
Bettina Stevanovic (née Keresztesi)
BPsych (Hons)

A thesis submitted for the Degree of Doctor of Philosophy
University of Western Sydney
November 2007
I hereby declare that this submission is my own work and, to the best of my knowledge, it contains no material previously published or written by another person, nor material which has been accepted for the award of any other degree or diploma at the University of Western Sydney, or any other educational institution, except where due acknowledgement is made in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project’s design and conception is acknowledged.

________________________
I dedicate this thesis to my soon-to-be-born child,

who has been ‘kicking’ me to submit this thesis over the last few months.
Acknowledgements

First I would like to thank my supervisor, Associate Professor Kate Stevens for her enduring support throughout my candidature. Thank you Kate for encouraging me to develop my own ideas, and for always being there to give me sound guidance and advice. You have set an example of a dedicated and persistent academic, which I will appreciate long after this thesis has been submitted.

I am also grateful to my co-supervisors, Professor Denis Burnham and Dr Agnes Petocz. Thank you Denis for your feedback and suggestions during the planning stages of my McGurk experiments, and for your advice on planned comparisons. Agnes, I am extremely grateful to you for all your analytical and insightful comments, which were instrumental in refining my arguments during the last few months of my candidature.

I would like to extend special thanks to Professor Catherine Best for her valuable feedback on my chapter concerned with speech-related perceptual learning, to Dr Nigel Nettheim for our stimulating discussions on Shepard tone perception, and to Associate Professor Dave Grayson for his expert advice on statistical analysis.

I would like to acknowledge a number of colleagues and friends who assisted me with the preparation of my speech-related experiments. Thank you to all the people whose speech I recorded: Brooke Adams, Morton Boyer, Tim Byron, Shaun Halovic, Jemma Harris, Justin Joseph, Linda Finger, and Nan Xu. Special thanks to Nicole Lees and Dogu Erdener for their advice on recording and constructing stimuli. My appreciation extends to Colin Schoknecht for his help with recording and for his general support around the
office; and to Arman Abrahamyan for his patient technical assistance, and for all the inspiring conversations.

It is during the hard times when the help of others is most appreciated. Being pregnant during the final stages of my candidature has given me an additional incentive to meet my planned submission date, which would not have been possible without a team of helpful hands. It is hard for me to find words to express my gratitude to Kate Stevens, Agnes Petocz, Mel Gallagher and Clare Howell for prioritizing my PhD during the last couple of weeks. Mel, thank you for your assistance with formatting, and for your general administrative support, which you always deliver with a smile; and thank you Clare for assisting with the reference list. I am also grateful to all my friends - in particular Linda Finger, Jemma Harris, Nicole Lees, Barbara Schwanhaeusser, and Manjusha Vijayan - for providing various kinds of support, such as proofreading, photocopying, car-pooling, listening, understanding, and making me laugh.

Finally I would like to thank all my family for their loving support. Most of all, I would like to express my heartfelt thanks to my husband, Goran. Goran, it is your enduring love, understanding, and encouragement that helped me through the tough times. Thank you for having confidence in me even when I doubted myself. You are a ‘tops’ husband.
Abstract

This thesis is concerned with describing and experimentally investigating the nature of perceptual learning. Ecological psychology defines perceptual learning as a process of educating attention to structural properties of stimuli (i.e., invariants) that specify meaning (i.e., affordances) to the perceiver. Although such definition comprehensively describes the questions of what humans learn to perceive, it does not address the question of how learning occurs. It is proposed in this thesis that the principles of classical and operant conditioning can be used to strengthen and expand the ecological account of perceptual learning. The perceptual learning of affordances is described in terms of learning that a stimulus is associated with another stimulus (classical conditioning), and in terms of learning that interacting with a stimulus is associated with certain consequences (operant conditioning).

Empirical work in this thesis investigated the effect of conditioning on pitch and speech perception. Experiments 1, 2, and 3 were designed to modify pitch perception in Shepard tones via tone-colour associative training. During training, Shepard tones were paired with coloured circles in a way that the colour of the circles could be predicted by either the F0 (pitch) or by an F0-irrelevant auditory invariant. Participants were required to identify the colour of the circles that was associated with the tones and they received corrective feedback. Hypotheses were based on the assumption that F0-relevant/F0-irrelevant conditioning would increase/decrease the accuracy of pitch perception in Shepard tones. Experiment 1 investigated the difference between F0-relevant and F0-irrelevant conditioning in a between-subjects design, and found that pitch perception in the two conditions did not differ. Experiments 2 and 3 investigated the effect of F0-relevant and F0-irrelevant conditioning (respectively) on pitch perception using a within-subjects (pre-test vs. post-test) design. It was found that the accuracy of pitch perception increased after F0-relevant conditioning, and was unaffected by F0-irrelevant conditioning. The differential trends observed in Experiments 2 and 3 suggest that conditioning played some role in influencing pitch perception. However, the question whether the observed trends were due to the facilitatory effect of F0-relevant conditioning or the inhibitory effect of F0-irrelevant conditioning warrants future investigation.
Experiments 4, 5, and 6 were designed to modify the perception of McGurk syllables (i.e., auditory /b/ paired with visual /g/) via consonant-pitch associative training. During training, participants were repeatedly presented with /b/, /d/, and /g/ consonants in falling, flat, and rising pitch contours, respectively. Pitch contour was paired with either the auditory signal (Experiments 4 and 5) or the visual signal (Experiment 6) of the consonant. Participants were required to identify the stop consonants and they received corrective feedback. The perception of McGurk stimuli was tested before and after training by asking participants to identify the stop consonant in each stimulus as /b/ or /d/ or /g/. It was hypothesized that conditioning would increase (1) /b/ responses more in the falling than in the flat/ rising contour conditions, (2) /d/ responses more in the flat than in the falling/ rising contour conditions, and (3) /g/ responses more in the rising than in the falling/flat contour conditions. Support for the hypotheses was obtained in Experiments 5 and 6, but only in one response category (i.e., /b/ and /g/ response categories, respectively). It is suggested that the subtlety of the observed conditioning effect could be enhanced by increasing the salience of pitch contour and by reducing the clarity of auditory/visual invariants that specify consonants.
Table of Contents

Acknowledgements .......................................................................................................... iv
Abstract ............................................................................................................................. vi
List of Figures ................................................................................................................. xiii
List of Tables ................................................................................................................... xv

CHAPTER 1: Introduction ................................................................................................ 1

CHAPTER 2: The Question of ‘What’ in Perceptual Learning: What Do Humans Learn to Perceive in their environments ....................... 8
  2.1 Introduction ........................................................................................................ 9
    2.1.1 The question of Indirect versus Direct Perception: Representationism versus Direct Realism ............................................ 9
    2.1.2 Experimental Investigations of Direct versus Indirect Perception ................................................................. 13
    2.1.3 Interpretations of the Experimental Findings ..................................... 15
  2.2 The Ecological Direct Realist Theory of Perception ....................................... 16
  2.3 Defining Perception as a Relation between the Subject and the Object of Perception .............................................................................................................. 17
    2.3.1 The Subject of Perception: The Perceiver .......................................... 17
    2.3.2 The Object of Perception: Environmental Stimuli.............................. 18
  2.4 What Properties of Stimuli are Perceived? ...................................................... 18
    2.4.1 The Concept of Invariance.................................................................. 18
    2.4.2 The Concept of Affordance ................................................................ 22
  2.5 Perception is Shaped by Learning.................................................................... 24
    2.5.1 The Innate Perceptual Ability to Perceive Invariants is the Result of Evolutionary Learning ................................................ 24
    2.5.2 The Perception of Complex Invariants is a Result of Developmental Learning............................................................ 25
  2.6 Conclusion........................................................................................................ 27

CHAPTER 3: The Question of ‘How’ in Perceptual Learning: How Do Humans Learn to Perceive in their environments ......................... 29
  3.1 Introduction ...................................................................................................... 30
  3.2 The Ecological Theory of Perceptual Learning ............................................... 30
    3.2.1 The Process of Differentiation ............................................................ 30
    3.2.2 The Role of Exploratory Activity in Perceptual Learning ...................... 32
    3.2.3 Limitations of the Ecological Theory of Perceptual Learning .............................................................................................. 33
  3.3 The Application of Conditioning to the Ecological Theory of Perceptual Learning ................................................................................. 34
    3.3.1 Associative Principles are Epistemologically Compatible with the Ecological Theory of Perceptual Learning .......................................................... 34
    3.3.2 Perceptual Learning through Classical Conditioning .......................... 36
    3.3.3 Perceptual Learning through Operant Conditioning ............................ 38
3.4 Neural Processes of Perceptual Learning .......................................................... 40
  3.4.1 Hebbian Synapses ....................................................................................... 40
  3.4.2 Artificial Neural Networks ........................................................................... 42
3.5 Experimental Investigation of Conditioning in Perception Research .............. 43
  3.5.1 Classical Conditioning Experiments ........................................................... 43
  3.5.2 Operant Conditioning Experiments ............................................................ 45
3.6 Conclusion ......................................................................................................... 47

CHAPTER 4: The Role of Conditioning in Pitch-Related Perceptual Learning ...... 49
  4.1 Introduction .................................................................................................... 50
  4.2 Stimuli in Musical Pitch Perception ............................................................... 50
  4.3 Pitch-Related Invariants ............................................................................... 52
    4.3.1 Invariant Frequency Relations in Harmonic Tones ................................ 52
      4.3.1.1 Virtual Pitch ...................................................................................... 53
      4.3.1.2 Sensory Consonance ........................................................................ 53
      4.3.1.3 Octave Equivalence ......................................................................... 55
    4.3.2 Invariant Frequency Relations in Melodies and Harmonies .................. 57
      4.3.2.1 Melodic Expectancies ...................................................................... 59
      4.3.2.2 Harmonic Expectancies .................................................................. 61
  4.4 Pitch-Related Perceptual Learning ................................................................. 62
    4.4.1 Educating Attention to Meaningful Invariants ...................................... 63
    4.4.2 Perceptual Learning Principles ............................................................... 64
      4.4.2.1 Classical Conditioning ...................................................................... 64
      4.4.2.2 Operant Conditioning ...................................................................... 65
  4.5 Conclusion ....................................................................................................... 68

CHAPTER 5: The Role of Conditioning in Speech-Related Perceptual Learning .... 70
  5.1 Introduction .................................................................................................... 71
  5.2 Stimuli in Speech Perception ......................................................................... 72
  5.3 Speech-Related Invariants ............................................................................. 74
    5.3.1 Invariants in Vowel Perception ............................................................... 74
    5.3.2 Invariants in Consonant Perception ......................................................... 75
  5.4 Speech-Related Perceptual Learning ............................................................... 77
    5.4.1 Educating Attention to Meaningful Invariants ...................................... 77
    5.4.2 Learning Principles ............................................................................... 80
      5.4.2.1 Classical Conditioning ...................................................................... 80
      5.4.2.2 Operant Conditioning ...................................................................... 82
  5.5 Perceptual Training Studies ............................................................................ 84
    5.5.1 Speech-Related Perceptual Learning induced by Classical Conditioning .... 84
    5.5.2 Speech-Related Perceptual Learning induced by Operant Conditioning ...... 85
  5.6 Conclusion ....................................................................................................... 88
CHAPTER 6: Experimental Set 1: The Effect of Conditioning on the Perception of Shepard Tones

6.1 Introduction

6.1.1 Previous Research Investigating the Perception of Shepard Tones

6.1.2 General Introduction to the Experiments

6.2 Pilot Experiment

6.2.1 Aim, Design, and Research Question

6.2.2 Method

6.2.2.1 Participants

6.2.2.2 Stimuli

6.2.2.3 Equipment

6.2.2.4 Procedure

6.2.3 Results

6.2.4 Discussion

6.3 Experiment 1

6.3.1 Aim, Design, and Hypotheses

6.3.2 Method

6.3.2.1 Participants

6.3.2.2 Stimuli

6.3.2.3 Equipment

6.3.2.4 Procedure

6.3.3 Results

6.3.4 Discussion

6.4 Experiment 2

6.4.1 Aim, Design, and Hypotheses

6.4.2 Method

6.4.2.1 Participants

6.4.2.2 Stimuli

6.4.2.3 Equipment

6.4.2.4 Procedure

6.4.3 Results

6.4.4 Discussion

6.5 Experiment 3

6.5.1 Aim, Design, Hypotheses

6.5.2 Method

6.5.2.1 Participants

6.5.2.2 Stimuli

6.5.2.3 Equipment

6.5.2.4 Procedure

6.5.3 Results

6.5.4 Discussion

6.6 General Discussion – Experimental Set 1

6.6.1 Summary and Interpretation of Results

6.6.2 Explanation of Results in Terms of the Ecological Theory of Perceptual Learning

6.6.3 Implications of the Results

6.6.3.1 Implications for Shepard tone research

6.6.3.2 Implications for Pitch Perception Research

6.6.3.3 Implications for Perceptual Training Studies
**CHAPTER 7: Experimental Set 2: The Effect of Conditioning on the Perception of McGurk Syllables**

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1 Introduction</td>
<td>136</td>
</tr>
<tr>
<td>7.2 Previous Research Investigating the Perception of McGurk Syllables</td>
<td>136</td>
</tr>
<tr>
<td>7.3 General Introduction to Experiments</td>
<td>139</td>
</tr>
<tr>
<td>7.4 Experiment 4</td>
<td>140</td>
</tr>
<tr>
<td>7.4.1 Aim, Design, and Hypotheses</td>
<td>140</td>
</tr>
<tr>
<td>7.4.2 Method</td>
<td>141</td>
</tr>
<tr>
<td>7.4.2.1 Participants</td>
<td>141</td>
</tr>
<tr>
<td>7.4.2.2 Stimuli</td>
<td>141</td>
</tr>
<tr>
<td>7.4.2.3 Equipment</td>
<td>143</td>
</tr>
<tr>
<td>7.4.2.4 Procedure</td>
<td>143</td>
</tr>
<tr>
<td>7.4.3 Results</td>
<td>145</td>
</tr>
<tr>
<td>7.4.3.1 Pre-Test Analysis</td>
<td>145</td>
</tr>
<tr>
<td>7.4.3.2 Training Effect Analysis</td>
<td>146</td>
</tr>
<tr>
<td>7.4.3.3 Reaction Time Analysis</td>
<td>148</td>
</tr>
<tr>
<td>7.4.4 Discussion</td>
<td>156</td>
</tr>
<tr>
<td>7.5 Experiment 5</td>
<td>157</td>
</tr>
<tr>
<td>7.5.1 Aim, Design, and Hypotheses</td>
<td>157</td>
</tr>
<tr>
<td>7.5.2 Method</td>
<td>159</td>
</tr>
<tr>
<td>7.5.2.1 Participants</td>
<td>159</td>
</tr>
<tr>
<td>7.5.2.2 Stimuli</td>
<td>159</td>
</tr>
<tr>
<td>7.5.2.3 Equipment</td>
<td>161</td>
</tr>
<tr>
<td>7.5.2.4 Procedure</td>
<td>161</td>
</tr>
<tr>
<td>7.5.3 Results</td>
<td>163</td>
</tr>
<tr>
<td>7.5.3.1 Pre-Test Analysis</td>
<td>163</td>
</tr>
<tr>
<td>7.5.3.2 Training Effect Analysis</td>
<td>167</td>
</tr>
<tr>
<td>7.5.4 Discussion</td>
<td>171</td>
</tr>
<tr>
<td>7.6 Experiment 6</td>
<td>172</td>
</tr>
<tr>
<td>7.6.1 Aim, Design, and Hypotheses</td>
<td>172</td>
</tr>
<tr>
<td>7.6.2 Method</td>
<td>173</td>
</tr>
<tr>
<td>7.6.2.1 Participants</td>
<td>173</td>
</tr>
<tr>
<td>7.6.2.2 Stimuli</td>
<td>173</td>
</tr>
<tr>
<td>7.6.2.3 Equipment</td>
<td>175</td>
</tr>
<tr>
<td>7.6.2.4 Procedure</td>
<td>176</td>
</tr>
<tr>
<td>7.6.3 Results</td>
<td>177</td>
</tr>
<tr>
<td>7.6.3.1 Pre-Test Analysis</td>
<td>178</td>
</tr>
<tr>
<td>7.6.3.2 Training Effect Analysis</td>
<td>180</td>
</tr>
<tr>
<td>7.6.4 Discussion</td>
<td>184</td>
</tr>
<tr>
<td>7.7 General Discussion – Experimental Set 2</td>
<td>185</td>
</tr>
<tr>
<td>7.7.1 Summary and Interpretation of Results</td>
<td>185</td>
</tr>
<tr>
<td>7.7.2 Explanation of Results in Terms of the Ecological-Conditioning Theory of Perceptual Learning</td>
<td>187</td>
</tr>
<tr>
<td>7.7.3 Implications of the Results</td>
<td>188</td>
</tr>
<tr>
<td>7.7.3.1 Implications for McGurk Effect Research</td>
<td>188</td>
</tr>
<tr>
<td>7.7.3.2 Implications for Speech Perception Research</td>
<td>189</td>
</tr>
</tbody>
</table>
CHAPTER 8: General Discussion .................................................................................. 190
  8.1 The Proposed Conditioning-Ecological Theory of Perceptual Learning .......... 191
  8.2 Demonstrating the Explanatory Power of the Theory ................................... 192
    8.2.1 The Application of the Theory to Musical Pitch Perception ................. 192
    8.2.2 The Application of the Theory to Speech Perception ....................... 193
  8.3 Investigating the Predictive Power of the Theory ........................................ 194
    8.3.1 Experimental Set 1 ......................................................................... 194
    8.3.2 Experimental Set 2 ......................................................................... 196
  8.4 Implications and Future Directions .......................................................... 198

CHAPTER 9: References ............................................................................................ 202

APPENDIX A: Frequencies and amplitudes of Shepard tones used in Experiments 1, 2, and 3 ................................................................. 222

APPENDIX B: Information sheet used in Experiment 1 ........................................ 224

APPENDIX C: Consent form used in Experiments 1, 2, and 3 ............................. 226

APPENDIX D: Questionnaire used in Experiment 1, 2, and 3 ............................. 228

APPENDIX E: Information sheet used in Experiments 2 and 3 ............................ 230

APPENDIX F: Duration and mean pitch of syllables used in Experiment 4 .......... 232

APPENDIX G: Information sheet used in Experiment 4 ...................................... 234

APPENDIX H: Consent form used in Experiments 4, 5, and 6 .............................. 236

APPENDIX I: Questionnaire used in Experiments 4, 5, and 6 ............................ 238

APPENDIX J: Duration and mean pitch of disyllables used in Experiments 5 and 6 ..... 240

APPENDIX K: Information sheet used in Experiment 5 ...................................... 242

APPENDIX L: Information sheet used in Experiment 6 ...................................... 244
List of Figures

Figure 6.1  Percentages of ascending pitch contour judgements in Shepard tones..... 97
Figure 6.2  Idealized data illustrating the expected trends in the percentages of ascending judgements after F0-relevant and F0-irrelevant training ................................................................. 101
Figure 6.3  Mean percentage of ascending pitch contour judgments in Shepard tritones after F0-relevant and F0-irrelevant training ...................... 106
Figure 6.4  Percentages of ascending pitch contour judgments in sinusoidal tritones after F0-relevant and F0-irrelevant training ............................ 108
Figure 6.5  Percentage of ascending pitch contour judgments for all Shepard tritones before and after training ......................................................... 117
Figure 6.6  Percentage of ascending pitch contour judgments for all sinusoidal tritones before and after training .............................................. 119
Figure 6.7  Percentage of ascending pitch contour judgments for all Shepard tritones before and after training ......................................................... 126
Figure 6.8  Percentage of ascending pitch contour judgments for all sinusoidal tritones before and after training .............................................. 128
Figure 7.1  Number of /ba/, /da/, and /ga/ responses for McGurk stimuli in the pre-test phase ................................................................. 145
Figure 7.2  The number of /ba/ responses as a function of time and pitch contour .................................................................................................... 146
Figure 7.3  The number of /da/ responses as a function of time and pitch contour .................................................................................................... 147
Figure 7.4  The number of /ga/ responses as a function of time and pitch contour .................................................................................................... 148
Figure 7.5  Mean reaction times for correctly identified training and catch-trial stimuli ............................................................................................. 150
Figure 7.6  Mean reaction times for McGurk stimuli in the /ba/ response category ................................................................................................. 151
Figure 7.7  Mean reaction times for McGurk stimuli in the /da/ response category ................................................................................................. 152
Figure 7.8  Mean reaction times for McGurk stimuli in the /ga/ response category ................................................................................................. 153
Figure 7.9  Mean reaction times for congruent /ba/ audio-visual syllables .............. 154
Figure 7.10 Mean reaction times for congruent /da/ audio-visual syllables .............. 155
Figure 7.11 Mean reaction times for congruent /ga/ audio-visual syllables .............. 155
Figure 7.12 Percentages of /b/ responses before training as a function of pitch contour and vowel context ................................................................. 163
Figure 7.13 Percentages of /d/ responses before training as a function of pitch contour and vowel context ................................................................. 164
Figure 7.14 Percentages of /g/ responses before training as a function of pitch contour and vowel context ................................................................. 165
Figure 7.15 Difference scores for the percentages of /b/ responses as a function of pitch contour and vowel context ...................................................... 167
Figure 7.16 Difference scores for the percentages of /d/ responses as a function of pitch contour and vowel context ...................................................... 168
Figure 7.17 Difference scores for the percentages of /g/ responses as a function of pitch contour and vowel context ...................................................... 169
Figure 7.18 Percentages of /b/ responses before training as a function of pitch contour and vowel context ................................................................. 177
Figure 7.19 Percentages of /d/ responses before training as a function of pitch contour and vowel context ................................................................. 178
Figure 7.20 Percentages of /g/ responses before training as a function of pitch contour and vowel context ................................................................. 179
Figure 7.21 Difference scores for the percentages of /b/ responses as a function of pitch contour and vowel context ...................................................... 180
Figure 7.22 Difference scores for the percentages of /d/ responses as a function of pitch contour and vowel context ...................................................... 181
Figure 7.23 Difference scores for the percentages of /g/ responses as a function of pitch contour and vowel context ...................................................... 182
List of Tables

Table 6.1  Chi-Square Statistics for Pitch Contour Judgements in Shepard Tritones ................................................................. 98
Table 6.2  The F0 contour of all possible Shepard tone-pairs excluding unisons........ 102
Table 6.3  Direction of the smallest frequency change in partials for all possible pairs of Shepard tones excluding unisons.......................... 104
Table 6.4  Mann-Whitney Statistics for the Percentage of Ascending Judgments in Shepard Tritones .................................................................................. 107
Table 6.5  Mann-Whitney Statistics for the Percentage of Ascending Judgments in Sinusoidal Tritones .................................................. 109
Table 6.6  Wilcoxon Statistics for the Percentage of Ascending Judgements in Shepard Tritones .................................................................................. 118
Table 6.7  Wilcoxon Statistics for the Percentage of Ascending Judgements in Sinusoidal Tritones .................................................. 120
Table 6.8  Direction of the frequency change in the loudest partial for all possible pairs of Shepard tones excluding unisons...................... 121
Table 6.9  Wilcoxon Statistics for the Percentage of Ascending Judgements in Shepard Tritones .................................................................................. 128
Table 6.10 Wilcoxon Statistics for the Percentage of Ascending Judgements in Sinusoidal Tritones .................................................. 129
Table 7.1  Planned comparison statistics for /b/ responses .................................. 167
Table 7.2  Planned comparison statistics for /d/ responses .................................. 168
Table 7.3  Planned comparison statistics for /g/ responses .................................. 169
Table 7.4  Planned comparison statistics for /b/ responses .................................. 180
Table 7.5  Planned comparison statistics for /d/ responses .................................. 181
Table 7.6  Planned comparison statistics for /g/ responses .................................. 182
CHAPTER 1

INTRODUCTION
The perception of an environmental stimulus is influenced by the perceiver’s prior experience with that stimulus. In general, the more humans encounter objects and events in the environment the faster and more attuned their perception becomes. The process by which experience changes the way humans perceive environmental properties is commonly referred to as perceptual learning. Much evidence for perceptual learning comes from developmental studies conducted in various domains, such as face perception (Young, 1992) and speech perception (Werker & Tees, 1984). Although the plasticity of human perceptual systems tends to reduce with age (Johnson, 2005), perceptual learning continues in adulthood, and can be experimentally demonstrated via perceptual training studies (Goldstone, 1998). The general conclusion of both developmental and perceptual training studies is that human perception is shaped by learning.

