ... 25
3.1 ARTICULATION AS AN AUDITORY SPEECH EVENT ...................... 26
3.2 THE PROBLEM OF SPEECH PERCEPTION ..................................... 26
  3.2.1 Psychological Reality of Levels of Speech Processing .................. 26
  3.2.2 Acoustic Cues, Co-articulation, and the Invariance Problem ......... 27
  3.2.3 Talker Variability ........................................................................ 28
  3.2.4 Segmentation ............................................................................. 29
  3.2.5 Categorical Speech Perception .................................................. 29
3.3 MODELS OF AUDITORY SPEECH PERCEPTION .......................... 30
  3.3.1 Motor Theory of Speech Perception .......................................... 31
  3.3.2 Perceptual Assimilation Model ................................................. 32
  3.3.3 Native Language Magnet Theory .............................................. 35
  3.3.4 PRIMIR ..................................................................................... 37
3.4 CROSS-LANGUAGE AND DEVELOPMENTAL STUDIES OF AUDITORY SPEECH PERCEPTION ................................................. 38
  3.4.1 Methods of Cross-Language and Developmental Study ............... 38
    3.4.1.1 The Ontogenetic Method .................................................... 39
    3.4.1.2 The Differential Method .................................................... 39
  3.4.2 Infancy: The Beginnings of Speech Perception ............................ 39
  3.4.3 Perception of Speech Segments over Age .................................... 41
    3.4.3.1 Perception of Consonants ............................................... 41
    3.4.3.2 Perception of Vowels ....................................................... 42
    3.4.3.3 Perception of Tones ........................................................ 43
3.5 A HEURISTIC OF SPEECH PERCEPTION DEVELOPMENT ............ 44

viii
3.5.1 The Psychoacoustic/Phonetic Stage (0-6 months)................................. 44
3.5.2 The Phonemic Stage (6-12 Months) ................................................... 45
3.5.3 The Semantic Stage (12-20 Months) ................................................... 46
3.5.4 Orthographic Stage (6-8 Years) .......................................................... 46
3.6 SUMMARY ................................................................................................. 48

CHAPTER 4 AUDITORY-VISUAL SPEECH PERCEPTION AND
DEVELOPMENT ................................................................................................ 50

4.1 Articulation as a Visual Speech Event ..................................................... 51
4.2 The Psychological Reality of ViseMES ..................................................... 51
4.3 Multimodal Speech Perception: Auditory-Tactile and Auditory-
Orthographic Speech Processing .................................................................. 53
4.4 Auditory-Visual Speech Processing ........................................................ 55
4.5 Major Models of Auditory-Visual Speech Perception ............................ 60
  4.5.1 Fuzzy Logical Model of Perception .................................................... 60
  4.5.2 Phonetic plus Post-Categorical (3PC) Model of Auditory-Visual Speech
      Perception .................................................................................................. 62
4.6 Other Issues in Auditory-Visual Speech Perception .............................. 63
  4.6.1 Auditory-Visual Speech Perception: Static vs. Dynamic .................. 63
  4.6.2 Early versus Late Integration of Auditory and Visual Speech Information  65
4.7 Ontogenetic Development of Auditory-Visual Speech Perception .......... 65
  4.7.1 Auditory-Visual Speech Perception in Infancy .................................... 66
  4.7.2 Auditory-Visual Speech Perception in Childhood ............................. 69
  4.7.3 Auditory-Visual Speech Perception in Adulthood to Late Adulthood .... 70
  4.7.4 Summary: Ontogenetic Studies of Auditory-Visual Speech Perception ... 72
4.8 Differential Development: Cross-Language Studies of Auditory-
Visual Speech Perception ............................................................................... 73
  4.8.1 Differences Between Languages ......................................................... 73
  4.8.2 Foreign Speaker Effects ...................................................................... 76
  4.8.3 The Benefit of Auditory-Visual Speech Information in Second and Non-
      native Language Contexts ...................................................................... 76
  4.8.4 Summary: Differential Studies of Auditory-Visual Speech Perception ... 77
CHAPTER 5  LANGUAGE SPECIFIC FACTORS IN THE DEVELOPMENT OF AUDITORY-VISUAL SPEECH PERCEPTION: SCHOOL CHILDREN & ADULTS  81

5.1 EXPERIMENT 1: LANGUAGE SPECIFIC FACTORS IN THE DEVELOPMENT OF AUDITORY-VISUAL SPEECH PERCEPTION IN SCHOOL CHILDREN AGED FIVE TO EIGHT AND ADULTS ........................................................................................................... 82

5.2 METHOD ........................................................................................................ 83
  5.2.1 Design ....................................................................................................... 83
  5.2.1.1 Auditory-Visual Speech Perception ....................................................... 83
  5.2.1.2 Language-Specific Speech Perception ................................................... 84
  5.2.1.3 Reading ................................................................................................ 85
  5.2.1.4 Articulation .......................................................................................... 85
  5.2.2 Recruitment of School Children ............................................................... 85
  5.2.3 Child Participants ..................................................................................... 86
  5.2.4 Adult Participants .................................................................................... 86

5.3 STIMULI MATERIALS, APPARATUS, AND PROCEDURE ...................... 87
  5.3.1 Auditory-Visual Speech Perception (McGurk) Test Stimuli and Apparatus 87
    5.3.1.1 Stimuli Material .................................................................................. 87
    5.3.1.2 Apparatus .......................................................................................... 88
    5.3.1.3 Procedure .......................................................................................... 88
  5.3.2 Speech Perception Test Stimuli, Apparatus and Procedure .................. 90
    5.3.2.1 Stimuli Material .................................................................................. 90
    5.3.2.2 Apparatus .......................................................................................... 92
    5.3.2.3 Procedure .......................................................................................... 92
  5.3.3 Articulation Test Stimuli, Apparatus and Procedure ......................... 93
    5.3.3.1 Stimuli Material .................................................................................. 93
    5.3.3.2 Apparatus .......................................................................................... 93
    5.3.3.3 Procedure .......................................................................................... 94
  5.3.4 Reading Test Stimuli material, Apparatus and Procedure ................. 95
5.3.4.1 Stimuli and Apparatus ................................................................. 95
5.3.4.2 Procedure ............................................................................ 95
5.4 GENERAL PROCEDURE .................................................................. 96
5.5 RESULTS ...................................................................................... 96
5.5.1 Auditory-Visual Speech Perception Test .................................... 96
5.5.2 Language Specific Speech Perception Test ............................... 100
5.5.3 Reading Test ........................................................................... 101
5.5.4 Articulation Test ..................................................................... 102
5.5.5 Regression Analyses ................................................................. 103
5.5.5.1 School Children: VSI Scores as Dependent Variable .......... 103
5.5.5.2 School Children: VO (Lipreading) Scores as Dependent Variable 105
5.5.5.3 Adults: VSI Scores as Dependent Variable .......................... 106
5.5.5.4 Adults: VO (Lipreading) as Dependent Variable ................. 107
5.6 DISCUSSION .................................................................................. 108

CHAPTER 6 AUDITORY-VISUAL SPEECH PERCEPTION IN PRESCHOOL CHILDREN 112

6.1 BACKGROUND .............................................................................. 113
6.2 LANGUAGE AND COGNITIVE DEVELOPMENT IN PRESCHOOL CHILDREN ..... 113
6.2.1 Language Development from Infancy to Childhood ............... 113
6.2.1.1 Vocabulary Development from Infancy to Childhood ....... 115
6.2.2 Cognitive Development and Executive Functions in Preschool Years ..... 117
6.2.3 The Current Experiment: Aims, Rationale, and Predictions ...... 118
6.3 EXPERIMENT 2: DETERMINANTS AND DEVELOPMENT OF AUDITORY-VISUAL SPEECH PERCEPTION IN 3- AND 4-YEAR-OLDS ......................................................... 119
6.4 METHOD ...................................................................................... 119
6.4.1 Design .................................................................................. 119
6.4.2 Participants ........................................................................... 119
6.5 DEPENDENT VARIABLES, STIMULI, APPARATUS AND PROCEDURE ........ 120
6.5.1 Auditory-Visual Speech Perception Test .................................. 121
6.5.1.1 Dependent Variable and Stimuli ...................................... 121
6.5.1.2 Apparatus and Experimental Setup ................................. 122
6.5.1.3 Procedure ....................................................................... 123
7.5 A NOTE ON LANGUAGE SPECIFIC SPEECH PERCEPTION DATA AND ITS ANALYSIS .................................................................................................................. 150
7.6 EXPERIMENT 3: AUDITORY-VISUAL SPEECH PERCEPTION IN CHILDREN WITH INCONSISTENT PHONOLOGICAL SPEECH DISORDER .......................................................... 150
7.7 METHOD .............................................................................................................. 150
7.7.1 Participants ........................................................................................................ 150
7.7.2 Stimuli, Apparatus and Procedure .................................................................... 151
7.8 RESULTS: CHILDREN WITH VERSUS WITHOUT SPEECH DISORDER .......... 151
7.8.1 Auditory-Visual Speech Perception Test ............................................................. 151
7.8.2 Executive Functions Test: Flexible Item Selection Task ................................. 152
7.8.3 Regression Analysis: VSI-AX as dependent variable ....................................... 153
7.8.4 Regression Analysis: VO (lipreading) as dependent variable ....................... 154
7.9 RESULTS: CHILDREN WITH SPEECH DISORDER ............................................. 155
7.9.1 Auditory-Visual Speech Perception Test ............................................................. 155
7.9.2 Executive Functions Test: Flexible Item Selection Task ................................. 156
7.9.3 Regression Analyses: VSI-AX as Dependent Variable .................................... 157
7.9.4 Regression Analyses: VO (lipreading) as Dependent Variable ....................... 158
7.10 DISCUSSION ....................................................................................................... 159

CHAPTER 8 GENERAL DISCUSSION AND CONCLUSION .................................. 161

8.1 SUMMARY AND INTERPRETATION OF THE FINDINGS: LANGUAGE SPECIFIC AND COGNITIVE FACTORS ...................................................................................... 163
8.1.1 Experiment1: School Children and Adults ....................................................... 163
8.1.2 Experiment 2: Preschool Children .................................................................. 168
8.1.3 Experiment 3: Preschool Children with Phonological Disorder ................. 170
8.2 CURRENT RESULTS, TWO MODELS OF SPEECH PERCEPTION, AND A SUGGESTION OF A FUTURE STUDY ....................................................................................... 171
8.2.1 Perceptual Assimilation Model ........................................................................ 171
8.2.2 Motor Theory of Speech Perception ............................................................... 171
8.2.3 A Future Test of Articulatory Models of Auditory and Auditory-Visual Speech Perception ................................................................................................................. 172
8.3 IMPLICATIONS OF THE CURRENT FINDINGS AND RESEARCH IN AUDITORY- VISUAL SPEECH PERCEPTION .............................................................................. 173
8.3.1 Implications for Teaching Literacy and Foreign Language .................. 173
8.3.2 Implications and Future Research Directions for Speech Therapy for Language-Impaired Children and Adults ...................................................... 174
8.4 Further Questions, Future Research and Conclusion .......................... 175

CHAPTER 9 REFERENCE LIST ......................................................................... 177

APPENDIX A1 QUEENSLAND ARTICULATION TEST ITEM LIST ............... 200

APPENDIX A2 COUNTERBALANCING ORDERS IN EXPERIMENTS 1, 2 AND 3 ............................................................................................................. 202

APPENDIX A3 THE CONVERSION OF LANGUAGE SPECIFIC SPEECH PERCEPTION DATA USING MATLAB® .................................................. 213

A3.1 Conversion of Language Specific Speech Perception Test Data Using MATLAB® ................................................................................................. 214
A3.2 The MATLAB® Script Used for Speech Perception Test Data Conversion .............................................................................................................. 216
CD-ROM Volume Table Of Contents

Appendix CD1 Participant Recruitment Documents……CD-ROM
CD1.1 Experiment 1 Parental Consent Form
CD1.2 Experiment 1 Information Sheet for Parents
CD1.3 Experiment 1 Information Letter for Principals
CD1.4 Experiment 3 Information Sheet for Parent
CD1.5 Sample Young Scientist Degree

Appendix CD2 DMDX Item Files and Queensland Articulation Test Items..........................CD-ROM
CD2.1a DMDX Code for Speech Perception Test (Version 1)
CD2.1b DMDX Code for Speech Perception Test (Version 2)
CD2.2 Queensland Articulation Test (QAT) Photo Items

Appendix CD3. Experiment 1: School Children & Adults..................................................CD-ROM
CD3.1 School Children and Adult Data File
CD3.2 Auditory-Visual Speech Perception Test Analysis of Variance Output Files
   CD3.2.1 The Auditory-Visual (AV) condition ANOVA results
   CD3.2.2 The Visual-Only (VO) condition ANOVA results
   CD3.2.3 The Auditory-Only (AO) condition ANOVA results
CD3.3 Speech Perception Test Analysis of Variance Output File
CD3.4 Reading Test Analysis of Variance Output File
CD3.5a School Children Articulation Test Error Matrices
CD3.5b Articulation Test Analysis of Variance Output File
CD3.6 School Children Correlation and Regression Analyses
CD3.6.1 SCHOOL CHILDREN GROUP CORRELATION ANALYSES
CD3.6.2 SCHOOL CHILDREN GROUP REGRESSION ANALYSES
CD3.7 ADULTS CORRELATION AND REGRESSION ANALYSES
   CD 3.7.1 ADULT GROUP CORRELATION ANALYSES
   CD 3.7.2 ADULT GROUP REGRESSION ANALYSES

APPENDIX CD4. EXPERIMENT 2: PRESCHOOL CHILDREN......... CD-ROM
CD4.1 LIST OF AUDITORY-VISUAL SPEECH CONTRASTS
CD4.2 McGURK AX TASK RESPONSE FORM
CD4.3 FIST INSTRUCTIONS SHEET
CD4.4 FIST RESPONSE SHEET
CD4.5 PRESCHOOL CHILDREN OVERALL DATA
CD4.6 AUDITORY-VISUAL SPEECH PERCEPTION TEST t-TEST ANALYSIS OUTPUT FILE
CD4.7 SPEECH PERCEPTION TEST ANOVA OUTPUT FILE
CD4.8 EXECUTIVE FUNCTION TEST (FIST) ANOVA OUTPUT FILE
CD4.9 RECEPTIVE VOCABULARY TEST t-TEST ANALYSES OUTPUT FILE
CD4.10 FIST FACTOR ANALYSIS
CD4.11 CORRELATION AND REGRESSION ANALYSES

APPENDIX CD5. EXPERIMENT 3: CHILDREN WITH SPEECH DISORDER.................................................CD-ROM
CD5.1 CHILDREN WITH SPEECH DISORDER OVERALL DATA FILE
CD5.2 ANOVA OUTPUT: AGE EFFECTS IN CHILDREN WITH SPEECH DISORDER
CD5.3 AUDITORY-VISUAL SPEECH PERCEPTION TEST ANALYSES
   CD5.3.1 t-TEST OUTPUT FILE: SPEECH DISORDER CHILDREN VS. NO-SPEECH DISORDER CHILDREN
   CD5.3.2 ANOVA OUTPUT FILE: SPEECH DISORDER CHILDREN
CD5.4 Executive Function Test (FIST) ANOVA Output File: Speech Disorder Children vs. No-Speech Disorder Children
CD5.5 Executive Function Test (FIST) ANOVA Output File: Speech Disorder Children by Age
CD5.6 FIST Factor Analysis: Disorder and No-Disorder Children
CD5.7 Regression Analyses: Speech Disorder and No-Speech Disorder Children
CD5.8 FIST Factor Analysis: Speech Disorder Children
CD5.9 Correlation and Regression Analyses: Speech Disorder Children

APPENDIX CD6: Conference Papers... .................................. CD-ROM

CD6.1b Eleventh Australasian Conference On Speech Science & Technology 2006 (Poster)
CD6.2a Seventh International Seminar On Speech Production 2006
CD6.2b Seventh International Seminar On Speech Production 2006 (Poster)

APPENDIX CD7: Bonus Material........................................ CD-ROM

CD7.1 Doğu Erdener’s MARCS Web Page

xvii
LIST OF FIGURES

FIGURE 1-1 THE COMPUTER HAL9000 LIPREADS ASTRONAUTS FRANK POOLE AND DAVE BOWMAN IN THE MOVIE 2001: A SPACE ODYSSEY........3

FIGURE 2-1 HUMAN SPEECH PRODUCTION SYSTEM AND ITS COMPONENTS.
FIGURE IS FROM HARDCASTLE (1976)..................................................7


FIGURE 2-3 THE IPA CLASSIFICATION OF VOWELS. THE ARRANGEMENT OF THE VOWELS ON THE VOWEL QUADRANT IS BASED ON QUALITATIVE REPRESENTATION OF THE TONGUE POSITION IN THE ORAL CAVITY. 17

FIGURE 2-4 THE CLASSIFICATION OF CONSONANTS ALONG VOICING, MANNER AND PLACE OF ARTICULATION DIMENSIONS. .......................19

FIGURE 3-1. DIFFERENCES BETWEEN ACTUAL (A) AND PERCEIVED Distances (B) AMONGST THE STIMULI USED BY IVerson AND KUHL (1994; 1996) WITH AMERICAN SUBJECTS. WITH PERCEPTUAL Distances AROUND THEM REDUCED, THE PROTOypical /R/ AND /L/ ACT AS PERCEPTUAL MAGNETS PULLING THE INSTANCES PERCEPTUALLY CLOSER TO THEM. THE LARGER THE CIRCLES ARE IN (A) THE BETTER THEY REPRESENT THE CATEGORY PROTOTYPE.
BLACK DOTS REPRESENT /R/ AND WHITE DOTS REPRESENT /L/. FIGURE ADAPTED FROM IVerson ET AL. (2003). ......................................................36

FIGURE 4-1 MEAN PHONEME ERRORS BY TURKISH AND AUSTRALIAN PARTICIPANTS IN THE AUDITORY-ORTHOGRAPHIC AND AUDITORY- VISUAL-ORTHOGRAPHIC CONDITIONS IN ERDENER AND BURNHAM (2005). .................................................................................. 55

FIGURE 4-2. THE MCGURK EFFECT. WHEN AUDITORY [BA], IS DUBBED ONTO THE LIP MOVEMENTS FOR [GA], THE RESULTANT PERCEPT IS USUALLY [DA] OR [THA]................................................................. 57

FIGURE 4-3 SCHEMATIC REPRESENTATION OF FLMP. THE INFORMATION SOURCES ARE GIVEN IN UPPERCASE LETTERS, AND PSYCHOLOGICAL FUZZY TRUTH VALUES ARE INDICATED IN SMALLCASE LETTERS. THE
EVALUATION PROCESS TRANSFORMS INFORMATION FROM SOURCES INTO FUZZY TRUTH VALUES, WHICH ARE THEN INTEGRATED INTO A DEGREE OF SUPPORT. IN THE FINAL DECISION PROCESS, THESE OUTPUTS ARE COMPARED AGAINST PROTOTYPES STORED IN LONG-TERM MEMORY. THE FIGURE WAS ADAPTED FROM MASSARO AND COLE (2000). .................................................................................................................................................. 61

FIGURE 5-1 AUDITORY-VISUAL SPEECH PERCEPTION TEST APPARATUS SETUP............................................................................................................................................................................ 90

FIGURE 5-2 THE LANGUAGE-SPECIFIC SPEECH PERCEPTION TEST SETUP... 92

FIGURE 5-3 THE ARTICULATION TEST PROCEDURAL SETUP. ................................. 95

FIGURE 5-4 AV-Congruent, AV-Incongruent and AO Mean Values Based on Auditory-Correct Responses in 5-, 6-, 7- and 8-Year-Old Groups. VSI Scores Were Computed by Subtracting the AV-Incongruent Scores from AV-Congruent Scores. Error Bars Represent the Standard Errors. ......................................................................................................................... 97

FIGURE 5-5 Adult VSI Scores (Calculated by AV+ minus AV-) by Background Noise and Stimulus Language Conditions. Error Bars Show Standard Error of the Mean. .................................................. 98

FIGURE 5-6 Visual Speech Index (AV+ minus AV-) Scores for Clear and Noise Conditions Across the Five Age Groups by Stimulus Language. Error Bars Represent the Standard Error of the Mean. ................................................................................................................................. 98

FIGURE 5-7 Lipreading (VO) Data in Four Child and Adult Groups. Error Bars Show the Standard Error of the Mean. .................... 99

FIGURE 5-8 AO (Auditory-Only) Scores Across Four Age Groups in the Clear and Noise Conditions. Error Bars Show the Standard Error of the Mean................................................................. 100

FIGURE 5-9 Experiment 1: Speech Perception Test Data in Four Child and Adult Perceivers by Stimulus Language. Error Bars Show the Standard Error of the Mean................................................. 101

FIGURE 5-10 Reading Scores Across Five Age Groups. Error Bars Show the Standard Error of the Mean......................................................... 102

xix
FIGURE 5-11 ARTICULATION SCORES ACROSS FIVE AGE GROUPS. ERROR BARS SHOW THE STANDARD ERROR OF THE MEAN ........................................ 103

FIGURE 6-1 THE EXPERIMENTAL SET-UP OF AUDITORY-VISUAL SPEECH PERCEPTION TEST. ........................................................................ 123

FIGURE 6-2 THE EXPERIMENTAL SETUP FOR SPEECH PERCEPTION TEST . .127

FIGURE 6-3 EXAMPLE OF A FLEXIBLE ITEM SELECTION TASK (FIST) TRIAL CARD. IN THIS TRIAL COLOUR (RED & YELLOW) AND SHAPE (BOAT & SHOE) ARE RELEVANT DIMENSIONS WHEREAS THE SIZE IS THE IRRELEVANT DIMENSION. ........................................................................ 128

FIGURE 6-4 MEAN VSIX-AX SCORES OF 3- AND 4-YEAR-OLDS. THE ERROR BARS SHOW THE STANDARD ERROR OF THE MEAN ........................................ 131

FIGURE 6-5 MEAN AUDITORY-ONLY (AO) AND VISUAL-ONLY (VO) SCORES OF 3- AND 4-YEAR-OLDS. THE ERROR BARS SHOW THE STANDARD ERROR OF THE MEAN ........................................................................ 131

FIGURE 6-6 MEAN NATIVE DI AND NON-NATIVE DI SCORES OF THE 3-, AND 4-YEAR-OLD GROUPS. ERROR BARS REPRESENT THE STANDARD ERROR OF THE MEAN .......................................................... 132

FIGURE 6-7 MEAN RULE ABSTRACTION, COGNITIVE FLEXIBILITY AND OVERALL FIST (RULE ABSTRACTION + COGNITIVE FLEXIBILITY) SCORES OF THE 3- AND 4-YEAR-OLD GROUPS. ERROR BARS REPRESENT THE STANDARD ERROR OF THE MEAN. ........................................................................ 133

FIGURE 6-8 MEAN RECEP TIVE VOCABULARY (PPVT) SCORES OF 3- AND 4-YEAR-OLD PRESCHOOL CHILDREN. ERROR BARS SHOW THE STANDARD ERROR OF THE MEAN .......................................................... 134

FIGURE 7-1 THE AUDITORY-VISUAL (TOP), AUDITORY-ONLY AND VISUAL-ONLY (BOTTOM) SCORES FOR CHILDREN WITH AND WITHOUT SPEECH DISORDER. ERROR BARS SHOW THE STANDARD ERROR OF THE MEAN. ........................................................................ 152

FIGURE 7-2 THE RULE ABSTRACTION AND COGNITIVE FLEXIBILITY SUBTEST SCORES OF THE FLEXIBLE ITEM SELECTION TEST FOR CHILDREN WITH AND WITHOUT SPEECH DISORDER. ERROR BARS SHOW THE STANDARD ERROR OF THE MEAN .......................................................... 153

xx
FIGURE 7-3 THE AUDITORY-VISUAL (TOP), AUDITORY-ONLY AND VISUAL-ONLY (BOTTOM) SCORES FOR CHILDREN WITH SPEECH DISORDER IN THREE AGE GROUPS. ERROR BARS SHOW THE STANDARD ERROR OF THE MEAN................................................................. 156

FIGURE 7-4 THE RULE ABSTRACTION AND COGNITIVE FLEXIBILITY SUBTEST SCORES OF THE FLEXIBLE ITEM SELECTION TEST FOR CHILDREN WITH SPEECH DISORDER. ERROR BARS SHOW THE STANDARD ERROR OF THE MEAN................................................................. 157

FIGURE 8-1 A MODEL TO DESCRIBE INTERRELATIONS AMONG READING, ARTICULATION, LIPREADING, AND AUDITORY-VISUAL SPEECH PERCEPTION. .................................................................................. 167

FIGURE 8-2 SCHEMATIC REPRESENTATION OF THE PHONETIC AND PHONOLOGICAL DEVELOPMENT OF AUDITORY-VISUAL SPEECH INTEGRATION OVER INFANCY, CHILDHOOD AND ADULTHOOD. ........ 175

xxi
LIST OF TABLES

TABLE 2-1 TONES IN CANTONESE .......................................................... 22
TABLE 4-1. DATA FROM THE ORIGINAL MCGURK AND MACDONALD (1976) STUDY. THE AUDITORY AND VISUAL RESPONSE CATEGORIES REFER TO RESPONSES BASED SOLELY ON A SINGLE MODALITY. THE ‘FUSED’ RESPONSE REFS TO THOSE RESPONSES IN WHICH INFORMATION FROM TWO MODALITIES ARE FUSED INTO A NEW ELEMENT, NOT PRESENT IN EITHER MODALITY (E.G. “DA” RESPONSE TO AUDITORY [BA] + VISUAL [GA]). THE ‘COMBINATION’ CATEGORY REPRESENTS RESPONSES WHICH EMBED ELEMENTS FROM BOTH MODALITIES (E.G. “BAGA” RESPONSE TO AUDITORY [GA] + VISUAL [BA]).............................. 56
TABLE 5-1 PARTICIPANTS’ AGE AND GENDER DATA ........................................... 86
TABLE 5-2 THE MCGURK STIMULI ................................................................ 88
TABLE 5-3 SYLLABLE DURATIONS IN THE LANGUAGE SPECIFIC SPEECH PERCEPTION TEST .................................................................................................................. 91
TABLE 5-4 NATIVE AND NON-NATIVE SPEECH CONTRASTS IN THE LANGUAGE SPECIFIC SPEECH PERCEPTION TEST. THE FIRST FOUR CONTRASTS WERE USED IN THE NATIVE BLOCK AND THE LAST FOUR CONTRASTS WERE USED IN THE NON-NATIVE BLOCK ....................... 91
TABLE 5-5 CORRELATIONS AMONG THE INDEPENDENT (AGE, AO, VO [LIPREADING], N-NN DI, ARTICULATION AND READING) AND DEPENDENT (VSI) VARIABLES FOR CHILDREN ....................................... 104
TABLE 5-6 SEQUENTIAL MULTIPLE REGRESSION OF AGE, AO, VO, N-NN DI, ARTICULATION AND READING SCORES AS PREDICTORS OF VSI SCORES .............................................................................................................. 104
TABLE 5-7 SEQUENTIAL MULTIPLE REGRESSION OF AGE, AO, VSI, NATIVE DI, NON-NATIVE DI, ARTICULATION AND READING SCORES AS PREDICTORS OF VO SCORES ............................................................................................................. 106
TABLE 5-8 CORRELATIONS AMONG THE INDEPENDENT (AO, VO, N-NN DI, ARTICULATION AND READING-WRAT) AND DEPENDENT (VO-LIPREADING) VARIABLES FOR ADULTS .......................................................... 107
TABLE 5-9 SEQUENTIAL MULTIPLE REGRESSION OF AO, VO, N-NN DI, ARTICULATION AND READING SCORES AS PREDICTORS OF VSI SCORES IN ADULTS. .......................................................... 107
TABLE 5-10 SEQUENTIAL MULTIPLE REGRESSION OF AO, VO, N-NN DI, ARTICULATION AND READING (WRAT) SCORES AS PREDICTORS OF VO SCORES IN ADULTS. .......................................................... 108
TABLE 6-1 NATIVE AND NON-NATIVE SPEECH CONTRASTS CREATED FOR THE LANGUAGE SPECIFIC SPEECH PERCEPTION TASK. .................. 126
TABLE 6-2 CORRELATION COEFFICIENTS AMONG THE INDEPENDENT (AO, VO [LIPREADING], N-NN DI, RULE ABSTRACTION, COGNITIVE FLEXIBILITY AND PPVT) AND DEPENDENT (VSI-AX) VARIABLES. ...... 135
TABLE 6-3 SEQUENTIAL MULTIPLE REGRESSION OF AO, VO, N-NN DI, FIST AND PPVT SCORES AS PREDICTORS OF VSI-AX SCORES .................. 136
TABLE 6-4 SEQUENTIAL MULTIPLE REGRESSION OF AO, VSI-AX, N-NN DI, RULE ABSTRACTION, COGNITIVE FLEXIBILITY AND PPVT SCORES AS PREDICTORS OF VO (LIPREADING) SCORES. .................. 137
TABLE 7-1 CORRELATION COEFFICIENTS AMONGST THE VARIABLES AGE, SPEECH DISORDER STATUS (SDS), AO, VO, VSI-AX, N-NN DI AND FIST SCORES ................................................................. 154
TABLE 7-2 MULTIPLE REGRESSION OF AGE, SPEECH DISORDER STATUS (SDS), AO, VO, N-NN DI AND FIST SCORES AS PREDICTORS OF VSI-AX SCORES ................................................................. 154
TABLE 7-3 MULTIPLE REGRESSION OF AGE, SPEECH DISORDER STATUS (SDS), AO, VSI-AX, N-NN DI AND FIST SCORES AS PREDICTORS OF VO (LIPREADING) SCORES. ................................................................. 155
TABLE 7-4 CORRELATION COEFFICIENTS AMONG THE VARIABLES AGE, AGE GROUPS, AO, VO, VSI-AX, N-NN DI AND FIST SCORES (SPEECH DISORDERED CHILDREN ONLY). ................................................................. 158
TABLE 7-5 MULTIPLE REGRESSION OF AGE, AGE GROUPS, AO, VSI-AX, N-NN DI AND FIST SCORES AS PREDICTORS OF VSI-AX SCORES (SPEECH DISORDERED CHILDREN ONLY). ................................................................. 158
TABLE 7-6 MULTIPLE REGRESSION OF AGE, AO, VSI-AX, N-NN DI AND FIST SCORES AS PREDICTORS OF VO (LIPREADING) SCORES. ................................................................. 159

xxiii
Table 8-1 Summary of the results of Experiments 1, 2 and 3.........................164
Table A2-1 Task orders in Experiment 1: School Children..................199
Table A2-2 Task orders in Experiment 1: Adults.................................203
Table A2-3 Task orders in experiment 2: Preschool Children..............205
Table A2-4 Task orders in Experiment 3: Children with speech disorder (D) 207
LIST OF FORMULAS

EQUATION 5-1 CALCULATION OF VISUAL SPEECH INDEX (VSI) SCORES...... 84
EQUATION 5-2 THE CALCULATION OF DISCRIMINATION INDEX SCORES FOR
  LANGUAGE-SPECIFIC SPEECH PERCEPTION TEST.............................. 84
EQUATION 6-1 THE CALCULATION OF MCGURK DISCRIMINATION TASK
  VISUAL SPEECH INDEX SCORE (VSI-AX). .......................................... 122
ACKNOWLEDGEMENTS

I am more than grateful to many people without whose support and help this doctoral project would have been much more complicated, longer and less enjoyable.

I do not know how I should thank my supervisor Professor Denis Burnham for his patience and belief in the first place. He has been available at the most unusual times (25/8) and places (B747s, edge of Grand Canyon, saving chapters from reindeers in Nara). Without his support and inspiration I don’t think I would’ve made this far. Denis is one of those rare people who make a real difference around them, and he made one here.

Assoc. Prof. Kate Stevens, one of the most brilliant minds I have met, deserves a special ‘thank you’ for creating MARCS as well as for provocative questions and comments regarding my work here.

I would like to thank Dr. Agnes Petocz, Prof. Herb Marsh, Dr. Bruno di Biase and Dr. Olivier Pascalis (University of Sheffield) for providing very insightful suggestions and comments at the start of the project. I am also grateful to Dr. Sophie Jacques (Dalhousie University) for allowing me to use Flexible Item Selection Test.

Thank you to Assoc. Prof. Robin Panneton (Virginia Tech) for beer in Melbourne, sushi in Kyoto, and extremely valuable feedback for one of my critical chapters while she was stranded at a café in London.

Thanks to Prof. Cathi Best for her valuable suggestions and feedback to help improve my conference presentations and to Dr. Christine Kitamura and Mel Gallagher for their help with MARCS BabyLab register. Also special thanks to Dr. Heather Winskel for her suggestions on how to deal with mid-PhD crisis, and always remembering me wherever she was.

A special ‘thank you’ goes to my co-supervisor Prof. Barbara Dodd for giving a significant momentum to the project and inspiration, and Dr. Sharon Crosbie and Mrs. Beth McIntosh (University of Queensland) for their valuable suggestions, and help with testing speech-disordered children as well as the great dinners in Brisbane.

xxvi
I would like to thank my co-supervisor Assoc. Prof. Chris Davis for his helpful advice and creative input. I also thank my other co-supervisor Prof. Sekiyama for advice and providing the McGurk stimuli.

I thank my proofreaders Lidija Krebs-Lazendič (plus for translating a paper from Serbian), Assoc. Prof. Chris Davis, Assoc. Prof. Robin Panneton, Iris-Corinna Schwarz, and Nan Xu. I also would like express my gratitude to the three anonymous reviewers for making this a better thesis.

On the technical and personal side of things very special thanks go to Mr. Colin Schoknecht for numerous reasons: building a response box, providing numerous computers and equipment, juggling my untimely technical needs, but more importantly his friendship and wonderful sense of humour over the past years.

Rua Haszard-Morris and Matt Frear wrote the original MATLAB codes, so bigthanks to them. Thank you Johnson (Yanqiang) Chen for helping me with God knows how many requests and especially for writing and revising the PsyScript and MATLAB codes, and of course the great Chinese food from both home and around the corner, 謝謝你.

My good friend Arman has always been there in times of computer troubles, whether he was in the middle of something important or not. So, 謝謝你!!

Brad McIntosh programmed the McGurk discrimination task with an admirable level of patience and understanding (of the subject matter to get the job done). Also thanks to Iris-Corinna Schwarz for lending the cartoon clips.

Mel Gallagher and Dr. Caroline Jones helped with the ethics applications. Special thanks should go to Gail Charlton, Darlene Williams, Mel Gallagher, Karen McConachie, and Ruth Goldsmith for helping me with indefinite number of administrative, travel and conference issues, most of which were complete mystery to me.

I would like to thank University of Western Sydney for providing an Australian Postgraduate Award and UWS Top-up scholarship, which gave me the financial freedom to do my PhD. In addition, I want to thank MARCS, ASSTA, ISCA, and UWS Research Office for providing me with funds to attend three international conferences. I also express my special xxvii
gratitude to Aprica Foundation (Osaka, Japan) for awarding me an Early Career Research Fellowship that enabled me to present a paper at the International Conference on Infant Studies in June 2006.

This work would not have been possible without the help of hundreds of children, their parents and teachers, who I cannot name here due to ethics regulations. I want to say a big thank you to all of them for helping me with this project. Also special thanks to NSW Department of Education for their permission to test at schools.

MARCS has been more than the place where I did my PhD. It is a living, breathing and vibrant social environment. One special friend is Iris, who, in short, has been there with me through thick and thin at every phase. I wish her a life long happiness in her new life with Ulf. Thank you Michael for your help with this, and a share of your talent and knowledge at times of need. Thanks to all of my friends here for making MARCS a cut above the rest in many ways: Brett Molesworth, Mel Gallagher, Ulf Kalla, Rudi Črnčec, Renee Glass, Clare Howell, Mark Antoniou, Shaun Halovic, Liz Francis-Beach, Mary Broughton, Rikke Bundgård-Nielsen, Tim Byron, Jess Hartcher-O’Brien, Jemma Harris, Graham Howard, Benjawan Kasisopa, Bettina Keresztesi, Lidija Krebs-Lazendić, Christa Lam, Nicole Lees, Nathan Perry, Khazriyati Salehuddin, Robert Sazdov, Barbara Schwanhäußer, Karen Mattock, Damien Smith, Wendy Vlismas, Nan Xu, and Anna Notley. Et à mon ami Aude: merci d’écouter dans la phase finale.

I want to thank my family in Australia and Turkey for their encouragement: in Australia my aunt Ayşe Kemikoğlu, my cousin Ceren Aşkın, the Gülsoys, and, of course my little big-hearted ‘Marvel’, and in Turkey my aunt Gülten Özb aş, my dad Özcan Erdener, and Savaş Kargin, who is missed very much.

I do not know how to thank my mother, Nermin Kemikoğlu, for enduring not just this with me, but many other things in the past three and a half decades. She has always been there for me, sacrificing a lot, being great, but more importantly, being my most honest friend: Teşekkürler Annem!

This thesis has been dedicated to the memory of my multilingual grandparents Nezihe and Vecih Kemikoğlu, whose love and linguistic gift were the setting stones for my interest in languages and the world. I miss you both.

xxviii
CHAPTER 1

INTRODUCTION
1.1 I COULD SEE YOUR LIPS MOVE

- Open the pod bay doors HAL.
- I’m sorry Dave. I’m afraid I can't do that.
- What’s the problem?
- I think you know what the problem is just as well as I do.
- What are you talking about HAL?
- This mission is too important for me to allow you to jeopardise it.
- I don’t know what you’re talking about HAL.
- I know you and Frank were planning to disconnect me. And I’m afraid that’s something I cannot allow to happen.
- Where the hell did you get that idea HAL?
- Although you took very thorough precautions in the pod against my hearing you, I could see your lips move. (Kubrick & Clarke, 1968, from the movie 2001: A Space Odyssey).

Just as Clarke (1945) had foreseen the birth of telecommunication satellites more than 60 years ago, Clarke, with Stanley Kubrick, also foresaw machines that can perfectly lipread in the absence of any auditory speech information in the 1968 movie 2001: A Space Odyssey (Kubrick & Clarke, 1968). Although HAL causes serious problems due to its splendid lipreading ability, the prospect of developing such lipreading systems presents a more optimistic future. There are numerous areas which can potentially benefit from the advancement of auditory-visual automatic speech recognition (AV-ASR) systems: speech therapy, audiovisual aids for disabled users and the hearing impaired and elderly, children with learning and language difficulties, foreign language instruction, telecommunication (particularly multimedia mobile telephones), human-machine interfaces, business and e-commerce, and media, film, and game industries.
Machines such as HAL9000 have not yet been developed and even the most
developed machines cannot lipread perfectly. However there are a number of
good systems being developed (Brooke, 1998), and by asking the correct
questions today, science is much closer to achieving this goal than in 1968 or
even 2001. Answering the correct questions is not only the duty of technical
areas of research such as information technology, computer science, and
engineering, but also (crucially) of areas that investigate how humans process
speech; speech science, psychology and psycholinguistics. Once the
mechanisms of human speech processing are understood and modelled AV-
ASR systems can be built that will make life easier for people who have little
or no access to information as a result of some form of communication
impediment. In order to understand how humans process speech through
understanding of speech as a multimodal phenomenon, research on how
auditory, visual and auditory-visual phenomena develop is required.

The goal of good AV-ASR system serves a dual purpose. The first is to allow
for the practical applications to run as smoothly as possible. The second
purpose is the provision of impetus for behavioural scientists to investigate
and specify how humans achieve auditory-visual speech perception, so we can teach machines how to do it. Studies show that this is the right way of developing better AV-ASAR systems such that inclusion of visual speech information significantly reduces the amount of speech recognition errors in AV-ASR systems (e.g. Petajan, 1985). Recent research has shown that the use of auditory-visual human speech databases such as CUAVE (Patterson, Gurbuz, Tufekci & Gowdy, 2002) help improve AV-ASR systems (Dean, Lucy & Sridharan, 2005).

Despite the research on auditory-visual speech perception in the past two decades (for reviews see Campbell, Dodd, & Burnham, 1998; Dodd & Campbell, 1987; Massaro, 1987, 1998), there is still a lot to be learnt about the development of auditory-visual speech perception. This thesis provides a small step in the continuing evolution of knowledge on auditory-visual speech perception development, especially in early childhood. Specifically, this thesis aims to answer the question of how auditory-visual speech perception develops in humans, particularly early in life between the ages of 3 and 8. The next three chapters present the fundamentals of and the general literature on auditory and auditory-visual speech perception. These are followed by three empirical studies of auditory-visual speech perception by children and adults, and a final chapter discussing and drawing implications from the results.
CHAPTER 2

ARTICULATION AND SPEECH PRODUCTION
This thesis is concerned with auditory-visual speech perception development and the factors that affect this development. As the phonological aspects of auditory-visual speech are integral to this investigation, an understanding of the fundamentals and levels of language processing is necessary. This chapter is dedicated to the articulatory process and the mechanics of speech production, and it is comprises five main sections: (i) anatomy of articulation, (ii) elements of speech, (iii) production of speech elements, (iv) articulation and its perceptual consequences, and finally (v) towards the problem of speech perception. This will pave the way for the description of auditory and auditory-visual speech perception, chapters 3 and 4, respectively.

2.1 Anatomy of Articulation

The vocal tract is a complex structure, made up of several elements that contribute to speech production. In this section the anatomy of the speech production system with particular emphasis on the visual speech articulators, namely, larynx, velum, tongue, teeth, mandible and lips are described. Visual speech articulators are those sections of the vocal tract and facial features that are visible to the perceiver attending an incoming speech signal in a face-to-face conversation. A general outline of vocal production is provided below and a schematic diagram of the vocal tract and specific visible articulators is shown in Figure 2-1.
2.1.1 Larynx & Vocal Folds

The larynx is used to control the airflow into and out of the lungs, to provide oxygen to the body, eliminate carbon dioxide, and prevent food, water and other substances from entering the lungs. As a by-product of these functions the larynx allows the phonation involved in speech production. The larynx is a tube-like structure covered by mucous membrane, made up of cartilages, interconnected by ligaments and membranes (Borden, Harris, & Raphael, 2003). In the larynx, there are two sets of folds: the ventricular folds (false vocal folds), and the true vocal folds or vocal cords. The ventricular folds form a constriction just above the true vocal folds. The true vocal folds are responsible for voicing in the production of speech as described in 2.3.1.
2.1.2 Velum

The velum (or the soft palate), is a flexible extension of the hard palate and functions to separate the nasal and oral cavities (Hardcastle, 1976). The separation between the nasal and oral cavities is particularly important in the production of nasal speech sounds (eg. [m] and [n]). The velum is controlled by three groups of muscles: lowering, raising and tensing and, it has two extents of movement, raised and lowered. It is raised in the rest position in normal breathing, in the production of vowels, and in the production of plosives, fricatives and approximants. It is lowered when nasal sounds are produced. The lowering movement adds length and resonance to the vocal tract, modifying the quality of the sound emitted (Zemlin, 1998).

2.1.3 Tongue

The tongue is the largest and most mobile of the articulators in the oral cavity. The principal functions of the tongue involve tasting, masticating, deglutition and transfer of food to the posterior digestive processes (Borden et al., 2003). However, the tongue functions in conjunction with other articulators such as alveolar ridge and palate to modify the shape of the vocal tract in order to inhibit or free the air flow (Zemlin, 1998), leading to hundreds of different variations of the vocal tract shape and volume, and changing the resonance characteristics. The tongue also functions as a noise generator and modulates laryngeal tone (Zemlin, 1998), which aids in the production of voiced consonants (Hardcastle, 1976; Zemlin, 1998).