The primary question regarding perceptual learning is the question of what humans learn to perceive as a result of experience (Gibson & Gibson, 1955b). This question has been thoroughly addressed by J. J. Gibson’s ecological theory of perception (Gibson, 1986). According to Gibson, people learn to perceive invariants and affordances in the environment. The concept of invariance refers to structural properties of stimuli that remain constant (i.e., invariant) while other properties of the stimulus change, and the concept of affordance refers to invariants that specify meaning to the perceiver. Thus, while invariants can objectively be described in terms of relations among stimulus variables, affordances can only be described in terms of relating the perceived stimulus property to the perceiver. The ecological concepts of invariance and affordance imply that the question of what people learn to perceive can only be completely answered by considering both the environment and the perceiver.

While ecological theories of perception comprehensively answer the question of what humans perceive, they only vaguely address the question of how perceptual learning occurs. According to Gibson and Gibson (1955a), perceptual learning is underlined by the mechanism of differentiation, which they initially defined as a process of “responding to variables of physical stimulation not previously responded to” (p. 34). The Gibsons’ proposal of the differentiation theory was followed by a detailed criticism by Postman (1955), who argued that it can only be maintained at the expense
of begging the question of learning. According to Postman, “the fact that the organism has learned to discriminate more qualities is the very fact that we need to explain” and “improvement in discrimination cannot be invoked to explain improvement in discrimination” (Postman, 1955, p.144).

In response to Postman’s criticism, Gibson and Gibson (1955b) noted that the theory of differentiation “is concerned with the question of what is learned in perceptual learning, not how it occurs, or at least not as yet” and it “is not a theory but only the promise of a theory, and its explanatory value remains to be seen” (p.447). Although the Gibsons admitted that the theory of differentiation is limited and could possibly be strengthened, it has not undergone any major theoretical development since the middle of the 20th century. A more recent definition of differentiation by Gibson and Pick (2000), similar to the previously noted definition, assumes rather than accounts for learning by referring to differentiation as a process of “narrowing down from a vast manifold of information to the minimal, optimal information that specifies the affordances of an event, object, or layout” (p. 150).

As an alternative to the differentiation theory, Postman (1955) defined perceptual learning as a change in perceptual responses (e.g., identification and discrimination) to environmental stimuli. Postman argued that the reformulation of perceptual learning in terms of stimulus-response associations provides a useful framework to explain the mechanisms of perceptual learning; however he did not elaborate on the conditions that change perceptual responses to stimuli. In fact, he noted that “The problem of mediating mechanisms, however, remains” and suggested that “One may, to be sure, rest content with the psychological associationism of stimulus and response, and pursue the specifications of the conditions of perceptual learning within this framework” (Postman, 1955, p. 442). Following the advice of Postman, this thesis is concerned with specifying the conditions of perceptual learning within an associative framework.

Associative principles have commonly been used to describe learning in various aspects of human behaviour (Domjan, 2005); however they have rarely been applied to perceptual learning (Davies, 1987). Most importantly, the rare applications of associative principles to perceptual learning (e.g., Davies, 1987; Hall, 1991; Razran,
1955) have not been conceptually related to the ecological theory of perception. The reluctance to relate associative learning principles to the ecological theory of perceptual learning is partly due to the misconception that associative principles imply *representationism* and are therefore incompatible with the epistemological view of ecological psychology (i.e., direct *realism*). As argued by Gibson and Pick (2000), “Perceptual learning is not properly described as association or as an addition of any kind, as a response to a stimulus, or as a ‘representation’ of ‘input’”. (p. 149).

It is argued in this thesis that the experimental analysis of stimuli and responses does not necessarily entail representationism and hence it is not incompatible with the epistemology of ecological perception. It is proposed that the associative principles of *classical* and *operant conditioning* can be used to strengthen and expand the ecological theory of perceptual learning. According to the principles of classical conditioning, humans learn to associate co-occurring stimuli in the environment (Rescorla, 1988). According to the principles of operant conditioning, humans learn to associate an environmental stimulus with the consequences of responding to that stimulus (Skinner, 1976). It is proposed in this thesis that both classical and operant conditioning can be interpreted as learning about the *affordances* of environmental stimuli.

One of the main advantages of conditioning principles is that they describe generic processes that can be readily applied across various domains (Rescorla, 1988). In fact, conditioning has been applied to explain a wide range of behavioural phenomena and has been successfully used to induce behavioural changes in clinical settings (Domjan, 2005). Since conditioning principles are considered as general principles of learning, their application to the ecological theory of perceptual learning increases its explanatory power. In order to demonstrate the explanatory power of the proposed ecological-conditioning theory of perceptual learning, the role of conditioning in perceptual learning has been reviewed in two perceptual domains, namely *musical pitch perception* and *speech perception*. These two auditory perceptual domains were chosen, since the ecological theory of perceptual learning is not commonly applied to auditory perception (but see Clarke 2005, Neuhoff, 2004). As noted by Clarke (2005), “Perhaps because of his own progressive deafness, Gibson developed his ideas much less in relation to the auditory perception than vision” (p. 5-6). Similarly,
Schmuckler (2004) observed that “although the ecological approach has been developed as a general theoretical framework, the primary applications of this theory have been to vision” (p. 271). Thus, by examining the perceptual learning of affordances in pitch and speech perception, the present thesis makes some contribution to theories of pitch and speech perception and strengthens the general theoretical framework of ecological perception.

Empirical work in this thesis consists of perceptual training experiments that were designed to test the effect of conditioning on pitch and speech perception. Although, the effect of conditioning on perception has previously been experimentally investigated (e.g., Davies, 1974a, 1974b, 1976; Howell, 1941; Morrot, Brochet, Dubourdieu, 2001), the number of studies that have been designed to condition perception is relatively sparse. As has recently been pointed out, “Future research will be required to better determine the relationship between conditioning and perceptual learning” (Seitz & Watanabe, 2005, p. 332). Thus experiments reported here contribute to understanding the role of conditioning in perceptual learning.

Stimuli used in the experiments involve ambiguous pitch and speech stimuli (i.e., Shepard tones and McGurk syllables, respectively). Ambiguous stimuli were used in the experiments because they generate less stable percepts than non-ambiguous stimuli, hence conditioning the perception of ambiguous stimuli presumably requires fewer training trials than conditioning the perception of non-ambiguous stimuli. Previous research involving Shepard tones and McGurk syllables focused on the perception of these stimuli per se, without investigating the extent to which the perception of these stimuli can be changed by perceptual training. Thus, the experimental work reported here not only contributes to the general understanding of the role of conditioning in perceptual learning, but also expands on previous research investigating the perception of Shepard tones and McGurk syllables.

Experiments 1, 2, 3 (i.e., Experimental Set 1) were conducted to modify the perception of Shepard tones using colour-tone associative training. During training, the presentation of Shepard tone-pairs was paired with the presentation of coloured circles (i.e., classical conditioning). The colour-tone associations were arranged in a way such that the colour of the circles could be predicted by a certain invariant
property of the Shepard tone-pair. It was hypothesized that participants’ attention would be directed to the invariant property of Shepard tone pairs that are relevant in the colour-tone associative arrangements. In order to further increase attention to the task relevant auditory dimension, participants received corrective feedback after their response (i.e., operant conditioning). It was hypothesized that training would improve participants’ perception of the invariant properties of Shepard tone pairs that predicted the colour of circles.

Experiments 4, 5, and 6 (i.e., Experimental Set 2) were designed to modify the perception of the McGurk syllables using pitch-consonant associative training. McGurk stimuli used in the experiment were constructed by pairing the visual articulation of /g/ consonant with the auditory signal of /b/ consonants presented in three different pitch contours (i.e., falling, flat, and rising). During training, participants were repeatedly presented with /b/, /d/, and /g/ consonants in falling, flat, and rising pitch contour contexts, respectively (i.e., classical conditioning). Pitch contour was paired with either the auditory signal (Experiments 4 and 5) or visual displays of articulation (Experiment 6). During training, participants were required to identify the stop consonants and they received corrective feedback after their response (i.e., operant conditioning). Thus participants were conditioned to learn that falling pitch contour affords consonant /b/, flat pitch contour affords consonant /d/, and rising pitch contour affords consonant /g/. The experiments were based on a within-subjects design, in which McGurk stimuli were presented in three different pitch contours (i.e., falling, flat, rising) and perception was tested before and after training. The perception of McGurk stimuli was tested by asking participants to identify stop consonants in each stimulus as /b/ or /d/ or /g/. Hypotheses were based on the assumption that the change in the perception of McGurk stimuli as a result of training would be influenced by pitch contour.

The thesis begins with reviewing the ecological perspective on the question of what humans perceive (Chapter 2) and then moves on to discussing the question of how perceptual learning occurs (Chapter 3). In Chapter 3, it is argued that the ecological theory of perceptual learning is limited to describing how perceptual learning occurs, and the limitations of the theory are addressed by the application of classical and operant conditioning principles. Chapters 4 and 5 focus on demonstrating the
explanatory power of the proposed ecological-conditioning theory of perceptual learning in the domains of musical pitch and speech perception, respectively. Chapters 6 and 7 review the experimental work that was conducted as part of this thesis to investigate the effect of conditioning on the perception of Shepard tones and McGurk syllables, respectively. In Chapter 8, the general experimental findings and their implications are evaluated in the light of the proposed ecological-conditioning theory of perceptual learning.
CHAPTER 2

THE QUESTION OF ‘WHAT’ IN PERCEPTUAL LEARNING: WHAT DO HUMANS LEARN TO PERCEIVE IN THEIR ENVIRONMENT?
2.1 Introduction

Answering the question what do humans perceive seems straightforward. Commonsense tells us that we perceive what is “out there” in our environment. For instance, we see, hear, touch, taste, and smell objects that surround us. However, a controversial issue in perception has been the question of how the objects in the environment “get into our mind” and/or in what sense, if any, they get inside our head. With regard to this question, there have been two basic positions: theories of indirect perception (or representationism) and theories of direct perception (or direct realism). A related controversial issue has been the question whether the choice between indirect and direct theories of perception can be determined via empirical investigation, or whether it is a matter of conceptual or logical analysis. Proponents of indirect perception have claimed experimental confirmation of their theory. In contrast, proponents of direct perception rest their case on conceptual grounds. It will be argued here that direct perception theorists are correct that the issue is a conceptual one. However, they have failed to address the experimental evidence sufficiently to persuade their opponents. In particular, they have failed to spell out the misunderstandings regarding the differences between the two theories and their implications. They have also failed to show how the supposed experimental confirmations of indirect perception are based on these misunderstandings.

2.1.1 The question of Indirect versus Direct Perception: Representationism versus Direct Realism

Indirect theories of perception propose a representationist epistemology, which is the view that humans get to know the world via internal mental images or representations. According to this view, we do not directly perceive the physical properties of objects in the world. Instead, our perception of environmental stimuli is mediated by mental states (sensations, images, ideas or some other kinds of token) and by mental processes (e.g., mental inferences or heuristics) operating on those internal states. We infer what is “out there” from our mental operations on what is “in here”. Therefore, the answer to the question what do we perceive is: internal representations of the external world. Helmholtz’s (1867, 1878) theory of unconscious inferences is a classic example of an indirect perceptual theory. According to Helmholtz, mental inference is based on a rule formed by earlier observations (major premise) being applied to sense impressions (minor
Indirect perception is the core of the major information-processing paradigm of cognitive psychology, and is widely accepted within psychology generally.

In contrast, direct realism is the epistemological corollary of the ontological thesis of realism. Realism is the thesis that the existence of reality is independent of its being perceived. According to the direct realist theory of perception, when we observe the objects in the environment, we observe them directly, without having to observe or know some mediating internal token or representation of them. That is, our access to objective reality is direct in the sense of being epistemologically unmediated. Therefore, the answer to the question what do we perceive is: the external world directly. The most influential theory of direct perception is found in the ecological approach of J. J. Gibson (1960, 1972, 1975, 1986). Gibson argues that the environment is the direct source of information pertaining to enduring features (invariants) and the meanings or functions of those invariants (affordances). Gibson’s supporters (e.g., Michaels & Carello, 1981) follow him in arguing that their choice of direct over indirect theories of perception is predominantly based on conceptual arguments.

These arguments have long been known and occur repeatedly in the literature (e.g., Armstrong, 1961; Bickhard, 1996; Hamlyn, 1971; Heil, 1981; Maze, 1991). The main point is that representationism cuts the ground from under its own feet. To begin with, in order to answer the question how the internal representation can itself be accessed or perceived, the representationist is forced to answer either that it is perceived directly, or that it is perceived indirectly. If the answer is that it is perceived directly, then the critic is justified in asking why, if representations can be perceived directly, cannot other things (objects in the world) be perceived directly. The critic is also justified in asking how the representation can be perceived directly, given that there are no perceptual organs in the brain. The theory is then forced to postulate some kind of internal perceiver or scanner (an homunculus or “little man”) in the brain. If, on the other hand, the answer is that the representation is perceived indirectly, then the homunculus must himself contain an internal scanner to perceive the representation of the representation. This leads to an infinite regress of homunculi. Moreover, this regress is a vicious regress, because the process of perceiving that is supposed to be explained by it never gets to be explained; instead, it is presumed at every step. A further problem is that of accounting for the connection between the representation and the external object that it supposedly represents.
This has been referred to as the “semantic problem” for representations (Fodor, 1985), and it is widely held to be as yet unsolved. However, once again, what is explicitly denied is being implicitly assumed, for the information that is supposedly given in the representation “would not be intelligible unless we already knew what it was to perceive a public world” (Hamlyn, 1971, p. 162). That is, in order to be able to infer the external object from the internal representation, we must have had prior access to the external object. This assumption is also needed to account for error. Ironically, all representationist explanations take as their starting point the fact of error. For example, when a straight stick partially immersed in water looks like a bent stick, there is no bent stick in water actually out there in the world; therefore, the perceiver must be looking at an internal mental image (of a bent stick). However, in order to mistake a straight stick for a bent stick, the perceiver must have had prior direct perceptual experience of a bent stick. Error is dependent on prior experience of veridical perception. The process of checking whether the representation matches the object in the external world requires access to both. Without that external access, and without the possibility of comparison, the fact of error could never have been conceived. In sum, representationism is both circular and self-contradictory. Furthermore, in presupposing the logically prior direct realism that it explicitly denies, it renders itself redundant.

Although direct realism is the only alternative to escape from the logical problems that are generated by representationism, direct theories of perception are often criticised for neglecting the process or mechanisms of perception, and for being unable to account for errors in perception (Fodor & Pylyshyn, 1981; Gyr, 1972; Hecht, 2000). However, such criticisms stem from three misconceptions: (1) that direct perception implies accurate or errorless perception; (2) that direct perception does not allow for inferences; and (3) that direct perception, in rejecting epistemological mediation, is thereby rejecting physiological mediation. The first two assumptions can be illustrated by Fodor and Pylyshyn’s (1981) criticism of Gibson’s theory of direct perception.

Probably the line that Gibson wants to take is if an affordance is correctly perceived, then it is perceived directly; and that is, of course compatible with the factivity of “directly perceive”. Notice, however, that such an approach does not help with the problem of misperception since it does not tell us how we are to describe the cases where the antecedent of the hypothetical is false. We will return
to this sort of difficulty. Suffice it at present to say that the problem of constraining “directly perceive” so as to provide a nonvacuous construal of the claim that perception is noninferential, and the problem of providing a coherent account of misperception without recourse to the notion of perceptual inference, are two sides of the same coin (p. 154)

Although the assumptions that direct perception implies veridical perception and that it does not allow for inferences are commonly held, it is important to realize that they are based on misunderstanding the notion of direct perception.

Firstly, “direct” in direct perception does not mean “correct”. It means that the perceiver is in direct epistemological contact with the external environment rather than with the mental representations of the environment. This theory of direct perception does not imply that the external properties of the environment are always correctly perceived. Perceptual errors can result from various factors (e.g., conflicting environmental cues, imperfections in the perceptual system, or aspects of the perceiver’s motivation) which do not necessarily imply the use of mental representations (i.e., indirect perception) (Rantzen, 1993).

Secondly, “direct” when used synonymously with “non-inferential” means that there is no need for the perceiver to infer the external environmental properties from the internal mental representation, since there is a direct epistemological contact between the environment and the perceiver. However, it is a mistake to conclude that direct realism does not leave room for perceptual inferences. As argued by White (1971), direct perception allows for inferences as long as they are based on the physical properties of the environment instead of internal mental entities.

Now, traditional accounts of perception often used the notion of inference from sensations, sense-data, etc., to the existence of material objects. One weakness of these accounts, however, was their assumption that such an inference was from the existence of observed objects of one sort, e.g., sense-data or sensations, to the existence of unobserved objects of a second sort, e.g., material objects, which were the alleged cause of the observation of the objects of the first sort. Hence, critics of the traditional view … have wrongly denied the place of inference in perception because they have rightly wished to deny that there are objects called sense-data or
sensations ... As I have mentioned, explanation - and, hence, explanation in inference - may be of the given data in terms, not of prior causal data, but of that of which the data are manifestations or exemplifications. (p.295)

In other words, inferences are recognised within direct realism. Indeed, any number of perceptual inferences may be made in direct perception provided only that they are inferences from directly observed external phenomena.

Finally, it is important to distinguish epistemological from physiological mediation. Direct realism rejects only epistemological mediation. It accepts physiological mediation. That is, it accepts that objects are perceived via a complex process of physiological mediation which includes the impact of energies on the sensory organs of the perceptual system, followed by neurophysiological processes in the nervous system, and so on. What is rejected is either that these mediating physiological processes are themselves perceived or that they give rise to mental entities (e.g., sensations) that then become the primary objects of our perception.

These misconceptions are most apparent in the attempts to empirically test between direct and indirect theories of perception, and in the conclusions which are drawn from such research.

2.1.2 Experimental Investigations of Direct versus Indirect Perception

The area of research which has attempted to empirically test theories of direct perception against theories of indirect perception concerns the visual perception of dynamic properties (Hecht, 1996, 2000; Runeson, Juslin, & Olsson, 2000). Dynamic properties of events generally refer to the forces acting on the movement of objects (e.g., mass) (Cohen, 2006). The ability to identify dynamic properties is investigated in point-light demonstrations of human movement that are created by filming people in action wearing small points of light on the main joints of their body. People are filmed in semi-darkness resulting in a collection of moving points on a dark background. This technique removes all information except the kinematic variables (e.g., angles and velocities). Studies using the point-light technique (e.g., Johansson, 1973; Runeson & Frykhol, 1981, 1983) have demonstrated that
kinematic variables allow people to perceive dynamic properties of the event, such as the mass of an object handled by the person (e.g., the weight of a box being picked up).

The perception of dynamic properties has also been systematically investigated in experiments that involve the event of two balls colliding (e.g., Cohen, 2006; Gilden & Profitt, 1989; Todd & Warren, 1982). In colliding-ball experiments, participants view two simulated balls colliding at different speeds and angles and they are asked to judge which one of them is the heavier or to make quantitative estimates of the mass ratios. Colliding-ball experiments provide a unique opportunity to investigate the nature of dynamic perception because the kinematic variables (i.e., angles and velocities) can be precisely described and experimentally manipulated (Gilden & Proffitt, 1989).

The collision of two balls is a multidimensional event, which involves kinematic variables varying simultaneously on the dimension of angles and the dimension of velocities. The mass-ratio of colliding balls is specified by the relative amount of velocity change (Jacobs, Runeson, & Michaels, 2001). This is expressed by the equation $\frac{m_B}{m_A} = \frac{|v_A - u_A|}{|v_B - u_B|}$, where $m_A$ and $m_B$ are the masses of the two balls, $u_A$ and $u_B$ are the velocities of the balls before impact, and $v_A$ and $v_B$ are the velocities of the balls after impact. According to Jacobs et al., the relative amount of velocity change is an invariant that specifies the mass-ratio of colliding balls.

The first experiments investigating the perception of mass-ratios of colliding balls were conducted by Todd and Warren (1982). They found that observers’ mass-ratio judgments are accurate in situations where the heavier object hits the lighter but the accuracy is reduced in situations where the lighter object hits the heavier object. Since Todd and Warren found perceptual errors in mass-ratio judgments, they concluded that participants do not always directly perceive the invariants that specify mass-ratios.

In order to further investigate the perceptual information used in mass-ratio judgments Gilden and Proffitt (1989) conducted two sets of experiments. In the first set, they found that when a lighter ball impinged on a heavier ball, participants were more accurate when they could see the incoming ball. However, when the heavier ball impinged on the lighter ball perceptual judgements were more accurate when the incoming ball was occluded. Since occluding the incoming ball reduces the kinematic information that is necessary to
perceive invariants specifying mass-ratios, Gilden and Proffitt suggested that perceptual judgments were based on mental inferences applied to elemental kinematic variables.

Gilden and Proffitt’s (1989) second set of experiments was designed to tease out the elemental kinematic variables and the mental inferences that participants use during the perception of mass-ratios. The salience of angle and vector information in the colliding events was systematically varied. The results of the experiment suggested that mass-ratio judgments are based on two simple heuristics (or mental rules) applied to velocity and angle information. The ball with the greater velocity in the post-collision is generally judged to be lighter than the other ball, and the ball that scatters at the greater angle is generally perceived to be lighter than the other ball. Gilden and Profitt pointed out that both the angle and the velocity heuristic correspond to general laws in nature, and that humans presumably learn them from everyday experience.

2.1.3 Interpretations of the Experimental Findings

These experiments investigating the perception of dynamic properties suggested that people do not always perceive invariant kinematic information that specifies the mass-ratios of colliding balls. Gilden and Profitt (1989) argued that such findings falsify the theory of direct perception and support heuristic models. In contrast, according to Hecht (1996), both direct and indirect models of perception are “immune to falsification” and “it is impossible to distinguish judgments that are based on heuristics from judgments that are based on (incomplete) invariants” (Hecht, 1996, p 68-69).

However, the experimental findings do not falsify the theory of direct perception. The fact that people may not perceive whatever invariants specify the to-be-perceived environmental property is not inconsistent with the basic tenet of direct perception (Withagen, 2004). Direct perception implies only that the perceiver is in direct epistemic contact with his/her environment, and that whatever aspects of any stimulus variables are perceived are done so without the intervention of mental representations. Hence, indirect perception is not implied by the perception of stimulus variables that only partially specify the to-be-perceived property, or by the failure to perceive those aspects which specify the information completely, or by the failure to make the correct inferences from the observed aspects of the stimulus variables.
Therefore, Hecht’s (1996) conclusion is correct, but the reason for it is not. The experimental results have no bearing on the question of direct versus indirect perception. However, that is not because of the impossibility of distinguishing between heuristics and incomplete inferences. Both heuristics and incomplete inferences are included in direct perception. In direct perception theory, the point-lights and movements are directly perceived, and experimental participants are required to make inferences from those observed partial stimuli to the size and weight of the objects. In indirect perception theory, the point-lights and movements are not directly perceived; instead, internal mental representations of those stimuli are directly perceived. Hence, the inferences which are made are inferences from those internal representations of point-lights and movements to the external point-lights and movements. The additional inferences which are then made to the size and weight of the objects are irrelevant to the epistemological question of how (viz. directly or indirectly) the point-lights and movements are themselves perceived.

Since experimental investigations cannot determine between direct and indirect perception, the issue must be decided on logical grounds. Conceptual analysis exposes the fatal logical flaws in theories of indirect perception, and reveals that such theories, in any case, implicitly assume direct perception. This leaves the theory of direct perception as the only coherent and viable theory of perception. However, none of the important features supposedly accommodated only by indirect perception is missing from direct perception. The theory of direct perception includes the processes of physiological mediation, accommodates perceptual inferences and mental heuristics, and accounts for the phenomenon of perceptual error in terms of breakdowns in the system or lack of clarity in the environment. Since the ecological direct realist theory of perception is the theory that is adopted in this thesis, it needs to be expounded in further detail.

2.2 The Ecological Direct Realist Theory of Perception

A theory of perception must address two of the major features of perception. The first is that humans perceive objects as relatively stable in the environment. The second is that perception is typically accompanied by an awareness of what the perceived object means to the perceiver. Perceiving object constancy is often considered the objective aspect of perception and perceiving the meaning of objects is regarded as the subjective aspect of
perception. Describing the objective and subjective aspects of perception is one of the greatest challenges in psychology. This challenge was taken up by Gibson, who borrowed a term from mathematics (i.e., invariance) and invented a new term (i.e., affordance) to develop his ecological theory of perceiving object constancies and meaning respectively.

2.3 Defining Perception as a Relation between the Subject and the Object of Perception

The direct realist approach to perception defines perception as a relation between an organism (i.e., the subject of perception) and an environmental stimulus (i.e., the object of perception) (Michaels & Carello, 1981). An essential point about any relation is that the existence of each term (i.e., subject and object of perception) is independent of the relation (i.e., perception) itself and therefore that they can be described separately. Gibson’s ecological theory of direct perception provides the following account of the nature of the subject and object terms in the perceiving relation.

2.3.1 The Subject of Perception: The Perceiver

Perception is impossible without a perceiver who is equipped with a perceptual system that is able to detect stimuli in the environment. Perception is underpinned by physiological changes in the sensory organs and the central nervous system as an effect of stimulus energies which impinge on the sense receptors. Each sensory organ contains receptors that convert the stimulus energies into electrical signals, which are then conveyed to the brain where they activate different cortical regions. Although physiological changes in the perceptual system cannot literally be observed by the human eye (i.e., we cannot see brain activities), they can be observed and measured by various neuroimaging techniques (e.g., PET, fMRI) that enable scientists to describe how the perceptual system is altered by various stimuli.

In addition to registering perceptual stimuli, perceivers are actively searching for meaningful environmental information, and they bring to that search their prior experiences. As Woodworth (1947) noted, “perception is always driven by a direct, inherent motive which might be called the will to perceive” (p. 123). According to Woodworth the will to perceive is always present in the use of senses. For instance, one cannot look and listen without trying to see and hear. Thus perception is affected not only by the stimuli but also by the perceiver’s prior experience and motivation to perceive.
Since perception involves both the environment and the perceiver’s motivation, a full account of what (and how) humans perceive should consider both the environment and the perceiver. This argument has been extended in ecological theories of perception that argue that perceptual information includes two aspects: information about the environment and information for the perceiver (Michaels & Carello, 1981).

### 2.3.2 The Object of Perception: Environmental Stimuli

In theories of perception generally, the object of perception is often referred to by the term stimulus, which has two general applications. On the one hand, the term stimulus is used to refer to objects and events in the environment, and on the other hand it is applied to the physical energies (e.g., light radiation, pressure waves, chemical features, and mechanical stimulation) that allow humans to perceive the objects and events in the environment. The dual application of the term stimulus was first made explicit by Koffka (1935), who formally distinguished between the two aspects. Koffka used the label distal stimuli to refer to objects and events, and the label proximal stimuli to refer to the energy patterns stimulating the sensory organs.

Although the distinction between the distal and the proximal aspects of the stimulus is useful as it highlights the fact that perception is physiologically mediated, it should be emphasized that both aspects play an equally important role in perception. Without distal stimuli there are no proximal stimuli and without proximal stimuli the distal stimulus cannot be perceived. In other words, it is argued here that perception involves both distal and proximal stimuli, and referring to one without the other only partially describes perception. Such argument is consistent with Gibson’s (1960) view that “every family of proximal stimuli arising from one object is, in a sense, one stimulus” (p. 697).

### 2.4 What Properties of Stimuli are Perceived?

#### 2.4.1 The Concept of Invariance

One of the primary functions of perception is to allow people to identify objects and events in the environment even though their point of perspective may change. For example, humans can generally identify a rectangular table top as having a rectangular shape from different perspectives. Similarly, humans can identify a certain melody even if it is played
in different musical keys or on different instruments. In psychology, the phenomenon of stimuli giving rise to the same perceptual response under different conditions of observation has been referred to as perceptual constancy (Walsh & Kulikowski, 1998). Research into perception has identified various environmental properties that humans tend to perceive in constancy that include shape (Day & McKenzie, 1973; Pizlo, 1994), size (Brunswick, 1939; Dukes, 1951), and colour (Brainard, Wandell, & Chichilinsky, 1993). Although, the experimental investigation of perceptual constancies is predominantly based on visual stimuli, perceptual constancies also exist in non-visual domains. Examples of perceptual constancies in pitch and speech perception will be discussed in Chapters 4 and 5, respectively.

The ability to identify objects and events regardless of change in perspective suggests that humans are able to perceive intrinsic structural properties of objects and events that remain constant or invariant irrespective of certain transformations in the viewing conditions. The structural properties of stimuli are usually described in terms of relations among stimulus variables. For example, the structural property of a rectangular table top consists of set relations among the four angles, and the structural property of a melody can be described in terms of the relations between individual notes. Since structure is based on interrelations between elements, the perception of constant structure (e.g., shape, melody) is based on the perception of relations among stimulus variables. The argument that humans perceive elemental stimulus variables in a relational manner (i.e., that they perceive stimuli as a structured whole) has been acknowledged by various influential theories of perception (e.g., Gibson, 1986; Kohler, 1967).

Gibson’s (1986) ecological theory of perception proposes that the structural properties of environmental stimuli are reflected by the physical energy patterns that impinge on the sensory organs (e.g., light stimulating the eye). The question how sensory stimulation reflects structural regularities in the environment is addressed by ecological physics – a discipline that describes how stimulus energy gets structured by environmental stimuli. For example, radiant light from the source (e.g., sun) becomes structured as it is reflected from the surfaces of objects. The structure of reflected light depends on various surface features, such as texture, size, and shape (Michaels & Carello, 1981). For instance, the way paper and coal reflects light greatly differs. While paper reflects more light than it absorbs, coal absorbs more light than it reflects. Similar to texture, other properties of
stimuli, such as size and shape contribute to how the energy is structured. Thus, in principle, the structure of energy patterns is lawfully predictable from the properties of the object.

The ecological assumption that the structure of energy allows perceivers to identify environmental properties is encapsulated in the concept of invariance. In order to clarify the concept of invariant structure, Gibson (1986) contrasted it with perspective structure. Perspective structure refers to the overall structure of stimulation on the sensory organs which changes when the point of observation changes (caused either by the movement of the perceiver or the movement of the object). For example, as the distance between the perceiver and the perceived object increases the size of the visual angle subtended on the retina decreases (i.e., projected image size decreases). Gibson argued that in spite of the change in perspective structure, there exists an invariant structure, which does not change with the changing perspective. The invariance is determined by the physical properties of the perceived object. This argument can be illustrated with Gibson’s (1986) description of how people perceive the shape of a rectangular tabletop to be constant under changing viewing conditions.

What are the invariants underlying the transforming perspectives in the array from the tabletop? What specifies the shape of this rigid surface as projected to a moving point of observation? Although the changing angles and proportions of the set of trapezoidal projections are a fact, the unchanging relations among the four angles and the invariant proportions over the set are another fact, equally important, and they uniquely specify the rectangular surface (p. 74).

Gibson’s comment illustrates that invariance is conceptually related to the invariant interrelations between stimulus features. Since the interrelations between stimulus features can be objectively measured, it is possible to describe and quantify invariance in terms of geometrical concepts.

A particularly useful type of geometry is the one that defines the group of transformations that leave certain properties invariant. For example, the perception of shape constancy in rotating objects has been conceptualised in terms of Euclidean geometry as the distance between two points on the object (Michaels & Carello, 1981). The distance between two
points on an object remains the same regardless of the rotation, and we see that it remains the same when we view the rotating object. It is important to note, however, that the application of Euclidean geometry to shape constancy (in terms of the relation between perspective and invariant structure) is sufficient only under constant viewing distances and constant polar projections. When the viewing distance varies or the object is slanted (i.e., perspective structure changes more), the description of shape constancy requires more complex geometries (e.g., similarity and projective geometry) (Michaels & Carello, 1981). Thus the description of invariance becomes mathematically more abstract under more complex transformations.

Although it was Gibson (1986) who directed the attention of psychologists to the concept of invariance, the perception of structural properties in the environment had been emphasized in perceptual theories prior to Gibson. For example, Gestalt theories of perception emphasized that humans perceive stimuli in a holistic rather than elemental manner (Kohler, 1941, 1967). The German term *Gestalt* refers to a property of a structure or configuration that is not revealed by the individual elements that make up the structure. The central principle of the Gestalt philosophy has been expressed by the phrase ‘the whole is something else than the sum of its parts’ (Koffka, 1935, p. 176). Gestalt psychologists describe several principles that explain how elemental informational variables are organised into a coherent and unified whole (e.g., according to the principle of proximity, elements that are near to each other are grouped together).

Although both ecological and Gestalt psychologists emphasize the perception of structure in the environment, they disagree on the epistemological question of how structural properties are perceived. Gestalt theories imply that structure is perceived by mentally integrating or grouping the stimulus features into a coherent whole. Gestalt theory was typically representationist. Ecological psychologists, on the other hand, argue that mental integration of the structural properties of stimuli is not required. According to Gibson (1986), invariant structure “is a unit, and components do not have to be combined” (p. 141). Thus, as discussed earlier, Gibson’s contribution to theories of perception was not the idea that humans perceive stimuli as structured entities but rather his argument that structures are perceived directly without the mental integration of elements.
2.4.2 The Concept of Affordance

During perception humans not only identify the intrinsic properties of objects (i.e., invariance) but also identify their value and meaning. Perceiving an environmental stimulus involves the awareness of what the stimulus can be used for and/or what it means to the perceiver. For example, when humans perceive a car they typically see not merely an object with four wheels but also the potential of that object for use as transportation. The question of how humans perceive the meaning and value of objects has long been addressed by psychologists. According to Koffka (1935) objects “tell us what to do with them” (p. 353) a phenomenon that he called demand character. This demand character or meaning of objects was also referred to by the German term Afforderungscharakter (Lewin, 1926), which later translated as “invitation character” (Brown, 1929) and as “valence” (Lewin, 1936). However, the first part of the German word is equivalent to the English “afford”, and it was this term that was adopted by Gibson (1986). To describe the perception of meaning, he introduced the concept of affordance:

The affordances of the environment are what if offers the animal, what it provides or furnishes, either for good or ill. The verb to afford is found in the dictionary, but the noun affordance is not. I have made it up. I mean by it something that refers to both the environment and the animal in a way that no existing term does. (p. 127)

The theory of affordances is a radical departure from existing theories of value and meaning. It begins with a new definition of what value and meaning are. (p. 140.) … An important fact about the affordances of the environment is that they are in a sense objective, real, and physical, unlike values and meaning, which are often supposed to be subjective, phenomenal, and mental. But, actually, an affordance is neither an objective property nor a subjective property; or both if you like. An affordance cuts across the dichotomy of subjective-objective and helps us to understand its inadequacy (p. 129)

As is emphasized here, Gibson (1986) introduced the concept of affordance to refer to values and meanings in a way that is different from the conventional use of these terms. As opposed to theories that conceptualize values and meaning as mental properties of the perceiver (e.g., Koffka, 1935), Gibson (1975) argued that affordances are “anchored in human beings at one pole and in things at the other” (p. 320). Thus the term affordance
relates the perceived object to the perceiver. Affordances are not subjective or private or mental. For example, a car’s usability for transportation remains even if it is not recognised by the perceiver.

The affordance of something does not change as the need of the observer changes. The observer may or may or may not perceive or attend to the affordance, according to his needs, but the affordance, being invariant, is always there to be perceived.” (Gibson, 1986, p. 138-139)

Here, by conflating the concept of affordance with the concept of invariance Gibson, perhaps inadvertently, implies that meaning is an intrinsic property of the object. However, he appears to want only to avoid suggesting that the affordance is imposed upon the object by the perceiver. In order to clarify the concept of affordance it may be useful to describe the perception of affordance as a three-term relation, according to which an intrinsic property of an object (i.e., invariance) affords or means something (i.e., affordance) to the perceiver. Such a way of describing the three terms in perception (i.e., invariance, affordance, and the perceiver) helps to clarify the notion that affordance relates to both the object and the perceiver but it is neither a physical property of the object nor a mental property of the perceiver. The claim that affordances are not mental properties is central to ecological theories of perception. Consistent with the theory of direct perception affordances, like invariants, are directly perceived.

There is only one kind of perception, the perception of the world with the meanings and values already in it. (Gibson, 1975, p. 320)

This is a radical hypothesis, for it implies that the ‘values’ and ‘meanings’ of things in the environment can be directly perceived. (Gibson, 1986, p. 127)

The theory of affordances is a radical departure from existing theories of value and meaning. It begins with a new definition of what value and meaning are. The perceiving of an affordance is not a process of perceiving a value-free physical object to which meaning is somehow added in a way that no one has been able to agree upon; it is a process of perceiving a value-rich ecological object. (Gibson, 1986, p. 140)
As discussed earlier, direct perception does not entail correct perception. Although affordances and invariants are perceived directly, this does not mean that they are inevitably perceived, or completely perceived, or necessarily perceived correctly. The direct realist theory of perception acknowledges (as do indirect theories) that the perception of complex environmental properties requires learning.

2.5 Perception is Shaped by Learning

One possibility for the failures of inference observed in the colliding-balls experiments which were discussed earlier is that the perception of incomplete invariants simply implies that the perceiver has not yet learned to discover the complex higher-order invariants that specify the stimulus property. Since mass-ratios of colliding balls are specified by higher-order invariants, it is possible that people require perceptual training before they learn to perceive those higher order invariants. In fact, it has been demonstrated that, after training with feedback, participants’ perception of mass-ratios improves, which suggests that people can learn to perceive more complex kinematic invariants that specify mass-ratios (Jacobs, Michaels, & Runeson, 2000; Jacobs, Runeson, & Michaels, 2001).

Perceptual learning is generally defined as a change in perception as a result of experience. As a result of experience, perceivers notice relations among stimulus variables that were previously unnoticed. That is, as a result of experience, people perceive more and more complex invariants in environmental stimuli. It is important to note that learning is not only considered in the sense of ontogeny (i.e., the development of the individual) but also in the sense of phylogeny (i.e., the evolutionary development of a species) (Michaels & Carello, 1981). From an evolutionary perspective, it is assumed that perceptual abilities that exist at the time of birth (i.e., innate abilities) are the product of learning that has occurred through the development of species.

2.5.1 The Innate Perceptual Ability to Perceive Invariants is the Result of Evolutionary Learning

Evolution is a process by which species adapt to their environment over a long period of time. At an evolutionary level, learning is reflected by a continuous interplay between genetic variation and natural selection. Natural selection ensures that genetic variations
that allow adaptive interaction with the environment are selected for the future population. If the individual survives (i.e., adapts to the environment), genetic material is passed along to offspring. Thus as a result of evolution, individuals at birth are genetically predisposed to perceive certain invariants within their environment.

The innate ability to perceive certain invariants is supported by various studies that involve young infants. For example, three to four month old infants were found to perceive invariants that comprise simple forms (Bomba & Siqueland, 1983), orientation (Bomba, 1984), and hue (Bronstein et. al., 1976). In addition, experimental evidence suggests that humans are predisposed to perceive biologically meaningful stimuli, such as faces. Newborn infants were found to prefer to attend to face-like patterns (Goren, Sarty, & Wu, 1975), and the ability to detect simple structural invariants in faces was found to be present at an early age of development (Younger, 1992). Further and more detailed examples of innate perceptual abilities will be provided in later chapters focusing on pitch and speech perception (Chapters 4 and 5, respectively).

It has been argued that young infants at the early stages of development are only able to perceive invariants that entail relatively simple relations (Younger & Gotlieb, 1988). As will be reviewed next, the perception of higher-order invariants is learned during development.

2.5.2 The Perception of Complex Invariants is a Result of Developmental Learning

Complex invariants are rarely perceived at birth, but rather humans need to learn to perceive them in the course of development. During development the exposure to objects and events increases, allowing infants and children to discover structural relations among stimulus variables. Face perception provides a good illustration of the developmental trend in the perception of structural properties. Developmental trends in face perception have been extensively investigated, presumably because faces are one of the most common visual stimuli that humans encounter from birth.

Although young infants can perceive simple facial invariants, in order to recognise and distinguish faces, humans need to develop the ability to perceive complex structural invariants. Simple structural invariants in face perception are regarded as structural
regularities that are common to all faces (e.g., two eyes above nose and mouth embedded in oval shape etc). Diamond and Carey (1986) referred to such relational properties as first-order facial relations. More complex facial invariants, referred to as second-order relations, depend on distinctive relations among elements that allow one to distinguish individual faces (Diamond & Carey, 1986). Developmental studies have found that, although young infants can perceive first-order invariants, it is not until the age of 10 months that infants start perceiving complex (second-order) facial invariants (Young, 1992).

Experimental evidence suggests that, during development, and as a result of exposure to faces, children progressively learn to recognise complex facial invariants that allow them to distinguish different faces. In a typical face-recognition task, participants are first given a picture of an unfamiliar face to inspect and then they are required to recognise the face from two or more test faces. Goldstein and Chance (1964, 1965) tested 6, 9, and 14 year old children and found a developmentally increasing trend in the accuracy of face recognition. Similar developmental trends were found in other studies investigating face perception in various age groups. For example, Carey and Diamond (1977) and Blaney and Winograd (1978) tested 6, 8, and 10 year old children; Carey, and Diamond and Wood (1980) tested 6 and 10 year olds. The result of these studies suggests that throughout childhood there is a general trend of improvement at face perception.

In addition to the age of the perceiver, face perception is influenced by the facial features that are perceived. For example, it has been found that adults are better at perceiving adult faces than child faces (McKelvie, 1981) or infant faces (Chance, Goldstein, & Andersen, 1986). It is possible that adults are more sensitive to adult facial invariants because due to social context (e.g., work) they encounter more adult than child faces. The idea that people are better at perceiving facial invariants that they encounter more often is also supported by cross-cultural studies. Goldstein and Chance (1980) found that Caucasian adults are better at perceiving Caucasian than Japanese faces. It is interesting, however, that in children (6-11 years) they found no difference in the ability to recognise own-race and other-race faces. These results further support the view that it is the result of developmental learning that humans become more sensitive to complex structural invariants that are relevant or meaningful in their everyday life. This view will be further discussed by reviewing developmental studies in pitch and speech perception (Chapters 4 and 5, respectively)
2.6 Conclusion

Generally, all theories of perception seek to explain how humans and other animals get to know their environment; however, various explanations are based on different epistemologies. To emphasize the fundamental difference between various theories of perception, they are usually classified as either direct or indirect. The term direct perception refers to a process of knowing, in which the organism is in epistemologically unmediated, but physiologically mediated, contact with the environment. The term indirect perception refers to a mediated process of knowing that involves mental entities, such as representations, and mental operations (such as inferences) from those internal representations to the external objects represented.

Although the difference between direct and indirect perception appears to be straightforward, perceptual researchers often misunderstand the assumptions and implications of direct and indirect perception. This misunderstanding has generated a number of perceptual studies that are based on false premises and therefore have resulted in mistaken claims regarding empirical evidence for or against the two theories. The question whether perception is direct or indirect can only be decided on conceptual grounds.

The fact that the question between direct and indirect perception is a conceptual issue has two important implications for perceptual research. Firstly, in order to provide a coherent theory of perception it is important to clearly describe and examine the epistemological assumptions and implications of the theory. Such considerations have led to the conclusion that the theory of direct perception is the only viable and coherent theory. Secondly, rather than attempting to test between direct and indirect perception, perceptual research can make genuine progress by focusing on aspects of perception that can be experimentally tested. One such aspect is perceptual learning, which is a well established and generally acknowledged phenomenon in psychology.

According to the ecological direct realist approach to perception, the answer to the question what is learned in perceptual learning is invariants and affordances. The next question to examine is the question of how these are learned. This question takes us to the details of
learning theory, and to the possibility of its combination with the ecological theory of direct perception.
CHAPTER 3

THE QUESTION OF ‘HOW’ IN PERCEPTUAL LEARNING: HOW DO HUMANS LEARN TO PERCEIVE IN THEIR ENVIRONMENT?
3.1 Introduction

Answering the question how perceptual learning occurs is challenging as there is no general agreement on what needs to be covered in addressing this issue. Perceptual learning can be described at different levels of analysis including mental operations, neural mechanisms, and behavioural laws that govern learning (Goldstone, 1998). According to Johnston (1981), when considering theories of learning one should distinguish between learning processes and learning principles as they are conceptually different. Learning processes are events that can be described in behavioural and/or physiological terms, and learning principles are formal or informal statements that explain and predict learning processes. According to Johnston, “The aim of theories is, in general, to explain or predict processes by means of principles” (p. 126).

The aim of this chapter is to highlight and address the limitations of the ecological theory of perceptual learning. The ecological theory of perceptual learning has been criticized for lacking learning principles that explain and predict how experience changes perception (McLaren & Mackintosh, 2000; Postman, 1955). This limitation will be addressed by applying associative learning principles (i.e., classical and operant conditioning) to explain perceptual learning within the ecological framework.

3.2 The Ecological Theory of Perceptual Learning

According to the ecological perspective, perceptual learning involves discovering invariants that convey meaning (i.e., affordances) to the perceiver. As Michaels and Carello (1981) noted, “the task for learning becomes the education of attention to those invariants …that specify the affordances of the local environment.” (p. 81). In order to address the question of how humans learn to perceive affordances, ecological psychologists have proposed the process of differentiation.