The tongue can be divided into two main sections: the blade and the root, and a further four regions in terms of its relationship with the roof of the mouth: Tip, blade, front, and back. The tongue tip is the fastest moving articulator with an average range of 7.2 to 9.6 cycles per second (Hz) (Hardcastle, 1976). According to Zemlin (1998) and Hardcastle (1976), there are two major classes of tongue muscle groups: intrinsic and extrinsic muscles. The complex coordination of these two muscle groups provides the tongue with a
substantial freedom of movement\(^1\). The intrinsic muscles are entirely located within the tongue, and their primary responsibility is to change the configuration of the tongue. The primary responsibility of the extrinsic muscles is to alter the position of the tongue in the oral cavity. These muscles have their origin external to the tongue and attach to the tongue at various points.

Hardcastle (1976) describes seven different articulatory parameters of the lingual movements; various positions and configurations of the tongue during speech, which essentially alter the shape of the vocal tract, and accompany, either alone or in conjunction with other configurations, the production of different phonemes as follows:

(i) Horizontal forward-backward movement (in the production of [h], [x], [a]).

(ii) Upward-downward movement of the tongue body, e.g. [i], and palatal stops (e.g. [t]), and clicks (e.g. [θ, ð]).

(iii) Forward-Backward Movement of the Tip-Blade (e.g. [l]).

(iv) Upward-downward movement of the tip-blade (e.g. [t], [l], [n], [s]).

(v) Concave-convex cross-sectional configuration (e.g. [s], and some vowels)

(vi) Degree of central grooving (e.g. [s]).

(vii) Spread-tapered surface plan configuration ([l], [s], [l], [i] [e]).

2.1.4 Teeth

The primary function of the teeth is biological; preparing the food for further digestive processes in the stomach. In addition the teeth are also important articulators, directly involved in the production of numerous consonants,

\(^1\) Hardcastle (1976), Zemlin (1998), and Borden et al. (2003) provide more detailed descriptions of the anatomy and physiology of speech articulators.
such as labiodentals, lingua-palatal, lingua-alveolar and interdentals (eg. [f], [v], [θ], and [ð]), and also crucial in the production of vowels (Zemlin, 1998). In addition to teeth being important articulators in consonant and vowel production, research indicates that they also provide information in terms of visual speech, although the amount of research on the role of teeth in visual speech perception is strongly limited. Summerfield, McLeod, McGrath, and Brooke (1989) report that as one follows a sequence of [i], [e] [a], and [u], the degree of separation between the inner margins of the lips increases. Additionally, as the jaw drops, the upper and lower teeth move further apart from each other, so there is a loss of visibility as the roundness of the vowels increases. Consistent with this, McGrath, Summerfield and Brooke (1984) found that when articulations are presented without teeth, particular vowels are confused with other vowels, that are normally articulated with the same lip shape but different interdental distance. With the teeth absent, perceivers also confused rounded vowels (teeth not visible under normal conditions) with unrounded vowels (teeth visible under normal conditions) significantly more than under normal full-face conditions. These results demonstrate the important role played by the visibility of the teeth in the auditory-visual perception of speech.

2.1.5 Mandible

Aside from its main duty of grinding food during mastication (Türker, 2002), the mandible has two major functions that assist speech production: Raising (elevation) and lowering (depression), and protrusion and retraction\(^2\). Raising the mandible moves the hyoid bone, alters the position of the tongue and, to lesser extent, the shape of the tongue, which aids the production of some vowels, (Hardcastle, 1976). In addition to this, lowering mandible increases vocal intensity by about 4-5 dB (Zemlin, 1998), and to a great extent, alters the

\(^2\) In addition, the jaw can also do protrusions and retractions, as well as lateral movements, which are more important for grinding food (Zemlin, 1998).
resonant characteristics of the vocal tract (Hardcastle, 1976; Zemlin, 1998). The range of movement of the mandible during speech is between 2-3 mm and 7-18 mm, and it can move up to 7.5 times per second. There are three groups of muscles performing these movements: Muscles of elevation, depression, and protrusion.

The role of mandible in visual speech is not very clear, though it appears that the degree of visual information it provides is rather dependent on the movements of other articulators. Brooke and Summerfield (1983) tested the role of various points on the face in speech perception. They used /aCV/³ syllables and found that vowels were not always captured accurately by mandible movements, and that they need to be accompanied by the movements of other articulators, such as the lips in order to provide visual speech information. To investigate the role of mandibular movements in auditory-visual speech perception, Guiard-Marigny, Benoît, and Ostry (1995) developed a lip model, based on five parameters derived from a native French-speaking talker’s utterances; and a jaw model based on six degrees of movement. Twenty native French-speaking listeners were tested on their perception of VCVCV stimuli in three conditions: auditory-alone, the lip model only and lip/jaw models combined. The results showed that there was a significant gain in intelligibility when stimuli were presented through the lip/jaw model, than the lip model, with the auditory-only condition yielding the poorest performance. More recently, Benoît (1997) compared these models to a full face model, and found that performance in the full face condition was superior to the lip/jaw model condition, lending support to the findings of Brooke and Summerfield (1983). Thus, in general it appears that the movements of mandible provide visible speech information most clearly in conjunction with the synchronous movements and visibility of adjacent articulators such as lips, teeth and tongue.

³ They used consonants /m/, /b/, and /p/ and vowels, /a/, /i/, and /u/. The vowels Brooke and Summerfield used were selected as producing maximal jaw opening, lip spreading and lip rounding, respectively.
2.1.6 Lips

Lips are very complex anatomical structures. The largest of the muscular structures in the lips is orbicularis oris, which is, in all directions, connected to the various muscles in the face, providing it with a substantial flexibility of movement. Upon contraction, the orbicularis oris closes the mouth and puckers the lips (Zemlin, 1998). Its contraction is vital for the labial closure required for bilabials such as [b], [p], and [m] and rounding for the production of some liquids and vowels, such as [w] and [u] (Borden et al., 2003). It is also responsible for the protrusion of the lips, which is necessary for the production of open rounded vowels, such as [ɔ] and [ɭ] (Hardcastle, 1976). The lips have seven basic movements, each important for speech production. These are closing, raising, lowering, rounding, retracting angles, raising angles, and lowering angles.

During speech, lips are highly visible and mobile; hence they are perhaps the most informative of the visible speech articulators. For example, Benoît and his colleagues (Benoît, 1997; Benoît, Guiard-Marigny, Le Goff, & Adjoudani, 1996; Guiard-Marigny et al., 1995) found that compared with auditory-alone and jaw model conditions, the lips provided additional information that renders the speech signal more intelligible (Sumby & Pollack, 1954). The movement of the lips also provides visual cues with respect to the rapidly changing low-intensity spectral detail on the mid- to high-frequency part of the spectrum; very robust cues with respect to place of articulation, as in the case of consonants [b] vs. [d] and various vowels (Summerfield et al., 1989).

2.2 Elements of Speech and Language

In this section the elements of speech production, namely morphemes, phonemes, allophones, and phones are described. Morphemes are the smallest linguistic units in a language that carry a meaning. For example, the word ‘bats’ is made up of two morphemes, ‘bat’ and ‘-s’, meaning a flying mammal, and more than one, respectively. Phonemes are sound categories in the repertoire of a given language that contrast in such a way that they signal
meaning differences but, unlike morphemes, phonemes do not mean anything on their own. For example, the words ‘bat’ and ‘pat’ differ in meaning because of the phonemic difference between /b/ and /p/. Phonemes are not single sounds but rather abstract representations of articulations; and each phoneme is actualized in the form of several variant articulations, referred to as *allophones*. For instance, the /p/ in ‘pat’, ‘spat’, and ‘sap’ involves different degrees of air burst release in their production (strong, minimal, and none, respectively), and each of these /p/ articulations, [pH], [p], and [p], respectively, are allophones of the phoneme category /p/. Standing alone, without reference to the phoneme class, these sounds are *phones*. The term phone is used when referring to a particular articulation, and in its written form is expressed in square brackets, e.g. [p], whereas phonemes are expressed in diagonal brackets e.g. /p/.

2.3 **Speech Production**

Up to this point the anatomy of the speech apparatus and the definitions of speech elements have been given. In this section the focus is the speech production process, specifically the processes of laryngeal and supralaryngeal variation.

2.3.1 **Laryngeal Variation and the Role of Vocal Folds in Speech Production**

The movement of the vocal folds in the larynx are responsible for voicing. Voicing is the process by which speakers transform air from the lungs into speech sounds. This is achieved using the air pressure from the lungs to bring the vocal folds together, resulting in a pattern of vibrations⁴ (Borden et al.,

---

⁴ This process is underlined by Bernoulli Effect, which states that in a horizontal fluid flow, an increase in the velocity of flow will result in a decrease in the static pressure. A typical example of Bernoulli Effect is demonstrated in aeroplanes. The aeroplane wings are designed in such a way that the airflow above the wing is faster than the airflow under the wing, hence less pressure is present on the wing than under it.
2003). Different vocal fold adjustments result in voiced vs. voiceless speech sounds and differences in fundamental frequency and voice quality. When at rest, the vocal folds are separated (abducted). The vocals are also abducted for the production of voiceless speech sounds, such as /s/ and /t/, but are adducted during production of voiced speech sounds, such as vowels and diphthongs, voiced consonants, such as /z/, and /d/.

The basic frequency or the lowest tone of complex sound streams, is called fundamental frequency (F₀). In human voices fundamental frequency corresponds to the basic formant of the vibration of the vocal folds (Borden et al., 2003). During speech, fundamental frequency often changes, and these changes are reflected in intonation patterns of the signal. In turn intonation is one of the primary cues in speech perception. Take, for example, the sentences “Did Jimbo go to Yowie Bay?” and “Jimbo went to Yowie Bay.” The former sentence has a rising intonation pattern, and indicates a question whereas the latter sentence can indicate a fact with a falling intonation. These intonational variations are produced via changes in the frequency of vocal fold vibration. According to the myoelastic aerodynamic theory of phonation⁵, frequency of vocal fold vibration is determined by the elasticity, tension, and mass of the vocal folds (Borden et al., 2003). Males have longer vocal folds (17-24 mm) than females (13-17 mm), and thus lower fundamental frequencies than females. The average fundamental frequency for males is around 125 Hz, and for children and females it is above 200 Hz. In general, the longer the vocal folds, the lower the fundamental frequency.

Fundamental frequency is the acoustic dimension on which the auditory perception of pitch is based (Borden et al., 2003). Thus fundamental frequency

---

⁵ The myoelastic aerodynamic theory of phonation was originally developed by Helmholtz and Müller in the nineteenth century. The theory suggests that vocal folds are vibrated by the airstream from the lungs, rather than nerve impulses, as had been thought prior to this view. The term “myoelastic” refers to the ability of the muscles to change their elasticity and tension in order to modulate the changes in frequency of vocal fold vibration (Borden et al., 2003).
is one of the major determinants of voice quality, which refers to differences among individual voices or among sounds produced by a single talker. Over and above voice pitch (based on \( F_0 \)), a major factor that determines voice quality is the degree to which vocals folds approximate each other during phonation. Paralysis is one form of anatomical irregularity that affects the approximation of vocal folds. Another example is hoarse voice, caused by irregularities in the vocal folds due to irritation or swelling. Under normal conditions, voice quality differences depend on the vocal fold vibrations. For example, breathy voice is produced by means of adducting the vocal folds close enough to vibrate but allowing the airflow to pass through (Borden et al., 2003).

### 2.3.2 Supralaryngeal Variation

Beyond the larynx and following the production of the basic speech source, supralaryngeal variation is important in the formation of different speech sounds, and variations in voice quality. *Supralaryngeal variation*, involves pharyngeal, oral, and nasal cavities and all articulators above the glottis. A principal source of supralaryngeal variation is the shape of the oral cavity and this shape is determined by movement of various articulators such as the mandible, lips, and tongue, which determine the resonances of the oral cavity.

### 2.4 Vowels, Consonants and Tones as Speech Segments and Their Production

#### 2.4.1 Periodic and Aperiodic Speech Sounds

Periodic speech sounds involve vocal fold vibrations with quasi-equal time intervals, and are signals in which component frequencies are the integral multiples of the fundamental frequency (Borden et al., 2003). The right side of Figure 2-2 illustrates an example of a periodic speech sound. Pure vowels are periodic speech sounds. Aperiodic speech sounds do not have equal time intervals, nor do they have component frequencies with frequencies as multiple of \( F_0 \).
2.4.2 Vowels as Speech Segments and Their Production

There are three factors that determine the production of a vowel: (i) the height of the tongue body, which is associated with the first formant (F1); (ii) the back-front position of the tongue, which is associated with the second formant transition (F2); and (iii) the degree of lip rounding. The combinations of these three parameters determine the shape of the vocal tract, hence the quality of the vowel articulated (see Figure 2-3). For example, for [i] as in ‘beet’, the front of the tongue is high, whereas for [u] as in ‘rude’, the back of the tongue is high. For the [a] in ‘car’, the front of the tongue is low, and for the [b] in ‘hot’ the back part of the tongue is low. On the other hand, for such vowels as [i] as in ‘tip’ and [e] in ‘red’, whose resonance is determined by the tongue height, whereas for such vowels as [u] as in ‘rude’ and [ɔ] as in ‘boy’, the articulation requires lip rounding.  

---

6 In human speech, formants are the peak values expressed in frequency spectra resulting from resonances emanating from vocal tract. The constriction of the airflow from the glottis by places and manner of articulation act as filters of the speech signal that is sourced from the glottis (see Fant, 1970 for a description of Source-Filter Model of Speech Production).

7 These and further examples are based on American English unless otherwise stated.

8 See Ladefoged (2000) for a detailed description of vowel articulations in English (pp.69-89).
There are two types of vowels: monophthongs (steady vowels) such as [a] and diphthongs in which two (or more) vowels are combined, e.g., [ai] as ‘mate’. Another class of sounds, semi-vowels is classified as consonants as they do not bear the sufficient formant energy to be positioned between two consonants.

![VOWELS Diagram](image)

**Figure 2-3** The IPA classification of vowels. The arrangement of the vowels on the vowel quadrant is based on qualitative representation of the tongue position in the oral cavity.

Unlike consonants, variations between vowels are more continuous than discrete: during vowel production, articulators do not meet or even approach contact, so the passage of air is uninterrupted. During airflow passage, the oral cavity takes on various shapes in order to create different vowels. The shape of the oral cavity is created by the relative position of highest point of the tongue (Ladefoged, 2000). Take, for example, the vowels /i/, /ɪ/, /ɛ/, /æ/, /a/, /u/ and /u:/, as in “bead”, “bed”, “kid”, “pad”, “father”, “put”, and “rude”, respectively. When producing the first four vowels, the highest point of the tongue is at the front of the mouth. These vowels are called front vowels. On the other hand, the latter three vowels are produced with the tongue close to the back of the mouth. These three vowels are called back vowels. Another important dimension of vowel production is lip position. Of the vowels mentioned above, /u/ and /u:/, are produced with a forward
movement of the corners of the lips (lip rounding), and these are called *rounded vowels*; whereas vowels such as /æ/ and /i/ are *unrounded vowels*, and do not require lip rounding for their production.

### 2.4.2.1 Diphthong Production

A *diphthong* is a vowel of changing resonance (Borden et al., 2003), i.e. a vowel with a change in quality within a single syllable (Ladefoged, 2000), *for example, /ɔi/* as in ‘boy’. In general, diphthongs in English end with [i] (e.g. /bɛɪ/, bay) or [u] (e.g. /bəut/, ‘bout’), which each require different vocal tract configuration transitions during production (Borden et al., 2003; Ladefoged, 2000), reflected in specific F₁-F₂ formant transitions.

### 2.4.3 Aperiodic Speech Sounds

In the production of consonants, the periodic signal which cross the supralaryngeal passages and is constricted by various formations of the vocal tract, causing varying degrees of turbulence in the airflow. Some consonants are a mixture of periodic (voiced) and aperiodic (unvoiced) components. Take, for example, the alveolar fricative consonants /s/ and /z/. /z/ is simply a mixture of an aperiodic /s/ and a voicing period (see the left panel of Figure 2-2). Other non-speech sounds do not originate from the glottis, but rather are formed as a result of turbulences created in the vocal tract and are not classified as speech sounds.

### 2.4.4 Consonants as Speech Segments and Their Production

Consonants can be described as speech sounds whose productions require relative constriction of the oral cavity. According to the 1993 classification of consonants in the International Phonetic Association (IPA) (see Figure 2-4) there are three parameters of consonant sounds: voicing, place of articulation, and manner of articulation. We now turn to description of these parameters, and consideration of consonant production.
CONSONANTS

<table>
<thead>
<tr>
<th></th>
<th>Bilabial</th>
<th>Labiodental</th>
<th>Dental</th>
<th>Alveolar</th>
<th>Postalveolar</th>
<th>Retroflex</th>
<th>Palatal</th>
<th>Velar</th>
<th>Uvular</th>
<th>Pharyngeal</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plosive</td>
<td>p b</td>
<td>t d</td>
<td>t d</td>
<td>c j</td>
<td>k g</td>
<td>q g</td>
<td>?</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nasal</td>
<td>m m̃</td>
<td>n ñ</td>
<td>ñ</td>
<td>ñ</td>
<td>ñ</td>
<td>ñ</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trill</td>
<td>b r</td>
<td>r r</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tap or Flap</td>
<td>r r</td>
<td>r t</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fricative</td>
<td>φ β f v</td>
<td>θ δ s z</td>
<td>s z</td>
<td>s z</td>
<td>c j</td>
<td>x y</td>
<td>χ ψ h f</td>
<td>h h</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lateral Fricative</td>
<td></td>
<td>l ʃ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approximant</td>
<td>v j ɹ j w</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lateral approximant</td>
<td>l l ξ l</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

**Figure 2.4** The classification of consonants along voicing, manner and place of articulation dimensions.

### 2.4.4.1 Voicing

Voicing refers to whether the vocal folds vibrate (voiced) or remain apart (voiceless) during speech production. The acoustic unit of measurement for degree of voicing is voice onset time (VOT), defined as the period between the obstruction of the airflow and the vibration of the vocal folds (phonation) (Lisker & Abramson, 1964), although there are other acoustic cues for voicing (Stevens & Klatt, 1974). For example, in English, a bilabial stop can be prevoiced e.g. [b] (as in “bat”), or voiceless e.g. [pʰ] (as in “pát”), or voiceless unaspirated [p] (similar to “spat”). If the phonation follows the release burst as in voiceless sounds, then the VOT value is positive. In prevoiced speech sounds, the phonation occurs before the release burst, hence the VOT value is negative such as the Thai prevoiced bilabial, [b] (Borden et al., 2003). In English, in voiceless unaspirated sounds, such as [p], the VOT is around 0 ms – adduction continues and vocal folds begin vibration at the same time as the release burst.

### 2.4.4.2 Place of Articulation

*Place of articulation* refers to the place of articulatory contact or airflow constriction in the vocal tract and the articulators used (Borden et al., 2003).
There are eight places of articulation in English (Ladefoged, 2000) as follows. Bilabials such as the /b/ involves articulation by both lips. Dentals involve the tongue tip or blade and upper front teeth, e.g. /ð/ as in ‘that’. Labiodentals involve the lower lip and upper front teeth, e.g. as in /f/ in ‘fat’. Palatals are articulations involving the front of the tongue and the hard palate, e.g. /j/ in ‘yoke’. Palato-alveolars involve the tongue blade and the back of alveolar ridge, e.g. /ʃ/ in ‘she’. Alveolars involve the tongue tip or blade and alveolar ridge such as /t/ in ‘tin’. Retroflexes involve tongue tip and the back of alveolar ridge. Retroflexes are frequently used in Hindi but are rarely used in English; in North American English, in some words retroflexes occur in initial position of some words, e.g. /s/ as in ‘rye’. Finally, velars involve the back of the tongue and soft palate, the farthest back place of articulation in English, as in the /ŋ/ in ‘swing’ or /k/ in ‘pick’. In addition to their acoustic manifestations, places of articulation provide significant visual speech information. This point is further elaborated in Chapter 4.

2.4.4.3 Manner of Articulation

Manner of articulation refers to the way the airflow is obstructed or channelled in the vocal tract. Ladefoged (2000) refers to five manners of articulation: stops, nasals, fricatives, laterals (approximants), and the group of trills, taps, and flaps. Stops are actualised by complete closure of articulators and blockage of the vocal tract preventing the air stream to escape, e.g., the bilabial stop /p/ and velar stop /k/. Nasals are produced by lowering the soft palate such that the sound is produced in the nasal cavity, e.g., /m/ and /ŋ/. Fricatives (e.g. /θ/, /ʃ/, and /v/) are produced by allowing the airflow passage through a narrow constriction resulting in turbulence. Laterals and approximants are produced by incomplete obstruction of the airstream at a point along the centre of the vocal tract, allowing partial airstream around the

9 Note that /ŋ/ does not occur in initial position in English
sides or over the roof of the tongue, e.g., /l/ and /r/, respectively. Trills are produced by rapid repetitive movement of the tip of the tongue by the air stream. Trills are found very rarely in English. In Edinburgh Scottish English there is a form of trilled /r/, which Ladefoged (2000) describes as a more tap-like articulation. Taps and flaps are more than a rapid articulation of a stop, and made by a single constriction of articulator muscles so that one articulator moves towards the other. Taps (e.g. /d/ in ‘ladder’) and flaps (e.g. /r/ in ‘dirty’) are frequently found in American English.

2.4.5 Tones as Speech Segments and their Production

Linguistic or lexical tone refers to the use of F₀ and other cues at the word level to convey meaning. The majority of the world’s languages employ tonal variations contrastively in their phonological repertoire. Tone is carried variously by fundamental frequency, duration, amplitude and voice quality; through usually the most prominent cues are the F₀ height and contour. Tones can be classified into two categories: static (or level) tones, which involve little or no pitch movement, and dynamic (or contour) tones, which involve a gliding movement with respect to F₀ (Ladefoged, 2000). The number of lexical tones varies across tone languages. The simplest form of a tone language has two tones: high and low in which every vowel simply carries a high or a low pitch (e.g., Shona spoken in Zimbabwe). Other tonal languages can have more sophisticated tone inventories, such as the Chinese dialect Cantonese (spoken predominantly in Hong Kong and Southern mainland China), which has six tones: high-high, low-low, mid-high, mid-low, high-rising, and low-rising. For example, the syllable ‘si’, can carry six tones, each of which has six different meanings, as depicted in Table 2-1 (Li, 2004). In this table, the first three tones are static and the last three are dynamic. The IPA system uses 5 levels (Chao values) within the speaker’s pitch range to represent Chinese tones on a scale of 5, in which 1 represents the lowest pitch and 5 represents the highest pitch (Ladefoged, 2000). Each tone is represented by two numbers; the first representing the onset and the second the offset pitch.
TABLE 2-1 TONES IN CANTONESE.

<table>
<thead>
<tr>
<th>No.</th>
<th>Description</th>
<th>Pitch</th>
<th>Tone</th>
<th>Syllable</th>
<th>Character</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>High-high</td>
<td>55</td>
<td>˨</td>
<td>[si]</td>
<td>האחרונה</td>
<td>Poet</td>
</tr>
<tr>
<td>2</td>
<td>Mid-high</td>
<td>33</td>
<td>˨</td>
<td>[si]</td>
<td>ဟွူခ်ူး</td>
<td>Hobby</td>
</tr>
<tr>
<td>3</td>
<td>Mid-low</td>
<td>22</td>
<td>˨</td>
<td>[si]</td>
<td></td>
<td>Matter</td>
</tr>
<tr>
<td>4</td>
<td>Low-rising</td>
<td>23</td>
<td>˩</td>
<td>[si]</td>
<td></td>
<td>City</td>
</tr>
<tr>
<td>5</td>
<td>High-rising</td>
<td>25</td>
<td>˩</td>
<td>[si]</td>
<td></td>
<td>History</td>
</tr>
<tr>
<td>6</td>
<td>Low-low</td>
<td>21</td>
<td>˩</td>
<td>[si]</td>
<td></td>
<td>Time</td>
</tr>
</tbody>
</table>

2.4.6 Prosodic Features of Speech Signal

The prosodic or suprasegmental features of a language are overlaid upon and independent of speech segments (consonants, vowels and tones). Here the prosodic features are presented in terms of stress, rhythm and intonation.

2.4.7 Stress

Stress results from extra respiratory energy placed on a single syllable during articulation. Functionally, the stress points to the syllable in an utterance that is most important (Borden et al., 2003). In addition, an important function of stress in English is to differentiate verbs and nouns. For instance, in the word ‘permit’, if the stress is placed on the first syllable, ‘permit’ it is a noun meaning ‘a written permission’, whereas if the stress is on the second syllable, per’mit, it is a verb meaning ‘to make possible’ or ‘to give permission’. In many languages the position of stress is consistent in the words, such as Czech, in which stress is almost always on the first syllable, and in Polish and Swahili stress is on the penultimate syllable (Ladefoged, 2000).

2.4.8 Rhythm

Linguistic rhythm refers to the alternation of strong and weak beats in speech. An important parameter in speech rhythm in stress-timed languages (e.g. English, Dutch, and Arabic) is the stress foot. Germanic languages (e.g.
English, German, Dutch, etc.) use stress more than other groups of languages, hence often they are referred to as stress-timed languages. For instance in English, a stress foot is trochaic, i.e. it is made up of a stressed syllable and an unstressed element adjacent to it (Echols, Crowhurst, & Childers, 1997; Nooteboom, 1997). On the other hand, languages such as Spanish, Italian, and French have relatively equal steps on each syllable and are called syllable-timed languages. One of the important functions of the rhythm of a language is the facilitation of word segmentation (Cutler, 1997; Nooteboom, 1997).

2.4.9 **Intonation**

Intonation is the pitch pattern in a sentence and marks the boundaries of syntactic units (Ladefoged, 2000), more specifically intonation can be defined as the use of suprasegmental phonetic features to convey sentence-level pragmatic meanings in a linguistically structured way (Ladd, 1996). In some languages, such as English, completion of a sentence is marked by a falling pitch contour, and a question is signalled by a rising pitch pattern (Borden et al., 2003). Incomplete utterances, in which the speaker pauses to convey some other message (e.g. “...and the winner is.........Sydney”), are marked by a rising pitch contour (Ladefoged, 2000). In addition to its syntactic function, intonation also carries information regarding emotion. For instance, high arousal states, such as anger or happiness are characterised by high degrees of fundamental frequency fluctuation, whereas low arousal states are marked by lower degrees of fundamental frequency variation (Borden et al., 2003).

2.5 **Articulation and its Consequences**

Articulation is simply the process by which means movements of speech organs and the shape of vocal tract are used to produce speech sounds. Up to this point the mechanics of speech production, the process by which the speech

---

10 However it should be noted that the definition and measurement of syllable-timed and stress-timed languages are ongoing debates (e.g. see White & Mattys, in press).
signal is produced, have been considered. The speech signal, as indicated in Chapter 1, does not convey information unimodally, but rather multimodally. Articulation has three main perceptual consequences; auditory, visual and tactile.

First, articulation, movement and contact between speech organs and articulators, which result in acoustic events—speech sounds, which, in turn, are transmitted to the perceiver. The issue of auditory speech perception and its development in humans is elaborated in detail in Chapter 3.

Second, articulation also has optic consequences available to visual perception. The movement of a number of articulators and speech organs provide significant visual speech information to perceivers in regarding the perceptual differences between speech sounds. The issue of auditory-visual speech perception and how it develops is reviewed in Chapter 4.

Thirdly, articulation also results in tactile information. A few studies show that the tactile sequelae of articulation can provide information regarding the speech sounds. This issue is also covered, albeit briefly, in Chapter 4.

This thesis is concerned with the development of auditory-visual speech perception. In the two chapters we turn our attention from the speech signal and how it is produced, to how speech signal is perceived auditorally and auditory-visually, and how such perception develops over age.
CHAPTER 3

AUDITORY SPEECH PERCEPTION AND DEVELOPMENT
3.1 **Articulation as an Auditory Speech Event**

Research has shown that there is a direct relationship between the movements of articulators and the formation of the acoustic signal. For instance, Borden et al. (2003) show that there is a significant correlation between (i) the frequencies of first two formants, (ii) the dimensions of the oral and pharyngeal cavities, and (iii) the position of tongue, jaw and lips and vowel articulation. This shows that at least some articulatory movements result in acoustic events that are auditorily salient to the perceiver. In this chapter, the problem of speech perception, adults‘ auditory speech perception, and the development of auditory speech perception are presented.

3.2 **The Problem of Speech Perception**

Here, the four fundamental issues of speech perception are discussed briefly; *levels of language processing, invariance problem, talker variability, and segmentation.*

3.2.1 **Psychological Reality of Levels of Speech Processing**

Three phonological levels of speech perception can be discriminated: Phonemic, phonetic, and acoustic. Evidence for these levels of processing comes from experimental studies. Werker and Tees (1984b) presented adult speakers of English with three types of contrasts: same phones with different acoustic variations (/tæ/ vs. /tə/ and /tɑ/ vs. /tɑ/), Hindi-native but English non-native voiceless unaspirated dental contrast /tə/ vs. /tɑ/, and English and Hindi-native dental contrast /tɑ/ vs. /dɑ/. These contrasts were presented with three different inter-stimulus intervals (ISI): 250 ms, 500 ms, and 1500 ms using an AX discrimination task. Results showed that English speakers were only able to perceive the difference when ISI was set at 250 ms, and 500ms but not at 1500 ms. The authors reasoned that when speech contrasts are presented at longer ISIs the acoustic information in short term memory would decay such that perceivers would be forced to process these
sounds in terms of phonetic characteristics up until 500 ms, and then beyond this period only in terms of their phonemic language-specific characteristics. Thus at 250 ms ISI all acoustic variations are available, at 500 ms ISI no acoustic variations are available as sounds have been perceived as speech and only phonetic/phonemic information is available, and at 1500 ms ISI no phonetic variations are available as sounds have been transferred to a long term memory store and coded in terms of their phonemic language-specific properties. In a subsequent study, Werker and Logan (1985) tested adult speakers of English and Hindi on the voiceless unaspirated dental Hindi stop consonant contrast /\textipa{\textipa{t}a}/ vs. /\textipa{\textipa{t}a}/ in three ISI conditions as in the Werker and Tees (1984b) study. These two sounds exist in English as allophones of [t]. It was found that Hindi speakers were able to discriminate this contrast at all three ISI levels, whereas English speakers were able to discriminate at 250 ms ISI condition, but not in 500 or 1500 ms ISI condition.

These (Werker & Logan, 1985; Werker & Tees, 1984b) and other studies show that humans can process language at three different levels which become apparent under different task requirements. Under normal listening conditions, however, humans would usually operate at the phonemic level.

### 3.2.2 Acoustic Cues, Co-articulation, and the Invariance Problem

In the mid-1950s researchers at Haskins Laboratories studied the identification of invariant speech units that would assist identification of phoneme classes by perceivers with the eventual aim of building a speech recognition machine. In one such study Delattre and his colleagues separated coarticulated consonant-vowel (CV) context speech segments in order to identify what part of the syllable would exclusively be perceived as the consonant. However, they were unable to find a point at which a coarticulated syllable could be split such that an isolated individual consonant would be perceived independent of the adjacent vowel (Delattre, Liberman, & Cooper, 1955). Thus no particular set of acoustic cues could be isolated which uniquely identified a particular consonant. These studies later on led to the
formulation of *Motor Theory of Speech Perception*, which claims that speech perception is an articulatory rather than an auditory event and in general predicts that articulatory movements (lip and tongue movements, vocal fold vibrations) of speakers provide perceptual and neuromotor feedback to perceivers, who in turn decipher the speech signal. A further detail of this theory is provided in 3.3.1.

3.2.3 *Talker Variability*

Talker variability refers to the normal set of affairs in which two or more speakers or the same speaker in different occasions differ in their articulations of the same target output. Despite such variability in terms of loudness, $F_0$, voice quality, etc., the same target productions are perceived to be the same across contents, talkers and listeners. Speech perception models have been set up to explain how rules are extracted and speech prototypes established from the native (ambient) language as a result of exposure to uncountable instances of speech input by countless talkers (e.g. Best, 1995; Kuhl, 1991), however consideration of these models with respect to talker variability issue is beyond the scope of this thesis.

Human auditory perception is sophisticated enough to deal with talker variability from very early in life (Jusczyk, Pisoni, & Mullenix, 1992). The issue of talker variability in the domain of auditory-visual speech perception is a more recent area of interest. In a second language acquisition study, Hardison (2003) investigated the role of talker variability in the acquisition of non-native speech contrasts and found an advantage of training perceivers by exposing them to auditory-visual presentations of multiple talkers uttering the target contrast versus a single talker$^{11}$.

There are various other sources of talker variability, which lead to phoneme variation, such as speech rate and physiological differences. For example, the rate at which a speaker talks varies considerably depending on the speech

---

$^{11}$ Hardison (2003) notes that the difference was only marginally significant and this may be due to the clarity of the speech input provided in both groups.
context (e.g. school teachers’ speech vs. casual conversation between adults).
In terms of physiological differences, one factor that contributes to talker
variability is the size of the vocal tract: the larger the vocal tract, the lower the
resonances, hence children and adult females, who have shorter vocal tracts,
have higher fundamental frequencies than adult males, who have longer
vocal tracts.
Despite the gradually increasing sophistication of automatic speech
recognition (ASR) systems, the issue of talker variability still poses great
challenges to engineers who design and try to advance these systems
(Gürbüz, Tüfekçi, Patterson, & Gowdy, 2001) and the challenge ahead of
researchers working on automatic speech recognition systems is to capture
and reproduce the unique human ability to cope with talker variability.

3.2.4 Segmentation
Speech is typically a continuous acoustic stream but it is perceived as
segmented series of words, phrases and sentences. Segmentation is not
apparent to native speakers listening to their own language, however, when
perceivers attend to a language with which they are not familiar, they find the
task of parsing or segmenting the incoming speech input into words or even
sentences quite challenging. Under such circumstances, cues such as stop
consonants might be erroneously treated as word boundary markers (Martin,
1970). Research has shown that humans employ sophisticated processes to
tackle the segmentation challenge in their native languages and languages
with which they are familiar such as attention to the prosodic requisites and
peculiarities of particular languages. Thus speech processing at the phonemic
level, i.e., in this case in the native language, allows segmentation whereas
speech processing at the phonetic level with an unfamiliar foreign language
does not allow segmentation (Cutler, 1997).

3.2.5 Categorical Speech Perception
Another way in which acoustic variability gives way to phonetic and
phonemic invariance is in the phenomenon of categorical speech perception.
In categorical speech perception, a continuum of speech sounds varying along a physical dimension is perceived in terms of distinct categories. In the original demonstration of the categorical perception of speech, Liberman, Harris, and Griffith (1957) presented a continuum of synthetic consonant-vowel (CV) syllables which varied in steps along the continuum in terms of their F2 transition; from /bV/, through /dV/, to /gV/in both an identification and a discrimination task. They found two remarkable results: First, response patterns from the identification task yielded abrupt boundary changes between phoneme categories despite the physical equal-step continuum, and second, the discrimination experiment revealed that discrimination accuracy was at chance level when stimulus pairs were within a category (with boundaries defined by the identification functions) but almost perfect when the pairs were on or near the category boundaries. The existence of this phenomenon led the authors to posit a special speech module, and later led to the development of the Motor Theory of Speech Perception (Liberman, Cooper, & Schankweiler, 1967). While it has subsequently been found that some discrimination within speech categories is possible, the general phenomenon of categorical speech perception provides another clue to how humans parse the variable acoustic stream into distinct segmental phonemes, words, and sentences.

### 3.3 Models of Auditory Speech Perception

In the past few decades a number of theories and models on speech perception have been advanced. Theories and models of speech perception have evolved out of two central questions (Mitterer & Cutler, 2005): whether there is an innate specialisation for speech perception (e.g. Native Language Magnet Theory), which is distinguished from auditory perception or whether speech perception is linked to speech production (e.g. the Motor Theory of Speech Perception). In this section the theories of speech perception considered are the Motor Theory of Speech Perception, the Perceptual Assimilation Model, and the Native Language Magnet Theory.
3.3.1 Motor Theory of Speech Perception

Following on from the studies of categorical speech perception described above (see 3.2.5), researchers at Haskins Laboratories (e.g. Liberman, Delattre, & Cooper, 1952; Liberman et al., 1957) developed the Motor Theory of Speech Perception (for a revised version of the Motor Theory, see Liberman & Mattingly, 1985). Motor theory refers to the articulatory movements of talkers as sources of speech signal, a process that requires a significant degree of visualisation of the speech. In its essence it is a theory that strongly links speech perception to production. Motor theory claims that the objects of speech perception are articulatory events rather than acoustic or auditory events and predicts that articulatory movements (lip and tongue movements, vocal fold vibrations) of speakers give rise to neuromotor commands to the articulators of perceivers. Liberman and his colleagues term these neuromotor commands intended gestures (Liberman et al., 1967; Liberman & Mattingly, 1985). Liberman et al. (1967) suggest that the process of speech production is characterised by a series of sequential steps: (a) phonemes; (b) neuromotor commands; (c) muscle contractions; (d) vocal tract shapes; and (e) acoustic signals.

Despite the original conception of a one-to-one relationship between phonemes/distinctive features and muscle contractions and vocal tract shapes because of the effects of co-articulation (see 3.2.2) the theory was later amended to reflect the rather more complex relationship among these factors. For example, when spectrographs of the utterances /di/ and /du/ are inspected, the values of the first two formants carry a significant amount of information with respect to the initial consonant identity. In both of these two syllables the rising F₁ value signifies that the initial consonant is a ‘stop’ consonant. In addition, the rising F₂ transition of /di/ and the falling F₂ transition of /du/ indicate the alveolar place of articulation. Such correlations between formant values and places of articulation led the proponents of the Motor Theory to conceive a robust perception-production link. That is the Motor Theory asserts that when people perceive speech they also ‘produce’ it
(Liberman & Mattingly, 1985) and that the perception of articulatory movements for speech perception is mandatory, which requires a high degree of visualisation of or seeing speech organs. Essentially, the theory asserts that both acoustic and visual speech inputs provide information about the same speech gesture throughout the speech perception process (Kerzel & Bekkering, 2000).

Findings from research in dialogue processing is consistent with the motor theory’s assertion that speech perceivers are not independent processors but speech perception and production are strongly aligned processes at many levels. This is such that interlocutors’ comprehension and production mechanisms are aligned and interact during the dialogue process (Pickering & Garrod, 2004). Recent data from neuroscience also support the assertion of the motor theory that a certain type of neurons, referred to as mirror neurons, located in Broca’s area are responsible for controlling otolaryngeal and orofacial movements needed to produce speech. According to this view, mirror neurons represent the link between speech production and perception as they were found to be active in response to speech input and during speech output (Rizzolatti & Arbib, 1998).

The motor theory has strong links with direct realist notions in which the objects of speech perception are contained in the stimulus array. In summary, motor and the direct-realist theories of speech perception state that speech perception and production are two processes which share a common metric, and strongly linked. Direct realism has links with Gibsonian notions of ecological perception, which underlines the Perceptual Assimilation Model, considered next in 3.3.2.

3.3.2 Perceptual Assimilation Model

The Perceptual Assimilation Model (PAM) has its theoretical origin in a very influential general perception theory, ecological perception (J. J. Gibson, 1979). According to ecological view, perceivers are endowed with integrated perceptual systems, which have evolved to pick up structural invariant
information from the environment. Thus the perceiver and environment work in unison: In a process known as perceptual learning, perceivers pick up information (or invariants in Gibsonian terms) from the environment as a result of changes in the stimulus flow (E. J. Gibson & Pick, 2000). In addition, species evolve to specialise to pick up certain kind of information, which afford the activities they require for their physical and social well-being: flying for eagles, foraging for bees, and speaking for humans. In humans, in the course of native language acquisition, infants are gradually attuned to the phonological structure of their native language, and by the time native speech production is fluent, phonological information pertaining to native speech is processed without effort (Best, 1995).

PAM uses gestural phonology to explain the structured nature of speech perception. Gestural phonology refers to the coordinated actions of the vocal tract resulting from the spatiotemporal movements of several articulators over space and time. Over the course of development certain patterns of gestural phonology provide the stimulation required for the formation of speech sound categories in the native language. PAM explains human speech perception phenomena on the basis of the perception of non-native speech sounds. In essence PAM suggests that non-native speech sounds are perceived in accordance with their similarities to and discrepancies from sounds in the native phonological space. There are six ways in which non-native speech sounds are processed:

Two-Category Assimilation (TC): In the case where one non-native sound is assimilated into a native category and a second non-native sound into a second native category, the two would be perfectly discriminated. In one study Best and her colleagues tested native English-speaking adults on various Zulu- and Tigrinya-specific contrasts and found that when two non-native sounds could be assimilated in separate categories, good discrimination resulted. For example, perceivers showed very good discrimination of the Tigrinya ejective contrast /p/-/t/, assimilating the two
sounds into the English-native voiceless bilabial /p/ and English-native voiceless alveolar /t/.

**Category-Goodness Difference (CG):** Here two non-native sounds are assimilated into the same native category but diverge markedly from the category prototype. Discrimination between them can be expected to be good.

**Single-Category Assimilation (SC):** This pattern is similar to CG, in that two non-native sounds are assimilated into the same category. However, they are both close phonetically to the prototype, resulting in poorer discrimination. An example is the poor discrimination of English word final position /t/ and /d/ by native Mandarin speakers, whose L1 does not allow word-final obstruents (Flege, 1989).

**Both Uncategorisable (UU):** Two non-native sounds are assimilated into the native phonological space, but are not placed in any native category. Discrimination of these non-native sounds would range from poor to excellent depending on their phonetic distance from each other and from native categorical prototypes.

**Uncategorised/Categorised (UC):** One non-native sound is categorised and falls in the native phonological space, and the other one does not. PAM predicts the discrimination to be very good.

**Nonassimilable (NA):** Both non-native sounds fall outside the native phonological space and are not categorised. They are perceived as non-speech sounds. Discrimination may range from very poor to very good depending on the salience of their phonetic difference. For example, English-speaking American adults perceive Zulu clicks as non-speech sounds, due to their double-stop articulations and suction release gestures (Best, McRoberts, & Sithole, 1988), but discriminate these very well.

Essentially PAM asserts that when an adult listener attends to non-native speech, (s)he detects native gestural invariants and the non-native speech sounds are assimilated in one of the six ways described above. What differentiates PAM from other speech models is that it gives special credence
to the role of articulation and indicates that perceptual categories/classes of speech are organised as a result of articulatory experience.

3.3.3 Native Language Magnet Theory

Native Language Magnet Theory also posits the build-up of perceptual classes via experience, but it is perceptual rather than perceptual/motor experience that is important. In the past two decades, Kuhl and her colleagues (Grieser & Kuhl, 1989; Kuhl, 1991, 1992, 1995, 1997; Kuhl, Williams, Lacerda, & Stevens, 1992) have maintained that, as a result of exposure to native speech sounds, native vowel categories are perceptually reorganized around prototypical instances, a process which begins around 4-6 months of age (Kuhl, 1991; Kuhl et al., 1992). Once prototypes are formed, sounds that are closer to the prototype are perceptually drawn to the prototype, whereas poor instances or non-prototypical sounds do not act in this fashion, an effect called the magnet effect. Kuhl (1991) tested this magnet effect with human infants, human adults and adult rhesus macaques on their perception of instances of the /i/ vowel with reference to a prototypical /i/ vowel, established on the basis of human adults’ judgments in the first part of the study. She found that adults’ perception of the instances of the vowel /i/ was strongly affected by the prototypical /i/. Additionally, infants’ perception of /i/ had a similar pattern as their adult counterparts by the age of 6 months; American English infants showed a magnet effect for the native language /i/, but not for the Swedish /y/, while Swedish infants showed a magnet effect for /y/ but not for /i/ (Kuhl et al., 1992). In tests with rhesus macaque monkeys, there was no indication of a magnet effect.

As a result of these studies, Kuhl and her colleagues developed the Native Language Magnet (NLM) Theory, which proposes a three-stage process for the development of vowel perception in the first year of life. In the first stage, infants’ perception of speech sounds is defined by language-general, acoustic properties, enabling them to discriminate most, if not all, of
the world’s consonant and vowel contrasts. In the second stage, at about 4 to 6 months of age, upon having heard thousands of instances of vowels, infants store this information in the form of prototypes for each vowel. It is at this stage when infants begin to show the magnet effect (Kuhl et al., 1992), in which sounds that are perceptually close to the prototypical vowel are not perceived to be different from the prototype. In the third stage, (6 months of age and onwards), the magnet effects become more apparent such that perceptual boundaries that once divided different vowel sounds are rearranged on the basis of native sounds. As a result, native phoneme categories, at the centre of which the prototypes are located, are formed.

Despite the fact that the NLM theory was designed to explain the development of vowel perception, Kuhl and her colleagues have investigated the development of liquids in the light of this theory. Using a multidimensional scaling (MDS) technique, Iverson and Kuhl (1994; 1996) generated a perceptual map of the American English /r/ and /l/ on the basis of the perception of a set of /r/ and /l/ stimuli generated by varying the F2 and F3 components. They found that perceived distances amongst these stimuli differed significantly from the actual physical distance. Perceived distances were rather distorted such that there were clear groupings around the prototypical /r/ and /l/ (See Figure 3-1).

![Figure 3-1](image_url)

**Figure 3-1. Differences between actual (A) and perceived distances (B) amongst the stimuli used by Iverson and Kuhl (1994; 1996) with American subjects. With perceptual distances around them reduced, the prototypical /r/ and /l/ act as perceptual magnets pulling the instances perceptually closer to them. The larger the circles are in (A) the better they represent the category prototype. Black dots represent /r/ and white dots represent /l/. Figure adapted from Iverson et al. (2003).**
Later Iverson et al. (Iverson et al., 2003) tested English-, Japanese-, and German-speakers and showed that language-specific processing alters the /r/ and /l/ category boundaries differently for these three speaker groups. While German speakers’ perceptual spacing approximates that of English speakers (despite a sharp distinction between the English and German /r/ prototypes), Japanese speakers’ perceptual spacing of the stimuli seems to rely more on acoustic cues such as F2, but insensitivity to F3 as opposed to other two speaker groups. Thus it can be seen that the Native Language Magnet Theory allows for and explains the way in which the different prototypes built up as a product of experience affect later speech perception.

3.3.4 PRIMIR

While to some extend they talk about the role of visual speech information in speech perception (e.g., motor theory) the data on which the above three models based are from auditory-only stimuli. A recent model, PRIMIR (developmental framework for Processing Rich Information from Multidimensional Interactive Representations), advanced by Werker and Curtin (2005) refers to the richness of linguistic input. For instance while PAM and NLM focus on gestural and acoustical information as predominant sources of speech perception respectively, PRIMIR takes all information relevant to the speech input into account, such as gender, prosody, visual speech, object-word pairings, and so on. The primary assumption of PRIMIR is that a wide array of rich information is available with the speech input and perceivers simply pick up this information along a number of multidimensional interactive planes. These interactive planes are general perceptual plane, word forms, and phonemic plane. The interaction of these processors with speech input is realized by three types of dynamic filters: initial biases, task demands, and developmental level. For example, PRIMIR predicts that young infants will detect any discriminable phonetic changes on the general perceptual plane, while older infants will be able to discriminate
language-specific categories that develop during the first year of life. Furthermore, in the auditory-visual speech domain, PRIMIR predicts that adults’ use of language-specific phonetic information would be evident in tasks that require phoneme identification and the integration of auditory-visual speech codes as in the McGurk effect (e.g., Werker, Frost, & McGurk, 1992). In developmental terms, at the initial stages of language development, infants will detect invariances in the general perceptual plane, and as a function of experience with native language, they will start forming meaning-sound associations in the process of building their lexicon. As the vocabulary expands and more words with common features are added, higher order processing on the phonemic plane will take place, e.g., adults’ sophisticated speech perception system will help integrate auditory and visual phonetic information on the phonemic plane.

In this section, ordered chronologically, the descriptions of four models that aim to account for auditory-only speech perception were given. In the next section we turn our attention exclusively to how human auditory speech perception develops as functions of age and language-specific experience.

3.4 CROSS-LANGUAGE AND DEVELOPMENTAL STUDIES OF AUDITORY SPEECH PERCEPTION

In this section the development of human auditory speech perception is considered. First the methods of developmental investigation are described. This is followed by a description of speech perception abilities based on native and non-native speech perception in the first year of life. The three final subsections review the development of three types of speech segments, consonants, vowels, and tones, from infancy through to adulthood. Finally, a stage-based heuristic of speech perception development is presented.

3.4.1 Methods of Cross-Language and Developmental Study

Developmental studies of human behaviour are generally conducted by means of comparing different groups of participants. Especially in the case of
psycholinguistic research, this is achieved in two ways: *ontogenetic* and *differential* methods (Burnham & Sekiyama, in press).

### 3.4.1.1 The Ontogenetic Method

In terms of psycholinguistic research, the use of the *ontogenetic* method involves the comparison of individuals of different ages who have been exposed to the same linguistic environment, and measurement of their performance on a specific task. Speech perception researchers who use the ontogenetic method are interested in the effect of the *amount* of experience on speech perception. For example, the ontogenetic method would be used by a researcher interested in the discrimination of Cantonese tone contrasts by native Cantonese speakers of four different age groups.

### 3.4.1.2 The Differential Method

The *differential* method, as opposed to the ontogenetic method, is concerned with the *type* of experience. In psycholinguistic research, this involves the comparison of individuals whose ages are similar but who were brought up in different linguistic environments. For example, a researcher may compare the effect of having been exposed to English versus French language environment on the perception of stress patterns peculiar to French.

### 3.4.2 *Infancy: The Beginnings of Speech Perception*

The human auditory system is fully functional by the beginning of the third trimester of gestation. Accordingly speech perception begins prenatally (Bredberg, 1985). Human speech sounds from the extra-uterine environment reach the human fetus via the low pass filter of the womb (Querleu, Renard, Versyp, Paris-Delrue, & Crepin, 1988), and in the case of the mother’s voice, via bone conduction. Despite the resultant degradation, the signal still carries sufficient prosodic information for the neonate’s preference for their mother’s voice (Cooper, Abraham, Berman, & Staska, 1997), and their native language (Mehler, Bertoncini, Barriere, & Jassik-Gerschenfeld, 1978).
While there are a very large number of speech sounds across the world’s languages, every language employs only a sub-set of these (Maddieson, 1984). Studies over the past four decades have consistently shown that neonatal human infants are capable of perceiving most, if not all, phonetic contrasts of the world’s languages. This appears to hold whether or not speech contrasts presented to infants are relevant to the language(s) spoken around them. As infants are exposed to the language(s) around them their seemingly universal, language-general ability to perceive speech contrasts is gradually reorganised, and becomes more language-specific. By adulthood perceivers are most sensitive mostly to speech sounds that are phonemically relevant in their native language (but also see Best et al., 1988) and least sensitive to phonemically irrelevant speech contrasts. A typical example of infants’ early speech perception abilities is the case in which Japanese infants can discriminate the non-Japanese [r] vs. [l] contrast with ease (Christophe & Morten, 1994), whereas their adult counterparts are unable to perceive this difference (Goto, 1971) unless they have specific experience or training (Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999). Such training studies show that the developmental reorganisation of speech perception occurs as a result of a change in perceptual strategy rather than any loss of sensorineural sensitivities (for reviews, see Burnham, Tyler, & Horlyck, 2002; Werker & Tees, 1992).

What purpose does the attenuation and reorganisation of early speech abilities as a result of exposure to native language serve? One answer is that a sophisticated phonemic processing ability, which is fine-tuned to the native language phonology, enables perceivers to segment the speech stream based on phonological and prosodic information. This ability is also used to develop knowledge of syntactic rules and semantic (word-sound) associations in the native language.

Research has shown that the attenuation and reorganisation of early speech perception abilities follows slightly different timeline and pattern for vowels.
and consonants. Developmental evidence is presented below in terms of the
perception of speech segments: consonants, vowels and also tones.

3.4.3 Perception of Speech Segments over Age

3.4.3.1 Perception of Consonants
Young infants are sensitive to consonant contrasts differing on single phonetic
features such as voicing (Eimas, Siqueland, Jusczyk, & Vigorito, 1971),
manner of articulation (Eilers & Minifie, 1975), even for consonant contrasts
that are irrelevant in their ambient language (Best et al., 1988; Werker, Gilbert,
Humphrey, & Tees, 1981). However, these seemingly universal speech
perception abilities for consonant perception is reorganised sometime
between 7 and 11 months of age (Werker et al., 1981; Werker & Tees, 1984a),
such that discrimination ability for non-native speech contrasts is attenuated.
Werker and her colleagues conducted a series of studies on the nature of the
perceptual attenuation of non-native speech contrasts and the level at which
native and non-native speech contrasts are processed. Werker and Tees
(1984a) tested three groups of English language environment infants (6-8-
month-olds, 8-10-month-olds & 10-12-month-olds), English language
environment adults and native Nthlakampx-speaking adults on the
Nthlakampx-native, but English-non-native glottalised velar vs. glottalised
uvular /kí/-/qí/ distinction. Results showed that the performance of
Nthlakampx-native speakers and English language environment infants was
better than English language environment adults, and the difference between
Nthlakampx-speaking adults and English language environment infants was
not significant. More specifically Werker and Tees (1984a) found that the
performance of English language environment infants on the Nthlakampx-
native contrast showed a decline over infant age. In other words, younger
infants were better than their older counterparts at discriminating a non-
native Nthlakampx-native contrast on the basis of phonetic categories.
Despite the general finding that adults’ perception of speech becomes more
language-specific over time, a number of studies have shown that adults can
discriminate some non-native consonants on the basis of their acoustic salience (Best et al., 1988; Werker & Logan, 1985). For example, Best et al. (1988), tested English-speaking infants and adults on the discrimination of non-native Zulu click contrasts and found that adults’ and infants’ performance were comparable. They suggest that adults make this non-native discrimination on the basis of salient acoustic properties of the speech sounds rather than language-relevant, phonemic characteristics. Thus it appears that some consonant contrasts do not undergo reorganisation when the cues on which they are based are lacking in linguistic salience for a particular language group. (see also Best’s PAM in 3.3.2).

Despite this reorganisation for speech sounds, studies in which ISI is manipulated (Werker & Logan, 1985; Werker & Tees, 1984b) or extensive training is given show that adults still have the ability to perceive these non-native speech sounds (Werker & Tees, 1984b) (see 3.2.1). This underlines the notion that speech perception reorganisation occurs at an attentional level rather than a sensorineural level.

3.4.3.2 Perception of Vowels

Infants as young as 4 weeks can discriminate vowels both in consonant contexts ([pa] vs. [pi] and [ta] vs. [ti]), and isolated ([a] vs. [i] and [i] vs. [u]) contexts (Trehub, 1973). Moreover, 5.5-month-old and 6.5-month-old infants can discriminate different occurrences of native vowel contrasts belonging to the same category (Kuhl, 1983). In addition, Swoboda, Morese and Leavitt (1976) have found that 2-month-old infants can discriminate between and within category vowel differences ([i] vs. [i]).

With regard to perceptual reorganisation, Polka and Werker (1994) tested two groups of English language environment infants aged 6-8 months and 10-12 months on two pairs of non-native German vowel contrasts, [v] vs. [U] and [y] vs. [u] in a dVt context. They found that discrimination performance of the younger group was superior to the older one. In a second experiment, they tested a group of 4-month-olds and another group of 6-month-olds on the
same stimuli, and found that the performance of 4-month-olds was significantly better than 6-month-olds. Together these results indicate that, 6 to 8 months is a transitory period for the decline of the language-general ability to discriminate vowel contrasts. Overall, findings on the perception and reorganisation of native and non-native vowels may rather be a more complex process than that for consonants. On the one hand a number of studies by Kuhl (1995) suggest that prototypical instances of speech sounds (both vowels and liquids) act as “perceptual magnets” (see 3.3.3), from as early as about 6 months of age (Kuhl et al., 1992). On the other hand it also appears that there are some perceptual magnets, which are independent of language background. Polka and Bohn (1996) have found that more extreme vowels in terms distance from the centre of F1-F2 vowel space are discriminated more easily irrespective of language background of the perceiver and phonological relevance of the sounds. Moreover, despite both consonants and vowels following a language-general to language-specific perceptual tuning pattern, the perceptual reorganisation of vowels takes place earlier than that for consonants (see Burnham, 1986; Kuhl et al., 1992). One possible reason for this is that consonants appear to be more language-universal, that is, more consonants are shared by world’s languages than are vowels, hence the perceptual system may be tuned sooner for vowels than for consonants. In addition, vowels carry significant amounts of prosodic and language-specific information, such as stress, speaker identity, accent, and emotion. Specific earlier perceptual tuning for vowels makes survival sense since affect and emotion, important communicative elements for infants, are carried by vowels, and caregivers teach their infants about the nature of the ambient language environment by means of their hyperarticulation of vowels (Burnham, Kitamura, & Vollmer-Conna, 2002).

### 3.4.3.3 Perception of Tones

Compared to consonants and tones, infants’ perception of tones is relatively uncharted territory. Using a Thai lexical tonal contrast ([bâ] vs. [bâ], both of
which are perceived as /ba/ by native adult English-speakers), Mattock and Burnham (2006) tested 6-and 9-month-old English language environment and Cantonese language environment infants. They found that at 6-months English language environment infants can discriminate Thai tonal contrasts, but that level of performance declined by 9-months of age. In contrast, Cantonese language environment infants were found to be able to discriminate tonal contrasts equally well at 6 and 9 months of age. This provides evidence that the perception of tones follows a similar pattern as consonants and vowels, such that their perception is attenuated around the second half of the first year if they are phonetically irrelevant in the ambient language.

In a number of studies Burnham and his colleagues found that native English-speaking adults can discriminate tonal contrasts with relative ease (Burnham, Francis, & Webster, 1996; Burnham & Torstensson, 1995). Burnham, Francis and Wester (1996) found that while school age children perceived tone distinctions less well than non-native vowel distinctions, the opposite was the case for adults. They explained this in terms of the alien nature of linguistic tone for children, and the psychoacoustic salience of F0 contours for adults who could treat the experimental task in a non-linguistic manner. Such psychoacoustic salience is consistent with other non-native contrasts perceived by adults, in particular click contrasts (Best & Avery, 1999; Best et al., 1988; Best, Traill, Carter, Harrison, & Faber, 2003).

3.5 A Heuristic of Speech Perception Development

Speech perception development can be described in terms of 4 stages (Burnham et al., 2002): Phonetic, Phonemic, Semantic, and Orthographic, a heuristic that provides a framework for developmental aspects of the current thesis.

3.5.1 The Psychoacoustic/Phonetic Stage (0-6 months)

Infants are born with sophisticated speech perception abilities: due to prenatal exposure to the mother’s voice through the womb and bone conduction,
neonates prefer their mothers’ voice (DeCasper & Fifer, 1980), their native language (Mehler et al., 1988) and can discriminate native language rhythm (e.g., stress-timed English) from non-native language rhythm (e.g., mora-timed Japanese) (Nazzi, Bertoncini, & Mehler, 1998). In addition, they can discriminate just about any speech contrast presented to them from the world’s languages: consonants (Werker et al., 1981), vowels (Trehub, 1973), or tones (Matlock & Burnham, 2006). They achieve this predominantly on the basis of acoustic characteristics of the speech sounds they hear. In other words, it is not the case that human auditory system has evolved to specially process speech sounds, but rather that human languages have evolved to make best use of the human auditory system (Burnham et al., 2002). Animals’ human-like perception of human speech sounds (Kuhl, 1981; Kuhl & Miller, 1978) and infants’ categorical perception of non-speech sounds (Jusczyk, Rosner, Cutting, Foard, & Smith, 1977) indicate that there is a psychoacoustic basis on which the language and its phonology is built. Hence, it appears that there is a phonetic basis of speech and language processing built upon a strong psychoacoustic basis in the first few months of language development.

3.5.2 The Phonemic Stage (6-12 Months)

After around the first half of the first year of life, infants become increasingly tuned to their native language, such that their earlier ability to discriminate non-native speech contrasts is attenuated and their performance for native speech contrast increases. As a result of experience with the ambient language, native language prototypes for vowels (Kuhl, 1992) are formed and consonant perception becomes more language-specific (Werker & Tees, 1992). Infants’ language-general speech perception is gradually attenuated and becomes more language-specific, ignoring phonemically irrelevant speech contrasts. Tuning attention towards native speech contrasts and away from non-native ones prepares infants for sound-meaning pairings, in other words vocabulary acquisition, a process which marks the third stage of this speech perception development heuristic.
3.5.3 The Semantic Stage (12-20 Months)

In the semantic stage children’s attention is shifted to associations that are formed between objects and the native speech sounds they hear around them. In order to investigate how speech perception abilities are invested in word acquisition tasks, Stager and Werker (1997; also see Werker, Cohen, Lloyd, Casasola, & Stager, 1998) tested 8- and 14-month-old infants. In a discrimination task the infants were presented with a minimally different syllable pair /bɪ/ vs. /dɪ/ in two contexts: word learning and perceptual. The results showed that 8-month-olds could discriminate the minimal pairs in both contexts, whereas 14-month-olds could discriminate the pair in the perceptual but not in the word-learning context. The authors indicate that 14-month-olds’ attenuated ability to process fine phonetic details is similar to infants’ gradual attenuation of perception of non-native speech contrasts. As the older infants move from speech perception to word learning, they shift their attention to sound-object associations, in essence word learning, and away from fine phonetic details. On the other hand, in 8-month-old infants the two contexts are functionally similar such that they process only the phonetic details of the minimal pairs presented. Vocabulary acquisition requires a large investment of attentional resources, and Stager and Werker (1997) argue that such cognitive investment in sound-object pairings temporarily disables older infants from attending fine-grained phonetic differences. A subsequent study has showed that 14-month-olds but not 8- and 12-month-olds can readily learn to make sound-object pairings (Werker et al., 1998). These results indicate that there appears to be a new functional reorganisation of speech perception abilities that paves the way for vocabulary acquisition at the start of second year of life.

3.5.4 Orthographic Stage (6-8 Years)

Burnham, Tyler and Horlyck (2002) indicate that there is a fourth stage of speech perception development with the onset of new formal language skills. The evidence generally comes from studies showing a relationship between
children’s speech perception abilities and formal language skills, particularly reading. Burnham, Earnshaw and Clark (1991) tested English language environment 10-month-olds, 2-year-olds, 6-year-olds, and adults on the voiceless-voiced bilabial native contrast /p/ vs. /pʰ/ and non-native (Thai-native) voiced-prevoced bilabial contrast /p/ vs. /b/. The results generally showed an age-related increase in native contrast performance but there was a trough at around age 6 on the performance of the non-native contrast, a pattern similar to the previous two stages of speech perception development during which certain aspects of speech perception abilities are attenuated as a result of certain type of linguistic experience. Burnham et al. reasoned that the decline in performance on the non-native contrast might be due to the influence of reading acquisition, which may cause a shift of attentional resources in order to master phoneme-grapheme correspondences in native language orthography.

In a follow-up study, Burnham (2003) tested 4-, 6-, and 8-year-olds on their perception of the same native and non-native speech contrasts and their reading ability. The results showed that 4-, and 8-year-olds’ performance on non-native speech contrast was better than their 6-year-old counterparts. Burnham also investigated the link between language specific speech perception and reading ability. Language specific speech perception is an index of the difference between performance on non-native speech contrasts and it is measured by the difference between perception for native speech contrasts and perception for non-native speech contrast. Burnham (2003) showed that language specific speech perception was positively related to children’s reading ability, that is children who were good readers showed a greater bias towards speech contrasts that are phonemically relevant in their language, and away from those contrasts which are phonemically irrelevant.

Later Horlyck and Burnham (under review) tested 5- and 6-year-olds with varying levels of school experience in order to control for and test the effect of amount of school experience: 5-year-olds with no school experience, 5-year-olds with 6 months of school experience, 6-year-olds with 6 months of school
experience, and 6-year-olds with 18 months of school experience. They found that language specific speech perception did not differ as a function of age but did as a function of experience. Horlyck and Burnham also found that children’s school experience and phonemic awareness ability predicted the level of their language specific speech perception.
Overall, these results suggest that the children’s limited attentional resources are shifted to phonemically relevant aspects of their native language as a result of exposure to certain linguistic challenges. In this stage of speech perception development the process of reading skills acquisition appears to call for or result in a bias in favour of native speech perception, and against non-native speech perception, hence the development of language specific speech perception.

3.6 Summary
In this chapter the basic issues of speech perception were described and a review of auditory speech perception development was presented. Overall, the research shows that at particular milestones of development, certain aspects of speech perception abilities are attenuated while others, particularly those relevant in the native language, are augmented. Towards the end of the first year of life the perception of speech segments go through a functional reorganisation such that the perception for non-native speech contrasts becomes poorer (e.g. Werker & Tees, 1984a). Similarly at the start of the second year of life children start to allocate their attention towards sound-meaning associations in order to create native vocabulary and away from fine phonetic details (e.g. Stager & Werker, 1997). Later with the onset of school children shift their attention away from phonemically irrelevant speech contrasts and towards phonemically relevant speech contrasts in order to tackle the challenges of phoneme-grapheme correspondences in reading. However, the challenges of speech perception for developing children are not just auditory, speech perception is not just an auditory phenomenon. As indicated at the end of Chapter 2, the articulatory process has at least three
kinds of perceivable consequences: auditory, visual and tactile. In the next chapter the auditory-visual (and tactile) aspects of speech perception and how it develops is provided, and the basic questions for the question of current thesis are advanced.
CHAPTER 4

AUDITORY-VISUAL SPEECH PERCEPTION AND DEVELOPMENT
4.1 Articulation as a Visual Speech Event

As indicated at the end of Chapter 2, articulatory information has various perceptual consequences for the perceiver: auditory, visual, and tactile. The auditory aspects of articulation have been covered in Chapter 3. This chapter focuses on the visual along with the auditory, consequences of articulation, in other words auditory-visual speech perception, and how it develops over age. Traditionally speech has been considered to be an auditory event. However, a good deal of information is also provided by other sources (visual and tactile). When speech is produced in a face to face situation, the movement of the speech articulators, such as the opening, closure or rounding of the lips, lowering of the jaw, or visibility of teeth provide a plethora of information to the perceiver, especially in noisy environments (Sumby & Pollack, 1954). In this chapter the multimodal aspects of speech perception, namely tactile and orthographic, and auditory-visual speech perception and its development are discussed. Towards the end of the chapter the research questions of the current thesis are provided.

4.2 The Psychological Reality of ViseMES

Phonemes are distinctive phonological units processed at the segmental level of language (Bernstein, in press). On the other hand, visemes can be defined as visual speech categories (Massaro, 1987). The aim of this section is twofold: to describe and investigate the perceptual reality of visemes. When perceivers are presented with talking faces with no auditory input (as in the case of speech perception by deaf perceivers in face-to-face communication), they are able to perceive to some extent what the talker utters. Categorical visual speech categories, which help perceivers to identify the speech signal, are called visemes (Fisher, 1968). Several studies showed that the number of viseme categories range from 4 to 7, and these depend on such factors as talker variability, stimuli and viewing conditions (Owens & Blazek, 1985). Woodward and Barber (1960) tested participants’ lipreading ability on visual-only CV-context syllables in a discrimination task. They
found four contrastive visual speech categories: labial, rounded labial, non-
labial, and labiodental; and that members within each category were homophenous (M. F. Woodward & Barber, 1960), that is they sounded
different, but were visually indistinguishable (Bernstein, in press). In a later
study, Walden, Prosek, Montgomery, Scherr, and Jones (1977) tested 31
hearing-impaired adults on their ability to identify English language
consonants. Their initial results revealed five different categories of visemes:
interdentals /ð θ/, labiodentals /f v/, bilabials /p b m/, palato-alveolars /ʃ z/, and approximant /w/, based on a 75% identification accuracy rate.
Subsequent extensive lipreading training resulted in increased ability to
discriminate within viseme category phonemes at post-test. In addition to the
above studies, more recently Owens and Blazek (1985) reported six different
viseme categories: bilabials /p b m/, labiodentals /f v/, dentals /θ ð/,
approximants /w r/, palato-alveolars /tʃ d ʃ/, and a mixed group of velars
and alveolars /t d s k n ɡ l/.
Perception of viseme categories also appear to be prone to individual
differences between speakers. In one study on the visual identification of
Quebec French consonants, participants were required to identify 85 syllables
produced by two speakers in a visual-only condition. Results revealed that
one of the speakers yielded six, and the other seven viseme categories (Jutras,
Gagnè, Picard, & Roy, 1998). The viseme categories common to both speakers
were bilabial /b p m/, labiodentals /f v/, and palato-alveolars /ʃ ʒ/, and one
speaker produced an extra approximant /r/ category.
There is also developmental evidence with respect to the psychological
existence of visual speech categories. For example, Mills (1987) found that
visually impaired children made more production errors defined by within
viseme category confusions, such as [m] versus [n], than children with no
visual impairment. This finding shows that at least some aspects of speech
perception are acquired by means of visual speech information.
4.3 Multimodal Speech Perception: Auditory-Tactile and Auditory-Orthographic Speech Processing

In this section, auditory-tactile and auditory-orthographic speech processing are discussed, followed by a more detailed discussion of auditory-visual speech perception, the core concern in this thesis.

The sense of touch can be instrumental in the perception of speech. For example, Fowler and Dekle (1991) found that when participants are presented with auditory syllables (/ba/ & /ga/) simultaneously with incongruent lip movements which were presented haptically – by touch – participants’ perception of syllables was affected by the haptically presented syllables, leading to an auditory-haptic McGurk effect. In a case study, Plant, Gnosspeilus, and Levitt (2000) found that the ability of their profoundly deaf native Swedish-speaking subject to identify vowel duration, consonant voicing, and manner of articulation was greater with tactile information (by means of placing the thumb on the talker’s throat) than by means of lipreading. In addition, his tracking speed was around 40 words per minute in a lipreading condition and around 60 words per minute in a tactile condition. As the participant was bilingual, Plant et al. also tested for native language effects and found that tracking rate was much faster for native language Swedish stimuli than for English stimuli.

The presentation of orthographic text information with the acoustic speech signal enhances speech perception. Frost, Repp and Katz (1988) tested undergraduate students on speech stimuli masked with white noise of the same amplitude level as the speech. When orthographic input was provided, subjects reported perceiving the words more clearly. Frost et al. (1988) suggested that printed words are decoded into some form of phonological representation, and consistent with this are findings that orthographic acquisition (literacy) affects phonological processing, in particular that alphabetic knowledge raises phoneme awareness (e.g. Morais, 1979).

There are a number of other studies showing that orthography aids speech perception. Massaro, Cohen and Thompson (1990) presented speech stimuli
in two conditions: Auditory-visual and auditory-text. It was found that while provision of visual speech information produced the greatest improvement over an auditory-alone condition, there was still perceptual improvement provided by the inclusion of text information. Dodd, Oerlemans and Robinson (1990) tested the cross-modal interaction of heard, lipread and text-read stimuli, and found that lipreading primed the other two modalities, and that orthographically presented stimuli also primed lipread stimuli—both lipreading and reading significantly enriched speech perception.

Finally, in a recent speech production study, Erdener and Burnham (2005c) tested Australian English (a language with an opaque orthography characterised by inconsistent phoneme-grapheme match-ups) and Istanbul Turkish speakers (a language with a transparent orthography characterised by consistent phoneme-grapheme match-ups) on speech stimuli in Irish and Spanish, which have opaque and transparent orthographies, respectively. Four conditions were tested: Auditory-only, auditory-visual, auditory-orthographic and auditory-visual-orthographic. The subjects’ task was to repeat the stimuli presented to them. It was found that visual speech information facilitated subjects’ productions compared with the auditory-only condition in both speaker groups. Interestingly, in the two orthographic conditions (auditory-orthographic and auditory-visual-orthographic), speakers of Turkish (a transparent orthography) performed better than speakers of Australian English (an opaque orthography) for the transparent Spanish stimuli, but worse for the opaque Irish stimuli.
Figure 4-1 Mean phoneme errors by Turkish and Australian participants in the auditory-orthographic and auditory-visual-orthographic conditions in Erdener and Burnham (2005).

Furthermore, there was much better performance by Turkish speakers on the Spanish than the Irish orthographic conditions, whereas their Australian counterparts’ performance was balanced between stimulus language conditions, showing less effect of the transparency/opacity of orthography (see Figure 4-1).

4.4 Auditory-Visual Speech Processing

Research has consistently shown that seeing the face of a talker increases perceived clarity and intelligibility of the message. This advantage of seeing the face on speech perception was first demonstrated by Cotton (1935), who presented participants with sentences embedded in a ‘buzzing’ background noise in two conditions and tested their comprehension of the sentences. In one condition, the speaker’s face was visible and this yielded better comprehension performance than a condition in which the speaker’s face was not visible. Sumby and Pollack (1954) presented their participants with speech stimuli in auditory-visual and auditory-only conditions across different background speech-to-noise (SN) ratios and showed a significant visual
contribution to speech intelligibility at low (up to 20 dB) SN ratios, (also see Erber, 1969; Summerfield, 1979).

The benefit of visual speech information is not limited to noisy conditions; it is also evident in clear listening conditions. McGurk and MacDonald (1976) were first to demonstrate the role of visual speech information in clear, undegraded speech conditions. They presented English-speaking participants in three age groups (3-5 year-olds, 7-8 year-olds, and 18-40 year-olds) with auditory-visual speech stimuli in which the auditory and visual components were incongruent. The stimuli consisted of a female talker’s face uttering the syllable [ga] onto which the syllable [ba] was dubbed, and vice versa. The procedure was repeated for the syllables [pa] and [ka] (see Table 4-1). The subjects were also presented with auditory-only stimuli. All three participant groups were highly accurate in their responses to the auditory-only stimuli, performing at above 90% accuracy. However, in auditory-visual condition, McGurk and MacDonald found that the perceivers reported phonemes that were not physically present. One of the most striking responses was to the auditory [ba] and visual [ga] presentation, which yielded a high percentage of “da” responses from all three age groups, and a smaller proportion of “tha” responses.

**Table 4-1. Data from McGurk and MacDonald (1976).** Auditory and visual response categories are based on a single modality. In ‘Fused’ responses input from two modalities are fused to a new element, absent in either modality. The ‘Combination’ category represents responses which embed elements from both modalities.

<table>
<thead>
<tr>
<th>STIMULI</th>
<th>Auditory</th>
<th>Visual</th>
<th>Response Categories (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Auditory</td>
</tr>
<tr>
<td>ba</td>
<td>3-5 yos</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>7-8 yos</td>
<td>36</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>18-40 yos</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>ga</td>
<td>3-5 yos</td>
<td>57</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>7-8 yos</td>
<td>36</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>18-40 yos</td>
<td>11</td>
<td>31</td>
</tr>
<tr>
<td>pa</td>
<td>3-5 yos</td>
<td>24</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>7-8 yos</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>18-40 yos</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>ka</td>
<td>3-5 yos</td>
<td>62</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>7-8 yos</td>
<td>68</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>18-40 yos</td>
<td>13</td>
<td>37</td>
</tr>
</tbody>
</table>

56
This effect, later termed the *McGurk Effect*, has been one of the most commonly used paradigms in auditory-visual speech perception research to measure the influence of visual speech information (Figure 4-2). A demonstration of McGurk effect can be found in the folder named “McGurk Effect Demo”, in the CD-ROM supplement.

McGurk and MacDonald (1976) explained their findings as follows. The consonant components of the syllables [ba] and [da] have common acoustic properties, such as similar formant frequencies; whereas they are very different in terms of visual characteristics, [ba] featuring a bilabial stop, and [da] a labio-dental place of articulation. In addition, lip movements for [ga] are similar to those for [da]. According to McGurk and MacDonald, the acoustic commonalities between ‘ba’ and ‘da’ (the manner of articulation) and the common visual characteristics of ‘ga’ and ‘da’ (place of articulation) results in a high proportion of ‘da’ (and also to a lesser extent ‘tha’) responses.

![Figure 4-2. The McGurk Effect. When auditory [ba], is dubbed onto the lip movements for [ga], the resultant percept is usually [da] or [tha].](image)

This manner-place hypothesis, in which acoustic cues provide information for the manner of articulation and optical cues provide information for place of articulation, was later tested by MacDonald and McGurk (1978) who found that the McGurk effect was more evident when the visual input was non-bilabial ([da] or [ta]) and the auditory input was a bilabial ([ba] or [pa]) sound. They also tested whether the reverse McGurk effect procedure would
produce a visual influence on speech perception. However, the effect was less pronounced in the reverse condition, in which visual input was bilabial and auditory input was velar. They interpret these findings to support the clear role of vision in speech perception and to partially support the manner-place hypothesis.

The manner-place hypothesis has also been supported in a number of subsequent studies (Dodd, 1980; 1977). Dodd (1980) tested hearing and hearing-impaired participants on a spelling task, which featured easy- and difficult-to-lipread stimuli. She found that hearing participants performed better on difficult items than their hearing-impaired counterparts, which she interpreted on the basis of the hearing subjects’ ability to retrieve stored auditory information in support of the visual speech information. In a second experiment Dodd tested congenitally blind and normal-sighted participants on a task that required them to identify errors on the basis of manner (e.g. corner vs. corder) and place of articulation (e.g. window vs. pindow). She found that sighted participants identified more place of articulation errors than their blind counterparts, whereas there was no group difference with respect to the identification of manner of articulation errors. Dodd interpreted this finding to show that identification of place of articulation might necessitate visual speech information, whereas for the identification of manner of articulation auditory only information may suffice, a result consistent with previous findings (Dodd, 1977; MacDonald & McGurk, 1978; McGurk & MacDonald, 1976).

Visual speech information has also been shown to aid speech intelligibility. For example, Reisberg, McLean and Goldfield (1987) have found that comprehension of a semantically complex text was enhanced by exposing the narrator’s face. More recently in three experiments Arnold and Hill (2001) tested native English speaking adults on the benefit of visual speech information in second language, non-native accent, and semantically complex speech stimuli. In all three experiments they found clear augmentation due to visual speech information.
One of the ongoing challenges of auditory-visual speech perception research is the issue of integration of auditory and visual speech information. In an attempt to investigate the level of integration of auditory and visual speech components, Summerfield (1979) tested participants on speech stimuli in five conditions: Auditory-only (A), auditory-visual-face (B), auditory-visual-lips (C), auditory-visual-dots (D; upper, lower, left and right corners of the lips were marked with white dots, and no other facial details were provided), and circle (E; circle’s diameter varied between 5 mm and 110 mm in proportion to variations of an amplitude envelope variation). Condition B resulted in superior performance to condition C, and condition D resulted in superior performance to condition E, showing that optical articulatory cues improve perception. In a second experiment perceivers were presented with three continua of synthetic syllables: [aba]-[ada], [ada]-[aga] and [aga]-[aba]. These syllables were presented in two conditions; audio-alone and audio-visual. Identification results showed that response categories emerged mainly as a function of provision of visual speech information. In the three syllable continua, all perceivers’ (except for one in the [ada]-[aga] continuum) judgements were heavily influenced by visual speech information. For example, the proportion of bilabial responses rose dramatically when visual bilabial closure information was provided irrespective of what the auditory component of the stimuli entailed. On the basis of these findings, Summerfield (1979) argues that optical articulatory information and auditory information presented separately to perceivers interact and these two kinds of speech information are co-perceived in a common metric.

These studies (MacDonald & McGurk, 1978; McGurk & MacDonald, 1976; Summerfield, 1979) suggest that there is integration of speech information sources at some point in speech perception although the actual level was difficult to specify. This paved the way for the question of the level of integration of auditory and visual speech components, and hence the development of various models, to which the following sections are addressed.
4.5 **Major Models of Auditory-Visual Speech Perception**

In this subsection, three approaches to modelling auditory-visual speech integration are described: the *integration* view, the *differentiation* view, and the *intensity hypothesis*.

According to the *integration view*, senses develop separately, and intersensory integration gradually develops over the course of development (Piaget & Inhelder, 1974). On the other hand, the *differentiation view* indicates that perception is intersensory, integrated at birth. Subsequently the general perceptual mechanism develops into separate perceptual systems (E. J. Gibson, 1969; Lewkowicz, 1994). Another model put forward by Lewkowicz (1994), the *intensity hypothesis*, suggests that intersensory perception develops as a function of the intensity and quantity of stimulation that takes over the course of development. In essence this view states that newborn infants are responsive to the quantitative aspects of stimulation and as a result of cumulative exposure to environmental stimulation, and maturation of the perceptual systems, they gradually become more attuned to the qualitative and more complex aspects of the environment, including intersensory perceptual processes (Lewkowicz, 1994, 2003).

Beyond these hypotheses, the debate on the auditory-visual speech integration is generally based whether the integration is phonologically/post-categorically or phonetically/amodally based. In this regard in the following two subsections two competing models are described.

### 4.5.1 Fuzzy Logical Model of Perception

Massaro (1987; 1998) has developed the Fuzzy Logical Model of Perception (FLMP), which, while a model of perception in general, specifically emphasises auditory-visual speech perception. According to this model, multiple sources of non-categorical information that are continuously available in order to recognise speech input. There are three processes involved in speech perception: evaluation, integration, and decision (see Figure 4-3), which operate on the basis of four basic assumptions as follows:
(i) Each source of information is evaluated to provide probability-based fuzzy truth values, (ii) and this evaluation is done independently for each of the information sources. (iii) The sources are integrated such that there is an overall degree of support for each response alternative, and (iv) the decision process maps the products of the integration process into a response alternative. In terms of auditory-visual speech perception, the model suggests that the auditory and visual sources of information are evaluated independently, and integrated by means of a probabilistic process. According to the FLMP the process of the integration of auditory and visual speech components occur at the phonological level rather than at lower levels of processing such as the phonetic level (for a contrasting account see Burnham, 1992). Finally, FLMP suggests that the perceptual verdict is made on the basis of the comparison of response alternatives to prototypes available in long-term memory storage. The model is schematically presented in Figure 4-3.

**Figure 4-3 Schematic representation of FLMP. The information sources are given in uppercase letters, and psychological fuzzy truth values are indicated in smallcase letters. The evaluation process transforms information from sources into fuzzy truth values, which are then integrated into a degree of support. In the final decision process, these outputs are compared against prototypes stored in long-term memory. The figure was adapted from Massaro and Cole (2000).**
4.5.2 Phonetic plus Post-Categorical (3PC) Model of Auditory-Visual Speech Perception

The Phonetic plus Post-Categorical (3PC) model proposes that auditory-visual speech integration first occurs at the phonetic level, prior to phonological influences, that basic speech perception mechanisms operate in the perception of auditory-visual speech (Burnham, 1998). Evidence for this model predominantly stems from developmental studies of auditory-visual speech perception with infants. It was reasoned that if infants perceive auditory-visual speech, then, to the extent that they do not have significant phonological experience, auditory-visual speech perception could be said to occur pre-phonologically (Burnham & Dodd, 2004) (see 4.7.1). These results suggest that infants may perceive McGurk effect at a phonetic level, and that phonological processing of auditory-visual speech information is not essential.

As for the role of phonological experience, the 3PC model suggests that the effects of learning a particular language are the same for both auditory and auditory-visual speech, such that phonological biases are established as a function of experience (Burnham, 1998). This affects the way auditory-visual speech is perceived- it is influenced by the phonological and phonotactactical rules of the particular language (e.g., Sekiyama, 1997a, 1997b). In order to investigate the role of phonology in auditory-visual speech perception, Burnham and Dodd (1996) tested native Thai and English-speaking adults on a variation of the original McGurk stimulus configuration: auditory [m]-visual [ŋ]. Note that [ŋ] (as in swing) occurs in initial, medial and final positions in Thai but only in medial and final positions in English. Participants were presented with the auditory [m]-visual [ŋ] stimulus either in initial or final positions in auditory-only, auditory-visual, and visual-only speech conditions. Results revealed that Thai perceivers reported more “ŋ” percepts to a visual-only [ŋ] than English-speaking perceivers (69.6% and 7.5%, respectively). On the other hand, there was no difference between the two
groups with respect to their responses to the matching auditory-visual [ŋ] condition and mismatching auditory [m] + visual [ŋ] stimuli despite the fact that initial [ŋ] is phonotactically legal in Thai and illegal in English in the initial condition. Research shows that auditory-visual speech perception is not affected by language specific constraints and that the integration of auditory and visual speech information occurs prior to phonological processing (Burnham & Dodd, 1996). A number of other studies have also found evidence showing that auditory and visual speech information are weighted prior to phonological processing and processed at the phonetic level (Green, 1998; Robert-Ribes, Piquemal, Schwartz, & Escudier, 1996; Shigeno, 2000).

4.6 OTHER ISSUES IN AUDITORY-VISUAL SPEECH PERCEPTION

4.6.1 Auditory-Visual Speech Perception: Static vs. Dynamic

Summerfield (1979) was one of the first to suggest that continuous time-varying information was crucial for effective auditory-visual speech perception. More recently, Rosenblum and Saldaña (1996) investigated whether temporal information is necessary and sufficient for auditory-visual speech perception. In their first experiment they tested two auditory-visual conditions—fully illuminated (FI) and a point-light (PL) condition along with an audio-alone condition. In the FI condition, the talker’s face was fully visible. In the PL face condition, 28 dots were placed on a talker’s face, lips, tongue and nose tip and the talker’s face was completely darkened except for these illuminated points. The purpose of the PL condition was to test whether basic kinematic movements of the face would be sufficient for visual speech perception. In addition, Rosenblum and Saldaña (1996) also tested whether dynamic information was necessary for auditory-visual speech perception. They presented auditory syllables simultaneously with static frames taken from the previous dynamic FI and PL conditions. The results showed that (a) both dynamic FI and PL conditions yield better performance than audio-alone
condition, showing that continuous and dynamic information was sufficient for accurate auditory-visual speech perception; (b) the dynamic FI and PL conditions resulted in better performance than in the static viewing conditions, showing that such dynamic visible speech information was necessary; (c) the FI condition produced stronger visual speech influence than PL condition.