3.2.1 The Process of Differentiation

Differentiation was initially defined as a process that increases the specificity of perception by “responding to variables of physical stimulation not previously responded to” (Gibson & Gibson, 1955a, p. 34). According to Gibson and Gibson, during differentiation people discriminate (i.e., differentiate) stimuli along various
dimensions. The Gibsons’ proposal of the differentiation theory was followed by a detailed criticism by Postman (1955). Postman argued that Gibson’s theory of perceptual learning can only be maintained at the expense of begging the question of learning. According to Postman, “the fact that the organism has learned to discriminate more qualities is the very fact that we need to explain” and “improvement in discrimination cannot be invoked to explain improvement in discrimination” (Postman, 1955, p.144). A similar issue was raised by McLaren and Mackintosh (2000) by posing the following question regarding differentiation. “What is the mechanism that drives this learning? J. J. Gibson and E. J. Gibson (1955) provided no hints” (p. 238)

In response to Postman’s criticism, Gibson and Gibson (1955b) noted that the theory of differentiation “is concerned with the question of what is learned [italics added] in perceptual learning, not how it occurs [italics added], or at least not as yet” and it “is not a theory but only the promise of a theory, and its explanatory value remains to be seen” (p. 447). Although Gibson and Gibson admitted that the predictive power of differentiation is limited and could possibly be strengthened, it has not undergone any major theoretical development since the middle of the 20th century. As illustrated below, a more recent definition of differentiation by Gibson and Pick (2000), similar to the previously noted definition, assumes rather than accounts for learning.

Perceptual differentiation can be characterized as a narrowing down from a vast manifold of information to the minimal, optimal information that specifies the affordances of an event, object, or layout. As the information is extracted, useless information is discarded. (p. 150)

It appears that the definition of differentiation provided by Gibson and Pick (2000) describes rather than explains perceptual learning. Perceptual learning, which is defined as the education of attention to invariants that specify affordances, can only be circularly defined as “narrowing down from a vast manifold of information to … information that specifies the affordances” (Gibson & Pick, 2000, p. 150). The question of how humans learn to perceive invariants that specify affordances is left unanswered by the theory of differentiation. The fact that the theory of differentiation in itself does not provide a full account of perceptual learning seems to be also
recognised by ecological psychologists since they elaborate on the conditions that result in perceptual learning. According to E. J. Gibson and colleagues (Gibson, 2000; Gibson & Pick, 2000) the conditions necessary for perceptual learning, as we discuss in the next section, involve the perceiver’s interaction with his/her environment (exploratory activity).

3.2.2 The Role of Exploratory Activity in Perceptual Learning

Ecological psychologists propose that perceptual learning is shaped by *exploratory activity*, which involves coordinated activity that leads to making contact with the environment (Gibson, 2000; Gibson & Pick, 2000). For example, young infants’ tendency to reach with their arms in an aimless fashion can be considered as exploratory activity. If reaching out with the arm results in grasping an object, then infants often carry the object to their mouth for further exploration. These exploratory behaviours, which have been shown to ‘calibrate’ infants’ bodies (van der Meer, van der Weer, & Lee, 1995), allow the infant to discover invariants that specify objects. When the child later starts crawling and walking he/she learns to discover invariants that afford locomotion.

An essential aspect of exploratory activity is that making contact with a certain object in the environment is followed by *consequences* that allow perceivers to learn about the affordances of the object (Gibson, 2000; Gibson & Pick, 2000). For example, learning about the affordances of a slope involves experiencing the consequences of descending the slope (Gibson, 2000). Descending on foot may result in falling, while descending by sitting can safely be done without painful consequences. As a result, a child learns that a slope with a certain angle can afford descending by sitting but it results in falling when descending on foot. Thus, the ecological theory of perception emphasizes that the perceptual learning of affordances is actively driven by the consequences of action.

Although the assumption that learning of affordances is shaped by action is central to the ecological theory, it should be noted that this assumption limits the notion of perceptual learning. Acting on the environment is not the only way that humans can discover affordances. Humans can learn the affordances of an object by *observing* its
relation to other objects and people. For instance, a child may initially learn the affordance of a cup by observing his/her parents using it for drinking. As pointed out by others (e.g., Costall, 1995, Heft, 1989), many objects and events are shaped by human intervention and their affordances are often the product of social and cultural conventions. The learning of such affordances involves observation and other social influences (e.g., use of language) that seem to be neglected by the proposed ecological principles of perceptual learning (i.e., exploratory activity).

3.2.3 Limitations of the Ecological Theory of Perceptual Learning

It could be argued that the ecological theory of perceptual learning is focused mainly on the question of what is learned and only partially addresses the question of how learning occurs. Firstly, the process of differentiation, as Gibson and Gibson (1955b) themselves acknowledged, “is concerned with the question of what is learned [italics added] in perceptual learning, not how it occurs [italics added]” (p. 447). Secondly, the proposed mechanisms of exploratory activity (i.e., interacting with the environment and evaluating the consequences of action) are only suited to explain perceptual learning that results in physically interacting with the environment. In addition to exploratory activity, there are other ways (e.g., observing environmental regularities) that allow humans to discover affordances in the environment. The ecological theory of perceptual learning proposes no principles to describe how humans can learn to discover affordances by simply observing their environment.

In the following section it is argued that the ecological theory of perceptual learning can be strengthened and expanded by theories of conditioning. Conditioning theories are based on associative learning principles which are widely used to explain a range of human behaviour (Domjan, 2005; Rescorla, 1988). Although associative principles are popular in behavioural research they are only rarely applied to perception. Most importantly, the rare application of associative principles to perception (e.g., Davies, 1987; Hall, 1991; Razran, 1955; Wolpe, 1981; Woodworth, 1947) has not yet been systematically related to the ecological theory of perceptual learning. As will be argued next, associative principles can be readily applied to explain how humans learn to perceive affordances and such an application strengthens and expands the ecological theory of perceptual learning.
3.3 The Application of Conditioning to the Ecological Theory of Perceptual Learning

Learning has long been described as a dynamic process which involves the formation or the detection of associations. In the 17th and 18th century, the British empiricists (e.g., Hobbes, Locke, Berkeley, Hume, and Hartley) described learning in terms of the association of ideas. The British empiricists’ view of associative learning is illustrated by the following example.

A man has suffered pain or sickness in any place; he saw his friend die in such a room: though these have in nature nothing to do one with another, yet when the idea of the place occurs to his mind [italics added], it brings (the impression being once made) that of the pain and displeasure with it: he confounds them in his mind, and can as little bear the one as the other. (Locke, in Watson, 1979, p. 41).

As the quote illustrates, Locke regarded associations as mental operations being applied to ideas (i.e., mental representations). Such a representational view of associative learning has changed over the centuries. Since the 19th century, associative learning has been described in terms of associations between stimuli (classical conditioning) and between stimuli and the consequences of responding to stimuli (operant conditioning) (Pearce & Bouton, 2001; Wasserman & Miller, 1997). The difference between theories referring to associations between ideas and theories referring to associations involving stimuli, responses and consequences is an epistemological one. As will be argued next, while the prior theories are clearly representational, the latter ones are compatible with direct realism and they can be readily applied to the ecological theory of perceptual learning.

3.3.1 Associative Principles are Epistemologically Compatible with the Ecological Theory of Perceptual Learning

The reluctance to apply associative principles to the perceptual learning of affordances partly appears to be due to the misconception that associative principles imply indirect perception. For instance, Gibson and Pick (2000) argue that
“Perceptual learning is not properly described as association or as an addition of any kind, as a response to a stimulus, or as a ‘representation’ of ‘input’”. (p. 149). It appears that Gibson and Pick regard the notion of associations as a representationist concept. The misconception that associative principles in themselves imply representationalism originates from Gibson and Gibson’s (1955a) criticism of associative perceptual theories that are in fact representationalist. One of the theories criticised by Gibson and Gibson was “Titchener’s context theory of meaning, which asserts that we learn to perceive objects when a core of sensations acquires by association a context of memory images” (p. 33). After their review of representationalist theories of perception, Gibson and Gibson concluded that:

It seems to us that all extant theories of the perceptual process, including those based on association [italics added],…have at least this feature in common: they take for granted a discrepancy between the sensory input and the finished percept and they aim to explain the difference. They assume that somehow we get more information about the environment than can be transmitted through the receptor system. In other words, they accept the distinction between sensation and perception. … Let us consider the possibility of rejecting this assumption altogether. (p. 33-34)

Although Gibson and Gibson were right to argue that associative theories that are based on the distinction between sensation and perception are representationalist, it is a mistake to conclude that associative principles in themselves imply representationalism. Associative principles simply refer to the learning of relationships and they do not in themselves imply indirect perception. It is the application of associative principles (specifically, the issue of what they are applied to), rather than the principles themselves, which have implications regarding the epistemological nature of perception. The application of associative principles to mental images (e.g., the British empiricists’ view of associations) implies indirect perception; however, the application of associative principles to observable stimuli, responses and their consequences without any reference to mental images is consistent with the notion of direct perception.
In the theoretical position adopted in this thesis, associations are considered as properties of the environment rather than properties of the organism’s mental world. It is assumed that associations exist in the real world and humans learn that they exist as a result of observing and interacting with the environment. In other words, associations are considered as *relations* that are discovered, rather than being mentally created by the observer. A similar realist view of associative learning has been put forward by Skinner (1976).

The runner’s heart is said to beat fast before the start of the race because he “associates” the situation with the exertion which follows. But it is the environment, not the runner, that “associates” the two features, in the etymological sense of joining or uniting them. (p. 43)… A person is changed by the contingencies of reinforcement under which he behaves; he does not store copies of the stimuli which have played a part on the contingencies. There are no “iconic representations” in his mind; there are not “data structures stored in his memory”; he has no “cognitive map” of the world in which he has lived. He has simply been changed in such a way that stimuli now control particular kinds of perceptual behavior. (p. 93-94).

These excerpts from Skinner illustrate that the description of behaviour in term of associative principles does not necessarily imply indirect perception. In fact, the above quote clearly shows that Skinner’s theory has the same epistemological basis as Gibson’s ecological theory of perception. The idea that Gibson’s perceptual theory and Skinner’s learning theory are underlined by the same epistemological assumption has been extensively argued by Costall (1988), who concluded that the two approaches “should keep a ‘sympathetic eye’ on one other’s progress” (p. 115). Thus, it appears that adopting a realist view of associations (i.e., associations exist in the world rather than in the mind) opens up the possibility of better explaining the ‘how’ of perceptual learning within the ecological framework.

### 3.3.2 Perceptual Learning through Classical Conditioning

The systematic study of stimulus-stimulus (S-S) associative learning originates in Pavlov’s (1927, in Windholz 1992) experiments that were designed to investigate the
digestive responses of animals. Pavlov was trained as a physiologist and investigated
the chemical composition of digestive juices that are secreted by animals to aid the
breakdown of ingested food. Pavlov’s pioneering experiment, which is generally
cited as a classical example of learning via S-S associations, involved presenting a
bell just before giving food to a hungry dog. After a number of ‘bell - food’ trials
Pavlov noted that the dog started salivating right after the presentation of the bell even
in the absence of food. The salivation as response to the bell indicates that the dog
learned the relationship between the bell and the food (i.e., that the bell predicts food).

Pavlov, being a physiologist, was interested in how stimuli were converted into neural
processes. However, his discovery had a greater impact in the field of psychology
than in that of neuroscience. In psychology, learning that is the result of discovering
the relationship between stimuli (i.e., S-S association) is generally referred to as
Pavlovian or classical conditioning (Domjan, 2005). The term conditioning refers to
the fact that the capacity of a stimulus (i.e., conditioned stimulus) to elicit a new
response (i.e., conditioned response) is conditional upon its pairing with a stimulus
(i.e., unconditioned stimulus) that naturally or reflexively elicits the response (i.e.,
unconditioned response). Pavlovian conditioning principles have been applied to
explain a wide range of involuntary human behaviour including phobic behaviour,
immune responses, and drug tolerance (Domjan, 2005; Siegel, Baptista, McDonald,
Weise-Kelly, 2000; Siegel & Ramos, 2002). In addition to describing behaviour,
classical conditioning principles are often used in clinical settings to modify human
behaviour (e.g., systematic desensitisation).

Although classical conditioning has been applied to explain and modify a wide range
of human behaviour, it is often done so without any reference to perceptual learning.
Nevertheless, it could be argued that perceptual learning is inherent in classical
conditioning.

The subject in a conditioning experiment does not have to learn how to
salivate or half-close his eyes. The new learning, the conditioning, is sensory
and not motor. The change that takes place in him during the process of
conditioning is a change in his way of receiving, or perceiving, the sequence
During classical conditioning, the perception of the conditioned stimulus changes as a result of it being paired with the unconditioned stimulus. As Garzia Hoz (2003) pointed out, conditioned responses are made “to the stimulus not so much for itself, for its particular and immediate property, but rather because that stimulus is the signal for another [italics added]” (p. 169). According to Garzia Hoz, classical conditioning can be interpreted in terms of signalization, in which the perceiver learns that the conditioned stimulus signals the unconditioned stimulus. It is important to realize that learning about what an object or an event signals can be understood as learning about the affordances of the stimulus. For instance, in Pavlov’s classic experiment, the dog learned to perceive that the bell (i.e., conditioned stimulus) affords or means food (i.e., unconditioned stimulus). Thus interpreting classical conditioning in terms of signalization is useful as it provides an insight about how certain affordances can be learned by simply observing (i.e., without physically interacting with) the environment.

3.3.3 Perceptual Learning through Operant Conditioning

The experimental investigation of operant conditioning, similar to classical conditioning, originates in research examining animal behaviour. While Pavlov was investigating classical conditioning in dogs, operant conditioning was first studied by Thorndike (1911, Wasserman & Miller, 1997) using cats as subjects. The aim of Thorndike’s classical experiment was to train a hungry cat to escape from a small chamber by pressing a lever which opened the door. In this experiment, the lever that opens the door is considered as a stimulus and pressing the lever is the response. In order to motivate the animal to learn the appropriate response Thorndike placed some food outside the chamber, which served as a reward when the animal performed the required behaviour.

Thorndike observed that initially cats moved around and explored the various parts of the box, and during their exploration they would accidentally (i.e., by trial and error) perform the required response that opened the door, which allowed access to food.
After performing the response, the animal was returned to the box, and the procedure was repeated many times. On each trial, Thorndike measured the amount of time the cat took to escape from the box, and observed that, as the number of trials increased, the amount of time to escape gradually decreased. The decrease in the amount of time to press the lever suggests that the cat learned to perform the appropriate response that allowed it to access the food. Thorndike concluded that actions that are followed by positive outcomes (such as food) are likely to be generated again in the future (law of effect).

Although it was Thorndike who pioneered the study of operant conditioning, it was Skinner (1935, 1966, 1976) who extensively elaborated on the principles that underlie such learning. Skinner (1976) emphasized that behaviour is not only shaped by environmental stimuli but also shaped by the consequences of responding to stimuli, arguing that “the environment affects an organism after, as well as before, it responds”. For example, when the cats in Thorndike’s experiment accidentally pressed the lever, their behaviour was followed by the positive outcome of receiving food which reinforced the cats to perform the same behaviour when returned to the box. Skinner conducted a number of experiments involving animal subjects (e.g., pigeons) to investigate how the consequences of responses shape future behaviour. Based on his experimental observations, Skinner proposed that if the response to the stimulus is followed by a positive outcome (reinforcement) the future probability of that response increases. On the other hand, if the response to the stimulus is followed by a negative outcome (punishment) the probability of the response to the stimulus decreases.

It is important to note that, although Skinner was predominantly focused on explaining behaviour without any reference to perceptual learning, he did not deny that the change in the behaviour may correspond to (or may be underpinned by) change in the perceived meaning (i.e., affordance) of the stimulus. For instance, Skinner (1976) noted that “if a rat is reinforced with food when it presses the lever in the presence of a flashing light but with water when the light is steady, then it could be said that the flashing light means food and the steady light means water” (p. 101). Although Skinner did not openly promote the idea that operant conditioning changes the perceived meaning of stimuli, operant conditioning has been interpreted as “a
process of meaning making that is governed largely by natural contingencies” (DeGrandpre, 2000, p. 721).

The argument that operant conditioning results in the perceptual learning of affordances is further supported by the fact that operant conditioning is based on the same principle as exploratory activity described by ecological psychologists. As reviewed earlier, ecological psychologists argue that observing the consequences of action is an essential component by which exploratory activity leads to perceptual learning. In fact, Gibson defined exploratory activity as “a perception–action sequence that has consequences [italics added]” (Gibson, 2000, p. 296). Considering the similarities between the notions of operant conditioning and exploratory activity, the principle of operant conditioning can be formally applied to explain action-related perceptual learning.

So far, classical and operant conditioning principles have been applied to explain the perceptual learning of affordances. It has been argued that affordances can be learned by discovering associations between stimuli and discovering the consequences of interacting with stimuli. According to ecological theory, the learning of affordances changes the perception of stimuli by educating attention to invariants that specify affordances. It is in turn argued that conditioning changes the perception of stimuli by educating attention to invariants that predict S-S associations (classical conditioning) and certain consequences of interacting with the stimulus (operant conditioning).

The process of educating attention to invariants that specify affordances is, of course, underpinned by physiological mechanisms. The following section describes the possible neural processes that underpin perceptual learning.

### 3.4 Neural Processes of Perceptual Learning

#### 3.4.1 Hebbian Synapses

The physiological aspect of learning is considered to be a structural change within the nervous system (i.e., plasticity). The nervous system consists of millions of neurons that are either directly or indirectly connected to each other. The region where the
neurons connect to each other is called the synapse, where neurons communicate with each other by the release of neurotransmitters. When a sending (i.e., pre-synaptic) neuron releases neurotransmitters, it produces excitation or inhibition of one or more receiving (i.e., post-synaptic) neurons. In the human brain, there can be up to 100000 synapses formed on a single post-synaptic neuron, which indicates the complexity of neural interaction (Rosenzweig, Leiman, & Breedlove, 1999).

The most influential theory which has linked learning to neural synapses was promulgated by Hebb (1949). Hebb proposed that learning is based on the increase in the number of synapses or an increase in the size of the already existing ones. Although Hebb’s theory was published in the middle of 20th century, it is a theory that holds a prominent place in the current field of neuroscience. Current theories in neuroscience and psychology often refer to ‘Hebbian synapses’ in their account of learning and scientific advances in neuroscience seem to confirm the ‘Hebbian learning rule’ (Cooper, 2005). According to Cooper (2005), “the Hebbian synapse is now ubiquitous in neuroscientific research, in a way Hebb himself could never have envisaged, and his neuropsychological postulate has assumed the statues of dogma, with instant recognition and acceptance” (p. 865).

In contemporary neuroscience, learning is described in terms of the change in the synaptic efficacy which refers to the ability of a pre-synaptic (sending) neuron to excite a post-synaptic (receiving) neuron (Cooper, 2005). The increase in synaptic efficacy is partly determined by the coordinated activity between neurons. Synaptic efficacy increases if pre and post-synaptic neurons are simultaneously activated. This process, which is referred to as long-term potentiation, is commonly described by the maxim ‘neurons that fire together, wire together’. Since long-term potentiation corresponds to a process where neural connections are established it has been suggested as underlying S-S associative learning (Stein, 1997). In addition to the coordinated activity between neurons, synaptic efficacy can be strengthened (i.e., learning occurs) by the release of certain dopamine neurotransmitters (Donahoe & Palmer, 1994). Since the release of dopamine is closely linked to reinforcement, the ability of dopamine to influence synaptic efficacy has been considered as a physiological correlate of operant conditioning (Donahoe & Palmer, 1994; Stein, 1997).
3.4.2 Artificial Neural Networks

Artificial neural networks are processing systems that are designed to model the functioning of the human brain. Neural networks consist of neuron-like computing elements (i.e., nodes or units), which are connected by weighted unidirectional links (Gluck & Bower, 1988). In general, the network consists of layers of sensory and response units often including one or more intermediate layers of hidden units. The sensory units are activated by the stimulus presentation and they pass their activation to the output or response units either directly or through the intermediate units. Artificial neural networks are useful analogies to explain the neural basis of perceptual learning because they can model the lawful neural changes that occur in the brain during learning. Changes in synaptic connections (i.e., neural basis of learning) are modelled in the network by altering the weights on the connections between units.

Learning in neural networks is often described as either supervised or unsupervised. The term supervised indicates that learning in the network is driven by a teaching signal, which is corrective feedback determined by the task that the network is designed to learn. Unsupervised learning, as opposed to supervised learning, is not driven by any teaching signal but rather is shaped solely by the input data. It is important to note that the concepts of unsupervised and supervised learning are conceptually similar to the notion of classical and operant conditioning. Since classical conditioning refers to learning about the structural property of the environment (i.e., S-S associations) it can be modelled by an unsupervised network. On the other hand, operant conditioning can be modelled by a supervised network since operant conditioning is shaped by the consequences of responding to stimuli (i.e., feedback). Thus, it appears that artificial neural networks of perceptual learning (provided that they do not lapse into implicit representationism) can be retained in the proposed ecological-conditioning theoretical framework.

So far, this chapter has been concerned with applying the principles of conditioning to the ecological theory of perceptual learning and with describing the possible neural basis of how conditioning changes perception. The next section will provide
empirical support for the argument that conditioning results in perceptual learning by reviewing some existing experiments, which demonstrated perceptual learning via classical and operant conditioning.

3.5 Experimental Investigation of Conditioning in Perception Research

3.5.1 Classical Conditioning Experiments

Most experimental demonstrations of classical conditioning in perception research involve cross-modal S-S associations. One of the first experiments demonstrating the effect of classical conditioning on perception involved the pairing of visual and auditory stimuli (Ellson, 1941). Ellson repeatedly presented participants with tone-light pairings and after the conditioning participants reported an auditory sensation of tone even when they were only presented with light. Later it was also shown that a conditioned percept can be induced in the visual domain as a result of auditory stimuli (Davies, 1974a, 1974b, 1976; Howells, 1941). These experiments demonstrated that classical conditioning can induce an auditory perceptual response to a visual stimulus and a visual perceptual response to auditory stimulus.

Perhaps the most impressive study inducing a conditioned visual percept in response to auditory stimulus was conducted by Howells in 1944. Howells trained participants to associate a lower tone (approximately 261 Hz) with the colour red, and a higher tone (approximately 392 Hz) with the colour green. In each training trial, participants heard a tone accompanied with a coloured patch appearing on a screen, and they were required to indicate the colour of the stimulus patch. After an extensive training period (i.e., total of 26,250 training trials presented over a number of sessions), participants’ colour perception was tested using a chromatoscope, which ranged from the colour red to the colour green through the neutral colour of white. Participants were required to adjust the chromatoscope to the best possible white while they were listening to either the lower or the higher tone. It was observed that if the lower tone (associated with red) was sounded during the adjustment phase, the point of adjustment was placed slightly towards the colour green. On the other hand, when the higher tone (associated with green) was sounded, participants adjusted the chromatoscope slightly towards red. These results indicate that participants learned
the colour-tone associations since in the presence of the lower and higher tones participants seemed to perceive the colour white as slightly red and green, respectively.