On the basis of these findings the authors suggest that for visual speech perception to occur dynamic visual speech information is both necessary and sufficient. In addition, the authors also argued that the mechanism responsible for the integration of auditory and visual speech information should be sensitive to modality-neutral, i.e., free from phonetic and phonological constraints, continuous kinematic patterns (given that the FI condition led to greater visual speech influence than PL condition), hence suggesting that integration occurs before phonetic labelling. Rosenblum, Johnson, and Saldaña (1996) also found that such visual speech movements raise intelligibility by an average of around 7 dB when the auditory component of the speech signal is degraded by a background noise. Research on deaf sign language communication using a similar point light technique showed a similar pattern of results. Poizner, Bellugi and Lutes-Driscoll (1981) placed point-light markers on a sign language user’s head, shoulders, elbows, wrists and index fingers. Results showed that deaf perceivers were able to extract linguistic information from the lexical movements of these point-light markers, showing the multimodal interference in speech perception (e.g., Fowler & Dekle, 1991). The authors also tested word identification by removing a point-light marker at a time. It was found that the index fingers were the most salient source of visual information required for comprehension by deaf perceivers (Poizner et al., 1981; Tartter & Knowlton, 1981).
4.6.2 Early versus Late Integration of Auditory and Visual Speech Information

There are two general competing models of auditory and visual speech integration in auditory-visual speech perception research: early and late integration models. Early integration models argue that the auditory modality is the source of dominant information source in auditory-visual speech perception and the visual input is translated into a representation of the auditory modality. According to early integration models the auditory and visual components of a speech signal are represented in a common metric at an early, pre-linguistic stage of speech perception (Schwartz, Robert-Ribes, & Escudier, 1998). On the other hand, the late integration models suggest that there are two distinct recognition processes for each modality, and the integration of auditory and visual components occurs post-phonologically, after the system obtains phonemic features from the two modalities (Massaro, 1987, 1998). An example of early integration models is Phonetic plus Post-categorical (3PC) Model of Auditory-Visual Speech Perception (Burnham, 1998) and an example of late integration models is Fuzzy Logical Model of Perception (Massaro, 1987, 1998).

4.7 Ontogenetic Development of Auditory-Visual Speech Perception

As can be seen in the above two models, one of the most important questions in the study of auditory-visual speech perception and processing is whether auditory-visual speech perception occurs late, at the post-categorical phonological level (Massaro, 1998), or early, at a phonetic level (Burnham & Dodd, 2004; Rosenblum & Saldaña, 1996). Developmental studies, both ontogenetic and differential (see 3.4.1), are crucial in order to answer such questions (Bernstein, Burnham, & Schwartz, 2002). In the following subsections, the development of auditory-visual speech perception is presented with reference to ontogenetic studies in three broad age periods: infancy, childhood, and adulthood. This is followed by a consideration of
language and culture on auditory-visual speech perception in differential developmental studies.

4.7.1 Auditory-Visual Speech Perception in Infancy

It is important to discover whether auditory-visual speech perception might operate in infancy before any significant language specific phonological experience. As was shown in section 3.4.2, the human auditory system is fully functional by the beginning of the third trimester of conception (Bredberg, 1985), while the visual system continues to develop postnatally during the first year of life and continues to develop as a function of learning and experience (Slater, 1999). Young infants can discriminate almost any speech sound contrast that researchers like to test them on (e.g. Burnham, 1986; Werker & Tees, 1984a). In addition, infants, as young as four days old, can discriminate their mother’s face from a stranger’s face (Pascalis, de Schonen, Morton, Deruelle, & Fabre-Grenet, 1995). Putting together both the components of speech, the auditory and visual, very little is known about auditory-visual speech perception at birth. However more is known about auditory-visual speech perception in older infants. Infants, as young as 3 months old, can learn arbitrary face-voice pairings on the basis of brief exposures (Brookes et al., 2001). Moreover, 1-, 3-, and 5-month-old infants’ recognition of their mother’s voice compared with a stranger’s voice is facilitated by presentation of the mothers’ face, showing that infants both perceive and combine to some extent have the ability to process two sources of perceptual information (Burnham, 1993)\textsuperscript{12}.

There are other studies of auditory-visual matching which more specifically investigate the actual phonemes of speech. With regard to temporal matching Dodd (1979) found that young infants aged between 10 and 16 weeks attend longer to asynchronous auditory-visual speech than in-synchrony auditory-

\textsuperscript{12} Burnham (1993) also found that 3- and 5-month-olds can recognise their mothers’ faces in a visual-only condition, whereas for 1-month-olds to do this, addition of the auditory information was required.
visual speech, showing that infants can process speech information at a very young age with regard to identity matching. Kuhl and Meltzoff (1982) tested 4½-month-old infants on their ability to perceive auditory-visual phoneme correspondences. In each trial they presented their participants with two side-by-side faces of a female talker uttering the vowels /i/ and /a/ visually and at the same time one of these vowels was presented aurally from a central location. The results showed that 73.6% of the time infants looked at the face that matched the auditory input. Kuhl and Meltzoff repeated this study with different vowels, /i/ and /u/, and found similar results. However, they did not obtain the same results when they replaced the auditory speech input with non-speech pure tones which excluded certain spectral information but retained the temporal characteristics of these vowels (Kuhl & Meltzoff, 1984; Kuhl, Williams, & Meltzoff, 1991). Similar results were also found by Walton and Bower (1993). They tested infants aged between 4 and 6 months of age using native and non-native (French) isolated vowel face-voice pairings. The infants predominantly preferred matching face-voice pairings in both native and non-native conditions. Again infants could not perform the pure tone-face matching task beyond chance levels, although adults were able to do this task easily. More recently Patterson and Werker reported two studies with 2-month-old (Patterson & Werker, 2003) and 4½-month-olds infants (Patterson & Werker, 1999) and found that infants of these ages match phonetic information in the voice and in the face. Together these results can be interpreted to show that infants’ ability to match face and voice correspondences was not based on simple temporal or amplitude cohesion, but rather on matching more complex spectral phonetic information with articulatory movements (Kuhl & Meltzoff, 1984). In addition, recently it was found that while 4- and 6-month-old monolingual and bilingual infants can discriminate English and French in silent speech, at eight months of age only bilingual infants can discriminate these two languages in a visual-only listening condition, showing that infants retain necessary visual components
of their ambient language(s) on a phonological basis (Weikum, Vouloumanos, Navarra, Soto-Faraco, Sebastián-Gallés, & Werker, 2007).

Over and above simple matching of auditory and visual speech components, it is of interest to know whether infants actually integrate speech components. To investigate this, studies using McGurk stimuli are useful. Desjardins and Werker (1996) tested two groups of 4-month-old infants, one habituated to matching auditory + visual [vi] and another to auditory + visual [bi]. In the dishabituation phase, both groups were exposed to auditory [bi] + visual [vi], which is generally perceived by adults as “vi”. They found that female infants who were initially habituated to auditory + visual [bi] showed a visually-based novelty response to the auditory [bi] + visual [vi], whereas female infants who were habituated to auditory + visual [vi] in the habituation phase did not show any interest. The male infants in both groups showed interest in neither stimulus. Desjardins and Werker interpreted these results as evidence that infants with little phonological experience are influenced by optically available articulatory movements, and that there is a perception-production link, similar to that suggested by the motor theory of speech perception suggests (see Liberman & Mattingly, 1985).

Using a habituation paradigm Rosenblum, Schmuckler and Johnson (1997) tested 5-month-old English language environment infants on McGurk type stimuli in which the visual component revealed the mouth area. Following a habituation phase with auditory + visual [va], they presented infants with three McGurk-type stimuli: auditory + visual [va], auditory [ba] + visual [va], which is generally perceived as “va” by adults, and auditory [da] + visual [va], which is generally perceived as “da” by adults. The results showed that infants gazed more when auditory [da] + visual [va] was presented than the other two presentations. Rosenblum et al (1997) interpreted these results as evidence that infants integrate auditory and visual speech information as adults do. Subsequently Burnham and Dodd (2004) found similar results but using the classic auditory [ba] + visual [ga] McGurk combination in which an emergent percept “da” or “tha” is expected. They tested two groups of 4½-
month-old English language environment infants using a McGurk paradigm using matched and mismatched auditory-visual speech stimuli. The experimental group was habituated to a mismatched McGurk stimulus (Auditory [ba] + Visual [ga]), and the control group to a matched stimulus (Auditory [ba] + Visual [ba]). Subsequently the infants were presented with three auditory-only speech stimuli, [ba], [da], and [ða]. The experimental group showed recognition of [da] and [ða] despite the fact that they had not been exposed to these stimuli in habituation trials, and treated [ba], which had been presented in habituation, as an unfamiliar stimulus. These results are supported by other studies of the McGurk effect in infancy with other stimulus configurations than the original auditory [ba] + visual [ga] configuration (e.g. Rosenblum et al., 1997).

4.7.2 Auditory-Visual Speech Perception in Childhood

In the original McGurk study (McGurk & MacDonald, 1976) three groups were compared: preschoolers (3- & 5-year-olds), school children (7- & 8-year-olds), and adults (18- to 40-year-olds). Performance in an auditory-only condition was comparable across ages, but there was an age-related increase in the amount of visual speech influence (see Table 4-1). Subsequently, Massaro (1984) tested children aged between 5 and 7 years and adults on a continuum of synthetic auditory syllables ranging between /ba/ and /da/ and matched with either videotaped /ba/, or /da/, or /ga/. He found that the probability of adults’ “da” responses to these stimuli was greater than that of children. In addition, Massaro, Thompson, Barron, and Laren (1986) tested children between 4 and 6 years and adults in a visual-only speech condition and found that adults were better lipreaders (96%) than young children (79%). On the basis of the latter finding, Massaro et al. (1986) infer that the difference between child and adult perceivers on auditory-visual speech perception stems from children’s poorer lipreading ability. Later Hockley and Polka (1994) tested four groups of children (5-, 7-, 9-, and 11-year-olds) and adults on auditory /ba/ paired with visual /va/, /ða/, /da/, or /ga/. They found
that the rate of fused responses increased between 5 and 11 years of age, supporting the other findings (Massaro, 1984; Massaro et al., 1986; McGurk & MacDonald, 1976) that auditory-visual speech perception performance generally increases over age. What might be the basis of this increase in auditory-visual speech perception performance over age? Desjardins, Rogers and Werker (1997) tested adults, and two groups of children aged between 3 and 5 years—those who made articulatory substitution errors and those who did not—with McGurk-type speech stimuli in auditory-only, visual-only, and auditory-visual listening conditions. The three groups performed comparably in the auditory-only condition. Adults were more prone to visual speech influence than the two groups of children, but of greater interest, children, who made less substitution errors, were found to be more influenced by visual speech information than were substituter children. The substituter children also performed poorer than the non-substituter group in the visual-only (lipreading) condition. Desjardins et al. conclude that experience in correctly articulating leads to a superior representation of visible speech (Desjardins et al., 1997). In a similar vein a study by Siva, Stevens, Kuhl, and Meltzoff (1995) showed that adults with cerebral palsy, (who, due to their condition, do not have normal articulation experience) were less prone to visual speech influence than normal-articulating adults. Together these results show that the ability or experience to produce speech sounds accurately seems to be related to the degree of an individual’s auditory-visual speech integration and/or visual speech influence.

4.7.3 Auditory-Visual Speech Perception in Adulthood to Late Adulthood

Adults’ auditory-visual speech perception was reviewed in 4.4. In general studies have shown that adults perceive speech in a relatively more auditory-visual manner and use more use visual speech information than children (Massaro et al., 1986; McGurk & MacDonald, 1976; Sekiyama, Burnham, Tam, & Erdener, 2003; Dupont, Aubin, & Ménard, 2005). These studies, though, are
mostly based on data obtained from participants aged up to their mid-forties, and there are very few studies investigating the status of auditory-visual speech perception in late adulthood. In this section, the few studies of auditory-visual speech perception with older perceivers are described to give a more complete picture of the development of auditory-visual speech perception over the life span. Few studies with older perceivers show that the use of visual speech information decreases with increasing age after around early 70s, showing a inverted U-shaped developmental pattern of visual speech benefit: infants and young children benefit less from visual speech input than adults, who later in life, the amount of visual speech information is attenuated.

Farrimond (1959) tested 179 participants with ages ranging from 20 to 79 years with a mean of 43.1 years, and found that age and vocabulary knowledge played important role in the ability to lipread. Later, Shoop and Binnie (1979) tested four groups of adults: 40-50-year-olds, 51-60-year-olds, 61-70-year-olds, and 71-plus-year-olds on their lipreading ability of 19 phonemes grouped in nine viseme categories in CV-syllable and sentence contexts. They found, in both contexts, that there was a progressive reduction in lipreading performance over age with especially poor performance overall in the 71-plus group. In particular the four groups of perceivers performed comparably in the recognition of /f-v/, /p-b-m/, /w/, /t/, /t-d-s-z/, /k-q/, but there was an age-related decrease in the recognition of the viseme categories, /l-n/, /ʃ-ʒ/, and /θ-ð/. Cienkowski and Carney (2002) tested perceivers aged 65 and over, younger adults (mean= 22.3 years) and a control of young adults (mean= 21.7 years) whose hearing thresholds were matched to those of the older group. The participants were presented sentence and CV-context syllable stimuli in three conditions: auditory-visual (AV), auditory-only (AO) and visual-only (VO, lipreading). Results showed that older perceivers performed comparably to their younger counterparts on the integration of auditory-visual speech components, and there was a benefit of visual cues, when provided. But interestingly the older group and the control
young adult group, who had diminished peripheral auditory sensitivity, utilised visual cues more than the young adult group. This supports the notion that when the auditory input is degraded in some way, visual cues are weighted more. Cienkowski and Carney (2002) propose that when competing auditory and visual speech components are presented (e.g. the McGurk effect) and the percept is a non-fused response, participants select a response alternative from the least ambiguous source, and this is visual for the hearing threshold control and the older groups, due to the diminished quality of their auditory input due to hearing loss. In addition, Cienkowski and Carney (2002) claim that the integration of auditory and visual speech signals occurs phonetically, not at a phonological level, on the basis of their finding of no relationship between auditory-visual speech integration and lipreading. These results also support the notion put forward earlier that auditory-visual speech perception occurs at a pre-phonological level (Burnham & Dodd, 2004; Rosenblum et al., 1997).

Overall, the studies of auditory-visual speech perception in late adulthood tend to show that the ability to speechread is gradually attenuated. In addition hearing acuity and appears to be related to lipreading performance in late adulthood. When hearing is impaired, visual cues still appear to be a robust response alternative.

4.7.4 Summary: Ontogenetic Studies of Auditory-Visual Speech Perception

Overall four conclusions can be drawn from the ontogenetic studies of auditory-visual speech perception. First, young infants integrate auditory and visual speech components, and it is speculated that this integration process is based on phonetic rather than language specific phonological processing.

Second, for English language environment perceivers the influence of visual information on speech perception increases over age. Third, it appears that the degree of visual speech influence in speech perception is related to experience with accurate production of speech sounds. Finally, at older ages
auditory-visual speech perception occurs and visual speech cues are utilised more in situations in which auditory signal is degraded due to hearing impairment. In order to obtain a complete picture of the development of auditory-visual speech perception, data from cross-language differential studies are also required, and this is provided in the next section.


4.8.1 Differences between Languages
Cross-language studies allow the investigation of the role of language-specific factors in speech perception. Recent studies have found differences with respect to processing of auditory-visual speech across numerous languages, including Dutch, German (Reisberg et al., 1987), Korean (Davis & Kim, 1998, 1999), and Spanish (Ortega-Llebaria, Faulkner, & Hazan, 2001). For example, using McGurk type stimuli, Werker and her colleagues (Werker, Frost, & McGurk, 1992) tested native Canadian English-speaking and Canadian French-speaking participants with varying levels of English language experience on the identification accuracy of auditory [ba], paired with matched visual [ba], and mismatched visual [va], [ða], [da], [3a], or [9a]. All these syllables are native in both English and French with the exception of [ða], which is native in English but not in French. They found that English-speakers identified more auditory [ba] + visual [ða] syllables than did their French-speaking counterparts. Interestingly, amongst French-speakers, the identification of [ða] instances increased as a function of English language experience. The authors interpret this finding in terms of a perceptual bias; that despite their apparent post-perceptual labelling of the [ða] syllable, French-speaking perceivers probably have a phonologically based perceptual bias against perceiving such non-native contrasts. In another study, Burnham and Dodd (2004) tested native Thai and English-speaking adults on McGurk-type mismatched auditory [m]-visual [ŋ], matched auditory -visual [ŋ], and
visual-only [ŋ] stimuli, in which initial-position [ŋ] is phonotactically legal in Thai but not in English. While there was no group difference with respect to the matched and mismatched McGurk stimuli, they found that Thai speakers lipread more accurately than did English speakers in the visual-only initial [ŋ] stimulus. These two studies show that the degree and nature of visual speech influence is related to phonotactical requirements of a language.

Sekiyama and her colleagues conducted a number of cross-language/cultural studies on auditory-visual speech perception recruiting native speakers of various languages. For example, Sekiyama and Tohkura (1993) tested English and Japanese speaking adults on McGurk-type stimuli produced by English and Japanese talkers. They found that, overall, English speaking participants were more prone to the McGurk effect than Japanese speakers. They surmised that this might occur because Japanese culture converters engage in relatively face contact during conversation than their Western counterparts; interpersonal communication in Japanese society is marked by relatively less of face-to-face interaction (Lebra, 1979) and process visual information more holistically rather than analytically (Miyamoto, Nisbett & Masuda, 2006). Interestingly, Yuki, Maddux and Masuda (2007) found that when judging emotions, Japanese people tend to look at the eye area, whereas Americans look at the mouth area, which may, in turn, be the cultural basis of relatively less visual speech influence among native Japanese speakers. In addition, Research has shown that compared to Western (i.e. United States) child-caregiver interactions, the Japanese child-caregiver interaction is usually characterized by more strictly-defined roles and less face-to face conversations (Dennis, Cole, Zahn-Waxler, & Mizuta, 2000). Interestingly Bloom and Masataka (1996) found that when judging infants’ social favourability and communicative intent, Japanese adults pay less attention to the visual speech information in the form of infants’ mouth movements than their Canadian counterparts.

Over and above this Japanese McGurk effect Sekiyama (1997a) found that Chinese language perceivers were even less prone to McGurk effect than their
Japanese counterparts, though for Chinese residents in Japan susceptibility to the McGurk effect increased as a function of length of residence up to seven years. The difference with respect to McGurk effect amongst English-, Japanese-, and Mandarin-speakers gives rise to a second possible reason for cross-language differences in the McGurk effect; it may be the case that these have to do with the relative use of tone and pitch accent in the languages. In this regard, Japanese is characterised by two pitch accents which differentiate otherwise equivalent lexical items. Chinese languages, on the other hand rely on tonal differences to signal meaning differences (Mandarin has four and Cantonese has six lexical tone contrasts). Sekiyama (1997b) suggests that speech information for tones is not instantiated visually, which thus results in less McGurk effect in Mandarin (four tones) than Japanese (two pitch-accents) perceivers (but see Burnham, Lau, Tam, & Schoknecht, 2001). However it should be noted that these studies do not necessarily mean Japanese and Mandarin speakers are not influenced by visual speech input. For example, Massaro and his colleagues tested native English, Spanish, and Japanese speakers on McGurk-type synthetic speech auditory-visual speech stimuli, consisting of various combinations of auditory and visual /ba/-/da/speech continua divided across five steps. They found that the degree of visual speech influence was comparable across speakers of English, Spanish and Japanese (Massaro, Cohen, Gesi, Heredia, & Tsuzaki, 1993). In addition, when presented with non-native speech stimuli Japanese and Mandarin speakers can in fact tune to visual speech information (Sekiyama, 1997a), so these studies do not show that Japanese and Mandarin cannot, but rather do not, under certain circumstances, pay attention to visual speech information. Above studies show that the degree to which visual information affects speech information is related to the phonological and phonotactical requirements of native language. One issue that surfaces from these studies is the foreign speaker effect, which refers to increased effect of visual speech information when listening to a non-native speaker. This issue is considered in the next section.
4.8.2 Foreign Speaker Effects

When Sekiyama and Tohkura (1993) tested native English and Japanese speakers, they found a strong foreign speaker effect, such that perceivers were more prone to visual speech information when they viewed a non-native talker. Similar results were obtained with Hungarian speakers perceiving an Austrian German talker (Grasenger, 1995), and Cantonese speakers perceiving a Dutch talker (de Gelder, Bertelson, Vroomen, & Chen, 1995). These results indicate that when perceivers are exposed to a talker speaking a non-native language, they pay more attention to visual speech cues. Such studies have significant implications for second language perception and production and the issue of second language perception and production issue is briefly addressed within auditory-visual speech perception framework in the next section.

4.8.3 The Benefit of Auditory-Visual Speech Information in Second and Non-native Language Contexts

When perceivers are presented with non-native speech input, they tend to use visual speech information more (Fuster-Duran, 1996; Sekiyama, 1997a, 1997b; but see Senema, Hazan, & Faulkner, 2003) and so cross-language auditory-visual speech perception research can profitably be applied to foreign language training settings. Hardison (2003) investigated the effect of presenting a talker’s face on the perception and production of a second language (L2). She tested native speakers of Japanese and Korean learning English in the United States on their ability to perceive and produce the /ɹ/ vs. /ɹ/ contrast in two training conditions (auditory-only vs. auditory-visual), varying the word position, and the adjacent vowel. In the first experiment, the participants (Japanese) were first pre-tested and then post-tested following a series of training sessions. The results revealed that auditory-visual training led to better performance than auditory-only training. In addition, there was a generalisation effect in which target consonants were perceived accurately when spoken by unfamiliar talkers and this was facilitated by auditory-visual
training. Ortega-Llebaria, Faulkner and Hazan (2001), on the other hand, found evidence for the contention that there are certain conditions under which auditory-visual speech information can be beneficial in an L2 learning environment, and provision of visual cues does not lead to a general improvement in the perception of non-native phonemes. Ortega-Llebaria et al. tested English-speaking and Spanish-speaking perceivers on the identification of 16 consonants and 9 vowels (British English) in CV, VCV, and VC context real word stimuli across two experimental conditions: auditory-visual and auditory-only. While there was no difference with respect to vowel identification, overall results revealed a benefit of auditory-visual training for consonants. Context feature analyses showed that the error patterns by Spanish speakers were predictable from phonological differences between English and Spanish. For instance, the Spanish listeners did not benefit from visual cues when presented with consonants that are phonemic in English but have allophonic status in Spanish (e.g. /b/ vs. /p/).

4.8.4 Summary: Differential Studies of Auditory-Visual Speech Perception

Cross-language studies of auditory-visual speech perception using the McGurk effect show that (a) visual speech influence in adults varies as a function of language background (Sekiyama, 1997a, 1997b), (b) when exposed to non-native speech or speakers, perceivers use more visual speech information. In addition second/foreign language studies on auditory-visual speech perception showed that perceivers use more visual speech information when they attend to a talker speaking a foreign language. A few recent studies investigated the development of auditory-visual speech perception using a combination of ontogenetic and differential methods. The next section describes these and paves the way for experiments in the current thesis.
4.9 CROSS LANGUAGE PLUS DEVELOPMENTAL STUDIES OF AUDITORY-VISUAL SPEECH PERCEPTION

Recently, using a combination of ontogenetic and differential methods Burnham and Sekiyama (2004; also see Sekiyama et al., 2003) tested English-speaking and Japanese-speaking 6-, 8-, and 11-year-old children and adults using McGurk-type stimuli in order to compare the developmental pattern of auditory-visual speech perception in these two language environments. The participants were presented with McGurk stimuli (matched and mismatched auditory-visual combinations of the syllables /ba/, /da/, and /ga/) spoken by Japanese and English talkers in three experimental conditions (auditory-only [A], auditory-visual [AV], and visual-only [VO]) and five auditory background noise conditions: -4, 0, +4, +8, and +12 dB signal-to-noise (SN) ratios, and a clear condition. It was found that for English speakers there was a low level of use of visual speech information at 6 years, but a significant increase from age 6 to 8, with use of visual information stable across 8 years, 11 years, and adulthood. The use of visual speech information by the Japanese-speaking 6-year-olds was comparable to that by the English-speaking 6-year-olds; however, visual speech influence in Japanese speakers remained at this low level, and did not change significantly over age, even in adults. The key question that these findings led to was, ‘What caused the shift in auditory-visual speech perception between 6 and 8 years of age in English-speaking environment?’ Sekiyama and Burnham (2004) discussed a number of possible reasons. One of these is cultural. In Japanese and many other cultures, looking at the talker’s face during conversation is considered to be inappropriate. As indicated earlier, developmental research has shown that Japanese mothers have less direct, face-to-face conversations with their toddlers (Dennis et al., 2000), and when judging infants’ social favourability and communicative intent they take visual face movements into consideration less than their North American counterparts (Bloom & Masataka, 1996).

Sekiyama and Burnham also provide a linguistic line of explanation. In Japanese, there are fewer visually identifiable speech elements (e.g.
labiodental fricatives, /f/ and /v/) and seemingly less extensive mouth movements. In addition, the use of pitch-accent differences to signify meaning differences in Japanese may also result in less need for visual speech information—see Japanese and Mandarin language studies in 4.8.1 (but also see Burnham et al., 2001; Sekiyama, 1997a).

In 4.7.1 evidence was presented showing that infants perceive discrepancies between auditory and visual speech information; that infants perceive visual speech. In light of recent ontogenetic, differential, and combined evidence that indicates respectively (a) increases in the use of visual speech information over age for English language children, (b) cross-language differences in adult auditory-visual speech perception in different languages, and (c) cross-language developmental differences between Japanese and English language children and adults in the use of visual information in speech perception, the question remains: What is it that leads to a developmental increase in the use of visual speech information in English speakers, particularly between 6 and 8 years of age? This is the question at which this thesis is directed. Before describing the aims and predictions of the current study, a short review of literature on auditory-visual speech perception in adults will be presented in the next section.

4.10 The Current Thesis

The argument for the current thesis was advanced as a three-stage process. First, studies show that native English-speaking adults use more visual information in speech perception than native Japanese-speaking adults (Sekiyama & Burnham, 2004). Second, there are differential changes in the use of visual speech information over age (increasing influence from visual speech information) for English but not Japanese language environment children (Sekiyama & Burnham, 2004). Third, the increased use of visual speech information in English speakers occurs sometime between 6 and 8 years of age. One of the most significant milestone events between these ages is the onset of schooling. One question that could be asked is: what is it in the
environment that influences auditory-visual speech perception? (Sekiyama & Burnham, 2004)

In this thesis possible linguistic bases for this developmental shift are considered. Burnham (2003) has shown that children’s reading ability is positively related to their language specific speech perception, measured by the difference between perceptual ability for native vs. non-native speech contrasts. In addition, as detailed in section 4.7.2, articulation ability seems to be related to the degree of visual speech influence in children (Desjardins et al., 1997; Siva et al., 1995). Experiments in the current thesis were conducted in order to test whether formal language skills (reading) and other linguistic variables (language-specific speech perception and articulation) might be related to the increase in visual speech influence in English-speaking children.
CHAPTER 5

LANGUAGE SPECIFIC FACTORS IN THE
DEVELOPMENT OF AUDITORY-VISUAL
SPEECH PERCEPTION: SCHOOL CHILDREN &
ADULTS
Sekiyama and Burnham (2004), using a combination of differential and ontogenetic methods, tested 6-, 8-, 11-year-olds and adults from English and Japanese language environments on McGurk type stimuli. They found that, for English speakers, there was a significant increase from age 6 to 8 years of age in the use of visual speech information, which was maintained through to 11 years and adulthood. On the other hand, while for Japanese language speakers at 6 years of age the level of visual speech influence was same as for English-speaking 6-year-olds, it remained at this level across age. Bernstein et al. (2002) suggest there are two important developmental questions in auditory-visual speech perception: when in the course of development that auditory-visual speech perception integration occurs and what factors might affect or cause any such modulation in auditory-visual speech integration. The Sekiyama and Burnham (2004) report pinpoints, to some extent, the first of these – the developmental locus of increased auditory-visual speech integration. The aim of the experiment reported in this chapter is to investigate what factors might affect auditory-visual speech perception development.

5.1 Experiment 1: Language Specific Factors in the Development of Auditory-Visual Speech Perception in School Children Aged Five to Eight and Adults

The aim of the experiment reported in this chapter is to investigate (i) the developmental locus of auditory-visual speech integration more precisely and (ii) the potential roles of a number of factors in the development of auditory-visual speech perception. The factors investigated here have previously been shown to be linked to the development of auditory speech perception (see Chapter 4). However the potential relevance of these factors to the visual and auditory-visual aspects of speech perception has not yet been investigated.

In this study, children aged between 5 and 8 years of age were given tests of reading, articulation, language-specific speech perception, and auditory-visual speech perception. It was hypothesized that, if auditory-visual speech
perception is subject to the same developmental influences as is auditory-only speech perception, then these factors should be good predictors of performance on an auditory-visual speech perception task. Specific hypotheses, as follows, were put forward:
(a) The degree of ability to articulate speech sounds correctly should be associated with the degree of auditory-visual speech perception (Desjardins et al., 1997);
(b) Previously it was found that reading ability predicts language specific speech perception, phonological perception; the better the reading ability, the greater the difference between auditory perception of native and non-native speech sounds (Burnham, 2003). If auditory-visual speech perception operates at the same level as auditory speech perception, then language specific speech perception ability should be linked to auditory-visual speech perception;
(c) in addition, if language specific speech perception ability predicts auditory-visual speech perception, then the level of reading ability should also predict auditory-visual speech perception ability on the basis of the findings by Burnham (2003).

5.2 Method

5.2.1 Design

A cross-sectional design was employed in which children of four ages (5, 6, 7, and 8 years) were tested on four separate tasks - auditory-visual speech perception (McGurk effect task), language specific speech perception, word reading, and articulation. The design in each of these four tasks is set out in turn below.

5.2.1.1 Auditory-Visual Speech Perception

The auditory-visual speech perception test had three factors, each with two levels: Noise level (clear vs. +4 dB SNR), speaker (native [English] vs. non-native [Japanese]), and the congruence of auditory and visual components of auditory-visual speech (congruent [AV+] vs. incongruent [AV-]). An
identification task was used to obtain the auditory-visual speech perception data, the dependent variable being the number of auditorally correct responses. From these, a visual speech index (VSI) was calculated. This takes into account the facilitative effect of auditory-visual congruent performance (AV⁺) over auditory-only (AO) performance (AV⁺-AO), plus the inhibitory effect of inconsistent AV performance (AV⁻) compared with AO performance (AO minus AV⁻). Thus

\[
VSI = [(AV⁺) - (AO)] - [(AO) - (AV⁻)].
\]

Simplifying, VSI is calculated by subtracting the AV⁻ score from AV⁺ (Equation 5-1).

\[
VSI = (AV⁺) - (AV⁻)
\]

**Equation 5-1 Calculation of Visual Speech Index (VSI) scores.**

### 5.2.1.2 Language-Specific Speech Perception

The language specific speech perception task had a single factor, native versus non-native speech sound contrasts, and these were tested in a same/different AX discrimination task. The dependent variable was the difference between discrimination index (DI) scores for native and non-native speech contrasts. These discrimination index scores were based on a variation of the d’ calculation, hits minus false positives divided by the number of trials, where the focus trials were the different (AB) trials. Thus discrimination indices were calculated via the formula given in Equation 5-2.

\[
DI = \frac{'Different'\ responses\ to\ AB\ trials - 'Different'\ responses\ to\ AA\ trials}{\sum AB\ Trials}
\]

**Equation 5-2 The calculation of discrimination index (DI) scores for language-specific speech perception test.**
This formula gives a maximum score of +1 for optimal responding with no false positives, a zero score for chance responding (equal number of hits and false positives), and negative scores for inverted responding—more false positives than hits.

5.2.1.3 Reading
The reading task [the reading test of Wide Range Activities Test, third version, (Wilkinson, 1993)] consisted of a list of 57 items (15 letters and 42 words) to be read. The dependent variable was calculated by dividing the correctly read items by the total number of items, giving a proportion between 0 (lowest possible performance) and 1 (highest possible performance).

5.2.1.4 Articulation
The articulation test (Queensland Articulation Test) consisted of a series of 64 pictures, which the child was required to name. The dependent variable was the proportion of correct utterances, which was calculated by the number of correct responses divided by the total number of items.

5.2.2 Recruitment of School Children
Following ethical clearance from the University of Western Sydney, and the New South Wales Department of Education twelve schools were approached to obtain permission to contact parents regarding their children’s participation in the study. Six of these schools13 responded positively and assisted in forwarding background information and consent forms to parents. The consent of children whose parents provided approval was obtained verbally in the presence of their teachers before the experiment began. Copies of ethical clearance permits, the letter to school principals, experiment

---

13 The ethical requirements of the New South Wales Department of Education stipulate that the schools involved in this study cannot be named.
background information for parents, and the consent form for children can be found in the CD-ROM supplement (Appendix CD1).

5.2.3 Child Participants

The children were recruited from six public schools in Sydney metropolitan area. A total of 98 Australian-English children of four age groups were tested: 5-, 6-, 7-, and 8-year-olds. Two kindergarten participants withdrew, as they did not wish to continue the experiment and their data were excluded from the analyses. All participants had normal hearing and division and none had reading problems. Descriptive data for age and sex for the remaining 96 children are shown in Table 5-1. The children were given a “Young Scientist” certificate (See Appendix CD1.5) and the choice of a toy for their participation.

<table>
<thead>
<tr>
<th>Groups</th>
<th>N</th>
<th>Females</th>
<th>Males</th>
<th>Mean Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-year-olds (Kindergarten)</td>
<td>24</td>
<td>12</td>
<td>12</td>
<td>5.40</td>
</tr>
<tr>
<td>6-year-olds (Grade One)</td>
<td>24</td>
<td>12</td>
<td>12</td>
<td>6.67</td>
</tr>
<tr>
<td>7-year-olds (Grade Two)</td>
<td>24</td>
<td>12</td>
<td>12</td>
<td>7.59</td>
</tr>
<tr>
<td>8-year-olds (Grade Three)</td>
<td>24</td>
<td>12</td>
<td>12</td>
<td>8.52</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>96</td>
<td>48</td>
<td>48</td>
<td>7.13</td>
</tr>
</tbody>
</table>

5.2.4 Adult Participants

Forty-eight native speakers of English (M_{age}= 21.77, SD=6.16, 7 males and 41 females) were tested. The participants were first year students enrolled in an introductory psychology unit at the Bankstown campus of the University of Western Sydney. All participants reported having normal hearing and normal or corrected-to-normal vision and none reported any reading problems. Their participation was compensated with 1% credit towards their final introductory psychology unit grade.
5.3 **Stimuli Materials, Apparatus, and Procedure**

5.3.1 *Auditory-Visual Speech Perception (McGurk) Test Stimuli and Apparatus*

5.3.1.1 Stimuli Material

The McGurk stimuli were originally created by Sekiyama and her colleagues and have been used in previous studies (Sekiyama & Burnham, 2004; Sekiyama et al., 2003). They consisted of /ba/, /da/, and /ga/ utterances by four talkers, two native English and two native Japanese speakers, with one male and one female for each language. The utterances were captured on videotape, digitised and edited (AVI format, 29.77 frames per second, 24-bit video sampling and 640 x 480 pixels) on a computer in order to create three types of stimuli: auditory-only (AO), visual-only (VO), and auditory-visual (AV). Two types of AV stimuli were created: *Congruent*, where audio and visual components were identical (e.g., audio [ba] and visual [ba]) and *incongruent*, where the audio and visual components were different (e.g. audio [ba], and visual [ga], as in the classic McGurk effect). There were also three kinds of incongruent stimuli, which were created by means of within-talker auditory and visual components as depicted in Table 5-2. VO stimuli were created by deleting the audio tracks of the [ba], [da], and [ga] recordings. For the AO stimuli, the audio tracts of the video recordings of [ba], [da], and [ga] were retained, but the video tracks were replaced by still face images of the appropriate talker. In the still images, the talkers’ mouths were closed. There were a total of 12 AO stimuli (3 consonants x 4 talkers), 12 VO stimuli (3 consonants x 4 talkers), and 24 AV stimuli (3 auditory consonants x 2 congruity types x 4 talkers). The noise level of the AV and AO stimuli was set at 65 dB.

In the experiment, the McGurk stimuli were presented in two background noise conditions: Clear and Noisy. In previous pilot studies (see Sekiyama et al., 2003) the average intelligibility of auditory and visual speech components were derived, and four levels of band noise were used (300 Hz-12000 Hz)
with signal-to-noise (SN) ratios of −4, 0, +4, +8 and +12 dB. As the effects of different SN ratios had already been investigated, and use of all these noise levels would result in too many trials to maintain children’s attention, the current experiments were limited to a single noise condition, along with a clear distinction. Based on these previous results, the midpoint SN ratio of +4 dB was adopted for the noisy condition in the current experiments.

5.3.1.2 Apparatus

The stimuli were presented on a notebook computer (Sharp Mebius MJ730R) equipped with a Pentium III microprocessor and 256 MB RAM, which was a suitable configuration in order to play video files. The notebook computer was attached to a 17-in flatscreen CTR monitor (Sony 17GS), which had a very higher refresh rate (86 Hz) than the notebook computer’s built-in LCD monitor (60 Hz). In addition, for better sound quality, a loud speaker (Aiwa SC-B10) was connected to the line out of the computer and this was positioned centrally in the participant’s sagittal plane. The responses were collected with a game controller connected to the computer via a universal serial bus (USB) port. Three stickers, labelled ‘BA’, ‘DA’, and ‘GA’, corresponding to the response alternatives were pasted on three buttons on the controller.

<table>
<thead>
<tr>
<th>McGurk Stimuli: Auditory-Visual (AV) Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CONGRUENT</strong></td>
</tr>
<tr>
<td><strong>AUDIO</strong></td>
</tr>
<tr>
<td>[ba]</td>
</tr>
<tr>
<td>[da]</td>
</tr>
<tr>
<td>[ga]</td>
</tr>
</tbody>
</table>

5.3.1.3 Procedure

The experimental set-up is shown in Figure 5-1. Participants sat facing a CRT monitor, 50 cm from them along their sagittal plane. The monitor was connected to a notebook computer, to which a game controller and a loud
speaker were also attached. The three buttons on the game controller were clearly labelled with the response alternatives, ‘BA’, ‘DA’, and ‘GA’. The loudspeaker was placed on top of the monitor facing the participants. The experimenter told participants that they would play a computer game where they were to say whether the person on the screen was saying the syllable ‘ba’, ‘da’, or ‘ga’. They were asked to look at the talker’s face on the screen in all of the experimental conditions and the experimenter checked this at every trial. Responses were indicated by pressing one of the three keys on the game controller. The time between responding and the onset of the next stimulus was 1.5 seconds. It was ensured that all participants, especially the 5-year-olds understood the task requirements. As an additional control, the 5-year-olds were asked to indicate their response orally before pressing the buttons in order to ensure button presses were consistent with their intended response.

There were three experimental conditions: auditory-visual (AV, featuring auditory-visual congruent [AV+] and auditory-visual incongruent [AV−] stimuli), auditory-only (AO), and visual-only (VO), and each of AV and AO was presented in clear and in noise conditions. In the noise condition, the speech stimuli were presented in band noise set at a 4 dB SN ratio. There were 24 AV trials (12 AV+ and 12 AV− items). Before the AV test trials, there were familiarization trials consisting of 12 AV+ and AV− trials. This phase was repeated if a child appeared not to have understood the task. There were 12 trials in each of AO and VO conditions, which were also preceded by familiarization phases, each of which consisted of 6 trials. The 24 AV and 12 AO trials were all presented twice; once in the noise, and once in the clear condition. The 12 VO stimuli were all presented in clear condition. Hence, there were a total of 84 experimental trials (48 AV, 24 AO, and 12 VO), and 24 familiarisation trials (12 AV, 6 AO, and 6 VO). In AV and AO conditions, the familiarization stimuli were presented with a background noise set at an 8 dB SN ratio. The experimental conditions were blocked on the basis of modality (AV, AO, and VO) and background noise condition. The presentation of trials was randomised within each block and the order of the presentation of blocks
was counterbalanced. In both orders the AV stimuli presented first: AV-AO-VO and AV-VO-AO. On average, the auditory-visual speech perception test took around 20-25 minutes for the 7- and 8-year-olds, and 30-35 minutes for the 5- and 6-year-olds as they usually required breaks between blocks in order to maintain their interest in the task.

**Figure 5-1 Auditory-Visual Speech Perception Test Apparatus Setup.**

5.3.2 *Speech Perception Test Stimuli, Apparatus and Procedure*

5.3.2.1 Stimuli Material

The language specific speech perception stimuli consisted of three Thai syllables spoken by a female native Thai speaker, the voiceless aspirated bilabial stop [pʰa], the voiceless unaspirated bilabial stop [pa], and the voiced bilabial stop [ba]. These syllables, originally produced, digitised and edited to NSP format for a number of earlier studies (e.g. Burnham, 2003), were transformed to Windows PCM (.wav) format for presentation on DMDX experimental environment (Forster & Forster, 2004).
There were three exemplars for each syllable, a total of nine syllables. Using these nine syllables a total of 36 native ([ba] vs. [pʰa]) and 36 non-native ([ba] vs. [pa]) speech contrasts were created. In order to keep the language specific speech perception test short enough for children’s attention span, two different versions of the language specific speech perception test were written using DMDX\(^{14}\). Two separate blocks were constructed for native and non-native contrasts (see Appendix CD2.1A and CD2.1B). The stimulus and contrast details are given in Table 5-3 and Table 5-4, respectively.

**Table 5-3 Syllable Durations in the Language Specific Speech Perception Test.**

<table>
<thead>
<tr>
<th>Syllables</th>
<th>File Name</th>
<th>Duration (msec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ba]-Exemplar 1</td>
<td>ba1.wav</td>
<td>693</td>
</tr>
<tr>
<td>[ba]-Exemplar 2</td>
<td>ba2.wav</td>
<td>718</td>
</tr>
<tr>
<td>[ba]-Exemplar 3</td>
<td>ba3.wav</td>
<td>752</td>
</tr>
<tr>
<td>[pa]-Exemplar 1</td>
<td>pa1.wav</td>
<td>653</td>
</tr>
<tr>
<td>[pa]-Exemplar 2</td>
<td>pa2.wav</td>
<td>633</td>
</tr>
<tr>
<td>[pa]-Exemplar 3</td>
<td>pa3.wav</td>
<td>577</td>
</tr>
<tr>
<td>[pʰa]-Exemplar 1</td>
<td>pha1.wav</td>
<td>649</td>
</tr>
<tr>
<td>[pʰa]-Exemplar 2</td>
<td>pha2.wav</td>
<td>642</td>
</tr>
<tr>
<td>[pʰa]-Exemplar 3</td>
<td>pha3.wav</td>
<td>608</td>
</tr>
</tbody>
</table>

**Table 5-4 Native and Non-native Speech Contrasts in the Language Specific Speech Perception Test. The first four contrasts were used in the native block and the last four contrasts were used in the non-native block.**

<table>
<thead>
<tr>
<th>Contrast</th>
<th>Native in English or Thai</th>
<th>Correct Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>ba vs. ba</td>
<td>English &amp; Thai</td>
<td>Same</td>
</tr>
<tr>
<td>pʰa vs. pʰa</td>
<td>English &amp; Thai</td>
<td>Same</td>
</tr>
<tr>
<td>pʰa vs. ba</td>
<td>English &amp; Thai</td>
<td>Different</td>
</tr>
<tr>
<td>ba vs. pʰa</td>
<td>English &amp; Thai</td>
<td>Different</td>
</tr>
<tr>
<td>ba vs. ba</td>
<td>English &amp; Thai</td>
<td>Same</td>
</tr>
<tr>
<td>pa vs. pa</td>
<td>Thai</td>
<td>Same</td>
</tr>
<tr>
<td>pa vs. ba</td>
<td>Thai</td>
<td>Different</td>
</tr>
<tr>
<td>ba vs. pa</td>
<td>Thai</td>
<td>Different</td>
</tr>
</tbody>
</table>

\(^{14}\) DMDX is an experimental programming environment designed for psycholinguistic experiments. It was developed by Kenneth Forster and Jonathan Foster at Monash University and the University of Arizona and derived from earlier code DMASTR developed by Kenneth Forster, Rod Dickinson and others at Monash University. It is a Win32 implementation that requires DirectX 7.0 or above. More information can be obtained and the program downloaded from [http://www.u.arizona.edu/~jforster/dmdx.htm](http://www.u.arizona.edu/~jforster/dmdx.htm).
5.3.2.2 Apparatus

The language specific speech perception task was presented on a notebook computer (Sharp Mebius MJ730R) equipped with a Pentium III microprocessor and 256 MB RAM, with a separate loudspeaker (Aiwa SC-B10) attached. To collect responses an external keyboard was attached to the computer, and two stickers labelled “SAME” and “DIFFERENT” were pasted on the left and right shift keys, the default same response and different response keys in DMDX. The language specific speech perception setup is depicted schematically in Figure 5-2.