In addition to hearing and vision, the effect of learned cross-modal associations on perception has also been investigated in other sensory modalities. Stevenson and colleagues (Stevenson, Prescott, & Boakes, 1995; Stevenson, Boakes, & Prescott, 1998) demonstrated that participants’ odour perception can be altered as a result of learning odour-taste associations. The experiments conducted by Stevenson et al. (1995) consisted of three main phases that were pre-test (Day 1), associative training (Days 2 -4), and a post-test (Day 5). During pre-test and post-test, participants sniffed and rated the sweetness and sourness of two target odours (i.e., lychee and water chestnut) that were chosen on the basis of being relatively unfamiliar and of moderate sweetness and sourness. During the associative training, participants sampled nine sucrose and nine citric acid solutions, to which the target odour had been added as a flavourant. Thus, training involved exposing participants to sucrose-lychee and citric acid-water chestnut pairings. The experiments found that after training (i.e., in the post-test phase), the perceived sweetness of lychee odour, and the perceived sourness of water chestnut odour increased.

Morrot, Brochet, and Dubourdieu (2001) demonstrated that perceived odour can also be manipulated by odour-colour associations. Morrot et al. changed the perceived odour of wine by experimentally manipulating the colour of wine (i.e., red and white). The experiment was carried out in two sessions separated by 1 week. During the first session, participants were presented with a red and a white wine. They were given a list of odour descriptors (e.g., lychee, honey, floral, pear, blackcurrant, pepper, strawberry, liquorice), and for each descriptor, they were required to indicate which of the two wines presented the character of the descriptor more intensely. During the second session, participants repeated the comparison test using the same descriptors. However, instead of the red wine they were presented with white wine coloured red (using a dye that had no perceptible odour). During the first session, participants used yellow or clear objects to describe the odour of the white wine. On the other hand, during the second session, when the same white wine was coloured red, participant generally used red or darker objects to describe the odour of the wine. Based on these
results Morrot et al. concluded that the wine’s colour provides visual information which significantly affects the olfactory perception of wine.

The experimental work regarding changing perception via classical conditioning is relatively sparse. Nevertheless, the experiments reviewed here provide empirical support for the claim that the perception of a stimulus can be changed by learning to associate it with another stimulus. From an ecological perspective, the change in the perception of a conditioned stimulus is explained in terms of learning to perceive that the conditioned stimulus affords the unconditioned stimulus.

### 3.5.2 Operant Conditioning Experiments

As noted earlier, the ecological theory of perceptual learning is based on the assumption that humans educate their attention to invariants that specify affordances. According to the principles of operant conditioning, humans can be trained to direct attention to certain invariants by manipulating the consequences of their behavioural responses. In perception research, responses to stimuli involve discrimination and identification judgments and the consequences of responding involve corrective feedback. Based on the principles of operant conditioning, corrective feedback in perception experiments can be used to reinforce people to attend to certain aspects of stimuli and ignore other aspects of stimuli.

The experimental investigation of operant conditioning in perception research involves categorization training, which has been referred to as a ‘many-to-one stimulus-response mapping’ (Nosofsky, 1987). During categorization training participants are trained (with the use of feedback) to learn that different stimuli are associated with (i.e., afford) the same category label. The most important aspect of categorization training is that participants learn to discover the invariants that specify the categories (i.e., affordances) by trial and error. That is, participants are not explicitly told what the category relevant dimension is. For instance, an experiment which was based on categorization of morphed faces included instructions as follows: "Your task will be to categorize faces as accurately as possible into different categories (clubs) that are based purely on physical appearance. If you believe the face belongs to Club A, press the A key. Press the B key for Club B" (Goldstone &
Steyvers, 2001, p. 121). It is the corrective feedback after the responses (e.g., key presses) that direct participants’ attention to the category-relevant features.

Goldstone (1994) conducted a series of experiments to examine the effect of categorization on perceptual learning. Stimuli used in one of his experiments were 16 squares that were constructed by factorially combining four values of size and four values of brightness. Participants in the experimental condition were given one hour of categorization training followed by a 40 minute perceptual discrimination task. Participants in the control condition completed the perceptual discrimination task without undergoing the prior categorization training phase. The experimental group was further assigned into one of the three categorization conditions, that were (1) size relevant, (2) brightness relevant, and (3) mixed (i.e., both size and brightness were relevant). Participants were told that some squares belonged to category A and others to category B, and their task was to learn which ones belonged to which categories. On each individual trial, participants were presented with one of the squares and then entered their response on the keyboard. After entering their response participants received feedback about whether the response was correct or incorrect. The perceptual discrimination task consisted of showing participants two squares and they were asked to judge whether the squares were the same or different. It was found that perceptual discriminations were more accurate in the experimental group than in the control group, and this was the case in both the size- and the brightness-relevant categorization conditions. In other words, participants made more accurate perceptual discriminations along a certain dimension if they engaged in categorization training in that dimension which was relevant for categorization.

Experimental support for the effect of categorization on perception includes various stimuli. Some studies used stimuli that differed on a clear-cut dimension such as squares varying in size and brightness (Goldstone, 1994), squares varying in brightness and saturation (Goldstone, 1994), and circles with different sized pie-like wedges (Cross, Lane, & Sheppard, 1965; Livingston, Andrews, & Harnad, 1998). Other studies used stimuli that differ on more complex dimensions, such as random dot patterns (Shin & Nosofsky, 1992; Thomas, Rhoads, Chambliss, 1979), drawings of chick cloaca (Livingston et al., 1998), drawings resembling micro-organisms (Livingston et al., 1998; Schyns & Rodet, 1997), pictures of identical twins.
(Stevenage, 1998), pictures of morphing between unfamiliar faces (Goldstone & Steyvers, 2001), unfamiliar ‘blobby’ shapes (Schyns & Murphy, 1991, 1994), and dermatosis pictures (Norman, Brooks, Coblentz, & Babcock, 1992). Although these studies do not explicitly link their procedures to operant conditioning, they demonstrate that the learning of categories (i.e., affordances) via corrective feedback (i.e., operant conditioning) changes the perception of the categorized stimuli.

3.6 Conclusion

This chapter has been concerned with the application of conditioning principles to the ecological theory of perceptual learning. The ecological theory of perceptual learning, as admitted by Gibson and Gibson (1955b) themselves, mainly focuses on the question of what humans learn to perceive as a result of experience, rather than the question of how learning occurs. Ecological psychologists define perceptual learning as the education of attention to invariants that specify affordances (i.e., meaning) in the environment (Gibson, 2000; Gibson & Pick, 2000; Michaels & Carello, 1981). Regarding the question of how humans learn to perceive affordances, ecological psychologists describe the role of action in perceptual learning (e.g., exploratory activity) (Gibson, 2000; Gibson & Pick, 2000), but they seem to neglect the role of observation in perceptual learning. It has been argued in this chapter that principles of conditioning, which have previously been used to explain a wide range of human behaviour (Domjan, 2005; Rescorla, 1988), can be used to strengthen and expand the ecological account of perceptual learning.

The theory of classical conditioning has been used to expand the ecological theory of perceptual learning by describing how humans learn affordances by observing certain regularities in the environment. Classical conditioning has been interpreted as learning about the affordances of stimuli by observing that a certain stimulus is associated with (i.e., affords) another stimulus. This interpretation expands the ecological theory of perceptual learning since it provides a learning mechanism by which humans can learn affordances without physically interacting with their environment. The theory of operant conditioning has been used to strengthen the ecological theory of perceptual learning by describing how humans learn affordances via acting on their environment. Operant conditioning has been interpreted as
learning about the affordances of stimuli by experiencing the consequences of interacting with the stimuli. Such an interpretation strengthens the ecological argument that action (i.e., exploratory activity) plays an important role in perceptual learning.

It has been argued in this chapter that conditioning changes the perception of stimuli by educating attention to invariants that predict S-S associations (classical conditioning) and the consequences of responding to stimuli (operant conditioning). This argument appears to be supported by the experimental work reviewed in this chapter. In order to further support the explanatory and descriptive power of the proposed ecological-conditioning theory, the next two chapters will describe the role of conditioning in two perceptual domains, namely musical pitch perception (Chapter 4) and speech perception (Chapter 5).
CHAPTER 4

THE ROLE OF CONDITIONING IN PITCH-RELATED PERCEPTUAL LEARNING
4.1 Introduction

The main purpose of this chapter is to demonstrate the descriptive and explanatory power of the proposed ecological-conditioning theory of perceptual learning by applying it to the domain of musical pitch perception. In addition, this chapter also serves as an introductory chapter to Chapter 6, in which the predictive power of the theory is investigated in the pitch perception domain.

According to the ecological theory of perceptual learning, humans educate their attention to meaningful invariants. This chapter describes pitch-related invariants and the question of how learning (musical training in particular) educates attention to musically meaningful invariants in pitch perception.

4.2 Stimuli in Musical Pitch Perception

As discussed in Chapter 2, the concept of stimulus in perception involves both distal properties (i.e., actual objects and events) and proximal properties (i.e., energies impinging on sensory organs). Although distal and proximal stimuli are conceptually distinguished they are inherently related, which can be clearly illustrated in pitch perception. This section will review how properties of sound energy (i.e., proximal stimuli) are determined by the properties of the actual sound source (i.e., distal stimuli).

In the physical sense, sound is described as a mechanical wave, which is created by vibrating bodies, such as a piano string or a vocal cord. Vibrating bodies result in sound waves that travel in a medium (e.g., gas, liquid, solid) by the longitudinal motion of the particles of the medium (i.e., particles move in the direction of the propagation). Since we cannot see the motion of molecules, an example of a coiled spring is often used to illustrate the propagation of a sound wave. A disturbance is created in the spring by the back and forth movement of the first coil of the spring. The first coil pushes and pulls on the second coil, which in turn pushes and pulls on the third coil, and so on. Thus the energy, which is introduced to the first coil, travels from coil to coil, that is from one location to the other.
The sound wave is propagated through a medium, typically through air, from one location to another by the interaction of particles. For example, a vibrating string pushes upon surrounding air molecules moving them toward the neighbouring molecules, which result in changes in the air pressure. The air pressure is high in regions where the air particles are compressed together (i.e., compression) and the air pressure is low in regions where the air particles are spread apart (i.e., rarefaction). Since a sound wave consists of a repeating pattern of high and low pressures it is often referred to as a pressure wave. The fluctuations in pressure occur at periodic intervals that are commonly represented by a sine curve, in which the crest and the troughs of the sine curve correspond to compressions and rarefactions, respectively.

Periodic fluctuations caused by the vibration of a sound source are typically described in terms of their *frequency* and *amplitude*. The frequency of the sound (measured in hertz, Hz) refers to the number of pressure changes that occur in a second. It is important to emphasize that the frequency of the periodic fluctuations in the air pressure (i.e., proximal stimulus) is determined by the actual vibration of the sound source (i.e., distal stimulus). For example, a sound source that vibrates at 500 Hz (i.e., 500 vibrations per second) results in a sound that has the frequency of 500 Hz. The amplitude of the vibration (measured in decibels, dB) corresponds to the amount of energy transferred by the sound wave. The frequency and the amplitude are perceived as functionally different properties of sound that are commonly referred to as the pitch and the loudness of the sound, respectively. Although pitch perception is conceptually linked to the frequency of the sound, it has been noted that amplitude and frequency interact in perception (Neuhoff & McBeath, 1996).

While pitch perception is generally linked to auditory frequencies, definitions of pitch differ with regard to their epistemological perspective. Indirect perceptual theories of pitch regard pitch as a perceptual correlate (i.e., mental representation) of auditory frequency, and direct perceptual theories define pitch as the actual property of sound, which is determined by acoustic frequencies. In this thesis, pitch is regarded as a frequency-related property of sound and it is assumed that pitch perception involves the direct perception of frequency-related invariants.
4.3 Pitch-Related Invariants

As noted in Chapter 2, perception is a holistic phenomenon, in that it is the invariant relation between the elements, rather than the elements in isolation, that are perceived. This tendency can be readily observed in pitch perception. Most naturally occurring sounds are complex (i.e., they consist of more than one frequency) and their perception seems to be determined by the relations among constituent frequencies rather than by individual frequencies per se. In addition to invariant frequency relations found in complex tones, musical pitch perception is also influenced by higher order invariants. Higher orders invariants are embedded in musical contexts such as melodies and harmonies. Melodies and harmonies are built on musical scale structures that involve invariant frequency relations among musical scale notes. Thus, as will be reviewed next, pitch perception is influenced by both invariant frequency relations embedded in complex tones and invariant frequency relations embedded in musical scales.

4.3.1 Invariant Frequency Relations in Harmonic Tones

Complex periodic sounds are commonly referred to as harmonic tones and the frequencies in harmonic tones are called harmonics. The invariant structural feature of a harmonic tone is based the numeric relations amongst its harmonics. The frequency of the second harmonic is double the frequency of the first harmonic, and the frequency of the third harmonic is triple the frequency of the first harmonic, and so on. In other words, higher harmonics are related to the first harmonic by whole number ratios (i.e., 1:2:3:4:5, and so on). Since higher harmonics originate from the first harmonic, it is the frequency of the first harmonic – referred to as the fundamental frequency ($F_0$) - that determines the pitch of a complex harmonic tone.

Although the pitch of a complex harmonic tone corresponds to the fundamental frequency, pitch perception is not solely based on the perception of the fundamental frequency. As noted earlier, humans do not perceive stimulus elements in isolation, but rather they perceive invariant relations among stimulus elements. This tendency in pitch perception is demonstrated by various pitch-related perceptual phenomena such as virtual pitch, sensory consonance, and octave equivalence. These pitch related phenomena - as will now be reviewed - suggest that pitch perception is
characterised by the perception of the invariant frequency relations embedded in harmonic tones.

4.3.1.1 Virtual Pitch

The perception of invariant frequency relations embedded in harmonic tones is often exemplified by the phenomenon of virtual pitch perception. Virtual pitch refers to the pitch of a complex tone in which the fundamental frequency is missing. The perceived pitch of complex tones with a missing fundamental tends to correspond to the fundamental frequency even though the fundamental frequency is missing (Terhardt, Stall, & Seewann, 1981a). An everyday example of virtual pitch is the perception of voice transmitted through the telephone. The fundamental frequency of male speech is not transmitted through the telephone (i.e., low frequencies are filtered out), yet we are able to perceive a pitch corresponding to it. Another example is the perception of complex tones produced by a clarinet. The lower notes of the clarinet consist of only the odd harmonics yet the perceived pitch corresponds to the fundamental frequency (Terhardt, Stall, & Seewann, 1981a).

It has been suggested that even in the presence of the fundamental frequency pitch perception is virtual in the sense that it is induced by the higher harmonics rather than by the fundamental frequency. Psychoacoustic data suggest that in the case of most harmonic tones which do include the fundamental frequency the perceived pitch is actually a virtual pitch (i.e., pitch derived from higher harmonics) (Terhardt, Stall, & Seewann, 1981a). According to Terhardt et al., it is only for tones with high pitch (i.e., when fundamental frequency is higher than 900 Hz) that pitch perception is directly influenced by the fundamental frequency. Thus it appears that perceiving the pitch of the harmonic tone is influenced by the invariant harmonic structure of the tone rather than the individual frequency components.

4.3.1.2 Sensory Consonance

In addition to virtual pitch, the structure of harmonic tones also influences the way pitch is perceived in musical settings. In a musical context, pitch perception is based on the perception of musical intervals (e.g., octave, perfect fifth, perfect fourth, major third, minor third, major sixth, minor sixth, major second, minor second, major
seventh, minor seventh, and tritone). Musical intervals are the product of either successive or simultaneous combination of two complex tones (i.e., melodic and harmonic intervals respectively) and they are defined based on the ratios formed by the fundamental frequencies of the two tones (i.e., 2:1, 3:2, 4:3, 5:4, 6:5, 5:3, 8:5, 9:8, 16:15, 15:8, 16:9, 45:32, respectively) (Krumhansl, 2000). Although intervals are commonly defined in terms of frequency-ratios, it is worth noting that Balzano (1980) suggested an alternative way of conceiving intervals. Balzano noted that every system of \( n \) ratios possesses a cyclic mathematical group structure (i.e., \( C_n \)), which allows the representation of intervals without recourse to ratios.

As Helmholtz (1863) noted, the perception of musical intervals is influenced by the structural property of harmonic tones. Specifically, musical intervals that correspond to the frequency ratios formed by the adjacent harmonics of harmonic tones produce a pleasant perceptual experience (e.g., smoothness, fusion), which is generally referred to as sensory consonance. For example, the octave - which is regarded as the most sensorially consonant interval - corresponds to the ratios formed by the first and the second harmonics (2:1). Similarly, the perfect fifth and the perfect fourth (i.e., the next most sensorially consonant intervals) correspond to the ratios formed by the second and third harmonics (3:2 ratio) and the third and the fourth harmonics (4:3 ratio), respectively. On the other hand, musical intervals that are perceived to be unpleasant or dissonant (e.g., tritone and minor seventh) are based on complex frequency ratios (i.e., 45:32, 16:9, respectively) that do not correspond to the simple frequency ratios formed by the adjacent harmonics of harmonic tones.

The psychoacoustic explanation of sensory consonance focuses on the interference of harmonics shared by the two tones that form a musical interval (Helmholtz, 1885/1954). As Helmholtz pointed out, tones that are related to each other by simple frequency ratios (e.g., 2:1, 3:2) have more harmonics in common than tones whose relation is based on complex frequency ratios (e.g., 16:9, 45:32). In addition to overlapping harmonics, sensory consonance requires that the non-overlapping harmonics are distant enough not to interfere with each other. Non-overlapping harmonics that are close to each other cause beats (i.e., oscillations in amplitude) that result in unpleasant or dissonant sensory experience. Although Helmholtz’s theory has been redefined in terms of novel psychoacoustic concepts (e.g., critical bands,
Plomp, 1966; Plomp & Levelt, 1965) the significant role of simple frequency ratios embedded in harmonic tones is generally recognised by theories of sensory consonance.

The notion of sensory consonance in musical practices is indicated by the prevalence of consonant musical intervals. The predominant use of consonant musical intervals can be observed in various musical cultures. Some cross-cultural studies have found that octaves, perfect fifths, and perfect fourths, are used across various places, such as India, China, Africa, and Western Europe (Burns, 1999; Meyer, 1956; Schellenberg & Trehub, 1994b; Trehub, Schellenberg, & Hill, 1997). Most importantly, the octave (2:1) which is considered the most prevalent ratio entailed in complex tones (i.e., ratio formed by the first and the second harmonic) plays a significant functional role in most (if not all) musical cultures (Burns & Ward, 1982; Dowling & Harwood, 1986).

4.3.1.3 Octave Equivalence

The invariant frequency ratio formed by the first and the second harmonics of harmonic tones (i.e., 2:1 or the octave) plays a special role in pitch perception. In general, two tones whose fundamental frequencies are related to each other by the 2:1 ratio are perceived to be similar. This phenomenon, which is generally referred to as octave equivalence, is explicitly indicated by the use of pitch chromas. Pitch chromas in Western music (i.e., C, C#, D, D#, E, F, F#, G, G#, A, A#, B) are the product of dividing an octave into 12 intervals called semitones, which are logarithmically equal in the well-tempered scale. Musical notes that are an octave (i.e., 12 semitones) apart have the same pitch chroma even though they have different fundamental frequencies. For example, a tone of 261.6 Hz has the same chroma (i.e., C) as a tone of double the frequency (i.e., 523.2 Hz). The notion of octave equivalence appears to be universal which is indicated by the fact that most musical cultures use pitch chromas (Dowling & Harwood, 1986).

Although pitch chromas are used across various cultures, the number of pitch chromas within an octave varies across different cultures. While Western music is built on 12 pitch chromas, other cultures use greater or fewer numbers of pitch chromas. For
example, in the North Indian and Arabic musical systems the octave is divided into 22 and 24 intervals, respectively (Justus & Hustler, 2005). However, regardless of the number of pitch chromas used in various cultures, the functional significance of pitch chromas and octave equivalence appears to be shared by all musical cultures (Dowling & Harwood, 1986). That is, pitch chromas are used to convey musical information (i.e., melodic and harmonic relations) and they enable playing and recognising melodies regardless of their absolute pitch (i.e., transposition).

The phenomenon of octave equivalence has been illustrated by visually representing pitch on a helix in a way that tones that are an octave apart are placed above each other (Revez, 1954; Shepard, 1964). Such a helical arrangement illustrates octave equivalence because it brings tones an octave apart close to each other in spatial proximity. The vertical (i.e., linear) dimension of the helix corresponds to pitch height and it is determined by the fundamental frequency. The horizontal (i.e., circular) dimension of the helix relates to pitch chroma. Although the helical model of pitch is conceptually useful to illustrate octave equivalence it has been suggested to be limited as it does not represent the importance of other sensorially consonant musical intervals (e.g., perfect fifth and perfect fourth) (Shepard, 1987).

The above mentioned perceptual phenomena - virtual pitch, octave equivalence, and consonance – demonstrate the perception of invariant frequency ratios that are naturally embedded in harmonic tones. Since the structure of harmonic tones occurs naturally in the environment it could be argued that the human perceptual system has evolved to innately perceive this structure. In fact, the innate, or at least developmentally early, ability to perceive the structural property of complex tones has been demonstrated by a number of experiments investigating pitch perception in infants. It has been observed that infants perceive the missing fundamental (Clarkson & Clifton, 1984) and they perceive tones that are an octave apart to be more similar than tones that are slightly more or less than an octave apart (Demany & Armand, 1984). Moreover, it was found that infants prefer consonant over dissonant intervals (Trainor & Heinmiller, 1998; Zentner & Kagan, 1998) and they can detect changes to consonant intervals but not to dissonant intervals (Schellenberg & Trehub, 1996). These studies suggest that humans are predisposed to perceive predominant frequency ratios that make up harmonic tones.
Although humans are predisposed to perceive invariant frequency ratios entailed in harmonic tones, this perceptual ability is further increased by perceptual learning. During development, humans are regularly exposed to harmonic tones (e.g., listening to speech and music) which assures that their sensitivity to invariant frequency ratios increases. In addition, during development humans learn to perceive higher-order frequency relations that occur in musical contexts, some of which will now be discussed.

4.3.2 Invariant Frequency Relations in Melodies and Harmonies

Humans not only perceive invariant frequency relations that are embedded in the harmonic tone structure but also invariant frequency relations that are embedded in musical contexts. Musical context is determined by the successive and simultaneous arrangement of musical intervals (i.e., melody and harmony, respectively). An important structural feature of melodies and harmonies is that they are both composed of a small set of discrete pitches, referred to as musical scales. The use of scales in music appears to be universal; however different musical systems use different scales. Most importantly, musical scales within specific cultures are invariant and their invariant structure is defined by the number of constituent notes and the relations among them. For example, the Western major scale is based on 7 of the 12 pitch chromas (e.g., the C major scale consists of the following notes: C, D, E, F, G, A, and B) and the 7 tones are separated by 2, 2, 1, 2, 2, 2, 1 semitones.