![Diagram of apparatus](image)

**Figure 5-2** The language-specific speech perception test setup.

5.3.2.3 Procedure

The experimental setup for the language-specific speech perception test is presented in Figure 5-2. In the language specific speech perception test participants were presented with 36 speech contrasts using an AX discrimination task (see 5.3.2.1). Stimuli were presented on a computer via DMDX (Forster & Forster, 2003), and native and non-native trials were blocked separately. Responses were collected via a keyboard with shift keys
labelled “Same” and “Different”. The keys were also labelled with two green squares and a red circle and a green square to make the task easier for 5-year-olds. Additionally, further instructions and explanations were given to 5-year-olds to clarify the tasks as required.

5.3.3 Articulation Test Stimuli, Apparatus and Procedure

5.3.3.1 Stimuli Material
The Queensland Articulation Test (QAT) was used to measure articulation ability (Kilminster & Laird, 1978). In the QAT, originally developed from the Edinburgh Articulation Test (Anthony, Bogle, Ingram, & Mclsaac, 1971), participants are asked to name 64 objects from 57 photographs and 3 non-photographical items corresponding to all the Australian English phonemes in all legal positions, initial, medial, and final. For example, [n] is found in all three positions, but [ŋ] can be found only in medial (e.g. bringing) and final (e.g. flying) positions but not in initial position.

For testing ease a computer version of the QAT was developed. For this purpose, digital photographs of the QAT items were taken (except nothing, that, and smooth) or appropriate graphic files created and stored in Windows Bitmap (BMP) format (See Appendix CD2 for the item photographs). One of the original QAT items, yes, was replaced by yellow, as obtaining a response for the latter item was easier than for the former. For those items that could not be photographed, nothing, that, and smooth, different strategies were used (see 5.3.3.3). The consonants tested and QAT items are presented in Appendix A1 and photo items can be found in Appendix CD2.2.

5.3.3.2 Apparatus
The QAT was administered using a notebook computer (Sharp Mebius MJ730R with Pentium III, 256 RAM) and a CTR monitor (Sony 17GS). The picture files were presented using IrfanView (Version 3.08) slideshow
program\textsuperscript{15} (Skiljan, 2002), which allows random presentation of picture files, thus helping to eliminate order effects. The oral responses were captured on digital audiotapes using a DAT recorder (Sony DAT Walkman TCD-D100), and a directional collar-attached microphone (Sony RM-ED100).

5.3.3.3 Procedure

The experimental setup for the articulation test is schematically presented in Figure 5-3. Participants were seated in front of a CRT screen 50 from them in sagittal plane. The monitor was attached to a notebook computer, which controlled the experiment. The participants also wore a non-intrusive collar microphone (Sony RM-100) in to collect verbal responses. The microphone was connected to a digital audiotape (DAT) recorder and oral responses were captured on DATs. The stimuli were presented via the IrfanView graphic display program (Skiljan, 2002). This program allowed the random presentation of test items. The general testing procedure for the Queensland Articulation Test (Kilminster & Laird, 1978) was that, where each child was simply asked to name the objects in the pictures on screen, except for \textit{that}, \textit{nothing} and \textit{smooth}. For \textit{that} the experimenter engaged children in an open-ended conversation, and prompted them to use the word ‘that’. For \textit{nothing}, the experimenter told children that he was going to place an item in his hand (in fact, he put nothing in his hand) and asked them to close their eyes. He told them when they were asked to open their eyes they were required to say what he was holding in his hand. For \textit{smooth}, children were asked to touch the desktop surface gently and prompted to use the word ‘smooth’. For all items, the children were cued and assisted as needed. The articulation test took around 10 minutes to administer, but usually longer for 5-year-olds, around up to 15 minutes.

\textsuperscript{15}IrfanView\textsuperscript{©} is a multiple-format graphics viewer software, which also supports various video and audio formats. It can be downloaded from http://www.irfanview.com/.
5.3.4 Reading Test Stimuli material, Apparatus and Procedure

5.3.4.1 Stimuli and Apparatus
The reading part of Wide Range Activities Test (WRAT-3) (Wilkinson, 1993) was developed for people aged between 5 and 18 to measure their reading levels. The test has two parts. The first is composed of 15 letters, and the second features 42 words, which are listed on a card in order of orthographic complexity (e.g. in → terpsichorean). The WRAT-3 provides two versions of the word reading task, a blue form version and a tan form version, each of which features a different list of words, but the same string of letters in the first part. As per test instructions, both of these versions were used and counterbalanced across participants.

5.3.4.2 Procedure
The participants’ task was simply to read aloud, on the experimenter’s prompt, the letters and words printed on a standard card provided by the test publisher. Testing was terminated upon five consecutive errors or no responses.
5.4 General Procedure

Children were tested individually. The testing sessions were conducted in quiet rooms provided by each of the six schools, which the participants attended. All participants started the experiment with the auditory-visual speech perception test, and then the order of the reading, articulation, and language-specific speech perception tasks was counterbalanced across participants to control for order effects (See Appendix A2). In addition, the order of experimental and background noise conditions (clear vs. noise) and the stimulus presentation conditions (auditory-visual [AV], auditory-only [AO], and visual-only [VO]) in the auditory-visual speech perception test and types of contrasts (native and non-native) in language-specific speech perception test were also counterbalanced.

Adults did the same tasks as children and they were tested individually in a quiet testing room at MARCS Auditory Laboratories.

5.5 Results

The results are presented in five sections; four sections in which analyses of variances for all four measures are presented and a final section in which overall regression and correlation analyses are presented. Raw data and statistical output files can be found in Appendix CD3.

5.5.1 Auditory-Visual Speech Perception Test

The aim of the auditory-visual speech perception task analysis was to investigate the extent to which native English-speaking children of different ages were influenced by visual speech information. AV-congruent (AV⁺), AV-incongruent (AV⁻), and AO proportion correct auditory responses are shown in Figure 5-4. Adults’ auditory-visual speech perception test data is presented in Figure 5-5. VSI scores are presented separately across the five age groups, by stimulus language and listening conditions in Figure 5-6.
Figure 5-4 AV-congruent, AV-incongruent and AO mean values based on auditory-correct responses in 5-, 6-, 7- and 8-year-old groups. VSI scores were computed by subtracting the AV-incongruent scores from AV-congruent scores. Error bars represent the standard errors.
Figure 5-5 Adult VSI scores (calculated by AV+ minus AV-) by background noise and stimulus language conditions. Error bars show standard error of the mean.

Figure 5-6 Visual Speech Index (AV+ minus AV-) scores for clear and noise conditions across the five age groups by stimulus language. Error bars represent the standard error of the mean.

VSI scores ([AV+] minus [AV-]) were subjected to a 5x (2 x 2) (age x noise/clear x stimulus language) analysis of variance (ANOVA) with repeated measures on the last two factors (see Appendix CD 3.2.1). The results showed that
adults gave more visually based responses than the four groups of children \[ F(1,139)= 47.809, MS_e=.065, p<.0001 \] (Figure 5-6). In addition, there were linear \[ F(1,139)= 24.858, MS_e=.065, p<.001 \], quadratic \[ F(1,139)= 7.257, MS_e=.065, p<.05 \] and cubic \[ F(1,139)= 5.388, MS_e=.065, p<.05 \] age-related increases in the VSI scores, which interacted with stimulus language \[ F(1,139)= 8.936, MS_e=.019, p<.01 \] and background noise \[ F(1,139)= 111.021, MS_e=.028, p<.001 \], indicating that there were more visually based responses to non-native stimuli and to stimuli presented with background noise, respectively. The VSI output files can be found in Appendix CD3.2.1.

The ANOVA of VO scores revealed both a linear \[ F(1,139)= 66.270, p<.001 \] and a cubic \[ F(1,139)= 3.978, MS_e=.065, p<.05 \] age-related increase in lipreading performance, showing that lipreading scores increase gradually between ages 5 and 8, revealing a significant overall difference between children and adults with respect to lipreading performance \[ F(1,139)= 80.833, MS_e=.065, p<.001 \] (Figure 5-7). In addition, lipreading performance with Japanese stimuli was better than that for English stimuli \[ F(1,139)= 10.802, MS_e=.024, p<.05 \] (see Appendix CD3.2.2).

![Figure 5-7 Lipreading (VO) data in four child and adult groups. Error bars show the standard error of the mean.](image-url)
The overall AO data are summarised in Figure 5-8 and the statistical output files can be found in Appendix CD3.2.3. The analysis of AO revealed that adults were better than the child groups in the overall AO condition performance \([F(1,139)=30.699, MS_e=.062, p<.001]\). There was a linear age-related improvement in the AO scores \([F(1,139)=45.884, MS_e=.062, p<.001]\) and there were more accurate identification in the clear condition \([F(1,139)=74.161, MS_e=.013, p<.001]\). There was also a quadratic interaction of age and clear/noise conditions \([F(1,139)=5.574, MS_e=.013, p<.05]\).

![Figure 5-8 AO (Auditory-only) Scores across Four Age Groups in the Clear and Noise Conditions. Error bars show the standard error of the mean.](image)

**Figure 5-8** AO (Auditory-only) scores across four age groups in the clear and noise conditions. Error bars show the standard error of the mean.

5.5.2 **Language Specific Speech Perception Test**

The language specific speech perception test data were subjected to a 5 x 2 (age x native/non-native) ANOVA. Overall statistical output files can be found in Appendix CD3.3. Figure 5-9 shows the overall speech perception data in five age groups by native and non-native language stimuli. The results revealed no overall difference among the four groups of children and the adult group. However, there was an overall effect of native language such
that the performance for native speech contrasts was significantly better than for non-native speech contrasts overall \[F(1,139)=194.879, p<.001\].

![Graph showing mean discrimination index score by age for native and non-native DI.](image)

**Figure 5-9** Experiment 1: Speech perception test data in four child and adult perceivers by stimulus language. Error bars show the standard error of the mean.

### 5.5.3 Reading Test

In the reading test analysis, the number of correct items was expressed as a proportion of the total number of items (1) to yield a score between 0 and 1. Each participant was given a maximum reading score of 57. ANOVA output file for reading test is given in Appendix CD3.4. The reading scores over age are schematically presented in Figure 5-10. The reading data were analysed by means of a simple one-way between-subjects ANOVA. The results showed a significant linear increase over age in the reading ability \[F (1,139)= 912.295, MS_e=.009, p< .001\], with the largest increase occurring between 5 and 6 years \[F_{\text{Quadratic}} (1,139)= 5.728, MS_e=.009, p< .05\]. In addition, adults’ reading ability was better than those of four groups of children \[F (1,139)= 598.338, MS_e=.009, p< .001\].
5.5.4 Articulation Test

The Queensland Articulation Test (QAT) was used to measure the ability to articulate English consonants correctly. The error scores were calculated by the experimenter. In the articulation test the number of correct items was divided by the total number of items to give a proportion score between 0 and 1. Child participants’ phoneme error matrices can be found in Appendix CD3.5A, and data analysis files are presented in Appendix CD3.5B. The articulation data are graphically presented in Figure 5-11. The articulation data were analysed by means of a simple between-subjects one-way ANOVA. It was found that there was a significant linear increase over age in the ability to articulate consonants correctly \[ F (1, 139) = 92.184, M_{S_{e}} = .001, p < .0001 \]. In addition, adults’ articulation of consonants was better than children’s \[ F (1, 139) = 7.855, M_{S_{e}} = .001, p < .05 \].
5.5.5 Regression Analyses

Regression analyses were conducted in two parts. First set of regression analyses was run using school children data, and the second adult data. In each set, there were two subsets of analyses. In the first subset the auditory-visual speech perception (VSI) scores and in the second visual-only (VO; lipreading) scores were entered as dependent variables.

5.5.5.1 School Children: VSI Scores as Dependent Variable

A sequential multiple regression analysis was performed with VSI scores as the criterion and six predictors entered in order of developmental (age) and linguistic (AO, VO (lipreading), N-NN DI, articulation, and reading) factors. Results of evaluations of assumptions were satisfactory after two outliers with a standardised residual greater than 3 standard deviations were omitted from the analysis. All individual Mahalanobis distances had $p < .001$, the critical value set for four independent variables (Tabachnik & Fidell, 2001). Output files of correlation and regression analyses are provided in Appendices CD3.6.1 and CD3.6.2, respectively. Table 5-5 shows correlations among
predictor and criterion variables. Correlations are very high between age, and scores on AO, VO, reading, and articulation. In addition, there is a small but significant correlation between VSI and N-NN DI scores. VSI scores are also significantly correlated with VO, but notably not with AO. Furthermore, a significant correlation exists between AO and VO scores. Both AO and VO scores are correlated with reading and articulation scores, which are also significantly correlated. It is notable that VO correlates with all other variables, while VSI correlates only with VO and N-NN DI.

**Table 5-5 Correlations among the Independent (age, AO, VO [lipreading], N-NN DI, Articulation and reading) and Dependent (VSI) Variables for Children.**

<table>
<thead>
<tr>
<th></th>
<th>VSI</th>
<th>Age</th>
<th>AO</th>
<th>VO</th>
<th>N-NN</th>
<th>Artic.</th>
<th>Read.</th>
</tr>
</thead>
<tbody>
<tr>
<td>VSI</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Age</td>
<td>.14</td>
<td>-</td>
<td>.07</td>
<td>.43**</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AO</td>
<td>.33**</td>
<td>.32**</td>
<td>.26**</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>VO</td>
<td>.21*</td>
<td>-.15</td>
<td>-.01</td>
<td>.05</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>N-NN DI</td>
<td>.07</td>
<td>.69**</td>
<td>.30**</td>
<td>.25**</td>
<td>-.15</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Articulation</td>
<td>.13</td>
<td>.86**</td>
<td>.46**</td>
<td>.35**</td>
<td>-.06</td>
<td>.68**</td>
<td>-</td>
</tr>
<tr>
<td>Reading</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

* Sig. at α=.05, **Sig at α=.01

**Table 5-6 Sequential Multiple Regression of age, AO, VO, N-NN DI, Articulation and Reading scores as predictors of VSI Scores.**

<table>
<thead>
<tr>
<th>Step</th>
<th>Variables</th>
<th>B at Step</th>
<th>β at Step</th>
<th>R² change at Step</th>
<th>F-value</th>
<th>Final B (Step 6)</th>
<th>Final β (Step 6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Age</td>
<td>.020</td>
<td>.140</td>
<td>.020</td>
<td>1.834</td>
<td>.026</td>
<td>.182</td>
</tr>
<tr>
<td>2</td>
<td>AO</td>
<td>.014</td>
<td>.012</td>
<td>.000</td>
<td>.011</td>
<td>-.050</td>
<td>-.042</td>
</tr>
<tr>
<td>3</td>
<td>VO</td>
<td>.314</td>
<td>.326</td>
<td>.094**</td>
<td>9.502</td>
<td>.298</td>
<td>.310**</td>
</tr>
<tr>
<td>4</td>
<td>N-NN DI</td>
<td>.089</td>
<td>.213</td>
<td>.044*</td>
<td>4.596</td>
<td>.089</td>
<td>.215*</td>
</tr>
<tr>
<td>5</td>
<td>Articulation</td>
<td>-.184</td>
<td>-.058</td>
<td>.002</td>
<td>.184</td>
<td>-.147</td>
<td>-.046</td>
</tr>
<tr>
<td>6</td>
<td>Reading</td>
<td>-.054</td>
<td>-.068</td>
<td>.001</td>
<td>.113</td>
<td>-.054</td>
<td>-.068</td>
</tr>
</tbody>
</table>

* Sig. at α=.05, **Sig at α=.01
Table 5-6 shows the unstandardised regression coefficients, (B), standardized regression coefficients (β), and $R^2$ change for the predictors at their step of entry and final B and β.

The introduction of variables age ($R=.140$, $R^2=.020$, $F(1,92)=1.834$, $p>.05$) in the first step, and AO scores ($R=.140$, $R^2=.020$, $F(1,91)=.011$, $p>.05$) in the second did not reliably predict VSI scores. Nevertheless, adding VO to the equation in the third step ($R=.337$, $R^2=.113$, $F(1,90)=9.502$, $p<.01$), and N-NN DI in the fourth step ($R=.396$, $R^2=.157$, $F(1,89)=4.596$, $p<.05$) revealed that these two variables reliably predict VSI scores. However, the addition of articulation ($R=.398$, $R^2=.159$, $F(1,88)=.184$, $p>.05$) in the fifth, and reading ($R=.400$, $R^2=.160$, $F(1,87)=.113$, $p>.05$) in the sixth and final step did not reliably improve $R^2$.

5.5.5.2 School Children: VO (Lipreading) Scores as Dependent Variable

The school children data were subjected to additional regression analysis in which VO scores were entered as the criterion variable with the VSI score added as a predictor variable. These analyses were conducted in order to investigate whether the predictive relationship between VO and VSI scores was one-way (i.e. VO predicts VSI, but not voice versa) or two-way (i.e. both variables predict each other). The independent variables were entered in the order of age, AO, VSI, N-NN DI, articulation, and reading. The results of this regression analysis showed that none of the independent variables reliably improve $R^2$. The regression and correlation analyses output files can be found in Appendices CD3.6.1 and CD3.6.2, respectively. In general, correlation analysis shows that VO scores are correlated with all variables except with N-NN DI scores. Interesting correlations were found between reading and VO, AO and articulation scores.
Results of evaluations of assumptions were satisfactory without any outliers and all individual Mahalanobis distance probability values were under (p<.001) the critical value set for four independent variables.

Table 5-7 shows the unstandardised regression coefficients, (B), standardized regression coefficients (β), and $R^2$ change for the predictors at their step of entry and final B and β. The results of this regression analysis showed that none of the independent variables reliably predict VO scores.

<table>
<thead>
<tr>
<th>Step</th>
<th>Variables</th>
<th>B at Step</th>
<th>β at Step</th>
<th>$R^2$ change at Step</th>
<th>F-value</th>
<th>Final B (Step 6)</th>
<th>Final β (Step 6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Age</td>
<td>.041</td>
<td>.277</td>
<td>.077</td>
<td>7.829</td>
<td>-.007</td>
<td>-.049</td>
</tr>
<tr>
<td>2</td>
<td>AO</td>
<td>.219</td>
<td>.174</td>
<td>.025</td>
<td>2.620</td>
<td>.168</td>
<td>.134</td>
</tr>
<tr>
<td>3</td>
<td>VSI</td>
<td>.182</td>
<td>.193</td>
<td>.036</td>
<td>3.873</td>
<td>.172</td>
<td>.182</td>
</tr>
<tr>
<td>4</td>
<td>N-NN DI</td>
<td>.043</td>
<td>.098</td>
<td>.009</td>
<td>.967</td>
<td>.036</td>
<td>.082</td>
</tr>
<tr>
<td>5</td>
<td>Articulation</td>
<td>.445</td>
<td>.133</td>
<td>.010</td>
<td>1.022</td>
<td>.301</td>
<td>.090</td>
</tr>
<tr>
<td>6</td>
<td>Reading</td>
<td>.191</td>
<td>.229</td>
<td>.013</td>
<td>1.345</td>
<td>.191</td>
<td>.229</td>
</tr>
</tbody>
</table>

* Sig. at α=.05, **Sig at α=.01

5.5.5.3 Adults: VSI Scores as Dependent Variable

A sequential multiple regression analysis was performed with VSI scores as the criterion and five predictors entered in order of developmental (age), linguistic (AO, VO, N-NN DI, articulation, and reading) factors. Table 5-8 shows correlations among predictor and criterion variables. Results of evaluations of assumptions were satisfactory and all individual Mahalanobis distance probability values were under .001, the critical value set for four independent variables (Tabachnik & Fidell, 2001). All three components of the McGurk task, AO, VO, and VSI, were significantly related to each other. The only other significant correlation was between reading ability and AO scores.
TABLE 5-8 CORRELATIONS AMONG THE INDEPENDENT (AO, VO, N-NN DI, ARTICULATION AND READING-WRAT) AND DEPENDENT (VO-LIPREADING) VARIABLES FOR ADULTS.

<table>
<thead>
<tr>
<th></th>
<th>VSI</th>
<th>AO</th>
<th>VO</th>
<th>N-NN</th>
<th>Artic.</th>
<th>Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>VSI</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AO</td>
<td>.39**</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VO</td>
<td>.27*</td>
<td>.28*</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N-NN DI</td>
<td>-.16</td>
<td>-.18</td>
<td>-.10</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Articulation</td>
<td>.07</td>
<td>.07</td>
<td>-.09</td>
<td>-.10</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Reading</td>
<td>.10</td>
<td>.25*</td>
<td>-.12</td>
<td>-.21</td>
<td>.23</td>
<td>-</td>
</tr>
</tbody>
</table>

*Sig. at α=.05, **Sig. at α=.01

As inspection of Table 5-9 shows, the results of regression analysis revealed that only AO scores significantly predict VSI scores and improve R² reliably.

TABLE 5-9 SEQUENTIAL MULTIPLE REGRESSION OF AO, VO, N-NN DI, ARTICULATION AND READING SCORES AS PREDICTORS OF VSI SCORES IN ADULTS.

<table>
<thead>
<tr>
<th>Step</th>
<th>Variables</th>
<th>B at Step</th>
<th>β at Step</th>
<th>R² change at Step</th>
<th>F-value</th>
<th>Final B (Step 6)</th>
<th>Final β (Step 6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AO</td>
<td>.871</td>
<td>.394</td>
<td>.155**</td>
<td>8.457</td>
<td>.713</td>
<td>.323**</td>
</tr>
<tr>
<td>2</td>
<td>VO</td>
<td>.519</td>
<td>.179</td>
<td>.030</td>
<td>1.630</td>
<td>.535</td>
<td>.184</td>
</tr>
<tr>
<td>3</td>
<td>N-NN DI</td>
<td>-.044</td>
<td>-.077</td>
<td>.006</td>
<td>.314</td>
<td>-.040</td>
<td>-.069</td>
</tr>
<tr>
<td>4</td>
<td>Articulation</td>
<td>.381</td>
<td>.052</td>
<td>.003</td>
<td>.142</td>
<td>.356</td>
<td>.049</td>
</tr>
<tr>
<td>5</td>
<td>Reading</td>
<td>.068</td>
<td>.019</td>
<td>.000</td>
<td>.015</td>
<td>.068</td>
<td>.019</td>
</tr>
</tbody>
</table>

*Sig. at α=.05, **Sig. at α=.01

5.5.5.4 Adults: VO (Lipreading) as Dependent Variable

The regression analyses presented in 5.5.5.3 were essentially repeated except that VSI scores were entered as a predictor variable and VO (lipreading) scores were entered as the dependent variable. Overall output files of these four analyses can be found in Appendices CD3.7.1 and CD3.7.2. The results of
regression analysis revealed that none of the independent variables predicted VO (lipreading) performance significantly (Table 5-10).

**Table 5-10 Sequential multiple regression of AO, VO, N-NN DI, articulation and reading (WRAT) scores as predictors of VO scores in adults.**

<table>
<thead>
<tr>
<th>Step</th>
<th>Variables</th>
<th>B at Step</th>
<th>β at Step</th>
<th>R² at Step</th>
<th>F-value</th>
<th>Final B</th>
<th>Final β</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AO</td>
<td>.211</td>
<td>.277</td>
<td>.077</td>
<td>3.819</td>
<td>.182</td>
<td>.240</td>
</tr>
<tr>
<td>2</td>
<td>VO</td>
<td>.067</td>
<td>.195</td>
<td>.032</td>
<td>1.630</td>
<td>.066</td>
<td>.193</td>
</tr>
<tr>
<td>3</td>
<td>N-NN DI</td>
<td>-.008</td>
<td>-.039</td>
<td>.001</td>
<td>.072</td>
<td>-.015</td>
<td>-.078</td>
</tr>
<tr>
<td>4</td>
<td>Articulation</td>
<td>-.302</td>
<td>-.121</td>
<td>.014</td>
<td>.706</td>
<td>-.205</td>
<td>-.082</td>
</tr>
<tr>
<td>5</td>
<td>Reading</td>
<td>-.239</td>
<td>-.192</td>
<td>.032</td>
<td>1.592</td>
<td>-.239</td>
<td>-.192</td>
</tr>
</tbody>
</table>

* Sig. at α=.05, **Sig. at α=.01

**5.6 Discussion**

As predicted, results for reading and articulation analyses show that there is an age-related increase in these two skills. Native *minus* Non-Native (N-NN) speech perception test results revealed no age related increase and native speech perception was better than for non-native speech perception in all age groups.

Analyses of auditory-visual speech perception test results showed that there was greater benefit of visual speech information in noise than clear conditions in all age groups, supporting earlier findings (Sekiyama & Burnham, 2004). Overall (combined VSI scores in clear and noise conditions and English and Japanese stimuli) results revealed a general age-related increase in the amount of visual speech influence, supporting previous results with English language environment children (Massaro, Thompson, Barron, & Laren, 1986; McGurk & MacDonald, 1976; Sekiyama & Burnham, 2004; Dupont, Aubin, & Ménard, 2005). In addition, results showed a foreign speaker effect such that when confronted with non-native speakers, perceivers use more visual speech information. Analysis of the auditory-visual speech perception test also revealed that lipreading ability increases significantly and that, over age,
children extract more accurate speech information from lip movements alone, thus supporting other age related results (Massaro et al., 1986; McGurk & MacDonald, 1976; Sekiyama & Burnham, 2004). Regression analyses show that language specific speech perception and lipreading reliably predict visual speech influence in English language environment school children aged between 5 and 8 years. However, no variable reliably predicted performance in lipreading. These differential results suggest that auditory-visual speech perception does not simply reflect lipreading per se. Auditory-visual speech perception rather appears to be an emergent ability in the auditory-visual domain, which appears to develop independently of lipreading. Previous developmental studies with infants and adults also support the view that auditory-visual speech perception is an emergent ability (Rosenblum et al., 1997; Sekiyama & Burnham, 2004).

Returning to the current results, why should greater relative attention to native than non-native speech contrasts predict visual influence in speech perception? Burnham (2003) showed that the degree of reading ability at the onset of reading instruction is related to language-specific speech perception in English-speaking children. It is reasoned that reading is a task involving high cognitive demand, especially in English with its opaque orthography and its one-to-many grapheme-phoneme correspondences (van den Bosch, Content, Daelemans, & De Gelder, 1994); thus the ability to attend to native contrasts and disregard non-native contrasts should aid reading (Burnham et al., 2002; Öney & Goldman, 1984). Given this, English-speaking children may also seek other extra information to establish phoneme-to-grapheme links and auditory-visual speech information might just be such an extra source. Moreover, it is possible that children who are good at attending to native speech contrasts and ignoring non-native speech contrasts are just those children who are good at using auditory-visual speech information.

In addition to the general linear increase in the amount of visual speech influence used over age, the data also point to a temporary decline in visual speech influence at around 7 years of age. This may also be related to the
reading acquisition process. It may be that age seven falls in the period in which reading starts to become an automatic skill and less controlled attention needs to be paid to phoneme-grapheme correspondences, so that visual speech information is not used as much at this age in reading acquisition. However, the data show that the 8-year-olds’ performance jumps back to the level of 6-year-olds’ so this cannot be the complete explanation of this cubic age trend. The reason for this increase in visual speech influence at (and after) 8 years of age may be due to other factors such as maturation and the gradual improvement in lipreading ability, which was found to predict auditory-visual speech perception, and was strongly correlated with age, articulation, and reading. In this way it may be the case that while auditory-visual speech perception develops as a separate construct, lipreading may act as a scaffolding factor in the development of various language-related abilities. Indeed VO scores are correlated with all other scores, but in a regression analysis VO scores are not predicted by any of these factors.

Adults’ data yield a different pattern. As opposed to the child data, no effect of lipreading or language specific speech perception on auditory-visual speech perception was found. The failure to find a significant link between lipreading or language specific speech perception and auditory-visual speech perception in adults may be because this link becomes redundant after sensitive periods of linguistic development. Lipreading ability and language specific speech perception may predict auditory-visual speech perception just during the orthographic stage (Burnham et al., 2002), in which a focus of attention on native speech sounds and away from non-native sounds aids reading development. In addition, in this stage, reading acquisition may also be facilitated by improved phoneme definition via visual information. Since by adulthood reading has presumably become an automatic skill with no active controlled matching of phoneme-to-grapheme, support from other sources is not required as much as in childhood.

So, what do these data tell us in terms of the difference between children’s and adults’ auditory-visual speech perception? As adults have already
established native speech prototypes, they may not need to process visual components of the speech signal in the way that children do, hence the link between language-specific speech perception and auditory-visual speech perception may be less salient in adulthood. In addition, the link between language specific speech perception and auditory-visual speech perception in reading age children might mark a transition from auditory-visual speech perception as a phonetic low level process present even in young infants (Rosenblum et al., 1997; Burnham & Dodd, 2004) to a more phonological, language specific process.

There might be other factors that lead to increased visual speech influence with age. The child participants tested in this experiment were monolingual English speakers, who came from ethnically homogenous home environments. At school, children are exposed to new people from various ethnic backgrounds, with various speech styles, accents (given the multicultural nature of Sydney metropolitan area), and even perhaps children with speech production disorders. On top of this, other distinct speech styles adopted by teachers such as ‘teacherese’ (Håkansson, 1987), might lead to increased use of visual speech information. The effect of such factors cannot be ascertained here, but warrant further investigation in future studies.

In sum, the current results show that language specific speech perception is linked to the degree to which visual information is used in speech perception, but only in childhood. In addition, against predictions, no link between auditory-visual speech perception and reading ability was found nor any between articulation and auditory-visual speech perception. Nevertheless, given that previous results show a link between articulation and auditory-visual speech perception (Desjardins et al., 1997; Siva et al., 1995), further investigation of the role of articulation in auditory-visual speech perception versus the role of language-specific speech perception is required.

The next question in this thesis regards the link between auditory-visual speech perception and language specific speech perception in children who have not yet been in a school environment.
CHAPTER 6

AUDITORY-VISUAL SPEECH PERCEPTION IN PRESCHOOL CHILDREN
6.1 BACKGROUND

It has been found that infants perceive both auditory and visual aspects of speech (Burnham & Dodd, 2004; Patterson & Werker, 1999, 2003; Rosenblum et al., 1997), and that young children’s auditory-visual speech perception abilities improve over age (see Chapter 4). In addition, we now know that school children’s, but not adults’ auditory-visual speech perception is linked to language specific speech perception (Experiment 1). The missing link in the research on the development of auditory-visual speech perception is the relationship between auditory-visual and language specific speech perception in toddlerhood between infancy and the school years. In the current study language specific speech perception, executive functions and receptive vocabulary knowledge data from preschool children aged 3 and 4 years are reported to investigate the role of these constructs in auditory-visual speech perception during this period. In the following section the background literature relevant to this study is given: first speech perception and its relationship to language and cognitive development, and then a description of language development, particularly the growth in receptive vocabulary, followed by cognitive development. Following this, the methodological details, results and discussion of the results of the current experiment are presented.

6.2 LANGUAGE AND COGNITIVE DEVELOPMENT IN PRESCHOOL CHILDREN

6.2.1 Language Development from Infancy to Childhood

Infants start life with a language-general speech perception capability, which is marked by the ability to perceive most, if not all, speech contrasts presented to them (see Chapter 3). Later in the first year of life, first the perception of vowels at around 4-6 months (Kuhl et al., 1992), and tones between 6 and 9 months (Mattock & Burnham, 2006) and consonants at around 7-11 months of age (Best, 1995) is reorganised as a result of experience with the ambient
language. There is a consensus that this reorganisation of speech perception abilities prepares the infant for native language development in the coming years (Burnham et al., 2002). A question of great interest is how cognitive and linguistic skills can be related to this perceptual reorganisation, especially in toddlers. For instance, in a longitudinal study, Tsao, Liu, and Kuhl (2004) tested 6-month-old English language environment infants on the /ti/ vs. /tu/ vowel contrast using a standard speech perception measure, the head-turn (HT) procedure. Later, they tested these infants at 13, 16, and 24 months of age on word understanding, word production, and phrase understanding, using the MacArthur-Bates Development Communicative Inventory (CDI), and found that speech perception ability predicted all the language abilities at all three ages, 13, 16, and 24 months, showing a strong link between phonological processing and cognitive abilities.

More recently, Kuhl, Conboy, Padden, Nelson, and Pruitt (2005) conducted another longitudinal study in which they tested 7-month-old ELE infants on a native (/ta/ vs. /pa/) and non-native speech contrast (Mandarin-native fricative-affricate /ci/ vs. /ʨʰi/), using an HT procedure. Later the same infants were tested at 14, 18, 24, and 30 months of age on the CDI, including measures of vocabulary production, sentence complexity, and mean length of utterance (MLU). The results showed that at 7 months, infants’ native and non-native speech perception abilities are negatively correlated, that is, good native speech perception is accompanied by relatively poor non-native speech perception, and vice versa. Additionally, the longitudinal data showed that native and non-native speech perception abilities at 7 months predict all three native language abilities tested at later ages. For example, children with better native speech perception at 7 months are those who have greater productive vocabulary.

Kuhl et al. (2005) suggest four possible mechanisms by which early speech perception abilities could affect later native language development. The first is that speech perception skills could assist infants to detect the phonotactic patterns that describe the legal phoneme combinations in the native language.
This knowledge should, in turn, lead to better word acquisition and recognition. Their second suggestion is that exposure to, and a good organisation of, native phonemes should lead to good native language phonemic representations, hence, better word recognition. A third proposal is that the ability to perceive differences among phonemes in running speech should lead to better word recognition. Finally, the fourth proposed mechanism is the association of sound patterns with objects, which assists in learning words.

Tsao et al. (2004) also discuss the relationship between early speech perceptual abilities and later language abilities. They suggested that infants’ cognitive and general auditory abilities are factors that might potentially account for the link between early speech perception and later native language abilities. They also suggested that speech perception measured by a paradigm such as the HT procedure is a good predictor, as adequate performance on this task requires a certain basic level of cognitive functioning (one in which the contingency between sound change and delivery of reward is cognised). Additionally, the HT procedure is based on general auditory skills that are relevant performance in speech perception.

The above studies show that early speech perception abilities are linked to the development of a number of linguistic abilities. The focus of the current experiment is on the possible link between auditory-visual speech perception and such linguistic abilities in the preschool years, particularly at ages 3 and 4. The next section outlines the development of word acquisition as a product of the development of phonological processing skills. Following this will be a brief discussion of the role of cognitive and executive functions.

6.2.1.1 Vocabulary Development from Infancy to Childhood

During what has been called the semantic stage of speech perception development (Burnham et al., 2002), it appears that the earlier reorganisation of early speech perception skills prepares young children for vocabulary (Nazzi & Bertoncini, 2003) and syntactic (Christiansen & Dale, 2001)
acquisition. During this stage young children start learning sound-object associations. That this involves significant cognitive resources is shown by Stager and Werker’s (1997) finding that between 8 and 14 months of age the ability to attend to fine phonetic detail in word learning tasks attenuates to allow increased attention to word-sound relationships. Productive vocabulary begins to emerge around 12 months, as evidenced by observational (Benedict, 1979) and experimental studies ( Werker et al., 1998). Initially productive vocabulary growth is relatively slow from at around 1 to 3 words to 50 words per week. Around 18 months of age there is a sudden increase in the rate of word acquisition to up to 9 words a day (Benedict, 1979), a phenomenon referred to as the vocabulary spurt (A. L. Woodward, Markman, & Fitzsimmons, 1994). Data from Nazzi and Bertoncini (2003) indicate that sound-object pairings prior to the vocabulary spurt are simple associative pairings of sound patterns and object categories, with more cognitively-oriented representation of the world via words and concepts acquired after the onset of the vocabulary spurt.

Several studies suggest that the vocabulary spurt is related to the acquisition of various cognitive skills (e.g., object permanence), the ability to remember the existence of objects that have been hidden, (Corrigan, 1978), object representation (Lifter & Bloom, 1989), and object categorisation (Gopnik & Meltzoff, 1987; Mervis & Bertrand, 1994). For instance, in a longitudinal study, Gopnik and Meltzoff (1987) tested 12 toddlers first at 16 months of age, and then 20 months of age and found a strong relationship among measures associated with object permanence, two-category grouping, and the rate of vocabulary acquisition.

In summary, the foundation of language acquisition appears to be laid down in the early stages of speech perception development. Several studies have shown that auditory speech perception development is linked to linguistic abilities, such as vocabulary development (Nazzi & Bertoncini, 2003; Stager & Werker, 1997) and that vocabulary development is related to object permanence and concept formation (Corrigan, 1978; Lifter & Bloom, 1989).
We now turn attention to how relevant cognitive abilities develop in the preschool years.

6.2.2 Cognitive Development and Executive Functions in Preschool Years

Human action may usefully be divided into two classes: automatic actions, which do not require the exercise of effortful mental processes (e.g. driving between home and work, playing piano professionally), and controlled actions, which require planned, effortful behaviour (e.g. driving in a foreign city, learning how to fly an aeroplane). Controlled actions can be viewed as arising from executive functions: cognitive functions that enable people to plan, initiate, and execute goal-directed behaviour (Oates & Greyson, 2004). Such functions are used in situations that require such things as new skill acquisition, planning, decision making, error correction and troubleshooting, initiating sequences of actions, attending dangerous or difficult circumstances, conscious control of behaviour and the overcoming of strong habitual responses. Perhaps the most relevant of these situations to language development in children is new skill.

Two important components of executive functions are rule abstraction and cognitive flexibility. Cognitive flexibility refers to the aspect of executive function that enables people to think and act according to the changing circumstances of a complex environment (Oates & Greyson, 2004) and is likely based on rule abstraction ability (Jacques & Zelazo, 2001).

A number of studies have shown that particularly, in the preschool period, there is improvement in executive functions as reflected in performance on the Stroop effect task\(^\text{16}\). It was shown that between 3 and 7 years of age children improve on cognitive tasks such as Stroop test (Gerstadt, Hong, &

---

\(^{16}\) In Stroop effect when a colour name such as blue, green, red, etc. is printed in a font colour which is different from the colour expressed by the word's semantic meaning (e.g. the word "green" printed in yellow ink), the time required to name the word's colour takes longer than naming the colour words printed in matching font colour.
Diamond, 1994) and tasks requiring the sorting of items interrelated on the basis of colour, shape, size and number (Frye, Zelazo, & Palfai, 1995), such as matching a red circle and a red square that share the same dimension of colour. In the latter case, 4- and 5-year-olds performed better than 3-year-olds on sorting tasks that require comprehending the link between objects which are colour-, shape-, size, or number-related. Similar findings were obtained by Jacques and Zelazo (2001): 3-year-olds were worse at tasks requiring rule abstraction than 4-, and 5-year-olds, and both 3-, and 4-year-olds were poorer than 5-year-olds in tasks requiring cognitive flexibility. In summary the above studies suggest a link between cognitive abilities and language development. Auditory-visual speech perception is an ability that requires the processing and integration of information from multiple sources. In this respect it could be predicted that cognitive abilities in the form of executive functions might be related to auditory-visual speech perception and lipreading as they also appear to develop with age during this period (e.g. Desjardins et al., 1997). The next section presents the aims, rationale and hypotheses of Experiment 2.

6.2.3 The Current Experiment: Aims, Rationale, and Predictions

In Experiment 1 a predictive relationship between language specific speech perception and auditory-visual speech perception was found in the early school years. In addition, lipreading ability was found to substantially improve between five and eight years and to be correlated with auditory-visual speech perception, reading, and articulation. It was suggested that at times of linguistic challenge, such as reading acquisition, and encountering unfamiliar speech styles, extra speech information is sought, in particular lipread and auditory-visual speech information. There are currently a number of studies that addressed infant auditory-visual speech perception (Burnham & Dodd, 2004; Patterson & Werker, 1999, 2003; Rosenblum et al., 1997), adult auditory-visual speech perception (Sekiyama, 1997a; Sekiyama & Burnham, 2004), and school-age auditory-visual speech
perception (Erdener & Burnham, 2005b; Sekiyama & Burnham, 2004). What remains to be investigated is auditory-visual speech perception in early childhood, especially with regard to how it interacts with other aspects of linguistic and cognitive development. In particular, it is predicted that since word learning and vocabulary development involve significant investment of cognitive resources (Stager & Werker, 1997), then such linguistic challenges would result in greater auditory-visual speech perception and in greater language specific speech perception.

In this experiment 3- and 4-year-old children were given tests of auditory-visual speech perception, language specific speech perception, cognitive flexibility, and receptive vocabulary knowledge. Over and above the general prediction that 4-year-olds should perform better than 3-year-olds on all measures it was hypothesised that auditory-visual speech perception should be predicted by lipreading, language specific speech perception, cognitive skills, and vocabulary growth in a predictive relationship.

6.3 Experiment 2: Determinants and Development of Auditory-Visual Speech Perception in 3- and 4-Year-Olds

6.4 Method

6.4.1 Design
Experiment 2 employed a two-factor, age (2 levels: 3-, and 4-year-olds) by task (4 levels: auditory-visual speech perception, language specific speech perception [N-NN], executive function, and receptive vocabulary knowledge) design, with repeated measures on the task factor. Details of how the levels of these factors were investigated are outlined below.

6.4.2 Participants
Twenty-four 3-year-old ($M_{age}$=3.08 years, $sd=0.17$ years) and 24 4-year-old ($M_{age}$=4.21 years, $sd=0.16$ years) preschool children were tested (N=48) with
equal number of males and females in each group. All children were from monolingual Australian English-speaking families. Parental reports indicated that all children had normal hearing and vision and none had corrected hearing or vision. All participants were born full-term and had no middle ear infection history according to parental report. The children were selected and recruited from the BabyLab register\textsuperscript{17} of MARCS Auditory Laboratories at the University of Western Sydney. Parents of children who met the selection criteria were initially contacted by telephone, and appointments were arranged. In addition to the initial written information sheet sent to parents, prior to the experiment the parents were also orally informed about the task requirements (Appendix CD1.4). Parents received $20 to reimburse travel costs and children received a soft toy of their choice and a “Young Scientist” certificate (Appendix CD1.5).