Although scale structures, just as the structure of harmonic tones, are invariant, they do not occur naturally but rather they are culturally constructed. Scales are artificially created systems of relationships. As Helmholtz (1863/1954) noted:

The construction of scales and of harmonic tissue is a product of artistic invention, and by no means furnished by the formation or natural function of our ear, as it has been hitherto most generally asserted. Of course the laws of the natural function of our ear play a great and influential part in this result; these laws are, as it were, the building stones with which the edifice of our musical system has been erected. But just as people with differently directed
It is important to note that, although scale structures and the structure of harmonic tones are conceptually distinguished, they are perceptually related. As Helmholtz noted, the construction of scales is greatly influenced by “the natural function of our ear” (i.e., the ability to perceive invariant frequency ratios embedded in harmonic tones). In other words, it could be argued that the scale structures are constrained by the notion of sensory consonance. In fact, the role of sensory consonance can be clearly demonstrated in Western musical scale structures. The 7 notes extracted from the 12 pitch chromas in Western musical scales have been regarded as the maximally self-consonant set of construction (Krumhansl, 2000). That is, when considering all possible intervals within the scale, the most frequently occurring intervals are sensorially consonant (e.g., perfect fifth) and the least frequently appearing intervals are sensorially dissonant (e.g., tritone). The prominence of sensorially consonant intervals has also been observed in other musical scales, such as the Japanese Ritsu mode, the common pentatonic scale, and the common ‘blues’ scales (Huron, 1994).

Since scale structures appear to be shaped by sensory consonance, they could be considered as emergent properties that dynamically arose from the interaction among harmonic tones. An important aspect of emergence is that there is a two-way causal link between elements (i.e., micro levels) and wholes (i.e., macro levels) (Wolf & Holvoet, 2005). That is, elements give rise to emergent structures (i.e., upward causations) and emergent structures constrain the behaviour of constituent elements (i.e., downward causation). The concepts of upward and downward causation in music psychology (and in psychology in general) are often described in terms of sensory and cognitive processes, respectively.

So far, we have discussed sensory processes with regard to musical scales (i.e., the role of sensory consonance in the construction of musical scales) and we will now consider the cognitive processes.
Cognitive processes in pitch perception have been related to the concept of musical consonance. As noted earlier, musical consonance is distinguished from sensory consonance in that it is related to higher-order, culturally constructed, melodic and harmonic relations (i.e., musical scale structure) rather than simply the structural property of harmonic tones. Musical scale structures in Western music are often described in terms of tonality, which refers to the tendency of centring a musical piece around a particular tone referred to as the tonic. In post 17th century Western music, the tonic is considered as a reference point in music because each tone in the scale is perceived in relation to the tonic. The tonic is regarded as the most stable and the most frequently occurring note in the scale and many Western music compositions are said to be in the key of a particular tonic. Other tones in a scale are ordered hierarchically depending on their relation to the tonic. For example, in Western musical scales the tonic, which is the first note of the scale, is most closely related to the third and the fifth tone of the scale.

The perception of scale structure (i.e., musical consonance) is often described as perceiving the expected movement of a melody or harmony. Expected movement of a melody concerns the direction and the size of pitch change of successive tones and expected movement of harmony concerns the progression of chords (i.e., simultaneous tone structures). Although the importance of scale structure has been noted in both melodic and harmonic perception, it is generally more emphasized in the latter. As will now be reviewed, harmonic expectancies are commonly explained in terms of the structural properties of musical scales, whereas melodic expectancies most often related to the statistical properties of scales. Such differences may be rooted in the fact that while the use of melodies is universal across cultures, the use of harmony is unique to Western musical culture (Dowling & Harwood, 1986). Therefore, theories of melodic expectancies tend to focus on statistical properties of scales that are universal, whereas theories of harmonic expectancies focus on the structural property of musical scales that is specific to Western music.

4.3.2.1 Melodic Expectancies

Pitch proximity has been suggested to be a fundamental principle that influences melodic expectancies (Narmour, 1990; Schellenberg, 1996, 1997). Pitch proximity is
regarded as a basic Gestalt grouping principle in audition which states that tones that are close in fundamental frequency are likely to be grouped together (Bregman, 1990). In a melodic context, pitch proximity is described as the listeners’ tendency to expect subsequent (or upcoming) tones in a melody to be close in pitch to the tones they have heard previously (Schellenberg et al., 2002). In general, listeners expect the size of the subsequent interval to be as small (small interval is defined as smaller than 5 semitones) as possible. That is, the size of the most expected interval is 0 semitones (i.e., unison, or the repetition of the last tone) followed by 1 and 2 semitones.

Besides the size of the intervals, listeners also learn to develop melodic expectancies regarding the direction of pitch change (i.e., pitch contour). Expectancies about the change in the pitch contour of the melody are referred to as pitch reversal. Pitch reversal is considered as a second order grouping principle, because it requires a more complex level of processing than pitch proximity. Whereas pitch proximity relates to expectancies based on the last perceived tone, pitch reversal involves the last two perceived tones (i.e., an interval) (Schellenberg et al., 2002). For example, after hearing a large interval (i.e., larger than 7 semitones), the listener tends to expect a reversal in pitch direction (i.e., from ascending to descending or vice versa).

Pitch proximity and pitch reversal have been related to the statistical properties of musical scales. Pitch proximity and pitch reversal are statistically constrained by the distribution of pitch within a melody (i.e., tessitura) and the melody’s freedom of movement (mobility) (Hippel, 2000a, 2000b; Hippel & Huron, 2000). Firstly, tessitura is determined by a limited pitch range which statistically predicts the predominance of small intervals. Secondly, the mobility of a melody is constrained by the statistical notion of “regression toward the mean” which refers to the notion that extreme values of a variable tend to be followed by less extreme (more moderate) values. This constraint sets up a prediction that the more extreme the pitch height of the antecedent tone, the more likely that the following tone moves towards the middle of the tessitura (pitch reversal). Thus pitch proximity and pitch reversal can be explained as general sensitivity to tessitura and mobility.

Although melodic expectancies are commonly explained in terms of the statistical properties of musical scales, experimental studies suggest that the perception of
melodies is clearly influenced by the structural properties of the musical scale. For example, Thompson, Cuddy, & Plaus (1997) found that melodic expectancies reflect the tonal stability of the tonic and preference of scale-notes over non-scale notes. Other studies demonstrated that people are more adept at processing melodies that are composed from scale structures of their culture than melodies that are not (Cuddy, Cohen, & Mewhort, 1981; Cuddy, Cohen, & Miller, 1979). These studies suggest that, in addition to the statistical constraints of musical scales, melodic expectancies are also influenced by the structural properties of musical scales.

4.3.2.2 Harmonic Expectancies

Harmonic structures in Western music involve chords that are the result of the simultaneous combination of three or more tones. The structurally most important chord in Western music is built on the tonic and referred to as the tonic chord. For example, the tonic chord in C major key consists of notes C, E, and G. Tonic chords are perceived to be stable (i.e. musically consonant) and complete without producing any expectation of movement. It is important to note that such a percept is induced by perceiving the chord embedded in the scale structure rather than the chord structure per se. A C major triad (C, E, G) sounds stable in the key of C, but it sounds unstable in the key of F# major. This happens because in the key of F# major the C major triad is not a tonic chord and therefore it produces the expectation of resolution to the tonic (i.e., musically dissonant). Thus the perception of chords is relative to the surrounding scale structure. A stable or musically consonant chord in one key can be perceived as unstable or musically dissonant in another key.

Harmonic expectancies have been investigated using various experimental methods one of which is harmonic priming (Bharucha & Stoeckig, 1986, 1987; Tekman & Bharucha, 1992). Harmonic priming experiments are based on the presentation of a chord (i.e., prime) followed by another chord (i.e., target). Participants are asked to judge whether the target chord was in tune with the prime chord. It is generally found that target chords that are harmonically related to the prime chord are processed faster (i.e., reaction time is shorter) than target chords that are harmonically unrelated to the prime chord (Bharucha & Stoeckig, 1986, 1987; Tekman & Bharucha, 1992). Such
findings suggest that humans perceive the invariant frequency relations embedded in the musical scale.

The perception of scale structures, as opposed to the perception of harmonic tone structures, is not innately given but rather appears to be the result of development and learning. This argument is supported by a study that compared adults’ and infants’ perception of melodies and harmonies (Trainor & Trehub, 1992). Trainor and Trehub (1992) tested the perception of scale structure in Western listeners using a well-structured Western melody. Experimental manipulations included the alteration of a single note of the melody so that it was either outside or remained within the key of the melody. Adults were found to detect the out-of-key change easier than the within-key change. By contrast, infants’ performance did not differ in the two conditions. These findings suggest that the perception of musical scales, unlike the perception of complex tone structure, is not innate but rather the product of learning.

4.4 Pitch-Related Perceptual Learning

As a result of being exposed to music, most adults learn to perceive the pitch of a tone relative to the pitch of another tone. Relative pitch perception involves the perception of both ordinal pitch relations (i.e., perceiving the direction of melodic contour) and interval pitch relations (i.e., perceiving the size of the interval). Relative pitch perception in musically trained populations can be explicitly measured by their ability to assign musical labels (e.g., ascending seventh) to pitch intervals. Although people without musical training are generally not familiar with musical labels, they can demonstrate their relative pitch perception ability by referring to musical intervals by the words of familiar songs (Smith, Kemler, Nelson, Grohskopf, & Appleton, 1994). For example, people without musical training may label an ascending fourth interval as the ‘Here Comes the Bride’ interval and they may label an ascending fifth interval as the ‘Twinkle, Twinkle Little Star’ interval. The ability of non-musicians to label familiar intervals suggests that listening to music educates attention to invariant frequency ratios, and learning in many cases is associative.

Although exposure to music naturally develops sensitivity to invariant frequency relations, pitch-related perceptual learning seems to be enhanced by formal musical
training. For instance, it has been demonstrated that musicians are better at detecting F0 manipulations than non-musicians (Schon, Magne, & Besson, 2004) and the ability to perceive octave equivalence is more enhanced in musically trained people than in musically untrained people (Allen, 1967; Krumhansl & Shepard, 1997; Thurlow & Erchul, 1977). Also, the perception of pitch intervals is better in musically trained people than in musically untrained people (Trainor, Desjardins, & Rockel, 1999; Peretz & Babai, 1992). Neuroimaging studies have revealed that musical training causes anatomical changes in areas in the brain that are associated with music perception. For instance, it has been found that primary auditory cortex (Schneider et. al, 2002) and planum temporale (Schlaug et. al, 1995) are larger in the musicians’ brain than in non-musicians’ brain. Thus empirical research together with neuroimaging studies suggest that that musically trained people are more accurate at pitch perception than people without musical training.

The observed difference between musically trained and musically untrained populations raises the question of what learning processes contribute to enhanced pitch perception in the musically trained population.

If singing in tune is a skill which can be improved by teaching, the question arises as to the best method to employ. The choice of method, however, must rely upon some understanding of the process of learning to sing in tune, assuming that this process can be defined and understood. Without such an understanding the purely pedagogic approach may rely too heavily on chance for success. (Welch, 1985, p. 4-5)

Although musical training often aims to improve singing in tune (i.e., pitch production), it should be noted that pitch production and pitch perception are inherently linked. Therefore perceptual learning processes (in addition to other processes such as muscle control) play a crucial role in learning to sing in tune.

4.4.1 Educating Attention to Meaningful Invariants

Perceptual learning, as defined in Chapter 3, is based on the process of educating attention to meaningful invariants (i.e., affordances). Meaningful invariants in
musical pitch perception involve musical intervals (i.e., frequency ratios) that are used in melodies and harmonies. As a result of being exposed to music, humans educate their attention to invariant frequency relations that are prominent in music. Such perception learning is often marked by shifting attention from the pitch properties of individual tones (i.e., absolute pitch) to pitch relations between tones (i.e., relative pitch). The developmental shift from absolute to relative pitch is supported by the observation that absolute pitch perception in adults is much less common than in infants (Saffran & Griepentog, 2001). In fact, it has been estimated that the population of adults who have the ability to identify or produce the pitch chroma of a tone (i.e., absolute pitch) without external reference is only 1 in 10000 (Bachem, 1955) or 5 in 10000 (Brown et al., 2003).

4.4.2 Perceptual Learning Principles

This section reviews the use of classical and operant conditioning principles in musical training practices.

4.4.2.1 Classical Conditioning

One of the features of musical training is to train students to assign musical labels to musical intervals. This can be viewed as a type of S-S association in which the musical interval is associated with a certain label. An important aspect of this association is that it involves many-to-one relations (or categorization); that is, different musical intervals (i.e. different in a way that they are composed of different tones) are associated with the same label. For example, all intervals in which the tones are separated by five semitones (e.g., C-F, E-A, A-D) are labelled ‘perfect fourth’. According to the proposed ecological-conditioning model of perceptual learning, associating different tone pairs with the same label directs attention to the invariant property of tone pairs (i.e., frequency ratios) that predict the musical interval label. Such perceptual learning results in relative pitch perception, since the invariants that reliably predict interval labels are determined by the pitch relations between the individual notes rather than their absolute pitch.

In addition to relative pitch perception, the development of absolute pitch perception can also be explained in terms of S-S learning principles. Absolute pitch perception
seems to be enhanced by learning to associate verbal labels with individual tones. It has been observed that musical training that is based on associating solfege syllables (i.e., do, re, mi etc.) with individual tones (i.e., ‘fixed do’ system) is more likely to result in absolute pitch perception than musical training in which the solfege syllables can be associated with any tones depending on the key being used in the training exercise (i.e., ‘movable do’ system) (Gregersen, Kowalsky, Kohn, & Marvin, 1999, 2000). Gregersen et al. found that absolute pitch perception was significantly higher in students who were educated by the ‘fixed do’ system as opposed to students that were educated by the ‘movable do’ system.

In addition to associating pitch with abstract verbal labels, musical training also enhances the perceptual learning of pitch relations by associating pitch with concrete visual representations of pitch height. For instance, visual signs used in sight-reading, the use of keyboard, and hand gestures used in solfege direct attention to pitch relations by associating pitch with a spatial position. According to Jones (1971, 1979), children’s understanding of musical concepts of pitch height and their pitch perception can be enhanced by the use of visual representations of pitch height. Jones argued that using the keyboard in a vertical position better represents the concepts of high and low than using the keyboard in a horizontal position. In fact, Jones (1971) experimentally demonstrated that musical training using a vertical-keyboard improves pitch perception and pitch productions skills more than musical training using horizontal-keyboard. The association of pitch with spatial position appears to be a useful teaching tool to direct attention to meaningful invariants in music, especially in children as they have limited understanding of abstract relational concepts.

4.4.2.2 Operant Conditioning

The role of operant conditioning in pitch-related perceptual learning can be linked to the use of feedback in musical training. Pitch related musical training involves student-teacher interaction, which involves the student producing (singing or playing) a musical interval and the teacher providing verbal feedback about the accuracy of the produced interval. It has been experimentally demonstrated that teachers’ verbal feedback improves the accuracy of pitch production (Joyner, 1969, Roberts, 1972). According to the ecological theory of perceptual learning, providing feedback
improves pitch production by directing attention to the frequency ratios that are associated with the production of the intended musical interval.

In the music education literature, the important role of feedback in improving musical skills has been widely acknowledged (e.g., Cobes, 1972; Gould, 1969; Jones, 1979; Joyner, 1969; Porter, 1977; Roberts, 1972; Welch, 1985; Wolner & Pyle, 1933). In order to explain the effect of feedback on pitch production, Welch (1985) proposed that feedback improves vocal pitch accuracy by allowing students to “label accurately any perceived pitch errors in their singing” (p. 14). It is important to realize that the so-called ‘error labelling’ mechanism proposed by Welch is compatible with the principles of operant conditioning; it could be argued that error labelling improves pitch production by directing attention to invariant frequency ratios that predict correct feedback during training.

Although musical pedagogy primarily relies on verbal feedback, it has been demonstrated that vocal pitch accuracy can be improved by providing feedback by other than verbal means. For instance, different studies have demonstrated that visual feedback (e.g., illuminated colours) can improve the accuracy of pitch production (Cobes, 1972; Jones, 1979; Welch, 1983, 1984). It is possible that utilizing both verbal and visual feedback is more beneficial especially in young children who have limited understanding of abstract concepts.

4.4.3 Neural Basis of Perceptual Learning

This section reviews evidence for a possible site in the auditory perceptual system which may allow the learning of invariant frequency relations.

It has been suggested that the perception of frequency ratios occur at the early stages of auditory processing, that is in the organ of Corti (Bell, 2002; Bell & Fletcher, 2004; Blinowska, Jedrezczak, & Konopka, 2005). The organ of Corti, which is regarded as the ‘seat of hearing’ (Rossing et al., 2002) contains a number of hairs cells (sensory receptors) that make synaptic connections to auditory neurons. Each hair cell has a number of cellular projections called stereocilia (Raphael & Altscheler, 2003), which move in a frequency that matches the frequency of the sound source. The movement
of the stereocilia stimulates the hair cells, which in turn excites auditory neurons. This process is based on frequency-related release of neurotransmitters in the synapses between inner hair cells and auditory neurons.

There are two groups of hair cells (inner and outer) that are distinguished based on their location within the cochlea. The inner hair cells are positioned on the inner part of the basilar membrane (i.e., zona arcuata) and the outer hair cells are located on the outer part of the basilar membrane. Most importantly, the functions of inner and outer hair cells are different, which is partly characterized by their different neural connections. The inner hair cells are mainly connected to *afferent* auditory neurons (i.e., sending information to the brain), and outer hair cells are connected to mainly *efferent* neurons (i.e., sending information from the brain). Since it is the inner hair cells that predominantly send information to the brain they are considered the ‘true auditory sensory cells’ (Raphael & Altscheler, 2003) and their role in pitch perception is generally recognised.

In addition to inner hair cells, outer hair cells have also been suggested to play a significant role in pitch perception. An important functional property of outer hair cells is their ability to modulate the activity of inner hair cells. The stereocilia of the outer hair cells have a unique structural feature that allows them to change length (i.e., motility). Length changes in the outer hair cells modify the mechanical vibrations arriving at the inner hair cells (Raphael & Altscheler, 2003) because outer hair cells push against the tectorial membrane, which in turn selectively amplifies the vibration of the basilar membrane. More specifically, outer hair cell motility provides a region of specific amplification that enhances transduction at the inner hair cells (Raphael & Altscheler, 2003). Thus length changes in the outer hair cells increase both the sensitivity and the specificity of inner hair cells.

According to some speculations (Bell, 2002; Bell and Fletcher, 2004; Blinowska, Jedrezjczak, & Konopka, 2005) the detection of frequency ratios occurs at the site of outer hair cells. Outer hair cells are arranged in three rows and they form a regular crystal-like lattice and, according to Bell (2002), such arrangement provides a basis for the perception of harmonic ratios. Bell considers the cochlea as a surface acoustic
wave resonator in which the resonating elements are the reverberation of ripples between the rows of outer hair cells. According to this theory, outer hair cells respond to the fluid pressure in the cochlea, and their movement creates ripples on the overlying tectorial membrane, which proceed to the inner hair cells where they are detected. Bell (2002) pointed out that, based on the geometrical arrangement of outer hair cells, frequency ratios could be signalled by the spatial firing pattern of outer hair cells.

Considering the characteristics of the outer hair cells, such as (1) they are able to modulate the activity of inner hair cells, (2) they are mainly connected to efferent neurons (i.e., receive feedback from the brain), (3) they are arranged in a way that allows the perception of harmonic ratios, suggests that they play an important role in the perceptual learning of invariant frequency ratios.

4.5 Conclusion

It has been argued in this chapter that pitch perception can be understood in terms of perceiving invariant frequency relations that are embedded in both the structure of harmonic tones and musical structures (i.e., melodies and harmonies). The perception of harmonic tone structure has been related to the phenomenon of virtual pitch, sensory consonance, and octave equivalence and the perception of musical structures has been related to melodic and harmonic expectancies.

It has been emphasized that while harmonic tones occur naturally in the environment as a result of pitch production, musical scales are constructed by musical cultures. Since harmonic tone structures are embedded in nature, the human perceptual system has evolved to innately perceive such structure. Although humans seem to perceive the harmonic tone structures at birth, this ability may be further shaped by perceptual learning during development. In addition to increasing sensitivity to harmonic tone structure, perceptual learning during development also increases sensitivity (i.e., educates attention) to higher-order frequency relations entailed in musical scales.

Although perceptual sensitivity to invariant frequency relations develops simply as a result of mere exposure (i.e., listening to music), associative learning further enhances
perceptual learning. Musical training utilizes both classical and operant conditioning to enhance the perception and the production of musical pitch. Classical conditioning typically involves that association of pitch stimuli (i.e., individual pitch or pitch intervals) with verbal labels (i.e., chroma names and musical interval names, respectively.). The associations of individual pitch stimuli with pitch chroma labels educates attention to the absolute properties of pitch and the association of pitch intervals with musical interval labels educates attention to the invariant relations between individual pitches (i.e., frequency ratios). Operant conditioning in musical training involves providing feedback about the musical interval produced by the student, which educates attention to frequency ratios that are considered musically meaningful.

This chapter provides support for the descriptive and explanatory power of the proposed ecological-conditioning theory perceptual learning. In order to provide further support for the descriptive and explanatory power of the theory, the next chapter will apply the principles of conditioning to perceptual learning in a different domain, namely the domain of speech perception.
CHAPTER 5

THE ROLE OF CONDITIONING IN SPEECH-RELATED PERCEPTUAL LEARNING
5.1 Introduction

The main purpose of this chapter is to demonstrate further the explanatory and descriptive power of the proposed ecological-conditioning theory of perceptual learning (see Chapter 3) by applying it to a domain beyond that of musical pitch, namely the domain of speech perception. In addition, this chapter also serves as an introductory chapter to Chapter 7, in which the predictive power of the theory is investigated via experimental work conducted in the speech perception domain.

A theory of perceptual learning should explain what humans learn to perceive and how that learning occurs. According to the ecological theory of perceptual learning, humans learn to perceive meaningful invariants (i.e., affordances), and they are learned by the process of the education of attention. This chapter includes the description of speech-related invariants and the education of attention to invariants that distinguish meaning in speech.

As argued in Chapter 3, a comprehensive theory of learning should not only include the description of learning processes (e.g., educating attention to affordances) but also the description of learning principles, which explain and predict the learning processes. This chapter will examine the extent that the learning principles of conditioning can explain speech-related perceptual learning. As noted in Chapter 3, the principles of conditioning are conventionally described in terms of learning the association between co-occurring stimuli (i.e., classical conditioning) and learning the association between a stimulus and the consequences of interacting with that stimulus (i.e., operant conditioning). In the context of the ecological theory, classical conditioning has been interpreted as learning to perceive that a stimulus affords another stimulus, and operant conditioning has been interpreted as learning to perceive that a stimulus affords certain behaviour (see Chapter 3). In order to demonstrate the descriptive and explanatory power of the proposed ecological-conditioning theory of perceptual learning, this chapter will apply the terminology of the theory to describe the principles of conditioning as they naturally occur during speech-related perceptual learning, and as they have been previously used to induce speech-related perceptual learning in experimental settings.
5.2 Stimuli in Speech Perception

Human speech sounds are produced by the means of articulation. During the production of speech, the vibrating vocal cords provide a harmonic sound source\(^1\), which is modified by certain articulatory gestures. Articulatory gestures, which involve the manipulations of the speech organs (e.g., tongue, jaw, lips, and velum), modify the spectral property of the harmonic sound source by accentuating certain frequencies and attenuating others (Ryalls, 1996). The acoustic properties of speech sounds are conventionally described in terms of the accentuated frequencies, which are referred to as formants (e.g., F1, F2, F3)\(^2\). It is important to emphasize that the frequencies of the formants are determined by the articulatory gestures. For example, moving the tongue body forward raises the frequency of F2, lowering the tongue and jaw raises the frequency of F1, and lip rounding lowers the frequency of F1 and F2 (Gelfand, 1998). These examples illustrate that the auditory properties of the speech sounds and the structural properties of articulatory gestures are intimately related.