6.5 Dependent Variables, Stimuli, Apparatus and Procedure

Children were tested individually. Parents were asked not to give feedback to their child during experiment, but they were asked to assist the experimenter during the experiment if needed. Prior to testing sessions, the experimenter spent at least 15 minutes in the MARCS BabyLab reception area with the parent playing and interacting with each child in order to establish good rapport and communication. All participants were tested with their parent present in the testing room (all but one with their mother) because (i) the presence of a responsible guardian was an ethical requirement, (ii) the presence of a parent assisted in providing a high sense of security for children, allowing them to concentrate on the tasks, and (iii) children may feel more comfortable asking for help from their parents. In the latter case, parents

\footnote{\textsuperscript{17}The BabyLab register is the infant and child participant database of MARCS Auditory Laboratories. The register contains the details of infants and children recruited via newspaper and baby magazine advertisements in the Sydney metropolitan area, recruitment days at local shopping centres, and distribution of pamphlets to Early Childhood Centres and private hospitals in Southern Sydney metropolitan region.}
were free to explain to their children what they should do to respond (e.g. “press the red button when you hear a sound change”) but not how or when they should respond.

Children were required to do four tasks: auditory-visual speech perception, language specific speech perception, flexible item selection and Peabody Picture Vocabulary tests. The task order was counterbalanced across participants (Appendix A2). The following sections describe the stimuli, apparatus and specific procedural details of these four tasks.

6.5.1 Auditory-Visual Speech Perception Test

6.5.1.1 Dependent Variable and Stimuli

Auditory-visual speech perception was measured using an AX discrimination task in which two stimuli were presented and participants are required to indicate whether the two were the “same” or “different”. The reason for using a discrimination task rather than identification was to reduce the number of response alternative and make the task easier for the 3- and 4-year-olds. A discrimination task visual speech index (VSI-AX) was calculated using only “different” auditory-visual incongruent (AV-) speech pairs, which differed either on both auditory and visual components or on the visual component alone. The “same” trials were not included in this calculation as any response (“same” or “different”) on these items would not reveal on which component, auditory or visual, of the stimulus pair the response was based. For the same reason the congruent (AV+) auditory-visual speech contrasts (e.g. same Auditory-[ba]+Visual[ba] vs. Auditory-[ba]+Visual[ba] and different Auditory-[ba]+Visual[ba] vs. Auditory-[ga]+Visual[ga]) were also not included in the analyses and were included in the study as same/different and congruent/incongruent counterbalancing items. For example, if perceivers responded ‘same’ to the auditory-visual speech contrast Auditory-[ba]+Visual[ga] vs. Auditory-[ga]+Visual[ga], then this response was deemed as visually based response, and given a raw visual score of ‘1’. The resultant total number of visually based responses to incongruent different trials was
divided by the total number of incongruent different trials (9), thus resulting in an overall VSI-AX score (Equation 6-1). The maximum possible VSI-AX score of ‘1’, indicates a very strong visual speech influence and the minimum possible VSI-AX score of ‘0’ indicates a very weak visual speech influence. The list of contrasts can be found in Appendix CD 6.2.

\[
\text{VSI-AX} = \frac{\sum \text{Visually-based Responses on Different (AV-) trials}}{\sum \text{Different (AV-) trials}}
\]

Equation 6-1 The calculation of McGurk discrimination task visual speech index score (VSI-AX).

In this experiment, only stimuli produced by native English speakers were used and no background noise was introduced. There were 12 auditory-only (AO), 12 visual-only (VO) and 36 auditory-visual speech stimuli, a total of 60 stimuli. Details of how these stimuli were created were given in 5.3.1, and the files for auditory-visual speech stimuli were identical to the corresponding stimuli in Experiment 1.

6.5.1.2 Apparatus and Experimental Setup

The auditory-visual speech perception stimuli were presented on a laptop computer (Compaq Evo N1000c) equipped with a Pentium 4 microprocessor (1.7GHz) and 256 MB RAM, which was a suitable configuration for playing video files. The computer was connected to a 17-in flatscreen CTR monitor (Sony 17GS), which had a very high refresh rate (86 Hz) in comparison with the notebook computer’s built-in LCD monitor (65 Hz). For better sound quality, a loud speaker (Aiwa SC-B10) was connected to the computer and positioned centrally on top of the monitor in the participant’s sagittal plane. The parent was seated behind their children. The experimental set up of the auditory-visual speech perception test is depicted in Figure 6-1.
6.5.1.3 Procedure
In order to ensure that children, especially 3-year-olds, understood the task requirements they were presented with as many practice items as required. The experimenter also showed a picture of a circle and star together and a picture of two circles together to ensure that they understood the concepts “different” and “same”, respectively. On each trial children were presented with two auditory-visual speech contrasts and asked whether the two items were same or different. To make the task more enjoyable and understandable to the children, the same and different concepts were put into a context. The first items were always produced by a male speaker and the second items were produced by a female speaker. The children were told that the man was to try to teach the lady how to say some funny words and that they should indicate whether the lady had said the same thing as the man. There were a total of 60 trials (36 AV, 12 AO and 12 VO). Children’s responses were recorded manually using a response form and later transferred to an Excel file for data analysis. The task took between 20 and 40 minutes depending on whether further instructions and/or a break was required.

**Figure 6-1** The experimental set-up of auditory-visual speech perception test.
6.5.2 Speech Perception Test

6.5.2.1 Dependent Variable and Stimuli

As in Experiment 1, the dependent variable for language-specific speech perception was the difference between discrimination index (DI) scores for native and for non-native speech contrasts. These discrimination index scores were calculated by taking the difference between the number of correct “different” responses on (AB) trials (hits) and the number of incorrect “different” responses on same (AA) trials (false positives) divided by the total number of trials (see Equation 5-2 in Chapter 5). This formula yields a score between −1 and +1, where −1 indicates incorrect association of trial types and response, +1, a high level of performance, and 0 a chance level responding. The same speech stimuli as in Experiment 1 were used.

Each correct trial in the speech perception test was rewarded with a 5-second duration cartoon clip, excerpts from short cartoon stories. To create these reward stimuli, small sections of each story were recorded on digital videotapes via a video camera. These were then transferred to a laptop computer (Compaq Evo N1000c) and edited into 5-second clips using the U-Lead video editing software. The cartoon clips were saved as video files in mpeg format at 26 frames per second rate.

6.5.2.2 Apparatus and Experimental Setup

The stimuli were presented on a Macintosh iBook laptop computer equipped with a 256MB built-in memory and Macintosh OS 9.2 operating system, using PsyScript software (Bates & D'Oliveiro, 2003) via a 17-in CRT flatscreen monitor (Sony 17GS). Responses to the speech stimuli were collected via a response box (modified game controller) with a large red button (7-cm diameter) which children pressed to record their responses. Auditory stimuli were presented via a loud speaker (Edirol M527) placed in the participants’

---

18 PsyScript is an experiment generating software developed by Timothy Bates at Macquarie University, Sydney. To download: [http://www.maccs.mq.edu.au/~tim/psyscript/](http://www.maccs.mq.edu.au/~tim/psyscript/)
sagittal plane and on top of the monitor. The experimental set up for the speech perception test is presented schematically in Figure 6-2.

6.5.2.3 Procedure

Children were seated in front of their parents, with the monitor placed around 50 cm away in their sagittal plane. A response box with which children’s responses were recorded was placed either in front, or beside, or on the children’s lap, depending on each child’s preference.

The language specific speech perception procedure for school children used in Experiment 1 was adapted to make it easier for preschoolers. A category change paradigm was used instead of a discrimination task. There were two reasons for modification. First, category change paradigm is essentially a discrimination task, so this version of the task is not any different than the speech perception task in Experiment 1. Second, correct responses were rewarded with the presentation of cartoon clips as described below. Continual presentation of the speech stimuli was thought to ensure that children would keep their attention on the task more than a simple discrimination task. In this paradigm, change (or different) trials one sound was presented from 2 to up to 6 times (randomly varying across trials) after which a second (the change) sound was presented and played from 4 times up to 8 times. In no-change (or same) trials the same sound was played 10 times. There were two phases in the test: familiarisation and test. The category change paradigm was used in all phases of the test.

The training phase consisted of three components (demonstration, pretask, and task competence). In the demonstration phase children were presented with the sound of a rooster “crow”, and were instructed to press the red button on the response box as soon as they heard a cow’s “moo”. There were 4 trials (2 change and 2 no-change), and these were repeated if necessary. The demonstration phase was followed by the pretask competence phase in which the rooster and cow sounds were replaced by the words ‘rag’ and ‘rug’, which differ only on the vowel component (/æ/ vs. /a/). These stimuli were spoken
by a female native Australian-English speaker. There were 8 trials in total in this phase (4 change and 4 no-change), and the children were required to respond correctly to 6 of the 8 items in order to proceed to the next phases of the test. In the task competence phase, children were required to discriminate randomly chosen speech contrasts used in the testing phase, and there were 8 task competence trials in total. In the testing phase, there were 18 native (English native [pa] vs. [pʰa]) and 18 non-native (Thai-native [ba] vs. [pa]) trials. The native and non-native speech contrasts are depicted in Table 6-1. The native and non-native speech contrasts were presented in separate blocks, and order of blocks was counterbalanced between subjects. In addition, the trials within blocks were also randomised. In all three training and the test phases, each correct response (button pressing for change and no button pressing for no-change trials) was followed by a 5-second cartoon clip that acted as a reward and as means to keep children’s attention on the task. The cartoon clips were stored in separate folders based on three stories. Every first cartoon clip was selected randomly from one of these folders, and the subsequent reward cartoon clips followed the first selected file in order to maintain a storyline. The children’s task was simply to press the big red button on the response box as soon as they heard the sound change. The responses were stored on the laptop computer for later conversion and data analysis.

**Table 6-1 Native and Non-native Speech Pairs Created for the Language Specific Speech Perception Task.**

<table>
<thead>
<tr>
<th>Speech Contrast</th>
<th>Native in English or Thai</th>
<th>Change or No-Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>ba vs. ba</td>
<td>English &amp; Thai</td>
<td>No change</td>
</tr>
<tr>
<td>pʰa vs. pʰa</td>
<td>English &amp; Thai</td>
<td>No change</td>
</tr>
<tr>
<td>pʰa vs. ba</td>
<td>English &amp; Thai</td>
<td>Change</td>
</tr>
<tr>
<td>ba vs. pʰa</td>
<td>English &amp; Thai</td>
<td>Change</td>
</tr>
<tr>
<td>ba vs. ba</td>
<td>English &amp; Thai</td>
<td>No change</td>
</tr>
<tr>
<td>pa vs. pa</td>
<td>Thai</td>
<td>No change</td>
</tr>
<tr>
<td>pa vs. ba</td>
<td>Thai</td>
<td>Change</td>
</tr>
<tr>
<td>ba vs. pa</td>
<td>Thai</td>
<td>Change</td>
</tr>
</tbody>
</table>

19 [ba] and [pa] are phonemic in Thai and but they are allophones of English /ba/.
6.5.3 Executive Functions Test

6.5.3.1 Dependent Variable and Stimuli

To measure executive function the *Flexible Item Selection Task (FIST)* was used (Jacques & Zelazo, 2001). This test was designed to measure two executive function constructs: rule abstraction, and cognitive flexibility. There was a total of 18 trials split into three sections: *item identification task*, *favourite item selection task*, and *flexible item selection task*. The procedural details of each task are explained in section 6.5.3.3.

The test results in two scores: a rule abstraction score (/15), and a cognitive flexibility score (/15), which combine to an overall score (/30). Responses were recorded on a response sheet provided by the test designers.

6.5.3.2 Apparatus

Stimulus items were printed on A4 size paper in landscape orientation and laminated. In addition there was a standard response record card provided by the test developers (Appendix CD6.4).
Figure 6.3 Example of a Flexible Item Selection Task (FIST) trial card. In this trial colour (red & yellow) and shape (boat & shoe) are relevant dimensions whereas the size is the irrelevant dimension.

6.5.3.3 Procedure

Every child was given the same standard set of instructions. (Appendix CD6.3). In order to teach children which finger to use to select response items in the tasks described below. A sticker was put on the index finger of the dominant hand of each child and the dominant index finger of the experimenter before each FIST session. The test consisted of three phases. In the first, the item identification phase (3 trials), children were asked to use their ‘stickered’ dominant index finger to identify three colours (yellow, blue, and red), three shapes (boat, shoe, and teapot), and three sizes (big, medium, and small). This was done to ensure that children were familiar with these three colour, shape, and size dimensions. Children were given feedback by the praising of a correct response or by the prompting of the correct answer by the provision of cues. In the favourite item selection task the experimenter demonstrated to children how to select two pictures out of three pictures. Each child was presented with a card with three pictures of an umbrella, a toaster, and a guitar. The experimenter first chose two items (e.g., guitar and umbrella) as his two favourite items using his ‘stickered’ dominant index finger. Then he indicated two other items as his favourite items (e.g. umbrella and toaster), again using his dominant index finger. Children were then asked to do the same over the next two trials. The test phase, flexible item selection
task began with two familiarisation tasks in which children were asked to point with their dominant index finger to two items that go together in one way (e.g., and with respect to Figure 6-3, yellow boat and yellow shoe). This first matching task in a trial was aimed to measure rule abstraction. Then the experimenter asked children to choose two other items that go together in another way (e.g., two boats). This second question was aimed to measure cognitive flexibility. Thus on each test trial there were two tasks, rule abstraction and a separate cognitive flexibility task; a separate score was given for each. In the test phase there were three familiarisation trials, one completed by the experimenter and two by the child, followed by 15 test trials. Each test trial resulted in two scores, a rule abstraction and a cognitive flexibility score.

6.5.4 Receptive Vocabulary Knowledge Test

6.5.4.1 Dependent Variable and Stimuli

Vocabulary knowledge was measured using Version 3 of the Peabody Picture Vocabulary Test (PPVT-III) (Dunn & Dunn, 1997). PPVT-III is a standard, commonly used receptive vocabulary test that requires minimal verbal instructions with responses extracted by means of a picture naming task. It features blocks of 12 words, which are presented in order of difficulty. The words in each block are assumed correspond to the vocabulary knowledge of child’s typical age. For instance the block labelled as “4 years” corresponds to the receptive vocabulary knowledge of a typical 4-year-old. The dependent variable for the vocabulary knowledge was the PPVT-III score for each participant based on their level of receptive vocabulary knowledge and age in years and months. Details pertaining to stimulus and procedural details of PPVT-III are presented in the following sections.

6.5.4.2 Apparatus

The PPVT-III items were presented using the standard picture cards and scoring sheets provided by the test publishers.
6.5.4.3 Procedure

Each child began with the block of items appropriate for their age group. That is, 3-year-olds began with the block labelled ‘Ages 2.6 - 3’ and 4-year-olds began with the block ‘Age 4’. Each block contained 12 trials. In each trial children were shown four pictures of animals or objects on a card (e.g., frog, fish, cat, dog) and asked to point to the target word named by the experimenter (e.g., dog). If children pointed at the correct picture, then the experimenter presented the next item. The test progresses until a child made 8 errors in a block or until the test has finished when all 14 blocks were completed, in any case none of the children continued beyond the tenth block. Responses were recorded on a standard response sheet, and a score for each child was computed using a conversion table provided with the PPVT package.

6.6 Results

Results are presented in two sections. In the first section the between-subject comparisons of two age groups of children (3- and 4-year-olds) are presented. In the second section, the results of a set of regression analyses to assess the degree to which auditory-visual speech perception can be predicted by the performance on the other tests are described.

6.6.1 Auditory-Visual Speech Perception

The VSI-AX, AO, and VO data are presented in Figure 6-4 and Figure 6-5 and output files for auditory-visual speech perception test analyses can be found in Appendix CD4.6. An independent t-test was conducted on the visual speech index discrimination (VSI-DX) scores of the two age groups, with α set at .05. There were no outliers and Levene’s test for equality of variances indicated that all assumptions of normality and homogeneity of variance were met. The mean VSI-AX score of 4-year-olds ($M=0.78$, $sd=0.23$) was significantly higher than the VSI-AX score of the 3-year-old group ($M=0.53$, $sd=0.18$), $t(46)=3.927$, $p<.001$. The t-test analysis of the AO data showed that
the 4-year-old group (M=0.79, sd=0.15) performed significantly better than their 3-year-old (M=0.59, sd=0.21) counterparts, t(46)=4.171, p<.001. The groups differed on the degree of VO (lipreading) ability, however this difference was just marginally non-significant t(46)=1.860, p=.07.

**Figure 6-4** Mean VSI-AX scores of 3- and 4-year-olds. The error bars show the standard error of the mean.

**Figure 6-5** Mean auditory-only (AO) and visual-only (VO) scores of 3- and 4-year-olds. The error bars show the standard error of the mean.
6.6.2 Speech Perception Test

The N-DI, NN-DI, and N-NN DI scores are presented schematically in Figure 6-6. ANOVA output files for the speech perception test can be found in CD4.7. Data were converted to a spreadsheet format for data analysis using MATLAB (for language specific speech perception data conversion procedural details see Appendix A3). The Native (N-DI) and non-native (NN-DI) discrimination index scores were subjected to a 2 x 2 (age x native/non-native) analysis of variance. This revealed a significant overall difference between the two groups of children, $F(1,46)= 7.709, MS_e=.045, p<.01$, but no native vs. non-native speech effect, $F(1,46)= 0.079, MS_e=.042, p<.001$, or interaction with age [$F(1,46)= 1.095, MS_e=.042, p>.05$].

![Figure 6-6 Mean Native DI and Non-native DI Scores of the 3- and 4-Year-Old Groups. Error bars represent the standard error of the mean.](image)

6.6.3 Executive Function Test

Figure 6-7 shows the mean scores on rule abstraction, cognitive flexibility, and overall FIST scores. The output files can be found in Appendix CD4.8. The two age groups’ rule abstraction and cognitive flexibility scores were subjected to a 2 x 2 (age x rule abstraction / cognitive flexibility) analysis of
variance. The results showed an overall statistically significant improvement between 3 and 4 years of age, $F(1,46)= 10.078$, $MS_e=56.361$, $p<.001$, and a main effect with rule abstraction/ cognitive flexibility scores, $F(1,46)= 131.324$, $MS_e=8.902$, $p<.001$, but no interaction with age, $F(1,46)= 0.338$, $p>.05$.

![Mean Executive Function Test Score (115)](image)

**Figure 6-7** Mean rule abstraction, cognitive flexibility and overall FIST (rule abstraction + cognitive flexibility) scores of the 3- and 4-year-old groups. Error bars represent the standard error of the mean.

### 6.6.4 Receptive Vocabulary Test

Mean receptive vocabulary scores of 3- and 4-year-old preschool child groups are presented in Figure 6-8. Overall output files can be found in Appendix CD4.9. Three- and 4-year-old preschool children’s receptive vocabulary data were subjected to an independent samples $t$-test analysis with $\alpha$ set at .05. Assumptions of normality and homogeneity of variance were met without any outliers. The results revealed a statistically significant improvement between 3 ($M=43.3$, $SD=14.3$) and 4 years of age ($M=67.6$, $SD=13.0$), $t(46)=-6.145$, $MS_e=3.95$, $p<.001$. 

133
**Figure 6-8** Mean receptive vocabulary (PPVT) scores of 3- and 4-year-old preschool children. Error bars show the standard error of the mean.

### 6.6.5 Regression Analyses

The regression analyses of the data from 3- and 4-year-old groups were run in two separate sets. In the first set of analyses, the VSI-AX scores were entered as the dependent variable and developmental (age) and linguistic (AO, VO, N-NN DI) and cognitive variables (executive function and vocabulary knowledge) were entered as predictors. In the second set of analyses, VO (lipreading) scores were entered as dependent variable and VSI-AX scores were added as predictors along with other independent variables as per in the first analysis.

#### 6.6.5.1 VSI-AX Scores as Dependent Variable

A sequential multiple regression analysis was performed with VSI-AX scores as the criterion and four predictors in the order of developmental (age), linguistic (AO, VO, N-NN DI, & PPVT scores) and cognitive (FIST scores). The factor analyses of two subsets of FIST scores revealed a single factor component score (Appendix CD4.10), and factor scores based on this were entered into the regression analysis. Overall correlation coefficients are

134
presented in Table 6-2, showing that VSI-AX scores are highly correlated with all factors except VO scores, which are correlated with PPVT and FIST scores. Age is correlated with AO, PPVT and FIST scores and N-NN DI scores are correlated with VSI-AX, PPVT and FIST scores. Table 6-3 presents the unstandardised regression coefficients, (\(B\)), standardized regression coefficients (\(\beta\)), and \(R^2\) change for the predictors at their step of entry and final B and \(\beta\). Output files of regression and correlation analyses are provided in Appendix CD4.11.

**Table 6-2 Correlation Coefficients Among the Independent (AO, VO [LIPREADING], N-NN DI, Rule Abstraction, Cognitive Flexibility and PPVT) and Dependent (VSI-AX) Variables.**

<table>
<thead>
<tr>
<th></th>
<th>VSI-AX</th>
<th>Age</th>
<th>AO</th>
<th>VO</th>
<th>N-NN</th>
<th>PPVT</th>
<th>FIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>VSI-AX</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>.45**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AO</td>
<td></td>
<td>.54**</td>
<td>.42**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VO</td>
<td></td>
<td></td>
<td>.21</td>
<td>.21</td>
<td>.37**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N-NN DI</td>
<td>.30*</td>
<td>.10</td>
<td>.21</td>
<td>.10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PPVT</td>
<td>.45**</td>
<td></td>
<td></td>
<td>.56**</td>
<td>.27*</td>
<td>.41**</td>
<td></td>
</tr>
<tr>
<td>FIST</td>
<td>.51**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Sig. at \(\alpha=.05\), **Sig. at \(\alpha=.01\)

Results of evaluations of assumptions were satisfactory after an outlier with standardized residual greater than 3 standard deviations was omitted from the analysis. Of the five predictors entered into the equation, AO at Step 2 and FIST scores at Step 6 reliably predicted VSI-AX scores and reliably increased the \(R^2\), whereas age at Step 1, VO scores at Step 3, N-NN DI at Step 4, and vocabulary scores at Step 5 did not reliably increase \(R^2\) and predict VSI-AX scores.
### Table 6.3 Sequential Multiple Regression of AO, VO, N-NN DI, FIST and PPVT Scores as Predictors of VSI-AX Scores.

<table>
<thead>
<tr>
<th>Step</th>
<th>Variables</th>
<th>B at Step</th>
<th>β at Step</th>
<th>R² change at Step</th>
<th>F-value</th>
<th>Final B (Step 6)</th>
<th>Final β (Step 6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Age</td>
<td>.180</td>
<td>.445</td>
<td>.198</td>
<td>11.390</td>
<td>.104</td>
<td>.259</td>
</tr>
<tr>
<td>2</td>
<td>AO</td>
<td>.493</td>
<td>.424</td>
<td>.346**</td>
<td>10.176</td>
<td>.467</td>
<td>.402**</td>
</tr>
<tr>
<td>3</td>
<td>VO</td>
<td>-.003</td>
<td>-.002</td>
<td>.346</td>
<td>.000</td>
<td>-.140</td>
<td>-.085</td>
</tr>
<tr>
<td>4</td>
<td>N-NN DI</td>
<td>.155</td>
<td>.189</td>
<td>.380</td>
<td>2.359</td>
<td>.092</td>
<td>.112</td>
</tr>
<tr>
<td>5</td>
<td>PPVT</td>
<td>-.001</td>
<td>-.041</td>
<td>.381</td>
<td>.045</td>
<td>-.002</td>
<td>-.152</td>
</tr>
<tr>
<td>6</td>
<td>FIST</td>
<td>.099</td>
<td>.340</td>
<td>.451*</td>
<td>5.196</td>
<td>.099</td>
<td>.340*</td>
</tr>
</tbody>
</table>

* Sig. at α=.05, **Sig at α=.01

### 6.6.5.2 VO Scores as Dependent Variable

The Experiment 2 data were subjected to a second regression analysis in which VO scores were entered as dependent variable and VSI-AX scores were added to the equation as an independent variable. These analyses were conducted in order to investigate whether VO scores were related to and predicted by VSI-AX and N-NN DI, FIST and PPVT scores. The independent variables were entered in the order of AO, VSI-AX, N-NN DI, rule abstraction, cognitive flexibility, and PPVT scores. Table 6-4 presents the unstandardised regression coefficients, (B), standardized regression coefficients (β), and $R^2$ change for the predictors at their step of entry and final B and β. Output files of regression and correlation analyses are provided in Appendix CD4.11. The results of this regression analysis revealed that none of the independent variables entered into the equation predicted VO (lipreading) scores reliably nor significantly improved the $R^2$, though it should be noted that FIST scores were very close to significance level ($p=.058$).
### Table 6-4 Sequential Multiple Regression of AO, VSI-AX, N-NN DI, Rule Abstraction, Cognitive Flexibility, and PPVT Scores as Predictors of VO (Lipreading) Scores.

<table>
<thead>
<tr>
<th>Step</th>
<th>Variables</th>
<th>B at Step</th>
<th>β at Step</th>
<th>R² at Step</th>
<th>F-value</th>
<th>Final B (at Step 5)</th>
<th>Final β (at Step 5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Age</td>
<td>.051</td>
<td>.207</td>
<td>.043</td>
<td>2.061</td>
<td>.005</td>
<td>.022</td>
</tr>
<tr>
<td>2</td>
<td>AO</td>
<td>.247</td>
<td>.349</td>
<td>.143</td>
<td>5.260</td>
<td>.246</td>
<td>.348</td>
</tr>
<tr>
<td>3</td>
<td>VSI-AX</td>
<td>-.002</td>
<td>-.003</td>
<td>.143</td>
<td>.000</td>
<td>-.074</td>
<td>-.121</td>
</tr>
<tr>
<td>4</td>
<td>N-NN DI</td>
<td>.009</td>
<td>.017</td>
<td>.143</td>
<td>.013</td>
<td>-.038</td>
<td>-.076</td>
</tr>
<tr>
<td>5</td>
<td>PPVT</td>
<td>.001</td>
<td>.085</td>
<td>.146</td>
<td>.139</td>
<td>.000</td>
<td>-.042</td>
</tr>
<tr>
<td>6</td>
<td>FIST</td>
<td>.062</td>
<td>.352</td>
<td>.219</td>
<td>3.811</td>
<td>.062</td>
<td>.352</td>
</tr>
</tbody>
</table>

* Sig. at α=.05, **Sig at α=.01

### 6.7 Discussion

The results yield that 4-year-olds showed more auditory-visual speech integration than the 3-year-olds, in line with improvements over age shown in Experiment 1 as well as in previous studies (Massaro et al., 1986; Sekiyama & Burnham, 2004; Sekiyama et al., 2003). Despite the marginal non-significant difference (p=.07), 4-year-olds also tend to be better lipreaders than their 3-year-old counterparts. These improvements may be maturational, but a different possibility is discussed below in the light of findings from the regression analyses. Four-year-olds also had higher scores on the language specific speech perception test. Finally, the 4-year-olds were better than 3-year-olds on the tests of executive function and receptive vocabulary knowledge. However, despite the significant difference, the performance of both preschool groups on language-specific speech perception test was quite low and there was a large individual variability in performance. So as task difficulty may partly be responsible for this despite the fact that the during the experiments lengthy ‘play breaks’ were taken to ensure that children were not exposed to information overload.
The regression analyses revealed that the degree of visual speech influence in the AV condition was predictable from AO scores and also overall executive function scores. In other words, the ability to extract auditory and visual speech information appears to depend on a child’s ability for auditory-only speech perception and their ability to extract common information between two objects or events. These results suggest that the way auditory-visual speech information is processed perceptually might be similar to the processing of two or more dimensions of an event, especially perhaps around the preschool period. In other words, the ability to process multimodal speech information might be related to the cognitive ability to process multidimensional events, and these abilities may co-develop and interact as a function of maturation, both cognitive and linguistic (Werker et al., 1998).

Despite a significant − but small − correlation, language specific speech perception did not reliably predict auditory-visual speech perception in 3- and 4-year-old preschool children. This is similar to the result obtained with adults in Experiment 1. Moreover, with both groups, adults and preschool children, it was the AO scores that reliably predicted auditory-visual speech integration. So why does language specific speech perception predict auditory-visual speech integration in school children but not in preschool children and adults? And why does AO performance predict auditory-visual speech perception in preschool children and adults, but not in school children? As discussed earlier, the integration of auditory and visual speech information might facilitate speech perception and phoneme-to-grapheme conversions in circumstances in which children face new linguistic challenges at school (Experiment 1). However by adulthood, these have become automatic skills and are no longer linguistically challenging (Experiment 1), hence extra supporting information is no longer required. On the other hand for preschool children, it appears that the linguistic challenges around 3 and 4 years do not constitute such a linguistic challenge that they require increased attention to native speech contrasts, as is useful in reading acquisition during
early school years (Burnham, 2003; Burnham et al., 2002). Presumably the
greatest linguistic challenge during the preschool years is the development of
grammar, vocabulary, and possibly articulation. Perhaps the already well-
established phonological categories established in infancy (Kuhl et al., 2005;
Kuhl et al., 1992; Werker & Tees, 1984a) are sufficient base for vocabulary
acquisition, without the need for additional visual speech information in this
period.
The results here show that there is no causal link between auditory-visual
speech perception and language specific speech perception in preschool
children, and that auditory-visual speech perception is best predicted by basic
auditory perception and cognitive flexibility. However, this conclusion is
based on the results of studies with intact speech production development. It
remains to be seen how auditory-visual speech perception develops when
phonological development is not progressing normally. In the final
experiment, an attempt was made to address this issue with a sample of
young children with inconsistent phonological speech disorder.
CHAPTER 7

AUDITORY-VISUAL SPEECH PERCEPTION IN CHILDREN WITH PHONOLOGICAL SPEECH DISORDER
7.1 INTRODUCTION

Results of Experiment 1 showed that language specific speech perception is positively related to auditory-visual speech perception between the ages of five and eight. The results were interpreted to show that when children are faced with new language-specific challenges, such as reading acquisition (see Burnham et al., 1991; Burnham et al., 2002), or even possibly encountering new faces and speech styles such as teacherese, they seek extra non-auditory sources of speech information, in particular visual information. This relationship between auditory-visual speech perception and language specific speech perception was not observed in adults (Experiment 1), nor was it observed in preschool children aged three and four years (Experiment 2). In fact, only auditory-only speech perception scores reliably predicted the degree of visual speech influence in adults, and interestingly a similar finding was observed in the 3- and 4-year-old preschool children. Additionally, ability for executive functions predicted auditory-visual speech perception in the preschool group. The adult data were interpreted to show that adults automatically use various individual speech perception skills and do not exclusively rely on one skill over, or in support of, others. The preschool children data were interpreted to show that the lack of a link among auditory-visual speech perception, receptive vocabulary and language specific speech perception was due to the efficiency of the link between language specific speech perception and the vocabulary acquisition. In other words, auditory-only speech was good enough for vocabulary acquisition, ruling out the need for augmentation by visual speech information.

In addition to the above variables, one factor that has been thought to be important in the development of auditory-visual speech perception is speech production ability (Desjardins et al., 1997; Siva et al., 1995). However, this was not found to be a significant predictor of auditory-visual speech perception here with either school (Experiment 1) or preschool (Experiment 2) children.
However, unlike the two previous studies on this issue (Desjardins et al., 1997; Siva et al., 1995), the child participants in Experiments 1 and 2 did not have large amounts of speech production errors (Desjardins et al., 1997) or motor speech disability (Siva et al., 1995). The final experiment to be described here was designed to investigate the link between auditory-visual speech perception and language specific speech perception in both speech disordered and non-disordered children. The next section defines and describes the types of speech production disorders. This is followed by consideration of the general development of speech perception in children with speech disorders. After this, predictions regarding the development of auditory-visual speech perception of children with the disorder of interest here, inconsistent phonological speech disorder (Dodd, 1995).

7.2 SPEECH DISORDERS

A communication disorder can be defined as an impaired ability to make use of expressed (spoken or written) language to convey thoughts to others. Most communication disorders are speech disorders characterised by either the complete inability to produce speech sounds or the inability to produce speech sounds accurately. In a recent incidence survey in the United Kingdom it was found that 6.4% of children without any physical, sensory, cognitive or any other type of developmental deficit were found to have a speech disorder (Broomfield & Dodd, 2004). Enderby and Philip (1986) report that in the United Kingdom the ratio of preschool and school children with non-congenital speech disorders in the absence of any other form of impairment range from 2 to 25%.

The principal problem with respect to the diagnosis and treatment of speech disorders is the heterogeneity of the target population. The term ‘speech disorder’ covers a wide range of speech production impairments, for example children with a lisp but intelligible articulation; children whose speech is completely unintelligible as a result of omission or misarticulation of speech sounds; children who have an anatomical abnormality from birth, such as
cerebral palsy; and children raised in language-impoverished environments without sufficient exposure to appropriate language input. In addition to this variety of speech production impairments, there are countless individual differences in almost each type of condition and these differ with respect to large number of parameters related to the underlying cause of the disorder such as the child’s awareness of their disordered speech, the degree to which syntactic, semantic and phonological errors are made, and the degree of severity of the disorder (Dodd, 1995). Due to such individual variation, the classification of speech disorders has been a challenging task for both researchers and practitioners. Dodd (1995) describes five criteria for the classification of speech disorders upon which she bases a four-category classification system of speech disorders. The criteria Dodd (1995) describes are as follows:

(a) *Age of Acquisition* refers to the age at which the speech disorder emerged. For example, some at-risk children are prone to speech disorders *congenitally* as a result of some other impairments or disabilities, such as Down syndrome. Some children are simply delayed and do not show the typical development of speech production; these are *developmental* speech disorders. Some are cases of *acquired* speech disorders due to accidents or illnesses, such as head injuries, or otitis media.

(b) *Severity* refers to the number of phoneme production errors that children with speech disorders make.

(c) *Aetiology* refers to the medical model of classification of speech disorders. This criterion takes the factors that led to the disorder into account, such as whether a given case of speech disorder is a result of organic or non-organic factors, whether the case features anatomical anomalies (e.g., cleft palate), or whether the disorder is associated with an intellectual disability.

(d) *Linguistic Symptomatology* refers to the degree to which an association or a pattern exists between speech errors and other factors such as
phonological processing, phonological rule abstraction and motor skills.

(e) Psycholinguistic Deficits refers to the degree to which there is a mapping between the stages of the speech processing chain (see below) and input, cognitive processes, organisation of verbal units and output.

Dodd (1995) indicates that the development of psycholinguistic models of the speech processing chain has enabled researchers to advance relevant models and map interactions between the stages of speech processing. One such model, advanced by Winitz (1969; 1975), suggests that speech disorders can emerge as a result of impairment at any one of five stages: (i) the auditory input stage, which a hearing impairment or impoverished language environment may affect the processing of the acoustic signal; (ii) the phonological stage, in which the abilities of phonological rule abstraction, attention, or memory can be affected; (iii) the systematic phonetic stage in which a breakdown between phonological processing and the articulatory/production system may be impaired; (iv) the articulatory planning stage, in which the formulation of speech sound sequences can be affected; and (v) the motor execution stage, in which motor execution of speech utterances is impaired as a result of peripheral neurological impairment (e.g., dysarthria) (see Dodd, in preparation, for a review of this model).

Winitz’s model is a powerful one as it refers to the broad categories of speech disorders: articulation, phonology, dyspraxia, and dysarthria. However, Dodd (in preparation) also indicates that most children with speech disorders have some impairment at the phonological level, and that there is a need for a model that explains and categorises the differences between children with various phonological speech disorders. Dodd (1995) describes four major subclasses of phonologically-based speech disorders aimed to assist in the diagnosis process. These four subclasses of speech disorders are presented in the next section.
7.3 TYPES OF SPEECH DISORDERS

The fundamental problem in relation to the diagnosis and classification of phonologically based speech disorders is that there exists a wide range of individual variation of speech errors made by children with speech disorders. Dodd and Iacono (1989) state that there is a pattern of developmental and non-developmental phonological processes in children with (and without) speech disorders. These patterns refer to which speech sounds are evident and which are not in a speech-disordered child’s speech and what rules are put into place in speech production by that child. On these bases speech therapists can classify the child’s speech disorder as one of four phonological speech disorder subtypes (Dodd, 1995) depending on the phonological rules used by the child. These subtypes are articulation disorder, delayed phonological acquisition, consistent phonological deviant disorder and inconsistent phonological disorder (Bradford & Dodd, 1996) and are described in the following sections.

7.3.1 Articulation Disorder

Articulation disorder is a condition characterised by an inability to produce a perceptually acceptable version of particular speech sounds in isolation or in context. Articulation disorder comprises 12.5% of all childhood speech disabilities (Dodd, Holm, Crosbie, & Broomfield, 2006). Speech production by children with articulation disorder is characterised by a specific distortion such as a lisp, in which the air escapes bilaterally rather than centrally in the articulation of /s/, or by phoneme substitution (Dodd, 1995). Some of frequently observed substitutions are [f] for [θ], [ð] for [v], and [w] for [r]. In cases of neurologically-based motor/articulation speech disorders (e.g. dysarthria), and disorders due to anatomical anomalies (e.g. cleft palate) the production of a wider range of speech sounds is affected (Dodd, 1995).

7.3.2 Delayed Phonological Acquisition

Having the largest frequency amongst speech disorders, delayed phonological acquisition accounts for 57.5% of childhood speech disabilities
CH.7 CHILDREN WITH SPEECH DISORDER

(Dodd et al., 2006). The diagnosis and classification of delayed phonological disorder is relatively more straightforward than the other subtypes. Children with delayed phonological acquisition display a normal pattern of phonological process development, but this development is typical of their chronologically younger counterparts. A delay of 6 months is considered to be sufficient for a child to be diagnosed with delayed phonological acquisition (Dodd, 1995), although two other factors are also taken into account in the diagnosis and treatment process: (a) the extent to which the phonological system of the child continues to develop similar to a typical developmental pattern of younger children; (b) the extent to which children with delayed phonological acquisition use chronologically mismatched phonological rules which are typical of both younger and older counterparts (Dodd, 1995).

7.3.3 Consistent Phonological Speech Disorder

Consistent phonological disorder is seen in around 21% of all child speech impairments (Dodd et al., 2006). The speech of children with consistent phonological disorder is characterised by the use of non-developmental but relatively consistent use of phonological rules. While some cases display speech errors marked by a single type of disordered phonological rule (e.g., consistent deletion of all syllable-initial fricatives consistently), most make errors characterised by a series of disordered phonological rules. Some of these rule violations can be straightforward (e.g., phoneme deletions in certain CV contexts) or others are more complex (e.g., no co-occurrence of nasals and plosives in the same syllabic structure). Usually a child should be considered as having consistent phonological disorder if at least one of the error patterns (s)he uses is atypical of a normally developing child. Children with consistent phonological disorder also have difficulty in mastering the phonotactic constraints of their native phonological system (Dodd, Hambley, & Leahy, 1989), as well as difficulty with on rhyme and phonological awareness tasks, possibly reflecting a cognitive-linguistic challenge (see Fox & Dodd, 2001).
7.3.4 Inconsistent Phonological Speech Disorder

Normally developing children display a degree of variability in their production of particular phonemes. For example, the initial /fl/ in flower can be produced as both [fauːə] and [lauːə] prior to the child eventually beginning to produce this cluster correctly. Such variability might indicate that a child’s phonological processing is in transition from an error form to correct usage (Dodd, 1995). However, in some cases inconsistent phonological errors may signify misperception and misrepresentation of native phonological categories (Dodd, 1995; Fox & Dodd, 2001). In some other cases, a whole class of phonemes is used interchangeably, each substituting for the other, similar to a kind of allophonic usage. For example, a child may use all fricatives indiscriminately, e.g. replacing [θ] with [f] or [s] or [ʃ]. Dodd (1995) suggests that if the inconsistency error rate exceeds 40%, a child should be classified as having inconsistent phonological disorder. Inconsistent phonological disorder represents around 9% of childhood speech impairment (Dodd et al., 2006), the lowest incidence rate amongst the four speech disorders described here.

7.3.5 Speech Disorder Subtypes: Summary and the Current Study

What generally differentiates these four subtypes of speech disorders are different patterns of surface speech errors (Dodd, 1995). Despite the general finding that children with articulation disorder and delayed phonological acquisition make fewer production errors than their counterparts with consistent and inconsistent phonological disorder, there is wide range of individual variety within each subclass of speech disorder.

In relation to the nature of the current thesis children with consistent phonological speech disorder are of most interest. These children display phonemic error patterns atypical of their own age group but typical of both their younger and older counterparts, reflecting a misrepresentation and/or misperception of the phonological system of their native language.

Speech production problems in children with phonological speech disorders do not stem from a motor deficit or anatomical deformity, but are rather due
to the misrepresentation of speech sound categories as a result of a failure in phonological processing (Dodd, 1995). The speech is characterised by indiscriminate use of the members of a phonological category (Fox & Dodd, 2001), e.g., substituting a given fricative target with /f/, /θ/, or /v/ inconsistently. Here, testing children with speech disorder provides an opportunity to study the link between speech perception and production from an auditory-visual speech perspective. For this purpose children with (SD) and without (no-SD) speech disorder were given auditory-visual speech perception, language specific speech perception tests and also tests of executive cognitive functions. Executive functions are sets of high cognitive functions enabling the planning, initiation, and execution of goal-directed behaviour (Oates & Greyson, 2004).

The next section sets out the predictions for the current experiment for speech disorder and no-speech disorder children based on the literature on speech production development in children with phonological speech disorder, and the findings of the first two experiments.

7.4 THE RATIONALE, AIM AND THE PREDICTIONS OF THE CURRENT STUDY

An important question in auditory-visual speech perception research is to understand the linguistic level at which the auditory and visual speech components are integrated: amodal / phonetic (Burnham & Dodd, 2004; Rosenblum & Saldaña, 1996; Rosenblum et al., 1997) or phonological (D. W. Massaro, Cohen, & Smeele, 1995). To answer this question developmental data are needed (Bernstein et al., 2002). As a part of this quest, Sekiyama and Burnham (2004) found that while visual speech influence in a McGurk task increases for English speakers between 6 and 8 years this was not the case for Japanese speakers. One explanation for these differential results is related to the relative paucity of certain classes of visually identifiable speech elements (e.g. labiodentals) in the Japanese phonological repertoire and less extensive mouth movements. Thus it may be reasoned that there are less extensive
articulations in Japanese, perceivers are not able to supplement auditory speech information with visual information effectively when faced with developmental linguistic challenges. On the other hand English language children have experience with producing a wide range of articulations and it may be the case that it is this experience that enables them to perceive and use the visible articulations of others, especially in linguistically-challenging situations. Indeed Burnham (2003) found that for 4-, 6-, and 8-year-old English-language children articulation ability for age significantly predicted language specific speech perception ability. In turn it was found here that language specific speech perception is related to auditory-visual speech perception in 5-, 6-, 7, and 8-year-old children (Experiment 1), but not in younger 3- and 4-year-old children (Experiment 2).

Given these results and the possible reasons for them the following predictions were advanced here. If articulation ability is related to language specific speech perception, or auditory-visual speech perception, then (a) language specific speech perception, (b) auditory-visual speech perception and (c) visual-only speech perception should be better in no-speech disorder children than in speech disorder children.

If language specific speech perception predicts auditory-visual speech perception in linguistically challenging situations (Experiment 1), and given the fact that language specific speech perception does not predict auditory-visual speech perception in 3- and 4-year-olds with normally-developing articulation (Experiment 2), then the linguistically-challenged speech disorder children may show the relationship between language specific speech perception and auditory-visual speech perception while no-speech disorder children do not.

If it is the speech disorder per se which affects auditory-visual speech perception, then there should be no improvement over age in auditory-visual speech perception ability for speech disordered children.
7.5 A Note on Language Specific Speech Perception Data and Its Analysis

Unfortunately there was a high attrition rate for speech disorder children in the language specific speech perception test (68.4%). This stemmed from the fact that these children were also tested on other measures unrelated to this study and were required to attend their speech therapy session on the same testing day, and it was not possible to invite these children back at the hospital to re-test them on the language specific speech perception test. For these reasons language specific speech perception data were not included in the following analyses and a reduced set of hypotheses were advanced as follows.

If articulation ability is related to auditory-visual speech perception and visual-only speech perception, then children with no speech disorder should perform better than children with speech disorder. In addition, no age differences among speech disorder are predicted for auditory-visual and visual-only speech perception. No difference is expected between children with and without speech disorder on cognitive skills.

7.6 Experiment 3: Auditory-Visual Speech Perception in Children with Inconsistent Phonological Speech Disorder

7.7 Method

7.7.1 Participants

Thirty-nine children with consistent phonological speech disorder ($M_{age} = 4.34$, $sd = 0.69$; 9 females & 30 males) and 18 children with normal speech development ($M_{age} = 4.12$, $sd = 0.46$; 8 females & 10 males) were recruited. The speech disorder children were assessed by experienced speech pathologists (Barbara Dodd, Beth McIntosh and Sharon Crosbie) at the Royal Brisbane Hospital. All children had normal or corrected-to-normal vision. The speech-
disordered children were recruited from a group of children coming to Royal Brisbane Hospital to attend their regular speech therapy sessions. The children with normal speech development were recruited from a number of childcare centres located in the Brisbane metropolitan area. Ethical clearances from the Universities of Queensland and Western Sydney and the Queensland Health Department were obtained as well as permissions from parents of children.

The speech disorder group was studied in two ways. First, in order to compare the disorder and no-disorder children, 18 of the speech-disordered (SD) children ($M_{age}= 4.06$, $sd=0.49$; 8 females & 10 males) were matched to the children with no speech disorder (no-SD) ($M_{age}= 4.12$, $sd=0.46$; 8 females & 10 males) on the basis of gender and age. Second, in order to test age effects within the speech disorder children, the full sample of 39 SD children was divided into three age groups: 3-year-olds: n=13, $M_{age}= 3.56$ & $sd=0.33$; and 4-year-olds: n=13, $M_{age}= 4.34$ & $sd=0.18$; 5-year-olds: n=13, $M_{age}= 5.13$ & $sd=0.28$.

(An analysis of variance (ANOVA) revealed that these three SD age groups significantly differed from each other in age, $F(1,36)=218.14$, $p<.0001$.)

### 7.7.2 Stimuli, Apparatus and Procedure

The stimuli, apparatus and procedural details employed in Experiment 3 were identical to those in Experiment 2 with the exception that the children in this experiment were not tested on receptive vocabulary knowledge due to time constraints - the SD children came in to the hospital for their speech therapy session on the day of testing, which limited the amount of time to be allocated for the experimental session. The tests were conducted by an experienced speech pathologist.

### 7.8 Results: Children With versus Without Speech Disorder

#### 7.8.1 Auditory-Visual Speech Perception Test

The visual speech discrimination index (VSI-AX), auditory-only (AO), and visual-only (VO) scores for SD vs. no-SD children were subjected to $t$-test
analyses. Mean values for each group over three measures of auditory-visual speech perception test are shown in Figure 7-1 and statistical output file can be found in Appendix CD5.3.1. Results showed that the two groups did not differ on the VSI-AX \( t(1,34) = .550, p>.05 \), or AO \( t(1,34) = .632, p>.05 \) scores, but that the no-SD group performed better than the SD group in the VO condition \( t(1,34) = 2.319, p<.05 \).

![Figure 7-1](image)

**Figure 7-1** The Auditory-Visual (top), Auditory-Only and Visual-Only (bottom) Scores for Children with and without Speech Disorder. Error bars show the standard error of the mean.

### 7.8.2 Executive Functions Test: Flexible Item Selection Task

The flexible item selection test (FIST) scores were analysed in a 2 x (2), Speech Disorder Status x (Rule Abstraction/Cognitive Flexibility) ANOVA. Figure 7-2 shows the mean data for each group schematically and the statistical output file can be found in Appendix CD5.4. Results showed a significant
difference between general rule abstraction and cognitive flexibility subtest scores \( F \left( 1, 34 \right) = 22.545, p<.001 \). However, there was no significant difference between the two groups of children \( F \left( 1, 34 \right) = 0.255, p>.05 \) nor was there a significant interaction between factors.

![Chart showing FIST score comparison between SD and no-SD children](image)

**Figure 7-2** The rule abstraction and cognitive flexibility subtest scores of the flexible item selection test for children with and without speech disorder. Error bars show the standard error of the mean.

### 7.8.3 Regression Analysis: VSI-AX as dependent variable

A sequential multiple regression analysis was run with VSI-AX scores as the dependent variable and five predictors as separate independent variables entered in the order of developmental (age and speech disorder status), linguistic (AO and VO) and cognitive (FIST factor scores) factors. Results of evaluations of assumptions were satisfactory after an outlier with standardized residual greater than 3 standard deviations was removed from the analysis. All individual Mahalanobis distance probability values were under .001, the critical value set for six independent variables after an outlier case was removed (Tabachnik & Fidell, 2001). The correlation coefficients are shown in Table 7-1. Table 7-2 shows the unstandardised regression coefficients, \( B \), standardized regression coefficients, \( \beta \), and \( R^2 \) change for the predictors at their step of entry and final \( B \) and \( \beta \). It can be seen that none of
the independent variables entered into the regression analysis predicted VSI-AX scores reliably (see Table 7-2).

**TABLE 7-1 Correlation coefficients amongst the variables age, speech disorder status (SDS), AO, VO, VSI-AX, N-NN DI and FIST scores.**

<table>
<thead>
<tr>
<th></th>
<th>VSI-AX</th>
<th>Age</th>
<th>SDS</th>
<th>AO</th>
<th>VO</th>
<th>FIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>VSI-AX</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>.21</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SDS</td>
<td>.02</td>
<td>.03</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AO</td>
<td>.32*</td>
<td>.18</td>
<td>-.11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VO</td>
<td>.18</td>
<td>.07</td>
<td>-.37*</td>
<td>.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FIST</td>
<td>-.13</td>
<td>.16</td>
<td>-.15</td>
<td>.00</td>
<td>-.14</td>
<td></td>
</tr>
</tbody>
</table>

* Sig. at α=.05, **Sig. at α=.01

**TABLE 7-2 Multiple regression of age, speech disorder status (SDS), AO, VO, N-NN DI and FIST scores as predictors of VSI-AX scores.**

<table>
<thead>
<tr>
<th>Step</th>
<th>Variables</th>
<th>B at Step</th>
<th>β at Step</th>
<th>R² change at Step</th>
<th>F-value</th>
<th>Final B (step 5)</th>
<th>Final β (step 5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Age</td>
<td>.072</td>
<td>.211</td>
<td>.044</td>
<td>1.531</td>
<td>.055</td>
<td>.160</td>
</tr>
<tr>
<td>2</td>
<td>SDS</td>
<td>.004</td>
<td>.011</td>
<td>.044</td>
<td>.004</td>
<td>.029</td>
<td>.095</td>
</tr>
<tr>
<td>3</td>
<td>AO</td>
<td>.332</td>
<td>.293</td>
<td>.126</td>
<td>2.893</td>
<td>.332</td>
<td>.293</td>
</tr>
<tr>
<td>4</td>
<td>VO</td>
<td>.309</td>
<td>.208</td>
<td>.163</td>
<td>1.337</td>
<td>.269</td>
<td>.181</td>
</tr>
<tr>
<td>5</td>
<td>FIST</td>
<td>-.018</td>
<td>-.115</td>
<td>.175</td>
<td>.421</td>
<td>-.018</td>
<td>-.115</td>
</tr>
</tbody>
</table>

* Sig. at α=.05, **Sig. at α=.01

### 7.8.4 Regression Analysis: VO (lipreading) as dependent variable

A second regression analysis was performed in which five predictors as independent variables entered in the order of developmental (age and speech disorder status), linguistic (AO and VSI-AX scores) and cognitive (FIST factor scores) factors. Table 7-3 shows the unstandardised regression coefficients, (B), standardized regression coefficients (β), and R² change for the predictors at their step of entry and final B and β. It can be seen that only speech
disorder status reliably predicted VSI scores and increase $R^2$ (Table 7-3); that is the absence speech disorder (no-SD group) predicted VSI-AX scores.

**Table 7-3 Multiple regression of age, speech disorder status (SDS), AO, VSI-AX, N-NN DI and FIST scores as predictors of VO (lipreading) scores.**

<table>
<thead>
<tr>
<th>Step</th>
<th>Variables</th>
<th>B at Step</th>
<th>β at Step</th>
<th>R² change at Step</th>
<th>F-value</th>
<th>Final B (step 5)</th>
<th>Final β (step 5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Age</td>
<td>.014</td>
<td>.061</td>
<td>.004</td>
<td>.125</td>
<td>.026</td>
<td>.111</td>
</tr>
<tr>
<td>2</td>
<td>SDS</td>
<td>-.076</td>
<td>-.373</td>
<td>.143</td>
<td>5.351*</td>
<td>-.081*</td>
<td>-.394</td>
</tr>
<tr>
<td>3</td>
<td>AO</td>
<td>-.025</td>
<td>-.032</td>
<td>.144</td>
<td>.037</td>
<td>-.045</td>
<td>-.059</td>
</tr>
<tr>
<td>4</td>
<td>VSI-AX</td>
<td>.073</td>
<td>.141</td>
<td>.162</td>
<td>.692</td>
<td>.049</td>
<td>.095</td>
</tr>
<tr>
<td>5</td>
<td>FIST</td>
<td>-.019</td>
<td>-.186</td>
<td>.193</td>
<td>1.148</td>
<td>-.019</td>
<td>-.186</td>
</tr>
</tbody>
</table>

* Sig. at $\alpha = .05$, **Sig. at $\alpha = .01$

**7.9 Results: Children with Speech Disorder**

As indicated in 7.7.1, the 39 SD children were split into three age groups: 3-year-olds, 4-year-olds, and 5-year-olds in order to test the effects of age-related differences within the SD group.

**7.9.1 Auditory-Visual Speech Perception Test**

The AO, VO and VSI-AX scores of three groups of SD children (see 2.1) were subjected to three sets of 3-level single factor (age) ANOVA (Age x AO, Age x VO, and Age x VSI-AX). The aggregate results are shown in Figure 7-3 and statistical output file can be found in Appendix CD5.3.2. Results showed that there were no age-based increase in VSI-AX, AO or VO scores [$F(1, 36) = 2.68, MS_{e}=.032, p>.05.$]. in addition the VSI-AX scores were higher than for AO and VO scores [$F(1, 36) = 9.443, MS_{e}=.032, p>.01,$] and AO scores were higher than VO scores [$F(1, 36) = 38.721, MS_{e}=.022, p>.001.$]
**CH.7 CHILDREN WITH SPEECH DISORDER**

![Bar charts showing VSI-AX and AO & VO scores for children with speech disorder at 3, 4, and 5 years of age.](image)

**Figure 7-3** The Auditory-Visual (top), Auditory-Only and Visual-Only (bottom) Scores for Children with Speech Disorder in three age groups. Error bars show the standard error of the mean.

### 7.9.2 Executive Functions Test: Flexible Item Selection Task

The FIST scores (rule abstraction and cognitive flexibility subtest scores) were subjected to a 3 x (2), age x (rule abstraction / cognitive flexibility) ANOVA. The mean FIST scores by age groups are shown in Figure 7-4 and the statistical output file can be found in Appendix CD5.5. Results showed that FIST scores increased linearly with age \([F (1, 36) = 52.47, MS_e=4.041, p<.001]\). In addition, rule abstraction scores were generally significantly higher than cognitive flexibility scores \([F (1, 36) = 78.305, MS_e=3.442, p>.05]\), with the difference between rule abstraction and cognitive flexibility scores being greater at 3 and 4 years than at 5 years \([F (1, 36) = 14.531, MS_e=3.442, p<.005]\).
Figure 7.4 The rule abstraction and cognitive flexibility subtest scores of the flexible item selection test for children with speech disorder. Error bars show the standard error of the mean.

7.9.3 Regression Analyses: VSI-AX as Dependent Variable

A regression analysis was performed in which four predictors as independent variables entered in the order of developmental (age), linguistic (AO and VO scores) and cognitive (FIST factor scores) factors. Results of evaluations of assumptions were satisfactory and except for one outlier case, all individual Mahalanobis distance probability values under .001, the critical value set for four independent variables. Table 7-4 shows the correlation coefficients among the dependent and independent variables and Table 7-5 shows the unstandardised regression coefficients, (B), standardized regression coefficients (β), and $R^2$ change for the predictors at their step of entry and final B and β. The correlation and regression analysis output files can be found in Appendix CD5.9. Results of the regression analysis showed that none of the independent variables predicted VSI-AX scores (see Table 7-5).
TABLE 7-4 CORRELATION COEFFICIENTS AMONG THE VARIABLES AGE, AGE GROUPS, AO, VO, VSI-AX, N-NN DI AND FIST SCORES (SPEECH DISORDERED CHILDREN ONLY).

<table>
<thead>
<tr>
<th></th>
<th>VSI-AX</th>
<th>Age</th>
<th>AO</th>
<th>VO</th>
<th>FIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>VSI-AX</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>.25</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AO</td>
<td>.23</td>
<td>.11</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VO</td>
<td>.11</td>
<td>.07</td>
<td>-.06</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>FIST</td>
<td>-.23</td>
<td>.02</td>
<td>.18</td>
<td>-.082</td>
<td>-</td>
</tr>
</tbody>
</table>

* Sig. at α=.05,  **Sig. at α=.01

TABLE 7-5 MULTIPLE REGRESSION OF AGE, AGE GROUPS, AO, VSI-AX, N-NN DI AND FIST SCORES AS PREDICTORS OF VSI-AX SCORES (SPEECH DISORDERED CHILDREN ONLY).

<table>
<thead>
<tr>
<th>Step</th>
<th>Variables</th>
<th>B at Step</th>
<th>β at Step</th>
<th>R² at Step</th>
<th>F-value</th>
<th>Final B</th>
<th>Final β</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Age</td>
<td>.063</td>
<td>.253</td>
<td>.064</td>
<td>2.462</td>
<td>.053</td>
<td>.213</td>
</tr>
<tr>
<td>2</td>
<td>AO</td>
<td>.207</td>
<td>.203</td>
<td>.105</td>
<td>1.587</td>
<td>.264</td>
<td>.258</td>
</tr>
<tr>
<td>3</td>
<td>VO</td>
<td>.178</td>
<td>.103</td>
<td>.115</td>
<td>.401</td>
<td>.146</td>
<td>.084</td>
</tr>
<tr>
<td>4</td>
<td>FIST</td>
<td>-.059</td>
<td>-.268</td>
<td>.184</td>
<td>2.801</td>
<td>-.059</td>
<td>-.268</td>
</tr>
</tbody>
</table>

* Sig. at α=.05,  **Sig. at α=.01

7.9.4 Regression Analyses: VO (lipreading) as Dependent Variable

A sequential multiple regression analysis was performed with VO (lipreading) scores as the dependent variable and four predictors as independent variables in the order of developmental (age), linguistic (AO and VO scores) and cognitive (FIST factor scores) factors. Results of evaluations of assumptions were satisfactory and all individual Mahalanobis distance probability values under .001, the critical value set for four independent variables. The results of this regression analysis showed that none of the variables significantly predicted VO scores (see Table 7-6).
TABLE 7-6 MULTIPLE REGRESSION OF AGE, AO, VSI-AX, N-NN DI AND FIST SCORES AS PREDICTORS OF VO (LIPREADING) SCORES.

<table>
<thead>
<tr>
<th>Step</th>
<th>Variables</th>
<th>B at Step</th>
<th>β at Step</th>
<th>R² at Step</th>
<th>F-value</th>
<th>Final B</th>
<th>Final β</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Age</td>
<td>.010</td>
<td>.072</td>
<td>.005</td>
<td>.190</td>
<td>.009</td>
<td>.062</td>
</tr>
<tr>
<td>2</td>
<td>AO</td>
<td>-.041</td>
<td>-.069</td>
<td>.010</td>
<td>.174</td>
<td>-.043</td>
<td>-.073</td>
</tr>
<tr>
<td>3</td>
<td>VSI-AX</td>
<td>.044</td>
<td>.091</td>
<td>.018</td>
<td>.276</td>
<td>.037</td>
<td>.076</td>
</tr>
<tr>
<td>4</td>
<td>FIST</td>
<td>-.006</td>
<td>-.045</td>
<td>.019</td>
<td>.060</td>
<td>-.006</td>
<td>-.045</td>
</tr>
</tbody>
</table>

*Sig. at α=.05, **Sig. at α=.01

7.10 DISCUSSION

Results in the auditory-visual speech perception test show that children with phonological speech disorder (SD) and children with no speech disorder (no-SD) are comparable in auditory-only speech perception (AO condition) and visual speech influence (VSI-AX scores) but that no-SD children had better VO scores than SD children. Similarly, regression analyses showed that speech disorder status did not reliably predict VSI-AX, but did reliably predict lipreading ability (VO scores) irrespective of age or executive function level. Together these results provide support for two conclusions. First, it appears that veridical experience with speech production (in the no-SD group) is associated with better lipreading whereas inaccurate speech production experience is not. This supports the results of Desjardins et al. (1997) that lipreading is better in non-substituters than substituter articulating children. Second, the experience of incorrect articulatory experience (SD group) has no measurable effect on the integration of auditory and visual speech components.

Secondly the results suggest that auditory-visual speech perception is a separate construct to auditory-only and visual-only speech perception. This indirectly supports previous studies showing that auditory and visual speech components are processed at an earlier, e.g., phonetic (Rosenblum & Saldaña, 1996) rather than a later, e.g., phonological (Massaro et al., 1995) stage. This is because the SD children in this study are speech disordered precisely because
they have a problem with executing phonological rules in their speech production.

One model that might explain the difference in lipreading performance is the motor theory of speech perception (Liberman & Mattingly, 1985). The fact that SD children both do not articulate as well and do not focus so much on visually discernable speech gestures consistent with the motor theory notion that there is a link between speech perception and the perception of intended articulatory gestures. This point warrants further research in relation to the link between auditory-visual speech perception and phonological processing (language specific speech perception) found in no-SD school children (Experiment 1) but which could not be tested here (see 7.5). Future research should address this issue by recruiting equal number of children with major specific phonological speech disorders of varying ages. This and other relevant future research directions are further discussed in 8.2.3.

Examination of the data of 39 SD children shows that, unlike the findings with no-SD children of the same age (Experiment 2), there were no age-based differences in auditory-visual speech perception. The only significant age difference in the SD group was an improvement in executive functions, similar to the age-related improvement in executive functions in Experiment 2 with no-SD children aged three and four. Thus SD children here and no-SD children in Experiment 2 are similar with respect to the effect of cognitive ability, but not with auditory-visual, auditory-only and visual-only speech perception. The speech disorder is then specifically speech related and not cognitively related, and one consequence of the speech disorder is that auditory-visual speech perception does not develop at the normal rate.

In relation to the hypotheses put forward earlier it can be concluded that no-SD children are better than SD children on VO, but not on auditory-visual speech perception. This result suggests that the ability to produce speech sounds correctly affects the degree of lipreading supporting earlier evidence with young children (Desjardins et al., 1997) and cerebral palsy adults (Siva et al., 1995).
CHAPTER 8

GENERAL DISCUSSION AND CONCLUSION
The finding that humans benefit from visual speech information not only in noisy conditions (Sumby & Pollack, 1954) but also in clear and uninterrupted listening conditions (McGurk & MacDonald, 1976) led to an exponential increase in research on how humans process visual speech information, and the role of visual information in speech perception in general. Despite the fact that the original McGurk effect study (McGurk & MacDonald, 1976) included a developmental investigation, most research focused on how adults rather than children process auditory-visual speech. However, recent reviews (Bernstein et al., 2002; Burnham, 1998) claim that in order to understand the dynamics of auditory-visual speech integration, further developmental studies are required over and above the then existing handful of research with infants and children (e.g. Burnham & Dodd, 2004; Desjardins & Werker, 1996; Massaro, 1984; Rosenblum et al., 1997; Sekiyama et al., 2003).

The studies reported here were motivated by the recent developmental findings by Sekiyama and Burnham (2004; Sekiyama & Burnham, under review), which showed that the onset of greater visual speech influence in English than Japanese listeners (Sekiyama & Tohkura, 1993) emerges between 6 and 8 years. Sekiyama and Burnham’s (2004; 2007) findings, and Sekiyama’s earlier results begged the question, what language and developmental factors are related to the increase in visual speech influence found in English children between 6 and 8 years of age?

This chapter presents the findings of this thesis in light of this question and the existing literature on the development of auditory-visual and auditory speech perception. The results are also discussed with respect to practical and theoretical issues: the auditory-visual speech integration debate, implications for speech therapy, and the advancement of new technologies to improve the quality of life for special populations. Before addressing these issues, an overall summary of the current results is presented.
8.1 Summary and Interpretation of the Findings: Language Specific and Cognitive Factors

Here, the findings of the research in this thesis are summarized by experiment and the overall picture of the development of auditory-visual speech perception in English-speaking children aged between three and eight is outlined. The findings are summarised by experiment in Table 8-1.

8.1.1 Experiment1: School Children and Adults

Experiment 1 revealed that, as predicted, auditory-only, visual-only, auditory-visual speech perception, non-native speech perception in the language specific speech perception task, reading, and articulation improved over age (5-, 6-, 7-, and 8-year-olds). Most important of these is auditory-visual speech perception: as hypothesised, it was found that in clear listening conditions visual speech influence increases over age between five and eight years. This supports and extends previous findings with English language perceivers (Massaro et al., 1986; McGurk & MacDonald, 1976; Sekiyama & Burnham, 2004). It is also of note that lipreading ability (measured by visual-only condition performance) also increased with age, again in accord with earlier findings (Desjardins et al., 1997; Hockley & Polka, 1994; Sekiyama & Burnham, 2004, 2007).

The most significant finding of the first experiment was that language specific speech perception was a significant predictor of visual speech influence. In other words, the children with relatively better perception of native speech than non-native speech sounds were those who were more influenced by visual speech information.

*Why should relatively greater attention to native than non-native speech contrasts predict visual influence in speech perception?* As discussed in chapter 5, one possible reason for this relationship is that it emerges when children encounter language-specific challenges that necessitate increased attention to the phonological characteristics of the language at hand.
<table>
<thead>
<tr>
<th>EXPERIMENT NUMBER</th>
<th>AGE GROUPS</th>
<th>FACTORS TESTED</th>
<th>RESULTS</th>
</tr>
</thead>
</table>
| EXPERIMENT 1      | 5-, 6-, 7- & 8-year-olds & Adults | • Auditory-Visual Speech Perception (AVSP)  
  • Auditory-Only (AO) Speech Perception  
  • Lipreading (VO)  
  • Language Specific Speech Perception (LSSP)  
  • Reading  
  • Articulation | Age-based increases in:  
  • AVSP (cubic trend)  
  • Lipreading  
  • AO Speech Perception  
  • LSSP  
  • Reading |
| EXPERIMENT 2      | 3- & 4-year-olds | • Auditory-Visual Speech Perception (AVSP)  
  • Auditory-Only (AO) Speech Perception  
  • Lipreading (VO)  
  • Language Specific Speech Perception (LSSP)  
  • Vocabulary Knowledge (VK)  
  • Executive Functions (EF)  
  (Rule Abstraction & Cognitive Flexibility) | • AVSP: 4-yos > 3-yos  
  • AO: 4-yos > 3-yos  
  • VO: 4-yos = 3-yos  
  • VK: 4-yos > 3-yos  
  • EF: 4-yos > 3-yos |
| EXPERIMENT 3      | 3-, 4- & 5-year-olds with (SD) and without (no-SD) inconsistent phonological speech disorder | • Auditory-Visual Speech Perception (AVSP)  
  • Auditory-Only (AO) Speech Perception  
  • Lipreading (VO)  
  • Executive Functions (EF)  
  (Rule Abstraction & Cognitive Flexibility) | **SD vs. no-SD Children:**  
  • AVSP: no-SD = SD  
  • AO: no-SD = SD  
  • VO: no-SD > SD  
  • EF: no-SD = SD  
  **SD Children:**  
  • AVSP: No age effects  
  • AO: No age effects  
  • VO: No age effects  
  • EF: 5-yos > 4-yos > 3-yos  
  **SD Children:**  
  • Speech disorder status predicts VO.

**Children:**  
• LSSP and VO predict AVSP  
• Nothing predicts VO

**Adults:**  
• AO predicts AVSP  
• Nothing predicts VO

**SD Children:**  
• No significant results
Reading appears to be one such challenge for young children (Burnham, 2003; Öney & Goldman, 1984). Burnham (2003) found that reading instruction is related to language specific speech perception in English-speaking children, such that those who have relative superiority for native over non-native speech perception are those children who have good reading skills. Reading is a task with a high cognitive demand, especially in English due to the complexity of its orthography, which makes English reading acquisition particularly challenging (Öney & Goldman, 1984). Given this, English-speaking children may seek all available extra information, including visual speech input, to assist in the development of phoneme-grapheme linkages.

Once stable phoneme-grapheme linkages are formed, the phonological focus of language specific speech perception and the visual speech focus of auditory-visual speech perception continue to exist, but independently – ability on one no longer predicts ability on the other. Such a conclusion is supported by the regression analyses with the adult perceivers in Experiment 1 - language specific speech perception did not reliably predict auditory-visual speech perception ability for adults. Rather it was basic auditory-only perception, which predicted auditory-visual speech perception.

The results of Experiment 1 also show strong correlations between reading, articulation and lipreading abilities. The articulation-lipreading correlation provides partial support for an earlier finding by Desjardins and her colleagues that young children’s ability to articulate speech sounds is positively related to the degree of visual speech influence in perception (Desjardins et al., 1997). These findings are also consistent with the motor theory of speech perception, which asserts that speech perception involves the perception of a pattern of articulatory movements (Liberman et al., 1967; Liberman & Mattingly, 1985), and is presumably enhanced by the perceiver’s own articulation (Siva et al., 1995). It is also of interest regarding the reading-lipreading correlation that around school-age children encounter a particular style of hyperarticulated speech called teacherese (Håkansson, 1987). However it was not possible to test the role of teacherese in the face of the methodological restrictions of the studies reported
here, and it would be beneficial to design future experiments that would test the possible function of the role of teacherese in the language development of school children. Specifically a comparison of the functions of infant-directed speech (motherese) (see Burnham, Kitamura & Vollmer-Conna, 2002) and teacherese would be beneficial in terms of understanding the age-specific roles of hyperarticulated speech styles directed to children and infants. While there are no studies of the phonetic and articulatory characteristics of teacherese, it is known that hyperarticulated speech, both its auditory and visual characteristics, improves speech detection considerably (Lees & Burnham, 2005). The use of this style of speech by teachers, especially when teaching reading, and especially given its distinct visual characteristics in what at times are noisy classrooms, might well give rise to the strong correlation between lipreading and reading. Finally, with regard to the articulation-reading correlation found here, Burnham found that articulation ability predicted language specific speech perception in young pre-school children, and that reading ability predicted language specific speech perception in school children. It is quite conceivable that good articulators are good readers because good articulation would result in superior knowledge of phonemes, and thus allow phoneme-grapheme links to be formed more easily. It can be seen that these three abilities, reading, articulation, and lipreading, are closely related and probably share a good deal of variance. Nevertheless, the regression analyses in Experiment 1 show that only lipreading, and neither reading nor articulation significantly predicts auditory-visual speech perception ability. On the basis of Experiment 1 here, and Burnham (2003), the relationship between the variables can be schematically represented as in Figure 8-1.
The model in Figure 8-1 is exhaustive in terms of the variables considered here, but it should be noted that other possible causes maybe also be at work. For example, when young children move from the home to the school environment they are socially deluged with new people with different facial physiognomy, speech styles, accents and sometimes even speech disorders. We know from previous research that when the auditory signal is degraded, reliance on the visual component of speech increases significantly (Sekiyama & Burnham, 2004; Sumby & Pollack, 1954). Thus it is viable to suggest that in order to perceptually tackle these new speech styles, new faces, and sometimes disordered speech, young children may use and benefit from visual speech information more when they are at school, at least in initial stages of schooling, than when they interact with their family members. Unfortunately, despite the logical legitimacy of this

---

Note that the participants in this project were from Australian English monolingual environments with no significant exposure to people from non-English speaking background before starting school. So they were not in general familiar with non-native accents or different speech styles prior to the onset of schooling.
proposal, no studies of this kind have yet been conducted and there would be several methodological problems to overcome in order to conduct such research. In summary, the results of Experiment 1 show that in the linguistically challenging period of school onset and learning to read, there is an increase in lipreading and auditory-visual speech perception over age; and, most importantly, that auditory-visual speech perception is facilitated by language specific speech perception, and so determined by the phonologically relevant aspects of the ambient speech. The results are consistent with the notion that auditory-visual speech is stored in a phonological code and that the stronger the phonological bias in the child (the more language specific speech is) the more auditory-visual information is used in speech perception. Given these results, the next question of interest was whether other language specific challenges before the onset of schooling lead to an increase in the use of visual speech information. This was the focus of Experiment 2, the results of which are discussed in the next section.

8.1.2 Experiment 2: Preschool Children

Experiment 1 showed that language specific speech perception is positively related to and predicts auditory-visual speech perception. This appears to indicate that visual (and auditory) speech information is stored in a phonological code and that the phonological focus provided by language specific speech perception assists auditory-visual speech perception via this phonological code. It was also suggested that the use of extra information, such as visual speech information, to establish and maintain phonological processing, is especially required in the face of linguistic challenges such as the onset of reading instruction. Experiment 2 was conducted in order to ascertain if the link between language specific speech perception and auditory-visual speech perception is maintained in the absence of reading instruction, before children go to school. Children aged three and four years were given tests of auditory-visual speech perception, language specific speech perception, vocabulary knowledge and executive functions. With the exception of the native component of language
specific speech perception and lipreading, all these abilities improved with age, as was the case in Experiment 1. However, contrary to the regression analysis results in Experiment 1 with children (but consistent with those for adults), it was found that the only variable that predicted auditory-visual speech perception was auditory-only speech perception. In addition, executive cognitive function was found to predict lipreading ability. Finally VO did not improve over age between three and four years in this sample. Nevertheless, it was found that over the whole sample of 3- and 4-year-olds, executive function ability almost predicted VO performance ($p=.058$). That is, despite the lack of age differences, children with greater cognitive flexibility had a tendency to be better lipreaders (although this did not flow on to auditory-visual speech perception ability).

For these preschool children it is of interest to consider the auditory-only (AO) and visual-only (VO) aspects of the auditory-visual speech perception task separately and then consider auditory-visual speech perception. With respect to the AO task, AO performance improved as a general function of age, and AO was the only variable that predicted auditory-visual speech perception.

With regard to VO ability in pre-schoolers it appears that auditory-visual speech perception is wholly accounted for by AO ability - while VO ability does vary between children as a function of executive function, it does not contribute to the variance between children in auditory-visual speech perception. On the basis of the interpretations of Experiment 1 with school children and adults, it might be suggested that this is because the 3- to 4-year-old period is not one of high linguistic challenge in the sense that reading onset is a linguistic challenge for older children. In this regard it is of interest to note that one of the main linguistic achievements in this period, rapid vocabulary growth, does not predict either lipreading or auditory-visual speech perception.

While the preschool period in general does not appear to be a period of high linguistic challenge, it may be the case that if or when there are linguistic challenges in this period, such as when there are articulation difficulties, children with better language specific speech perception and thus stronger phonological
codes may have better auditory-visual speech perception. This was the subject matter of Experiment 3.

The finding that there is an increase from age three to four in auditory-visual speech integration supports other similar developmental findings (Experiment 1 here; Massaro et al., 1986; McGurk & MacDonald, 1976; Sekiyama & Burnham, 2004). One possible question regarding this age-related increase concerns the status of the link between phonological processing and auditory-visual speech perception. It is possible that for children with phonological speech disorders, such an increase in auditory-visual speech perception may not occur due to insufficient or degraded articulatory input. The results and theoretical implications of Experiment 3 are discussed in the next section.

8.1.3 Experiment 3: Preschool Children with Phonological Disorder

Experiment 3 was conducted in order to investigate further the status of the link between phonological processing and auditory-visual speech perception in the presence of disordered phonological processing.

Comparison of speech disordered (SD) and non-speech disordered (no-SD) children revealed better lipreading ability by the no-SD group, but no differences in auditory-visual speech perception, auditory-only speech perception, or executive function. The fact that children without speech disorder are better lipreaders suggests that they have a stronger representational code for visual speech information than SD children.

Analysis of data from three age groups of SD children revealed no age-based differences in any of the auditory-visual speech perception test components (VSI-AX, AO, and VO), yet the groups differed on executive function performance. The finding that there were no age differences on the auditory-visual speech perception test measures is at odds with comparable no-SD child age groups (see Experiment 2). Presumably due to their articulation difficulty and possibly their limited phonological representations (Dodd, 1995) SD children do not improve significantly on auditory-visual speech integration, AO speech perception, or lipreading. Considering these results in relation to those of Experiments 1 and 2,
one of the factors that should be investigated is whether there is a connection
between speech perception and vocabulary size in children with children speech
disorders. Unfortunately it was not possible to test these speech-disordered
children for their receptive vocabulary for several practical reasons.

8.2 CURRENT RESULTS, TWO MODELS OF SPEECH PERCEPTION, AND A
SUGGESTION OF A FUTURE STUDY

8.2.1 Perceptual Assimilation Model
Details of the perceptual assimilation model (PAM) were given in 3.3.2. In
essence, PAM asserts that native language speech perception affects the
perception of non-native phonetic perception. PAM is based on auditory
perception but, due to its underpinnings of motor and gestural theory, the role of
visual speech information could well be integrated into the theory. The results
here show that young children use visual information in speech perception and
this improves over age. Moreover there is a system of correlations between
articulation, reading, and lipreading (see Figure 8-1) with inferior lipreading in
children with articulation difficulties (speech disorders). Given these findings,
PAM might profitably accommodate the role of visual speech in perceptual
assimilation in order to explain more fully the perception and development of
speech.

8.2.2 Motor Theory of Speech Perception
The position of the current results in relation to the motor theory of speech
perception has been elaborated, in particular regarding the results of Experiment
3 (see 8.1.3). Essentially the motor theory of speech perception claims that speech
perception relies upon the perception of the pattern of articulatory movements
(Liberman & Mattingly, 1985). Given evidence for the role of visual information
in speech perception in this thesis, it would appear that apprehending intended
speech gestures would be aided by visual information even in young children.
Indeed, it was found that auditory-visual speech perception is not as good in SD
children, suggesting that they have less traffic along the articulation-perception
link, and the motor theory may benefit from testing its predicts with special populations such as children and adults with phonologically based speech production problems.

### 8.2.3 A Future Test of Articulatory Models of Auditory and Auditory-Visual Speech Perception

Articulation is a significant linguistic challenge for children and future studies of auditory-visual speech perception could focus on the role of articulatory aspects during this period. This could be done in two ways, with normally articulating children and speech disordered children. For normally articulating children it would be fruitful to investigate the relationship between auditory-visual speech perception and articulation along with other age related linguistic abilities. In retrospect, and the given the overall results, articulation could have been included here, and could be profitably included in future studies.

In Experiment 3, due to a large amount of missing data for language specific speech perception, it was not possible to test the link between auditory-visual and language specific speech perception, a link found in Experiment 1. Testing consistent and inconsistent phonological speech disorder children would provide a good test of accounts of auditory-visual speech perception by articulatory models of speech perception. A future study should recruit children with inconsistent and consistent phonological speech disorders, and control children with normal speech development. The best time to test these children would be at five years or the first year of school (kindergarten). This is the start of a linguistically challenging period due to an increase in exposure to different dialects, accents, teacherese, etc., but free from reading instruction, which appear to increase the use of visual speech information (see 5.6). Children with these two types of speech disorder make more speech errors than children with normal speech, with inconsistent speech disorder children making as many or more errors than children with consistent phonological speech disorder (Dodd, 1995). If auditory-visual speech perception is phonologically-oriented at five years, then
in the face of new linguistic challenges at school, it is hypothesised that for five-year-old children with normal speech development and five-year-old children with consistent phonological speech disorder, there should be a predictive relationship between language specific and auditory-visual speech perception, as for these two groups there is a consistent perception-production link for phonemic categories. However, for five-year-old children with inconsistent phonological speech disorder, it is predicted that the link between language specific and auditory-visual speech perception would not be as strong as that for children with normal speech.

8.3 Implications of the Current Findings and Research in Auditory-Visual Speech Perception

8.3.1 Implications for Teaching Literacy and Foreign Language

Perhaps the most important findings of this thesis come from the results of Experiment 1 with school children: auditory-visual integration, lipreading, articulation, and reading skills increase with age, and language specific speech perception predicts the degree of auditory-visual speech perception. In addition, we know from research in auditory-visual speech perception and non-native and second language processing that perceivers benefit from the addition of visual speech information either alone (Erdener & Burnham, 2005c; Fuster-Duran, 1996; Hardison, 1999), especially when visual cues are made more salient (Hazan, Sennema, Iba, & Faulkner, 2005), or together with orthographic information (Erdener & Burnham, 2005c). Moreover, it has been found that articulatory training of unfamiliar phonemes which involves a good degree of visualisation and awareness of the position of articulators, yields more accurate production of target phonemes than training methods solely based on auditory presentation and repetition (Catford & Pisoni, 1970). In addition, there is also evidence of a link between reading and lipreading skills (see de Gelder & Vroomen, 1998), and that children with dyslexia use visual speech information less than their normal-reading counterparts (Cavé, Stroumza, & Bastien-Toniazzo, 2007), which calls for further scrutiny of the relationship between different types of dyslexia and
auditory-visual speech perception. Further to this, teaching and teacher training methods that may be developed in light of research findings could include emphasis on visual speech information in the form of video-recorded aids, computer programs that can exploit this knowledge in literacy teaching or teachers’ emphasis on visual speech cues to enrich the teaching process. Though this thesis was mainly concerned with the development of auditory-visual speech perception, on the basis of the Experiment 1 finding that auditory-visual and language specific speech perception are related in the face of linguistic challenges during childhood, it can hypothesised that the same link might be re-activated when adults learn a foreign language. Therefore further research on the relationship between auditory-visual and language specific speech perception and other language skills in adults learning a foreign language would be beneficial.

8.3.2 Implications and Future Research Directions for Speech Therapy for Language-Impaired Children and Adults

Unfortunately empirical research on the role of auditory-visual speech perception in disordered communication has been limited to a handful of studies on people with hearing problems. Other studies on disordered communication are on the auditory-only aspect of speech perception. Based on the current results, it appears that speech therapy techniques for people with speech disorders, in particular phonological speech disorders, would benefit from research on auditory-visual speech perception. Therapies could emphasise the visual aspects of the speech signal by the provision of ‘viseme training’ in order to compensate for limited naturally occurring lipreading or the auditory-visual integration. Further research is required with children and adults with and without speech disorders in order to increase our knowledge of the multimodal speech processing system of these populations (see 8.2.3).
8.4 Further Questions, Future Research and Conclusion

Overall this research has found that, in addition to early-stage phonetic auditory-visual speech integration (Burnham & Dodd, 2004; Rosenblum et al., 1997), there are also language specific factors operating at a phonological level that affect auditory-visual speech integration. So, in addition to infants’ early integration of auditory and visual speech information phonetically, over age and, as a function of linguistic experience, this becomes phonologically based as schematically portrayed in Figure 8-2.