The correlation between speech sounds and articulatory gestures has generated a debate about the question of what should be considered as the object (i.e., stimulus) of speech perception. Some researchers (e.g., Fowler, 1986; Liberman & Mattingly, 1985) argue that the object of speech perception is the articulatory event; they consider the acoustic signal as a medium, which allows the detection of the articulatory event. On the other hand, other researchers (e.g., Diehl & Kluender, 1989) argue that the object of speech perception is the acoustic signal of speech and articulatory gestures are the means by which the acoustic signal is produced. The ecological theory offers material for the resolution of this debate. The debate itself originates in the separation of *distal* and *proximal* aspects of the stimulus (i.e., environmental events and energy patterns stimulating the sensory organs, respectively), which has been addressed in Chapter 2. It was argued in Chapter 2 that distal and proximal aspects of stimuli are intrinsically related and referring to one without the other only partially describes perception. Thus, the debate about the

---

\(^1\) As described in the pitch perception chapter (i.e., Chapter 4), a harmonic sound consists of frequency partials that are related by whole number ratios (i.e., 1:2:3:4:5, and so on).

\(^2\) In addition to F1, F2, and F3 speech sounds also contain higher formants; however, it is the first three formants that play the most important role in speech perception.
object of speech perception can be resolved by acknowledging that both the articulatory gesture (i.e., the distal aspect of speech) and the speech sound (i.e., proximal aspect of speech) combine to form the object of speech perception.

An interesting aspect of speech perception is that articulatory gestures can be perceived visually, as well as acoustically. That is, not only can spoken words be heard but they can also be seen (i.e., lip-read). The visual perception of speech (i.e., lip-reading) has been found to aid speech perception, especially in people with hearing problems (e.g., Walden, Prosek, & Worthington, 1974) and in situations when the auditory signal is degraded in any way (Binnie, Montgomery, & Jackson, 1974; Grant, 2001; Grant & Seitz, 2000; Kim & Davis, 2003, 2004; Sumby & Pollack, 1954; Summerfield, 1979). For instance, an experiment conducted by Summerfield (1979) demonstrated that lip-reading increased the recognition of sentences embedded in noise from 23% to 65%. Such experiments support the notion that speech perception is a multi-modal phenomenon, which involves both hearing and seeing articulatory events.

Although during speech perception humans use both auditory and visual speech information (provided that the perceptual system is equipped to do so), the human auditory system is better adapted to perceive speech than is the visual system. The human auditory system is highly developed to process variations in speech sounds that occur (1) in spectral amplitude over frequency at a fixed time, (2) in amplitude over time, and (3) in spectral peaks and valleys over time (Stevens & Blumstein, 1981). On the other hand, the visual system, due to its low temporal acuity, is less effective in perceiving speech information. In fact, experimental evidence suggests that the accuracy of lip-reading is below chance for normal-hearing adults (MacLeod & Summerfield, 1987). Thus, although visual articulation cues enhance speech perception, they are generally considered secondary to auditory speech information.

---

3 The visual system is more effective in resolving two events separated in space than two events separated in time (Handel, 1988; Hirsh & Sherrick, 1961).
5.3 Speech-Related Invariants

Invariants in speech perception involve more complex relations than invariants in pitch perception. While invariants in pitch perception can be mathematically described in terms of frequency ratios (see Chapter 4), invariants in speech perception involve complex relationships among various dimensions of articulatory features (Gracco & Abbs, 1986; Saltzmann & Munhall, 1989). According to Saltzmann and Munhall (1989), articulatory invariants have to be functionally defined in a task-dynamic manner, such as the creation and release of constrictions along different regions of the vocal tract. Such constrictions have been described in terms of tract variables that specify the location and the degree of constriction along the vocal tract (Salzmann & Munhall, 1989). Since articulatory structures determine the properties of speech sounds, speech-related invariants are often described in terms of formants.

The following sections will review some of the invariants that contribute to the perception of speech sounds (i.e., vowels and consonants). Rather than providing an exhaustive review of speech sounds, the focus is on those vowels and consonants which were used in the experiments reported in Chapter 7.

5.3.1 Invariants in Vowel Perception

Vowels are speech sounds that are produced with the mouth open and with the vocal cords vibrating (Ryalls, 1996). The articulatory invariants of vowels are generally described in terms of (1) where the sound is produced in the oral cavity (i.e., front, centre, back), (2) the tongue position (i.e., high, middle, low), and (3) lip-rounding (i.e., rounded and unrounded). As noted before, this section will only review the articulatory and acoustic invariants that specify the vowels that were used in Experiment Set 2, that is, /i/, /a/, /u/.

Vowels /i/, /a/, and /u/ are referred to as the point vowels and they are the three most commonly occurring vowels in the world’s languages (Maddieson, 1984). These vowels are highly different from each other in their articulatory features. Tongue and jaw position during the production of point vowels suggests that the production of point vowels requires more articulatory effort (i.e., greater muscular force) than the production of less commonly used vowels (MacNeilage & Sholes, 1964). During the
production of the vowel /i/, the tongue body is pushed forward and raised close to the palate. During the production of the vowel /u/, the tongue body is similarly raised but maximally retracted. The vowel /a/ is produced by lowering both the tongue and jaw and retracting the tongue body towards the back.

Since articulatory gestures highly differ in point vowels, they have different acoustic properties. The vowel /i/ has relatively low frequency F1 and high frequency F2, vowel /u/ has relatively low frequency for both F1 and F2, and vowel /a/ has relatively high frequency F1 and low frequency F2. As a result, in the acoustic signal of /i/, /u/, and /a/ the energy is concentrated in different frequency regions, that are high (2000-4000Hz), low (200-900 Hz), and mid (700-1200 Hz) frequency ranges, respectively (Diehl & Kluender, 1989). Thus, the invariants that specify point vowels are characterised by maximal dispersion on both articulatory and acoustic dimensions.

5.3.2 Invariants in Consonant Perception

Consonants are speech sounds that are produced with the mouth completely or partially closed (Ryalls, 1996). While vowels can be produced in isolation, consonants require vowels for production. Without vowels, some consonants sound like a chirp-like noise. Since the production of consonants depends on vowels, the perception of consonants is generally investigated using consonant-vowel syllable combinations (e.g., /ba/)

During the production of consonants the vocal tract is constricted causing the sound to be aperiodic or noise-like. Consonants are generally distinguished based on (1) where the obstruction occurs (i.e., place-of-articulation), (2) how the obstruction occurs (i.e., manner of articulation), and (3) whether there is vocal fold vibration (i.e., voicing). As mentioned earlier, this section is not aimed to provide a general review of consonants but rather focuses on the articulatory and acoustic invariants that specify consonants used in Experiment Set 2 (i.e., /b/, /d/, and /g/ consonants).

Consonants /b/, /d/, and /g/ are referred to as stop consonants and they differ from each other based on where the obstruction of the vocal tract occurs. Consonant /b/ is articulated with a closure at the lip end (forward obstruction), /g/ is articulated with a
closure far back on the tongue, and the closure during the articulation of \( /d/ \) is at the
tongue tip, thus lying between these two points of articulation. During the production
of stop consonants the air pressure is first built up behind the obstruction and then
released with a burst of noise, followed by a transition into the following vowel
(Ryalls, 1996).

It has been suggested that the brief period of burst provides invariant information in
stop consonant perception (Liberman et al., 1956). In different stop consonants, the
energy in the burst of noise is concentrated in different frequency regions. For
consonant \( /b/ \) the energy is concentrated at relatively low frequencies (500-1500 Hz).
Compared to \( /b/ \), the energy concentration in consonant \( /d/ \) is at high frequencies
(4000 Hz and higher), and for consonant \( /g/ \) the energy is concentrated at intermediate
frequencies (1500-4000 Hz) (Liberman et al., 1956). It has also been noted that the
gross spectrum shapes at consonant release (sampled over the initial 10-20 msec)
differ by place of articulation in stop consonants (Stevens & Blumstein, 1981). In
consonant \( /d/ \) the spectrum rises, while in consonant \( /b/ \) the gross shape of the
spectrum is falling (although in both consonant \( /b/ \) and \( /d/ \) the spectral energy is
spread over a wide frequency range). The spectral configuration of consonant \( /g/ \) is
characterised by a prominent spectral peak in the mid-frequency region (Stevens &
Blumstein, 1981). These spectral variables are considered as static invariants in stop
consonant perception.

In addition to the onset spectrum shape, dynamic properties of formant transitions
have also been suggested to be invariant in stop consonants (Liberman et al., 1954).
An interesting aspect of formant transitions is that they vary as a function of vowel
context. This phenomenon, which is referred to as co-articulation, has been illustrated
by describing F2 transitions for stop consonants in three different vowel contexts (/a/,
/i/, /u/) (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). In general, the
same consonants followed by different vowels differ in the starting frequency and
slope of the formants. For example, in /di/ the starting frequency of F2 is relatively
high followed by a rising slope, whereas in /du/ the F2 starts relatively low and then
falls to the steady state of the vowel. However, regardless of these differences, it has
been observed that in each stop consonant the locus of F2 transitions is the same (i.e.,
invariant) across different vowel contexts (i.e., 700 Hz for \( /b/ \), 1800 Hz for \( /d/ \), and
3000 Hz for /g/) (Liberman et al., 1954). Thus it is suggested that the locus of F2 transition provides invariant information to identify stop consonants.

Since the main difference among stop consonants lies in the higher frequency regions (i.e., second and third formant transitions), stop consonants are often confused in noisy conditions or over a limited bandwidth transmission channel (e.g., telephone) (Miller & Nicely, 1955). In order to better distinguish stop consonants and other speech sounds that have similar acoustic properties, humans rely greatly on visual articulation cues.

Visual articulation cues seem to enhance speech perception by complementing auditory speech information. Since auditory information is more informative about temporal changes in the environment and visual information is more informative about spatial changes in the environment (Handel, 1988; Hirsh & Sherrick, 1961), speech sounds that are easily confused based on the acoustic cues can usually be easily distinguished based on visual articulation cues (MacDonald & McGurk, 1978; Rober-Ribes, 1998), and vice versa. For example, the acoustic properties of /b/ and /d/ are very similar, whereas the visual cues to their place of articulation are notably different (e.g., consonant /b/ is relatively easy to lip-read because it is articulated at the front of the vocal tract). Thus, as will also be demonstrated in Experiment Set 2, lip-reading is often used to distinguish acoustically similar consonants, such as /b/ and /d/.

5.4 Speech-Related Perceptual Learning

5.4.1 Educating Attention to Meaningful Invariants

In this thesis perceptual learning is defined as educating attention to invariants that specify meaning (i.e., affordances) for the perceiver. In the research literature on speech perception, speech sounds that convey meaning (i.e., afford speaking and understanding speech) are referred to as phonemes. For instance, in the English language, /l/ and /r/ are regarded as different phonemes because they distinguish the meaning of words (e.g., ray and lay have different meanings). Although languages share certain phonemes (e.g., point vowels and stop consonants), some speech sounds only occur in a certain language or languages. For example, the English /r/ - /l/
contrast is absent in many Asian languages (e.g., Japanese and Korean), and the English language lacks some speech sounds used in other languages (e.g., click consonants used in some African languages and dental-retroflex contrasts present in Hindi stop consonants). Speech-related perceptual learning then involves the education of attention to invariants that specify the phonemes of a given language.

Speech-related perceptual learning is often investigated by comparing the perception of native and non-native speech sounds as a function of development (Best, 1994; Trehub, 1976; Werker et. al, 1981; Werker & Tees, 1999). It is generally observed that while infants can discriminate between both native and non-native speech contrasts, adults show significant difficulty in discriminating some non-native speech contrasts. For example, Werker et al., (1981) found that Hindi speech contrasts can be discriminated by English-learning infants aged between 6-8 months but the majority of English-speaking adults are unable to discriminate between the Hindi speech contrasts. These results suggest that humans are born with the ability to distinguish a variety of speech sounds (both native and non-native); however, that ability is modified as a result of linguistic development. During development, humans educate their attention to speech sounds that are relevant in the language that they are learning, and at the same time their attention to irrelevant speech dimensions recedes.

The question of when perception becomes attuned to native phonetic categories has been investigated in developmental studies (e.g., Werker & Tees, 1983; Werker & Lalonde, 1988). Werker and Tees (1983) reported that children as young as 4 years-old have difficulty discriminating non-native speech contrasts, suggesting that speech-related perceptual learning commences at the early stages of language development. Subsequently, Werker and Tees (1984) investigated the perception of phonetic contrasts in 6-12 month-old English-learning infants. Stimuli used in their experiment included the English /ba/-/da/ phonetic contrast, and two non-English phonetic

---

4 Hindi speech contrasts investigated by Werker et. al (1981) are the retroflex vs. dental place-of-articulation contrast (i.e., /Ta/-/ta/) and the voiceless aspirated vs. breathy voiced dental stops (i.e., /tʰa/-/dʰa/).
It was found that 6-8 months-old infants could discriminate all three sets of contrasts, but 10-12 months-old infants could only discriminate the English /ba/-/da/ contrasts. Werker and Lalonde (1988) found similar developmental trends using a different non-native contrast (i.e., Hindi voiced retroflex-dental contrast). These results suggest that the education of human infant attention to invariants that specify their native phonetic categories starts around the age of 8-10 months.

As a result of educating attention to invariants that specify native phonetic categories, adults’ perception of non-native speech sounds is greatly influenced by the extent to which the non-native speech sound resembles a native phonetic category (Best, 1994; Flege, 1995). According to Best (1994) non-native speech sounds that are similar to a native phonetic category are ‘perceptually assimilated’ to that native phonetic category. For example, English-speakers perceive the Hindi dental /t/ and retroflex /t/ to be assimilated to (i.e., phonetically the same as) the English /t/ consonant. On the other hand, non-native speech sounds that are not similar to any particular native phonetic category are either perceived as uncategorizable speech sounds or they are perceived as nonspeech. For example, native English speakers perceive click consonants used in African languages as nonspeech events that resemble ‘cork popping’ or ‘finger snapping’. Best (1994) found that the perceptual discrimination of non-native speech sounds is influenced by their perceptual assimilation to native phonetic categories. In general, non-native speech sounds that are assimilated to different native phonetic categories or to no particular native category are easier to discriminate than non-native speech sounds that are assimilated to the same native phonetic category.

So far, speech-related perceptual learning has been described in terms of the process of educating attention to invariants that specify phonemes. The next section will focus on the role of conditioning in speech-related perceptual learning.

---

5 The non-English contrasts used in the experiment were Hindi voiceless retroflex-dental and Nthlakampx glottalized velar-uvular contrasts.
5.4.2 Learning Principles

The aim of this section is to demonstrate that conditioning principles understood as the learning of affordances provide a useful theoretical framework to understand how speech-related perceptual learning occurs. It will be argued that one example of learning to perceive affordances in speech (i.e., phonemes) is the result of learning to understand speech and learning to produce speech, and that these processes involve classical and operant conditioning.

5.4.2.1 Classical Conditioning

Classical conditioning in speech-related perceptual learning involves learning the relation (i.e., the association) between an invariant phonetic sequence and the object or event it relates to (i.e., affords) in the environment. Such learning is often described in terms of ‘word learning’ which refers to the process of learning the relation between a word and its referent (i.e., what the word refers to, what it means) (Akhtar & Tomasello, 2000).

Word learning begins in the early stages of development and it is driven by the interaction between infants and their caregivers. Caregivers often use gestures to direct infants’ attention to the object or the event to which their speech refers. Zukow-Goldring (1997) reviewed a number of gestures that caregivers use to establish the association between words they say and their referents. The gestures described by Zukow-Goldring involve (1) ‘act-ons’ where caregiver puts the infant through some action (e.g., pulling infants up while saying ‘up’); (2) ‘shows’ where caregivers move the objects towards the infant (e.g., saying ‘ribbit’ while holding and moving a toy frog); (3) ‘demonstrations’ where the caregiver motivates the infant to carry out an action (e.g., inviting the infant to feel the texture of bristles in a broom while saying the word ‘prickly’); and (4) ‘points’ where caregivers point towards the object that their speech refers to. In addition to the gestures initiated by the caregivers, caregivers also respond to gestures initiated by the infant. For instance, caregivers tend to automatically name objects that the infant looks at, points at, picks up, and gives to them (Horne & Lowe, 1996). Thus during the early stages of development, gestures - used by both infants and caregivers – play an important role in establishing the association between spoken words and their referents.
In addition to the use of gestures, word learning is also facilitated by the prosodic features of the speech addressed to infants. The prosody of speech addressed to infants (i.e., infant-directed speech), in comparison to speech addressed to adults (i.e., adult-directed speech), is characterised by slower tempo, higher overall pitch, more distinctive pitch contours, increased amplitude, and longer pauses (Berman, 1990; Fernald, 1992; Gleitman, Newport, & Gleitman, 1984; Grieser & Kuhl, 1988; Snow, 1977; Spieker, Barnett, & McKain, 1983). Several studies demonstrated that prosodic features of infant-directed speech capture infants’ attention more than that of adult-directed speech (e.g., Cooper & Aslin, 1990; Fernald, 1985; Werker, Pegg, & McLeod, 1994). It is possible that increased attention to infant-directed speech plays a crucial role in speech-related perceptual learning. For example, Karzon (1985) demonstrated that 1-4 months old infants can only discriminate certain speech sounds if they are accompanied by an increase in pitch, intensity, and duration (i.e., acoustic features common to infant-directed speech).

When talking to infants, adults not only modify the prosody of their speech but also the articulatory structure of their speech. Articulatory gestures in infant-directed speech are exaggerated, which makes the acoustic properties of phonemes more distinct from each other (Burnham, Kitamura, & Volmer-Conna, 2002; Kuhl et al., 1997). It is possible that enhanced differences in phonemes helps infants learn to identify acoustic invariants that convey meaning (i.e., phonemes) in their language. In fact, it has been demonstrated that well articulated or phonetically-enhanced speech is positively correlated with infants’ ability to discriminate speech sounds (Liu, Kuhl, & Tsao, 2003). Thus, the articulatory structure of infant-directed speech assists infants in establishing the association between phonetic segments and their referents.

In addition to the prosodic and articulatory structure of infant-directed speech, its linguistic properties also play an important role in assisting infants in word learning. Linguistic properties of infant-directed speech include simplified sentence structure, repetition of words, and placing focused words at the end of utterances (Fernald & Mazzie, 1991; Snow, 1977). It has been suggested that such properties of infant-directed speech facilitate word-segmentation (Thiessen, Hill, & Saffran, 2005). In fact, experimental studies demonstrated that placing words at the end of utterances
(Aslin, 2000; Fernald, Pinto, Swingley, Weinberg, & McRoberts, 1998), and repeating words within a sentence (Fernald & Cummings, 2003) enhances infants' word-recognition. Since the linguistic properties of infant-directed speech help infants identify words in a sentence, it is possible that infant-directed speech plays an important role in establishing the association between a word and its referent.

According to the conditioning-ecological theory of perceptual learning, learning the association between a word and its referent (i.e., classical conditioning) directs attention to speech sounds that specify the meaning of the word (i.e., phonemes). This section reviewed some of the environmental conditions (e.g., use of gestures, and infant-directed speech) that enhance such perceptual learning in speech. The next section focuses on the role of operant conditioning in speech related perceptual learning.

### 5.4.2.2 Operant Conditioning

As noted in Chapter 3, operant conditioning refers to learning by experiencing the consequences of behaviour. In this thesis, operant conditioning has been interpreted as a type of perceptual learning in which humans discover affordances as a result of receiving feedback contingent on their behaviour. In speech-related perceptual learning, as will be reviewed now, operant conditioning plays an important role in early development and it involves the interaction between infants’ production of speech and their caregivers’ feedback.

The production of speech develops as a result of being exposed to a particular linguistic environment. Prelinguistic vocalization can be observed in the form of ‘babbling’ when infants produce simple consonant-vowel speech sounds (Dale, 1976). The production of such speech sounds, which is largely based on imitating caregiver’s speech (Kuhl & Meltzoff, 1996), is often followed by feedback from the caregivers. It has been observed that parents’ reactions to their infants’ vocalization is largely influenced by the quality of speech sounds produced by the infant (Beaumont & Bloom, 1993; Bloom & Lo, 1990; Goldstein & West, 1999; Gros-Louis, West, Goldstein, & King, 2006). In general, infants’ vocalization that resembles speech sounds used in their linguistic community is followed by positive social feedback,
which involves smiling at the infant, touching the infant, and repeating what he or she had said. Such feedback educates infants’ attention to speech sounds that are commonly used in their linguistic community.

Experimental studies investigating social interaction between parents and infants support the notion that prelinguistic vocalisation is actively shaped by parental feedback (Bloom, Russel, & Wassenberg, 1987; Goldstein, King, & West, 2003; Hsu, Fogel, & Messinger, 2001; Moerk, 1982; Poulson, Kymissis, Reeve, Andreatos, & Reeve, 1991; Rheingold, Gewirtz, & Ross, 1959; Routh, 1969; Weisberg, 1963). Poulson et al., (1991) have demonstrated that 9-to 13-month-old infants increased their vocal imitation behaviour after they were reinforced for producing an utterance that approximated the utterance produced by an adult. Other studies have found that not only vocal imitation but also the spontaneous production of speech sounds is increased by social reinforcement. Goldstein et al. (2003) investigated the effect of social reinforcement, such as smiling and touching, on 8-month-old infants’ production of speech sounds. Mothers in the experimental condition were instructed to smile and touch infants in response to their infants’ vocalization. Mothers in the control condition were also asked to smile and touch infants, but the timing of these behaviours was determined by the experimenter so as to not be contingent on the infants’ vocalization. It was found that both the quantity and the quality of produced syllables increased more when parents’ smiling and touching behaviour was contingent on infants’ vocalization. These experiments suggest that social reinforcers play an important role in the operant conditioning of prelinguistic verbal behaviour.

So far, classical and operant conditioning have been described as they naturally occur in speech-related perceptual learning. It is argued that describing naturally occurring associations between words produced by the caregiver and their referents (i.e., classical conditioning), and associations between speech sounds produced by the infant and the caregivers’ feedback (i.e., operant conditioning) provides an explanatory framework to understand how speech-related perceptual learning occurs. In order to further support the argument that classical and operant conditioning can be used to induce perceptual learning, the next section will review perceptual training studies, in which conditioning has been artificially manipulated by the experimenter to induce speech-related perceptual learning.
5.5 Perceptual Training Studies

5.5.1 Speech-Related Perceptual Learning induced by Classical Conditioning

Speech-related perceptual training studies using classical conditioning principles are often designed to train participants to use novel perceptual variables in speech perception (Brooks & Frost, 1983; Brooks, Frost, Mason, & Gibson, 1986a, 1986b; Plant, 1998; Stephens & Holt, 2006). Such studies are usually based on associative training in which a perceptual variable that is normally irrelevant or not used in speech perception is consistently paired with speech-related invariants. As a result of such pairings, participants learn (often implicitly) that the initially irrelevant perceptual variable affords, or allows for, the identification of speech-related invariants.