![Figure 8-2 Schematic representation of the phonetic and phonological development of auditory-visual speech integration over infancy, childhood and adulthood.](image)

As indicated in Chapter 1, developmental data are required to understand auditory-visual speech integration fully, and this knowledge may well assist engineers to develop AV-ASR systems. The findings of this thesis both reinforce and more clearly specify that the integration of auditory and visual speech information is both phonetically (Burnham & Dodd, 2004; Rosenblum et al., 1997) and phonologically processed, and that auditory-visual speech perception is aided by the gradual sophistication of auditory-only and visual-only speech processing. Some of the areas which may exploit the findings of developmental
and non-developmental auditory-visual speech processing research include online interactive transaction systems, automatic auditory-visual speech recognition systems, interactive systems that provide information for the elderly, those with sensory (especially auditory or visual) disabilities, telecommunications, and computer interfaces. The main challenge is to understand how humans perceive and process speech input in a multimodal fashion. However, this challenge cannot only be overcome by psychological or engineering research alone. In order to understand the multimodal nature of human speech mechanisms, a multi-disciplinary approach is required. Such efforts require researchers and professionals from industry to collaborate to study multimodal human speech processing from all angles: psychological, linguistic, engineering, and more.}

21 An example of such a multidisciplinary program is The Thinking Head Project centred at MARCS Auditory Laboratories in collaboration with several other research institutes in Australia, the United States, Denmark and Germany. For more information see http://thinkinghead.uws.edu.au.
CHAPTER 9

REFERENCE LIST


APPENDIX A1

QUEENSLAND ARTICULATION TEST ITEM LIST
# Table A1-1 The Consonants and Test Items Used in the Queensland Articulation Test

<table>
<thead>
<tr>
<th>Sound</th>
<th>Initial</th>
<th>Medial</th>
<th>Final</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>Pig</td>
<td>Apple</td>
<td>Cup</td>
</tr>
<tr>
<td>b</td>
<td>Ball</td>
<td>Baby</td>
<td>Web</td>
</tr>
<tr>
<td>m</td>
<td>Mouse</td>
<td>Hammer</td>
<td>Drum</td>
</tr>
<tr>
<td>w</td>
<td>Window</td>
<td>Flower</td>
<td>Window</td>
</tr>
<tr>
<td>f</td>
<td>Fish</td>
<td>Elephant</td>
<td>Knife</td>
</tr>
<tr>
<td>v</td>
<td>Vacuum</td>
<td>Television</td>
<td>Stove</td>
</tr>
<tr>
<td>θ</td>
<td>Thumb</td>
<td>Nothing</td>
<td>Mouth</td>
</tr>
<tr>
<td>δ</td>
<td>That</td>
<td>Feather</td>
<td>Smooth</td>
</tr>
<tr>
<td>t</td>
<td>Tap</td>
<td>Letter</td>
<td>Boat</td>
</tr>
<tr>
<td>d</td>
<td>Dog</td>
<td>Spider</td>
<td>Bed</td>
</tr>
<tr>
<td>s</td>
<td>Sock</td>
<td>Glasses</td>
<td>House</td>
</tr>
<tr>
<td>z</td>
<td>Zip</td>
<td>Scissors</td>
<td>Toes</td>
</tr>
<tr>
<td>l</td>
<td>Leaf</td>
<td>Lollies</td>
<td>Ball</td>
</tr>
<tr>
<td>n</td>
<td>Knife</td>
<td>Banana</td>
<td>Spoon</td>
</tr>
<tr>
<td>j</td>
<td>Shoe</td>
<td>Washing</td>
<td>Fish</td>
</tr>
<tr>
<td>tj</td>
<td>Chair</td>
<td>Matches</td>
<td>Watch</td>
</tr>
<tr>
<td>dz</td>
<td>Jam</td>
<td>Vegemite</td>
<td>Cage</td>
</tr>
<tr>
<td>j</td>
<td>Yellow</td>
<td>Onion</td>
<td>-</td>
</tr>
<tr>
<td>r</td>
<td>Red</td>
<td>Carrot</td>
<td>-</td>
</tr>
<tr>
<td>k</td>
<td>Car</td>
<td>Bucket</td>
<td>Book</td>
</tr>
<tr>
<td>g</td>
<td>Gun</td>
<td>Tiger</td>
<td>Dog</td>
</tr>
<tr>
<td>η</td>
<td>N/A</td>
<td>Swinging</td>
<td>Swing</td>
</tr>
<tr>
<td>h</td>
<td>House</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
APPENDIX A2

COUNTERBALANCING ORDERS IN EXPERIMENTS
1, 2 AND 3
### APPENDIX A2 - TASK ORDERS

#### Table A2-1: Task Orders in Experiment 1: School Children

<table>
<thead>
<tr>
<th>Subject Code</th>
<th>1st Task</th>
<th>2nd Task</th>
<th>3rd Task</th>
<th>4th Task</th>
<th>“different” button</th>
</tr>
</thead>
<tbody>
<tr>
<td>5M01</td>
<td>McGurk1</td>
<td>Reading B</td>
<td>QAT</td>
<td>LSSP1</td>
<td>Left</td>
</tr>
<tr>
<td>5M02</td>
<td>McGurk2</td>
<td>Reading T</td>
<td>QAT</td>
<td>LSSP1</td>
<td>Right</td>
</tr>
<tr>
<td>5M03</td>
<td>McGurk1</td>
<td>Reading B</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Left</td>
</tr>
<tr>
<td>5M04</td>
<td>McGurk2</td>
<td>Reading T</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Right</td>
</tr>
<tr>
<td>5M05</td>
<td>McGurk1</td>
<td>LSSP1</td>
<td>Reading B</td>
<td>QAT</td>
<td>Left</td>
</tr>
<tr>
<td>5M06</td>
<td>McGurk2</td>
<td>LSSP1</td>
<td>Reading T</td>
<td>QAT</td>
<td>Right</td>
</tr>
<tr>
<td>5M07</td>
<td>McGurk1</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Reading B</td>
<td>Left</td>
</tr>
<tr>
<td>5M08</td>
<td>McGurk2</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Reading T</td>
<td>Right</td>
</tr>
<tr>
<td>5M09</td>
<td>McGurk1</td>
<td>QAT</td>
<td>Reading B</td>
<td>LSSP1</td>
<td>Left</td>
</tr>
<tr>
<td>5M10</td>
<td>McGurk2</td>
<td>QAT</td>
<td>Reading T</td>
<td>LSSP1</td>
<td>Right</td>
</tr>
<tr>
<td>5M11</td>
<td>McGurk1</td>
<td>QAT</td>
<td>LSSP2</td>
<td>Reading B</td>
<td>Left</td>
</tr>
<tr>
<td>5M12</td>
<td>McGurk2</td>
<td>QAT</td>
<td>LSSP2</td>
<td>Reading T</td>
<td>Right</td>
</tr>
<tr>
<td>5F01</td>
<td>McGurk1</td>
<td>Reading B</td>
<td>QAT</td>
<td>LSSP1</td>
<td>Left</td>
</tr>
<tr>
<td>5F02</td>
<td>McGurk2</td>
<td>Reading T</td>
<td>QAT</td>
<td>LSSP1</td>
<td>Right</td>
</tr>
<tr>
<td>5F03</td>
<td>McGurk1</td>
<td>Reading B</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Left</td>
</tr>
<tr>
<td>5F04</td>
<td>McGurk2</td>
<td>Reading T</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Right</td>
</tr>
<tr>
<td>5F05</td>
<td>McGurk1</td>
<td>LSSP1</td>
<td>Reading B</td>
<td>QAT</td>
<td>Left</td>
</tr>
<tr>
<td>5F06</td>
<td>McGurk2</td>
<td>LSSP1</td>
<td>Reading T</td>
<td>QAT</td>
<td>Right</td>
</tr>
<tr>
<td>5F07</td>
<td>McGurk1</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Reading B</td>
<td>Left</td>
</tr>
<tr>
<td>5F08</td>
<td>McGurk2</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Reading T</td>
<td>Right</td>
</tr>
<tr>
<td>5F09</td>
<td>McGurk1</td>
<td>QAT</td>
<td>Reading B</td>
<td>LSSP1</td>
<td>Left</td>
</tr>
<tr>
<td>5F10</td>
<td>McGurk2</td>
<td>QAT</td>
<td>Reading T</td>
<td>LSSP1</td>
<td>Right</td>
</tr>
<tr>
<td>5F11</td>
<td>McGurk1</td>
<td>QAT</td>
<td>LSSP2</td>
<td>Reading B</td>
<td>Left</td>
</tr>
<tr>
<td>5F12</td>
<td>McGurk2</td>
<td>QAT</td>
<td>LSSP2</td>
<td>Reading T</td>
<td>Right</td>
</tr>
<tr>
<td>Subject Code</td>
<td>1st Task</td>
<td>2nd Task</td>
<td>3rd Task</td>
<td>4th Task</td>
<td>LSSP “different” button</td>
</tr>
<tr>
<td>---------------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>6M01</td>
<td>McGurk1</td>
<td>Reading B</td>
<td>QAT</td>
<td>LSSP1</td>
<td>Left</td>
</tr>
<tr>
<td>6M02</td>
<td>McGurk2</td>
<td>Reading T</td>
<td>QAT</td>
<td>LSSP1</td>
<td>Right</td>
</tr>
<tr>
<td>6M03</td>
<td>McGurk1</td>
<td>Reading B</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Left</td>
</tr>
<tr>
<td>6M04</td>
<td>McGurk2</td>
<td>Reading T</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Right</td>
</tr>
<tr>
<td>6M05</td>
<td>McGurk1</td>
<td>LSSP1</td>
<td>Reading B</td>
<td>QAT</td>
<td>Left</td>
</tr>
<tr>
<td>6M06</td>
<td>McGurk2</td>
<td>LSSP1</td>
<td>Reading T</td>
<td>QAT</td>
<td>Right</td>
</tr>
<tr>
<td>6M07</td>
<td>McGurk1</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Reading B</td>
<td>Left</td>
</tr>
<tr>
<td>6M08</td>
<td>McGurk2</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Reading T</td>
<td>Right</td>
</tr>
<tr>
<td>6M09</td>
<td>McGurk1</td>
<td>QAT</td>
<td>Reading B</td>
<td>LSSP1</td>
<td>Left</td>
</tr>
<tr>
<td>6M10</td>
<td>McGurk2</td>
<td>QAT</td>
<td>Reading T</td>
<td>LSSP1</td>
<td>Right</td>
</tr>
<tr>
<td>6M11</td>
<td>McGurk1</td>
<td>QAT</td>
<td>LSSP2</td>
<td>Reading B</td>
<td>Left</td>
</tr>
<tr>
<td>6M12</td>
<td>McGurk2</td>
<td>QAT</td>
<td>LSSP2</td>
<td>Reading T</td>
<td>Right</td>
</tr>
<tr>
<td>6F01</td>
<td>McGurk1</td>
<td>Reading B</td>
<td>QAT</td>
<td>LSSP1</td>
<td>Left</td>
</tr>
<tr>
<td>6F02</td>
<td>McGurk2</td>
<td>Reading T</td>
<td>QAT</td>
<td>LSSP1</td>
<td>Right</td>
</tr>
<tr>
<td>6F03</td>
<td>McGurk1</td>
<td>Reading B</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Left</td>
</tr>
<tr>
<td>6F04</td>
<td>McGurk2</td>
<td>Reading T</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Right</td>
</tr>
<tr>
<td>6F05</td>
<td>McGurk1</td>
<td>LSSP1</td>
<td>Reading B</td>
<td>QAT</td>
<td>Left</td>
</tr>
<tr>
<td>6F06</td>
<td>McGurk2</td>
<td>LSSP1</td>
<td>Reading T</td>
<td>QAT</td>
<td>Right</td>
</tr>
<tr>
<td>6F07</td>
<td>McGurk1</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Reading B</td>
<td>Left</td>
</tr>
<tr>
<td>6F08</td>
<td>McGurk2</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Reading T</td>
<td>Right</td>
</tr>
<tr>
<td>6F09</td>
<td>McGurk1</td>
<td>QAT</td>
<td>Reading B</td>
<td>LSSP1</td>
<td>Left</td>
</tr>
<tr>
<td>6F10</td>
<td>McGurk2</td>
<td>QAT</td>
<td>Reading T</td>
<td>LSSP1</td>
<td>Right</td>
</tr>
<tr>
<td>6F11</td>
<td>McGurk1</td>
<td>QAT</td>
<td>LSSP2</td>
<td>Reading B</td>
<td>Left</td>
</tr>
<tr>
<td>6F12</td>
<td>McGurk2</td>
<td>QAT</td>
<td>LSSP2</td>
<td>Reading T</td>
<td>Right</td>
</tr>
<tr>
<td>Subject Code</td>
<td>1st Task</td>
<td>2nd Task</td>
<td>3rd Task</td>
<td>4th Task</td>
<td>LSSP “different” button</td>
</tr>
<tr>
<td>--------------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>------------------------</td>
</tr>
<tr>
<td>7M01</td>
<td>McGurk1</td>
<td>Reading B</td>
<td>QAT</td>
<td>LSSP1</td>
<td>Left</td>
</tr>
<tr>
<td>7M02</td>
<td>McGurk2</td>
<td>Reading T</td>
<td>QAT</td>
<td>LSSP1</td>
<td>Right</td>
</tr>
<tr>
<td>7M03</td>
<td>McGurk1</td>
<td>Reading B</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Left</td>
</tr>
<tr>
<td>7M04</td>
<td>McGurk2</td>
<td>Reading T</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Right</td>
</tr>
<tr>
<td>7M05</td>
<td>McGurk1</td>
<td>LSSP1</td>
<td>Reading B</td>
<td>QAT</td>
<td>Left</td>
</tr>
<tr>
<td>7M06</td>
<td>McGurk2</td>
<td>LSSP1</td>
<td>Reading T</td>
<td>QAT</td>
<td>Right</td>
</tr>
<tr>
<td>7M07</td>
<td>McGurk1</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Reading B</td>
<td>Left</td>
</tr>
<tr>
<td>7M08</td>
<td>McGurk2</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Reading T</td>
<td>Right</td>
</tr>
<tr>
<td>7M09</td>
<td>McGurk1</td>
<td>QAT</td>
<td>Reading B</td>
<td>LSSP1</td>
<td>Left</td>
</tr>
<tr>
<td>7M10</td>
<td>McGurk2</td>
<td>QAT</td>
<td>Reading T</td>
<td>LSSP1</td>
<td>Right</td>
</tr>
<tr>
<td>7M11</td>
<td>McGurk1</td>
<td>QAT</td>
<td>LSSP2</td>
<td>Reading B</td>
<td>Left</td>
</tr>
<tr>
<td>7M12</td>
<td>McGurk2</td>
<td>QAT</td>
<td>LSSP2</td>
<td>Reading T</td>
<td>Right</td>
</tr>
<tr>
<td>7F01</td>
<td>McGurk1</td>
<td>Reading B</td>
<td>QAT</td>
<td>LSSP1</td>
<td>Left</td>
</tr>
<tr>
<td>7F02</td>
<td>McGurk2</td>
<td>Reading T</td>
<td>QAT</td>
<td>LSSP1</td>
<td>Right</td>
</tr>
<tr>
<td>7F03</td>
<td>McGurk1</td>
<td>Reading B</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Left</td>
</tr>
<tr>
<td>7F04</td>
<td>McGurk2</td>
<td>Reading T</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Right</td>
</tr>
<tr>
<td>7F05</td>
<td>McGurk1</td>
<td>LSSP1</td>
<td>Reading B</td>
<td>QAT</td>
<td>Left</td>
</tr>
<tr>
<td>7F06</td>
<td>McGurk2</td>
<td>LSSP1</td>
<td>Reading T</td>
<td>QAT</td>
<td>Right</td>
</tr>
<tr>
<td>7F07</td>
<td>McGurk1</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Reading B</td>
<td>Left</td>
</tr>
<tr>
<td>7F08</td>
<td>McGurk2</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Reading T</td>
<td>Right</td>
</tr>
<tr>
<td>7F09</td>
<td>McGurk1</td>
<td>QAT</td>
<td>Reading B</td>
<td>LSSP1</td>
<td>Left</td>
</tr>
<tr>
<td>7F10</td>
<td>McGurk2</td>
<td>QAT</td>
<td>Reading T</td>
<td>LSSP1</td>
<td>Right</td>
</tr>
<tr>
<td>7F11</td>
<td>McGurk1</td>
<td>QAT</td>
<td>LSSP2</td>
<td>Reading B</td>
<td>Left</td>
</tr>
<tr>
<td>7F12</td>
<td>McGurk2</td>
<td>QAT</td>
<td>LSSP2</td>
<td>Reading T</td>
<td>Right</td>
</tr>
<tr>
<td>Subject Code</td>
<td>1st Task</td>
<td>2nd Task</td>
<td>3rd Task</td>
<td>4th Task</td>
<td>LSSP “different” button</td>
</tr>
<tr>
<td>--------------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>8M01</td>
<td>McGurk1</td>
<td>Reading B</td>
<td>QAT</td>
<td>LSSP1</td>
<td>Left</td>
</tr>
<tr>
<td>8M02</td>
<td>McGurk2</td>
<td>Reading T</td>
<td>QAT</td>
<td>LSSP1</td>
<td>Right</td>
</tr>
<tr>
<td>8M03</td>
<td>McGurk1</td>
<td>Reading B</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Left</td>
</tr>
<tr>
<td>8M04</td>
<td>McGurk2</td>
<td>Reading T</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Right</td>
</tr>
<tr>
<td>8M05</td>
<td>McGurk1</td>
<td>LSSP1</td>
<td>Reading B</td>
<td>QAT</td>
<td>Left</td>
</tr>
<tr>
<td>8M06</td>
<td>McGurk2</td>
<td>LSSP1</td>
<td>Reading T</td>
<td>QAT</td>
<td>Right</td>
</tr>
<tr>
<td>8M07</td>
<td>McGurk1</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Reading B</td>
<td>Left</td>
</tr>
<tr>
<td>8M08</td>
<td>McGurk2</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Reading T</td>
<td>Right</td>
</tr>
<tr>
<td>8M09</td>
<td>McGurk1</td>
<td>QAT</td>
<td>Reading B</td>
<td>LSSP1</td>
<td>Left</td>
</tr>
<tr>
<td>8M10</td>
<td>McGurk2</td>
<td>QAT</td>
<td>Reading T</td>
<td>LSSP1</td>
<td>Right</td>
</tr>
<tr>
<td>8M11</td>
<td>McGurk1</td>
<td>QAT</td>
<td>LSSP2</td>
<td>Reading B</td>
<td>Left</td>
</tr>
<tr>
<td>8M12</td>
<td>McGurk2</td>
<td>QAT</td>
<td>LSSP2</td>
<td>Reading T</td>
<td>Right</td>
</tr>
<tr>
<td>8F01</td>
<td>McGurk1</td>
<td>Reading B</td>
<td>QAT</td>
<td>LSSP1</td>
<td>Left</td>
</tr>
<tr>
<td>8F02</td>
<td>McGurk2</td>
<td>Reading T</td>
<td>QAT</td>
<td>LSSP1</td>
<td>Right</td>
</tr>
<tr>
<td>8F03</td>
<td>McGurk1</td>
<td>Reading B</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Left</td>
</tr>
<tr>
<td>8F04</td>
<td>McGurk2</td>
<td>Reading T</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Right</td>
</tr>
<tr>
<td>8F05</td>
<td>McGurk1</td>
<td>LSSP1</td>
<td>Reading B</td>
<td>QAT</td>
<td>Left</td>
</tr>
<tr>
<td>8F06</td>
<td>McGurk2</td>
<td>LSSP1</td>
<td>Reading T</td>
<td>QAT</td>
<td>Right</td>
</tr>
<tr>
<td>8F07</td>
<td>McGurk1</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Reading B</td>
<td>Left</td>
</tr>
<tr>
<td>8F08</td>
<td>McGurk2</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Reading T</td>
<td>Right</td>
</tr>
<tr>
<td>8F09</td>
<td>McGurk1</td>
<td>QAT</td>
<td>Reading B</td>
<td>LSSP1</td>
<td>Left</td>
</tr>
<tr>
<td>8F10</td>
<td>McGurk2</td>
<td>QAT</td>
<td>Reading T</td>
<td>LSSP1</td>
<td>Right</td>
</tr>
<tr>
<td>8F11</td>
<td>McGurk1</td>
<td>QAT</td>
<td>LSSP2</td>
<td>Reading B</td>
<td>Left</td>
</tr>
<tr>
<td>8F12</td>
<td>McGurk2</td>
<td>QAT</td>
<td>LSSP2</td>
<td>Reading T</td>
<td>Right</td>
</tr>
</tbody>
</table>
## TABLE A2-2 TASK ORDERS IN EXPERIMENT 1: ADULTS

<table>
<thead>
<tr>
<th>Subject Code</th>
<th>1st Task</th>
<th>2nd Task</th>
<th>3rd Task</th>
<th>4th Task</th>
<th>“different” button</th>
</tr>
</thead>
<tbody>
<tr>
<td>FA01</td>
<td>McGurk1</td>
<td>Reading</td>
<td>QAT</td>
<td>LSSP1</td>
<td>Left</td>
</tr>
<tr>
<td>FA02</td>
<td>McGurk2</td>
<td>Reading</td>
<td>QAT</td>
<td>LSSP1</td>
<td>Right</td>
</tr>
<tr>
<td>FA03</td>
<td>McGurk1</td>
<td>Reading</td>
<td>LSSP1</td>
<td>QAT</td>
<td>Left</td>
</tr>
<tr>
<td>FA04</td>
<td>McGurk2</td>
<td>Reading</td>
<td>LSSP1</td>
<td>QAT</td>
<td>Right</td>
</tr>
<tr>
<td>FA05</td>
<td>McGurk1</td>
<td>LSSP1</td>
<td>Reading</td>
<td>QAT</td>
<td>Left</td>
</tr>
<tr>
<td>FA06</td>
<td>McGurk2</td>
<td>LSSP1</td>
<td>Reading</td>
<td>QAT</td>
<td>Right</td>
</tr>
<tr>
<td>FA07</td>
<td>McGurk1</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Reading</td>
<td>Left</td>
</tr>
<tr>
<td>FA08</td>
<td>McGurk2</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Reading</td>
<td>Right</td>
</tr>
<tr>
<td>FA09</td>
<td>McGurk1</td>
<td>QAT</td>
<td>Reading</td>
<td>LSSP1</td>
<td>Left</td>
</tr>
<tr>
<td>FA10</td>
<td>McGurk2</td>
<td>QAT</td>
<td>Reading</td>
<td>LSSP1</td>
<td>Right</td>
</tr>
<tr>
<td>FA11</td>
<td>McGurk1</td>
<td>QAT</td>
<td>LSSP2</td>
<td>Reading</td>
<td>Left</td>
</tr>
<tr>
<td>FA12</td>
<td>McGurk2</td>
<td>QAT</td>
<td>LSSP2</td>
<td>Reading</td>
<td>Right</td>
</tr>
<tr>
<td>MA01</td>
<td>McGurk1</td>
<td>Reading</td>
<td>QAT</td>
<td>LSSP1</td>
<td>Left</td>
</tr>
<tr>
<td>FA13</td>
<td>McGurk2</td>
<td>Reading</td>
<td>QAT</td>
<td>LSSP1</td>
<td>Right</td>
</tr>
<tr>
<td>FA14</td>
<td>McGurk1</td>
<td>Reading</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Left</td>
</tr>
<tr>
<td>FA15</td>
<td>McGurk2</td>
<td>Reading</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Right</td>
</tr>
<tr>
<td>FA16</td>
<td>McGurk1</td>
<td>LSSP1</td>
<td>Reading</td>
<td>QAT</td>
<td>Left</td>
</tr>
<tr>
<td>FA17</td>
<td>McGurk2</td>
<td>LSSP1</td>
<td>Reading</td>
<td>QAT</td>
<td>Right</td>
</tr>
<tr>
<td>FA18</td>
<td>McGurk1</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Reading</td>
<td>Left</td>
</tr>
<tr>
<td>FA19</td>
<td>McGurk2</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Reading</td>
<td>Right</td>
</tr>
<tr>
<td>FA20</td>
<td>McGurk1</td>
<td>QAT</td>
<td>Reading</td>
<td>LSSP1</td>
<td>Left</td>
</tr>
<tr>
<td>FA21</td>
<td>McGurk2</td>
<td>QAT</td>
<td>Reading</td>
<td>LSSP1</td>
<td>Right</td>
</tr>
<tr>
<td>FA22</td>
<td>McGurk1</td>
<td>QAT</td>
<td>LSSP2</td>
<td>Reading</td>
<td>Left</td>
</tr>
<tr>
<td>FA23</td>
<td>McGurk2</td>
<td>QAT</td>
<td>LSSP2</td>
<td>Reading</td>
<td>Right</td>
</tr>
<tr>
<td>Subject Code</td>
<td>1st Task</td>
<td>2nd Task</td>
<td>3rd Task</td>
<td>4th Task</td>
<td>LSSP “different” button</td>
</tr>
<tr>
<td>-------------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>FA24</td>
<td>McGurk1</td>
<td>Reading</td>
<td>QAT</td>
<td>LSSP1</td>
<td>Left</td>
</tr>
<tr>
<td>FA25</td>
<td>McGurk2</td>
<td>Reading</td>
<td>QAT</td>
<td>LSSP1</td>
<td>Right</td>
</tr>
<tr>
<td>FA26</td>
<td>McGurk1</td>
<td>Reading</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Left</td>
</tr>
<tr>
<td>FA27</td>
<td>McGurk2</td>
<td>Reading</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Right</td>
</tr>
<tr>
<td>FA28</td>
<td>McGurk1</td>
<td>LSSP1</td>
<td>Reading</td>
<td>QAT</td>
<td>Left</td>
</tr>
<tr>
<td>FA29</td>
<td>McGurk2</td>
<td>LSSP1</td>
<td>Reading</td>
<td>QAT</td>
<td>Right</td>
</tr>
<tr>
<td>MA02</td>
<td>McGurk1</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Reading</td>
<td>Left</td>
</tr>
<tr>
<td>FA30</td>
<td>McGurk2</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Reading</td>
<td>Right</td>
</tr>
<tr>
<td>FA31</td>
<td>McGurk1</td>
<td>QAT</td>
<td>Reading</td>
<td>LSSP1</td>
<td>Left</td>
</tr>
<tr>
<td>FA32</td>
<td>McGurk2</td>
<td>QAT</td>
<td>Reading</td>
<td>LSSP1</td>
<td>Right</td>
</tr>
<tr>
<td>FA33</td>
<td>McGurk1</td>
<td>QAT</td>
<td>LSSP2</td>
<td>Reading</td>
<td>Left</td>
</tr>
<tr>
<td>FA34</td>
<td>McGurk2</td>
<td>QAT</td>
<td>LSSP2</td>
<td>Reading</td>
<td>Right</td>
</tr>
<tr>
<td>MA03</td>
<td>McGurk1</td>
<td>Reading</td>
<td>QAT</td>
<td>LSSP1</td>
<td>Left</td>
</tr>
<tr>
<td>MA04</td>
<td>McGurk2</td>
<td>Reading</td>
<td>QAT</td>
<td>LSSP1</td>
<td>Right</td>
</tr>
<tr>
<td>FA35</td>
<td>McGurk1</td>
<td>Reading</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Left</td>
</tr>
<tr>
<td>FA36</td>
<td>McGurk2</td>
<td>Reading</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Right</td>
</tr>
<tr>
<td>FA37</td>
<td>McGurk1</td>
<td>LSSP1</td>
<td>Reading</td>
<td>QAT</td>
<td>Left</td>
</tr>
<tr>
<td>FA38</td>
<td>McGurk2</td>
<td>LSSP1</td>
<td>Reading</td>
<td>QAT</td>
<td>Right</td>
</tr>
<tr>
<td>FA39</td>
<td>McGurk1</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Reading</td>
<td>Left</td>
</tr>
<tr>
<td>FA40</td>
<td>McGurk2</td>
<td>LSSP2</td>
<td>QAT</td>
<td>Reading</td>
<td>Right</td>
</tr>
<tr>
<td>FA41</td>
<td>McGurk1</td>
<td>QAT</td>
<td>Reading</td>
<td>LSSP1</td>
<td>Left</td>
</tr>
<tr>
<td>MA05</td>
<td>McGurk2</td>
<td>QAT</td>
<td>Reading</td>
<td>LSSP1</td>
<td>Right</td>
</tr>
<tr>
<td>MA06</td>
<td>McGurk1</td>
<td>QAT</td>
<td>LSSP2</td>
<td>Reading</td>
<td>Left</td>
</tr>
<tr>
<td>MA07</td>
<td>McGurk2</td>
<td>QAT</td>
<td>LSSP2</td>
<td>Reading</td>
<td>Right</td>
</tr>
</tbody>
</table>
### APPENDIX A2 – TASK ORDERS

#### TABLE A2-3 TASK ORDERS IN EXPERIMENT 2: PRESCHOOL CHILDREN

<table>
<thead>
<tr>
<th>Subject Code</th>
<th>1(^{st}) Task</th>
<th>2(^{nd}) Task</th>
<th>3(^{rd}) Task</th>
<th>4(^{th}) Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>3M01</td>
<td>McGurk1</td>
<td>LSSP</td>
<td>PPVT</td>
<td>FIST</td>
</tr>
<tr>
<td>3F01</td>
<td>McGurk2</td>
<td>LSSP</td>
<td>FIST</td>
<td>PPVT</td>
</tr>
<tr>
<td>3M02</td>
<td>McGurk1</td>
<td>PPVT</td>
<td>LSSP</td>
<td>FIST</td>
</tr>
<tr>
<td>3F02</td>
<td>McGurk2</td>
<td>PPVT</td>
<td>FIST</td>
<td>LSSP</td>
</tr>
<tr>
<td>3F03</td>
<td>McGurk1</td>
<td>FIST</td>
<td>LSSP</td>
<td>PPVT</td>
</tr>
<tr>
<td>3M03</td>
<td>McGurk2</td>
<td>FIST</td>
<td>PPVT</td>
<td>LSSP</td>
</tr>
<tr>
<td>3F04</td>
<td>LSSP</td>
<td>McGurk</td>
<td>PPVT</td>
<td>FIST</td>
</tr>
<tr>
<td>3F05</td>
<td>LSSP</td>
<td>McGurk</td>
<td>FIST</td>
<td>PPVT</td>
</tr>
<tr>
<td>3F06</td>
<td>LSSP</td>
<td>PPVT</td>
<td>McGurk</td>
<td>FIST</td>
</tr>
<tr>
<td>3M04</td>
<td>LSSP</td>
<td>PPVT</td>
<td>FIST</td>
<td>McGurk</td>
</tr>
<tr>
<td>3M05</td>
<td>LSSP</td>
<td>FIST</td>
<td>McGurk</td>
<td>PPVT</td>
</tr>
<tr>
<td>3F07</td>
<td>LSSP</td>
<td>FIST</td>
<td>PPVT</td>
<td>McGurk</td>
</tr>
<tr>
<td>3M06</td>
<td>PPVT</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
</tr>
<tr>
<td>3F08</td>
<td>PPVT</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
</tr>
<tr>
<td>3F09</td>
<td>PPVT</td>
<td>LSSP</td>
<td>McGurk</td>
<td>FIST</td>
</tr>
<tr>
<td>3M07</td>
<td>PPVT</td>
<td>LSSP</td>
<td>FIST</td>
<td>McGurk</td>
</tr>
<tr>
<td>3M08</td>
<td>PPVT</td>
<td>FIST</td>
<td>McGurk</td>
<td>LSSP</td>
</tr>
<tr>
<td>3F10</td>
<td>PPVT</td>
<td>FIST</td>
<td>LSSP</td>
<td>McGurk</td>
</tr>
<tr>
<td>3F11</td>
<td>FIST</td>
<td>McGurk</td>
<td>LSSP</td>
<td>PPVT</td>
</tr>
<tr>
<td>3M09</td>
<td>FIST</td>
<td>McGurk</td>
<td>PPVT</td>
<td>LSSP</td>
</tr>
<tr>
<td>3M10</td>
<td>FIST</td>
<td>LSSP</td>
<td>McGurk</td>
<td>PPVT</td>
</tr>
<tr>
<td>3M11</td>
<td>FIST</td>
<td>LSSP</td>
<td>PPVT</td>
<td>McGurk</td>
</tr>
<tr>
<td>3F12</td>
<td>FIST</td>
<td>PPVT</td>
<td>McGurk</td>
<td>LSSP</td>
</tr>
<tr>
<td>3M12</td>
<td>FIST</td>
<td>PPVT</td>
<td>LSSP</td>
<td>McGurk</td>
</tr>
</tbody>
</table>
## APPENDIX A2 – TASK ORDERS

<table>
<thead>
<tr>
<th>Subject Code</th>
<th>1st Task</th>
<th>2nd Task</th>
<th>3rd Task</th>
<th>4th Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>4F01</td>
<td>McGurk1</td>
<td>LSSP</td>
<td>PPVT</td>
<td>FIST</td>
</tr>
<tr>
<td>4F02</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
<td>PPVT</td>
</tr>
<tr>
<td>4M01</td>
<td>McGurk</td>
<td>PPVT</td>
<td>LSSP</td>
<td>FIST</td>
</tr>
<tr>
<td>4F03</td>
<td>McGurk</td>
<td>PPVT</td>
<td>FIST</td>
<td>LSSP</td>
</tr>
<tr>
<td>4M02</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
<td>PPVT</td>
</tr>
<tr>
<td>4M03</td>
<td>McGurk</td>
<td>FIST</td>
<td>PPVT</td>
<td>LSSP</td>
</tr>
<tr>
<td>4M04</td>
<td>LSSP</td>
<td>McGurk</td>
<td>PPVT</td>
<td>FIST</td>
</tr>
<tr>
<td>4M05</td>
<td>LSSP</td>
<td>McGurk</td>
<td>FIST</td>
<td>PPVT</td>
</tr>
<tr>
<td>4M06</td>
<td>LSSP</td>
<td>PPVT</td>
<td>McGurk</td>
<td>FIST</td>
</tr>
<tr>
<td>4M07</td>
<td>LSSP</td>
<td>PPVT</td>
<td>FIST</td>
<td>McGurk</td>
</tr>
<tr>
<td>4F04</td>
<td>LSSP</td>
<td>FIST</td>
<td>McGurk</td>
<td>PPVT</td>
</tr>
<tr>
<td>4M08</td>
<td>LSSP</td>
<td>FIST</td>
<td>PPVT</td>
<td>McGurk</td>
</tr>
<tr>
<td>4M09</td>
<td>PPVT</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
</tr>
<tr>
<td>4M10</td>
<td>PPVT</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
</tr>
<tr>
<td>4M11</td>
<td>PPVT</td>
<td>LSSP</td>
<td>McGurk</td>
<td>FIST</td>
</tr>
<tr>
<td>4F05</td>
<td>PPVT</td>
<td>LSSP</td>
<td>FIST</td>
<td>McGurk</td>
</tr>
<tr>
<td>4F06</td>
<td>PPVT</td>
<td>FIST</td>
<td>McGurk</td>
<td>LSSP</td>
</tr>
<tr>
<td>4F07</td>
<td>PPVT</td>
<td>FIST</td>
<td>LSSP</td>
<td>McGurk</td>
</tr>
<tr>
<td>4F08</td>
<td>FIST</td>
<td>McGurk</td>
<td>LSSP</td>
<td>PPVT</td>
</tr>
<tr>
<td>4F09</td>
<td>FIST</td>
<td>McGurk</td>
<td>PPVT</td>
<td>LSSP</td>
</tr>
<tr>
<td>4M12</td>
<td>FIST</td>
<td>LSSP</td>
<td>McGurk</td>
<td>PPVT</td>
</tr>
<tr>
<td>4F10</td>
<td>FIST</td>
<td>LSSP</td>
<td>PPVT</td>
<td>McGurk</td>
</tr>
<tr>
<td>4F11</td>
<td>FIST</td>
<td>PPVT</td>
<td>McGurk</td>
<td>LSSP</td>
</tr>
<tr>
<td>4F12</td>
<td>FIST</td>
<td>PPVT</td>
<td>LSSP</td>
<td>McGurk</td>
</tr>
<tr>
<td>Subject Code</td>
<td>1st Task</td>
<td>2nd Task</td>
<td>3rd Task</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td></td>
</tr>
<tr>
<td>C86</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
<td></td>
</tr>
<tr>
<td>C88</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
<td></td>
</tr>
<tr>
<td>D125</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
<td></td>
</tr>
<tr>
<td>D128</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
<td></td>
</tr>
<tr>
<td>D129</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
<td></td>
</tr>
<tr>
<td>D131</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
<td></td>
</tr>
<tr>
<td>D133</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
<td></td>
</tr>
<tr>
<td>D542</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
<td></td>
</tr>
<tr>
<td>E162</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
<td></td>
</tr>
<tr>
<td>E163</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
<td></td>
</tr>
<tr>
<td>E164</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
<td></td>
</tr>
<tr>
<td>E166</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
<td></td>
</tr>
<tr>
<td>E167</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
<td></td>
</tr>
<tr>
<td>F203</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
<td></td>
</tr>
<tr>
<td>F204</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
<td></td>
</tr>
<tr>
<td>F205</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
<td></td>
</tr>
<tr>
<td>F207</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
<td></td>
</tr>
<tr>
<td>F208</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
<td></td>
</tr>
<tr>
<td>D504</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
<td></td>
</tr>
<tr>
<td>D505</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
<td></td>
</tr>
<tr>
<td>D508</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
<td></td>
</tr>
<tr>
<td>D510</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
<td></td>
</tr>
<tr>
<td>D512</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
<td></td>
</tr>
<tr>
<td>D513</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
<td></td>
</tr>
<tr>
<td>D514</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
<td></td>
</tr>
<tr>
<td>D515</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
<td></td>
</tr>
<tr>
<td>D519</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
<td></td>
</tr>
<tr>
<td>D522</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
<td></td>
</tr>
</tbody>
</table>
## APPENDIX A2 – TASK ORDERS

<table>
<thead>
<tr>
<th>Subject Code</th>
<th>1st Task</th>
<th>2nd Task</th>
<th>3rd Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS23</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
</tr>
<tr>
<td>DS28</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
</tr>
<tr>
<td>DS29</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
</tr>
<tr>
<td>DS31</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
</tr>
<tr>
<td>DS32</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
</tr>
<tr>
<td>DS33</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
</tr>
<tr>
<td>DS35</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
</tr>
<tr>
<td>DS39</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
</tr>
<tr>
<td>DS43</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
</tr>
<tr>
<td>DS44</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
</tr>
<tr>
<td>DS45</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
</tr>
<tr>
<td>DS47</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
</tr>
<tr>
<td>DS56</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
</tr>
<tr>
<td>DS57</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
</tr>
<tr>
<td>DS59</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
</tr>
<tr>
<td>DS60</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
</tr>
<tr>
<td>DS61</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
</tr>
<tr>
<td>DS63</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
</tr>
<tr>
<td>DS65</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
</tr>
<tr>
<td>DS66</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
</tr>
<tr>
<td>DS67</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
</tr>
<tr>
<td>DS70</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
</tr>
<tr>
<td>DS71</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
</tr>
<tr>
<td>DS72</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
</tr>
<tr>
<td>DS74</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
</tr>
<tr>
<td>DS75</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
</tr>
<tr>
<td>DS76</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
</tr>
<tr>
<td>DSXX</td>
<td>McGurk</td>
<td>LSSP</td>
<td>FIST</td>
</tr>
<tr>
<td>DSXY</td>
<td>McGurk</td>
<td>FIST</td>
<td>LSSP</td>
</tr>
</tbody>
</table>
APPENDIX A3

THE CONVERSION OF LANGUAGE SPECIFIC SPEECH PERCEPTION DATA USING MATLAB®
A3.1 CONVERSION OF LANGUAGE SPECIFIC SPEECH PERCEPTION TEST DATA USING MATLAB

In Experiment 3 a special experimental design software called PsyScript (Bates & D'Oliveiro, 2003) was used in order to present the speech perception stimuli to preschool children. The program produces a number of output files for each participant in a txt format which necessitated the manual entry of each individual data set on an Excel-format (.csv) worksheet. Instead of following this time-consuming process software developers (Rua Haszard-Morris & Johnson Chen) at MARCS Auditory Laboratories wrote a data conversion script using MATLAB, a high-level flexible programming language and interactive environment that enables algorithms for various computationally intensive tasks, such as statistics and data analysis, signal processing, image processing, and financial modelling. First, a target folder was created which contained the language specific speech perception data from PsyScript to be analysed (C:\DoguData\PreschoolKids). Second, the data processing script was placed in a folder which MATLAB can easily process under the same directory root as the MATLAB program (C:\MATLAB\N-NNScripts).

When MATLAB was started it was ensured that the script featured the network path (U:\Rua\m-public) as the processing script accesses critical batch files located in this directory. After the path is set, at the MATLAB prompt in the command window the directory under which data are located is set by the following command:

```
cd C:\DoguData\PreschoolKids\Preschooldata
```

After that, the processing script was run by typing the following command:

```
summarisedoguNNNresults('C:\DoguData\PreschoolKids\Preschooldata')
```

Screenshots of the conversion process in MATLAB) and a sample output file are shown in Figure A3-1 and Figure A3-2. The MATLAB script that converted the data is presented in the next section.
Figure A3-1 A screenshot of the processing of the speech perception test data using a MATLAB-based script.

Figure A3-2 The screenshot of a sample output file in Excel format after the MATLAB processing of PsyScript raw data files.
A3.2 The MATLAB® Script Used for Speech Perception Test Data Conversion

function info = summarisedoguNNNresults(resultsfolder, csvfile)
  
  \texttt{SUMMARISEDOGU}NNNRESULTS Generate summary spreadsheet, one line per participant, for Dogu's N-NN experiment.
  
  \texttt{info = summarisedoguNNNresults(resultsfolder[, csvfile])}
  
  \texttt{arguments:}
  \texttt{\hspace{1cm} resultsfolder - Top level folder, containing folders for each participant.}
  \texttt{\hspace{1cm} csvfile - The name & path of the file to save the summary info in.}
  \texttt{\hspace{1cm} If unspecified or empty defaults to 'summary.csv'.}
  
  \texttt{return values:}
  \texttt{\hspace{1cm} info - A hierarchical structure array. Each element represents a single participant's data, and within each element the 'phases' member stores an array of structures, each containing the summary stats for that particular phase.}
  
  \texttt{see also}
  \texttt{READPHASERESULTSFILE}

if ~exist('resultsfolder')
  dispusage;
  return
end

if ~exist('csvfile') | isempty(csvfile)
  csvfile = 'summary.csv';
end

parpfolders = findsubdirs(resultsfolder);

% the spec for the results file for each interesting phase
% one is probably N, the other NN?
APPENDIX A3 – MATLAB CODE FOR LSSP DATA

phasenames = {'Experiment1*.txt', 'Experiment2*.txt'};
phasenames = {'Experiment1', 'Experiment2'}; % can substitute 'N' 'NN'
     later or make these parameters?

csvheadings = {...
    'Participant ID', ...
    phasenames{1}, [], [], [], [], [], ...
    phasenames{2}, [], [], [], [], [], []; ...
    'C/R', 'F/P', 'hits', 'misses', 'DI', 'mean correct RT'...
    'C/R', 'F/P', 'hits', 'misses', 'DI', 'mean correct RT'...
};

% for each participant
for i=1:length(parpfolders)
    info(i).parpfolder = parpfolders(i).name;
    basepath = fullfile(resultsfolder, info(i).parpfolder);

    % for each interesting phase
    for j=1:length(phasenames)
        tmp = dir(fullfile(basepath, phasenames{j}));
        info(i).phases(j).found = length(tmp);

        if info(i).phases(j).found < 1
            disp(['  Warning: no data file found for phase ' phasenames{j} ' for participant ' num2str(i) ', " info(i).parpfolder "'; skipping']);
            continue
        elseif info(i).phases(j).found > 1
            disp(['  Warning: multiple data files found for phase ' phasenames{j} ' for participant ' num2str(i) ', " info(i).parpfolder "'; only processing the first file, ' tmp{1}.name]);
        end
        tmp = tmp{1};
        info(i).phases(j).fname = fullfile(basepath, tmp.name);

        disp(['Processing data for phase ' phasenames{j} ' for participant ' num2str(i) ', " info(i).parpfolder "...']);

    % get the stuff out of the file, if there's an error don't
    freak out,

    217
APPENDIX A3 - MATLAB Code for LSSP Data

% just move on
try
    [a   b   c   d   e   f] = 
    readphaseresultsfile(info(i).phases(j).fname);
    info(i).phases(j).CRcount = a;
    info(i).phases(j).FPcount = b;
    info(i).phases(j).hitcount = c;
    info(i).phases(j).misscount = d;
    info(i).phases(j).DI = e;
    info(i).phases(j).meanRT = f;
catch
    disp(['  Warning: an error occurred reading data file
          '' info(i).phases(j).fname '' for participant ' num2str(i) ', ''
          info(i).parpfolder '' : ' ']);
    disp(['    ' lasterr]);
end
end

% now generate the csv
bigcsv = {};
ro = 1;

% for each participant
for i=1:length(info)
    col = 1;

    bigcsv(ro, col) = info(i).parpfolder;
    col = col + 1;

    % for each interesting phase
    for j=1:length(info(i).phases)
        if info(i).phases(j).found > 0
            bigcsv(ro, col) = info(i).phases(j).CRcount;
            col = col + 1;
            bigcsv(ro, col) = info(i).phases(j).FPcount;
            col = col + 1;
            bigcsv(ro, col) = info(i).phases(j).hitcount;
            col = col + 1;
            bigcsv(ro, col) = info(i).phases(j).misscount;
            col = col + 1;

218
APPENDIX A3 - MATLAB CODE FOR LSSP DATA

bigcsv{ro, col} = info(i).phases(j).DI;
col = col + 1;
bigcsv{ro, col} = info(i).phases(j).meanRT;
col = col + 1;
else
    % for some reason this guy is lacking some results,
better skip over it so subsequent results are in correct columns
    col = col + 6;
endend
ro = ro + 1;
end

cell2csv(bigcsv, csvheadings, csvfile);

%---------------------------------------------------------------------
---
function dispusage
disp('info = summarisedoguNNNresults(resultsfolder[, csvfile])');