One of the studies, which used classical conditioning principles to induce speech-related perceptual learning, was based on training participants to associate acoustic speech signals with non-speech related visual signals (Stephens & Holt, 2006). The auditory speech signals used in the experiment included /b/, /d/, and /g/ consonants presented as vowel-consonant-vowel tokens and the visual signals were computer generated animated movements that did not resemble articulatory gestures. Participants were presented with consistent consonant-animation pairs across multiple sessions of training. It was found that after training, participants were able to identify consonants based on the animation alone. The results demonstrate that classical conditioning is an effective method to train participants to perceive non-articulatory visual stimuli as affordances of phonetic categories.

Studies based on classical conditioning are often motivated to develop a training device and method that would improve speech perception in the hearing impaired population (Brooks & Frost, 1983; Brooks, Frost, Mason & Gibson, 1986a, 1986b). As described earlier, people with normal hearing and vision rely on both auditory and visual variables to perceive speech. Utilizing variables in both auditory and visual modalities simultaneously results in more accurate speech perception than relying solely on unimodal perceptual variables. Since people with hearing problems have a reduced ability to perceive auditory variables their speech perception can be markedly
deficient. Establishing an additional sensory channel, such as touch, through which phonemic message can be conveyed and perceived would benefit the hearing impaired population.

It has been experimentally demonstrated that, through classical conditioning, people can be trained to use tactile stimuli to perceive speech (Brooks & Frost, 1983; Brooks, Frost, Mason & Gibson, 1986a, 1986b). Tactile stimuli used in such experiments are conveyed by devices that transform the acoustic speech signal into vibrational patterns on the skin. Tactile devices transmit phonemic information by filtering the acoustic waveform into different frequency channels with each channel activating a vibrator on the skin. Tactile training is generally based on wearing tactile devices while lip-reading and/or listening to speech. A single participant study, using a 16 channel tactile device, demonstrated that after 80.5 hours of training, the participant learned to identify 250 words using tactile stimulation and was also able to generalize to words that were not presented during the training (Brooks et al., 1986).

Although tactile stimulation does not replace auditory speech perception, it has been found to benefit speech perception when it is used in conjunction with lip-reading (Brooks et al., 1986; Plant, 1998). Brooks et al. (1986) tested speech perception after tactile training in two conditions; lip-reading only (i.e., unimodal) and lip-reading combined with tactile stimulation (i.e., bimodal). They found that word identification was more accurate in the bimodal condition (79.6%) than the unimodal condition (57.8%). Similarly, Plant (1998) found that after training, participants’ speech perception was 30% better when lip-reading was supplemented with tactile stimulation compared with the lip-reading only condition. These experiments suggest that although the sense of touch, unlike the sense of vision and hearing, has not evolved to perceive speech, people can be conditioned to perceive phonetic sequences presented by tactile devices.

5.5.2 Speech-Related Perceptual Learning induced by Operant Conditioning

Operant conditioning in perceptual training studies is mainly used in the form of giving corrective feedback to participants based on their task performance. As will be reviewed in this section, experiments that have demonstrated speech-related
perceptual learning induced by the use of feedback involve training participants to perceive non-native phonetic contrasts or synthetically created speech sounds.

Several studies demonstrated that Japanese speakers, who are normally not sensitive to the English /r/-/l/ contrast, can be successfully trained to discriminate and/or identify the English /r/ and /l/ phonemes (Bradlow, Pisoni, Yamada, & Tohkura; 1997; Logan, Lively, & Pisoni; 1991; Lively, Pisoni, Yamada, Tohkura, & Yamada; 1994; Strange & Dittmann, 1987). Participants in these experiments underwent extensive training, which was based on the identification of the /r/ and /l/ consonants in a number of naturally produced /r/-/l/ word pairs (e.g., rock-lock) followed by corrective feedback. For instance, perceptual training used in the experiment conducted by Bradlow et al. (1997) involved 45 one-hour sessions over a period of three to four weeks and feedback was given the form of auditory signals (i.e., correct responses were followed by a chime signal and incorrect responses were followed by a buzzer signal). Bradlow et al. found that the accuracy of /r/ and /l/ identification increased from 65% (pre-test) to 81% (post-test) as a result of perceptual training with feedback. These results together with results obtained in similar studies (e.g., Logan et al., 1991; Lively et al., 1994; Strange & Dittmann, 1987) suggest that systematic feedback educated participants’ attention to /r/-/l/ contrasts.

Although perceptual training studies using feedback demonstrate an improvement in perception, the role of feedback in such improvement can only be claimed if it is demonstrated that perceptual training without feedback (i.e., mere exposure) results in less improvement than perceptual training with feedback. In order to explore this question, McCandliss, Fiez, Protopapas, Conway and McClelland (2002) investigated the perceptual learning of /r/-/l/ contrasts (presented in ‘rock-lock’ and ‘road-load’ word-pair contexts) both with and without feedback. In both feedback and no-feedback conditions, participants were presented with a number of discrimination trials using either exaggerated or non-exaggerated /r/-/l/ contrasts (i.e., relatively easy or difficult discriminations, respectively). Perceptual training consisted of three training sessions (each consisting of 480 training trials), which were presented to participants on separate days. It was found that participants presented with the exaggerated stimuli showed improvements in their perception of /r/-/l/ contrasts in both the feedback and the no-feedback conditions. On the other hand, participants
presented with normal exemplars of /r/-/l/ contrasts (i.e., non-exaggerated stimuli) only showed improvement in the feedback condition. These results suggest that providing feedback in perceptual training is most effective when participants are trained to discriminate stimuli that they normally find hard to discriminate.

In addition to training Japanese speakers to discriminate the /r/-/l/ contrast, several studies have been conducted to improve English-speaking adults’ perception of non-native Hindi phonetic contrasts (Werker et al., 1981; Golestani & Zatorre, 2003; Pruitt, Jenkins, and Strange, 2006). Werker et al., (1981) found that just 25 training trials are sufficient to improve English-speaking participants’ ability to discriminate Hindi speech contrasts. The effectiveness of perceptual training to change the perception of non-native speech sounds is also supported by neuroimaging evidence (Golestani & Zatorre (2003). Golestani and Zatorre (2003), who conducted an fMRI study to investigate English-speaking participants’ brain activation during the classification of non-native Hindi contrasts, demonstrated neural plasticity after extensive training (i.e., 5 hours). They found that, after training, the perception of non-native Hindi contrasts involved the activation of brain areas that were not activated before training.

Another study which demonstrated neural plasticity as a result of perceptual training was based on training English-speaking adults to perceive pre-voicing (Tremblay, Kraus, McGee, Ponton, & Otis, 2001). Pre-voicing, which refers to the voice onset time (VOT) that precedes the release of the consonant, is used in some languages (but not in English) to distinguish phonetic categories. Since pre-voicing does not distinguish linguistic meaning in the English language, English-speaking adults generally find it difficult to distinguish pre-voiced /ba/ sounds from non pre-voiced /ba/ sounds. Tremblay et al., (2001) trained English-speaking adults to perceive differences on the pre-voicing dimension and they examined neural changes in the brain that occur during training. Perception of the pre-voicing cue was measured by a two-alternative force-choice identification test, in which participants were required to identify stimuli with -10 and -20 msec VOT as ‘ba’ and ‘mba’, respectively. It was found that perceptual training with feedback increased identification accuracy form 56 % (pre-test) to 82 % (post-test). In addition, it was demonstrated that the increase in
the accuracy of performance was accompanied by changes in neural activity (i.e.,
plasticity).

In addition to training people to perceive non-native phonetic contrasts, speech-related
perceptual training has also been demonstrated using synthetic speech sounds as
stimuli (e.g., Carney, Widin, & Viemesiter, 1977; Samuel, 1977). Carney, Widin, and
Viemesiter (1977) trained people to discriminate synthetically created speech sounds
that ranged along the /ba/-/pa/ continuum. During training, participants were
presented with two consecutive stimuli and were required to report whether the two
stimuli were the same or different. Feedback, which was an essential component in
training, was given after each response. It was found that the discrimination of
stimuli improved after training. Similar findings were obtained in another study using
stimuli ranging on the /da/-/ta/ continuum (Samuel, 1977). These studies, similar to
perceptual training studies using non-native phonetic contrasts, suggest that people
can be conditioned with the use of feedback to educate their attention to speech-
related invariants to which they are normally insensitive.

5.6 Conclusion

The invariants that humans learn to perceive as a result of being exposed to speech
involve articulatory structures that are acoustically mediated by the formant
frequencies (i.e., resonances in the oral cavities). Speech-related perceptual learning
involves the education of attention to articulatory structures that convey meaningful
linguistic messages (i.e., affordances). The smallest meaningful units in speech
perception are called phonemes, which refer to speech sounds that distinguish
linguistic meaning. One form of perceptual learning in speech development is
characterised by an increased ability to discriminate native phonetic contrasts, which
co-occurs with a decreased ability to discriminate non-native phonetic contrasts.
Developmental studies suggest that such learning begins at the early stages of
development (i.e., 8-10 months of age).

In this chapter it has been shown that, at the early stages of development, the
principles of conditioning play an important role in speech-related perceptual learning
related to phonemes and word learning. Classical conditioning involves learning that
a certain word uttered by others refers to (i.e., affords) a certain object or event in the environment. Such learning provides an example of how the attention of infants is educated or attuned to speech sounds that convey meaningful messages in their language. Operant conditioning involves learning about the affordances of speech sounds by producing them (e.g., babbling) and observing others’ reaction to it. Since infants receive social reinforcements for the production of speech sounds that are used in their linguistic environment, the operant conditioning of vocal behaviour provides another example of how human infants educate their attention to native speech sounds.

This chapter together with the previous chapter provides support for the descriptive and explanatory power of the proposed ecological-conditioning theory perceptual learning. The next two chapters will review two sets of experiments that were conducted as part of this thesis to examine the predictive power of the proposed ecological-conditioning theory of perceptual learning.
CHAPTER 6

EXPERIMENTAL SET 1: THE EFFECT OF CONDITIONING ON THE PERCEPTION OF SHEPARD TONES
6.1 Introduction

As noted in Chapter 3, experimental studies that are designed to investigate the effect of conditioning on perception are relatively sparse. In order to better understand the role of conditioning in perception and the extent that perception can be changed via conditioning, more experimental work is needed. The need for more conditioning studies in the perceptual domain has also been pointed out by others suggesting that “Future research will be required to better determine the relationship between conditioning and perceptual learning” (Seitz & Watanabe, 2005, p. 332).

In this chapter a set of experiments is described that were designed to investigate the effect of conditioning on the perception of ambiguous pitch stimuli (i.e., Shepard tones). Shepard tones were chosen as stimuli for the conditioning experiments as it is believed that due to their ambiguous nature they would require fewer training trials to induce changes in perception than non-ambiguous tones. Previous research involving Shepard tones, as reviewed next, has focused on the perception of Shepard tones per se, without investigating the effect of learning on the perception of Shepard tones. Since no previous studies have been designed to modify the perception of Shepard tones via perceptual training, the present experiments not only contribute to the general understanding of the role of conditioning in perceptual learning, but also expand on previous research investigating Shepard tone perception.

6.1.1 Previous Research Investigating the Perception of Shepard Tones

Shepard tones are computer-generated complex tones with ambiguous pitch (Shepard, 1964). Ambiguous pitch in Shepard tones is attained by modifying the invariant structural properties of harmonic tones. As described in Chapter 4, the structural property of harmonic (i.e., non-ambiguous) tones is based on the numeric ratios between the fundamental frequency and the higher harmonics (i.e., 1:2:3:4:5:6, and so on). This structure is modified in Shepard tones by including only the harmonics that are related to the fundamental frequency by an octave. More specifically, Shepard tones consist of 10 sinusoidal frequencypartials spaced at octave intervals (i.e., 1:2:4:8:16:32:64:128:256:512). Excluding harmonics that are normally present in harmonic tones reduces the perceptual cues that specify the pitch (i.e., fundamental
frequency). In addition, Shepard tones are characterized by a fixed bell-shaped spectral envelope of sound pressure level, which directly attenuates the fundamental frequency.

The structure of Shepard tones is often graphically described in terms of collapsing the pitch helix across pitch height and mapping it onto the chroma circle. This analogy is based on the fact that, originally, Shepard (1964) created 12 octave-related complex tones (i.e., Shepard tones) that corresponded to the 12 pitch chroma. Shepard (1964) observed that the perceived pitch contour in two successively presented Shepard tones is generally determined by the shortest distance between the two tones on the chroma circle. For example, an 11 semitone descending shift in the frequency partials is perceived as an ascending pitch contour because the ascending complementary interval is smaller (i.e., 1 semitone). This proximity bias results in a puzzling auditory experience (i.e., Shepard illusion), which is induced by playing a sequence of Shepard tones repeatedly in 1 semitone steps. When the sequence of tones restarts (i.e., F0 jumps back 11 semitones) listeners generally continue to perceive the pitch as increasing or decreasing depending on whether the tones are played in a forward or backward sequence.

The Shepard illusion has been replicated using other octave-related complex tones (e.g., Deutsch, 1987; Repp, 1994), suggesting that the illusion has its origin in octave equivalence. However, Burns (1981) demonstrated that the illusion is also induced by complex tones whose partials are separated by equal ratios other than an octave. Therefore Burns argues that the Shepard illusion should be explained in terms of the principle of proximity rather than the notion of octave equivalence. According to the principle of proximity, the pitch contour of Shepard tones separated by a half-octave (i.e., tritones) is ambiguous because both the ascending and descending intervals are equal (i.e., six semitones).

Several studies investigating the perception of Shepard and other octave-related tones found that different tritones are not perceived equally ambiguously (Deutsch, 1987; Deutsch & Henthorn, 2004; Deutsch, North, & Ray, 1990; Giangrande et al., 2003; Ragozzine, 2002; Repp, 1994, 1997). These findings suggest that, besides the proximity cues, there are other variables that influence the perceived pitch of octave-
related complex tones. Some of these variables that have been suggested to influence the perception of Shepard tritones include the frequency range of the listener’s speech (Deutsch, 1987; Deutsch & Henthorn, 2004; Deutsch et al., 1990), the spectral envelope of the stimuli (Repp, 1994, 1997), and the preceding percept (Giangrande et al., 2003).

Although the role of the above mentioned variables in Shepard tone perception is acknowledged, the present experiments were not designed to test the effect of these variables on Shepard tone perception. Experiments reported in this chapter were designed to investigate the effect of associative learning (i.e., conditioning) on the perception of Shepard tritones.

6.1.2 General Introduction to the Experiments

As argued in Chapter 3, both classical and operant conditioning result in perceptual learning, during which the learner educates his/her attention to invariants that specify affordances in the environment. For instance, as described in Chapter 4, pitch-related perceptual learning results in increased attention to frequency ratios that specify musical intervals. Similarly, as described in Chapter 5, speech-related perceptual learning results in increased attention to invariants that specify native phonetic categories. In addition, developmental studies suggest that perceptual learning is characterised by a decrease in attention to invariants that do not specify affordances. For instance, as a result of learning to perceive and to produce musical intervals, humans’ attention to invariants that specify absolute pitch recedes (See Chapter 4); and as a result of learning to perceive and to produce native speech sounds, humans’ attention to invariants that specify non-native speech sounds recedes.

The aim of the present experiments was to modify pitch perception in Shepard tones using classical and operant conditioning. Meaningful invariants in laboratory conditions (i.e., experimental studies) are set up by the experimenter and they are often referred to as task-relevant stimulus dimensions (i.e., they depend on the task that the experimenter has designed). In the present experiments, participants were
trained to educate their attention to either the F0 contour of Shepard tones (i.e., F0-relevant training) or an auditory dimension that is inconsistent with the F0-contour (i.e., F0-irrelevant training). The task-relevant auditory dimensions were defined by pairing red and green coloured circles with the Shepard tone-pairs (i.e., classical conditioning). In all experiments, tone pairs and the coloured circles were presented simultaneously to participants and they were asked to identify the colour of the circles. In Experiments 2 and 3 participants were also presented with the tone pairs without the coloured circles and they were asked to indicate the colour that they thought the tone pair was associated with. Responses were followed by corrective feedback (i.e., operant conditioning).

Hypotheses were based on the assumption that colour-tone conditioning would direct participants’ attention to the invariant auditory dimension that reliably predicted the colour categories, which would in turn change participants’ perception of Shepard tones. It was expected that F0-relevant conditioning would improve participants’ perception of F0 contour and F0-irrelevant training would impede participants’ perception of F0 contour.

Shepard tones used in the experiments were constructed and their perception was initially investigated by Stevens, Keller, Tyler, and Pressing (1999, unpublished). Thus, prior to conducting the experiments, the raw data collected by Stevens et al. were analysed to investigate the perception of Shepard tritones.

6.2 Pilot Experiment

6.2.1 Aim, Design, and Research Question

The aim of the experiment conducted by Stevens et al. (1999, unpublished) was to replicate Shepard’s (1964) work. The experiment conducted was based on a 2 x 11

---

6 Since pitch has been defined as a holistic rather than elemental phenomenon (See Chapter 4) the term F0 will, in turn, be used in a holistic rather than elemental manner. That is, any reference to the F0 also entails the structural property of the complex tone (i.e., harmonic relations between the fundamental and other frequency partials) which allows perceptual inferences to the F0. This is important to emphasize since in Shepard tones, the fundamental frequency is attenuated thus pitch perception is based on the harmonic partials of a tone.
repeated measures factorial design, in which the independent variables were tone type (Shepard, sinusoidal\(^7\)) and steps clockwise on the circular representation of chromatic pitches (1 step to 11 steps) between the first and second tone of pairs of sequentially presented tones. The perception of tone-pairs was measured by recording whether the second tone was judged to be higher or lower in pitch than the first tone. For the purpose of the current investigation, only the perception of Shepard tritones (i.e., tones separated by six semitones) was analysed. The following research question was proposed: To what extent does the perceived ambiguity of pitch contour differ across different Shepard tritones?

6.2.2 Method

6.2.2.1 Participants

Participants were 24 (16 female and 8 male) first year psychology students from the University of Western Sydney with self reported normal hearing. The mean age of participants was 22.13 years (SD=5.43). The majority of the participants (i.e., 19 participants) had no musical training. Three of the participants had relatively little musical training (i.e., 1-3 years) and two of the participants had more significant musical training (i.e., 6-11 years).

6.2.2.2 Stimuli

Stimuli were 132 pairs of Shepard tones and 132 pairs of sinusoidal tones, that were constructed by using 12 different Shepard tone and 12 different sinusoidal tones, respectively. Shepard tones were generated in CSound. Frequencies and amplitudes for each Shepard tone are shown in the Appendix A. For convenience, tones were numbered from 1 to 12 according to their order based on their F0 (with Shepard tone 1 having the lowest F0). The duration of each tone was 120 msec, and the two tones within a pair were separated by 120 msec.

6.2.2.3 Equipment

The experiment was conducted using a Macintosh computer running SuperLab 1.74 experimental software. Instructions were given verbally by the experimenter and also

\(^7\) Sinusoidal tones were included to establish a baseline of pitch perception in an unambiguous context.
appeared on the computer screen outlining what participants were required to do at each stage. Participants listened to tone-pairs through headphones at a comfortable listening level (approximately 65 dB SPL). Pitch contour judgements were indicated by pressing marked keys on the keyboard. Ascending judgments were indicated by pressing the ‘.’ button on the keyboard, which was marked with a letter ‘U’ (i.e., pitch moves \textit{upward}). Descending judgements were indicated by pressing the ‘0’ button on the keyboard, which was marked with a letter ‘D’ (i.e., pitch moves \textit{downwards}).

\subsection*{6.2.2.4 Procedure}

Participants were presented with each tone-pair once during the experiment. On each trial, participants were asked to indicate whether the pitch of the tone-pair moved upwards or downwards. Responses were entered by pressing marked keys on the keyboard. No feedback was given after responses. The presentation of stimuli was arranged into 4 blocks, with each block consisting of 66 tone-pairs of the same tone type (i.e., either Shepard or sinusoidal tone-pair). The presentation order of blocks was counterbalanced across participants.

\subsection*{6.2.3 Results}

For the purpose of the current investigation, only the perception of Shepard tritones was analysed. The mean percentage of ascending judgments to each Shepard tritone is shown in Figure 6.1.
As Figure 6.1 illustrates, not all tritones yielded pitch contour judgments that could be interpreted as ambiguous (i.e., 50% ascending judgement). In order to measure the extent to which each tritone was perceived ambiguously, the difference between the percentages of ascending and descending\(^8\) pitch contour judgments was analysed using Chi-square non-parametric test\(^9\). The analysis was based on the assumption that if the difference between the percentages of ascending and descending judgments is not significant then the tritone was perceived ambiguously. The Chi-Square results for each Shepard tritone are presented in Table 6.1.

---

\(^{8}\) The percentages of descending judgments can be obtained by subtracting the percentages of ascending judgments from 100.

\(^{9}\) Chi-square non-parametric test was chosen to analyse results because each participant was presented with each tritone once and pitch contour judgments were measured using a binary variable (i.e., either ascending or descending).
Table 6.1

<table>
<thead>
<tr>
<th>Shepard Tritones</th>
<th>1-7</th>
<th>2-8</th>
<th>3-9</th>
<th>4-10</th>
<th>5-11</th>
<th>6-12</th>
<th>7-1</th>
<th>8-2</th>
<th>9-3</th>
<th>10-4</th>
<th>11-5</th>
<th>12-6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2$</td>
<td>8.17</td>
<td>13.50</td>
<td>10.67</td>
<td>10.67</td>
<td>13.50</td>
<td>1.50</td>
<td>8.17</td>
<td>6.00</td>
<td>16.67</td>
<td>13.50</td>
<td>2.67</td>
<td>1.50</td>
</tr>
<tr>
<td>df</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>N</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>p</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.22</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.10</td>
<td>0.22</td>
</tr>
</tbody>
</table>

With alpha set at .05, statistically significant differences between the percentages of ascending and descending judgments were obtained for tritones 1-7, 2-8, 3-9, 4-10, 5-11, 7-1, 8-2, 9-3, and 10-4. Non-significant results were only obtained for tritones 6-12, 11-5, and 12-6. Thus, it could be concluded that only 3 out of the 12 tritones were perceived ambiguously.

It is worth noting that the direction of pitch contour judgments for tritones that were not perceived ambiguously seems to correspond to the direction of F0 contour. That is, for tritones with ascending F0 contour (i.e., 1-7, 2-8, 3-9, 4-10, 5-11), the percentage of ascending judgments was above 50% and for tritones with descending F0 contour (i.e., 7-1, 8-2, 9-3, and 10-4) the percentage of ascending judgment was below 50%.

6.2.4 Discussion

The results suggest that different Shepard tritones are not perceived equally ambiguously. In addition, it appears that pitch contour judgments in Shepard tritones are, to some extent, affected by the F0 contour. It is, however, possible that the observed F0 contour effect was confounded with other variables that were previously reported to affect pitch perception in octave-related tritones. Such variables include the frequency range of the listener’s speech (Deutsch, 1987; Deutsch & Henthorn, 2004; Deutsch et al., 1990), the spectral envelope of the stimuli (Repp, 1994, 1997), and the preceding percept (Giangrande et al., 2003). Since the experiment did not control for these variables, it cannot confidently be concluded that the perception of Shepard tritones is influenced by F0 contour. Nevertheless, the results have an important implication regarding the experimental work reported in this chapter. That
is, since tritones are not perceived equally ambiguously, the extent that perceptual training changes pitch perception may vary across different Shepard tritones. It is possible that conditioning is more effective on Shepard tritones that are initially perceived more ambiguously.

Although Shepard tritones used in conditioning experiments were the same as those used by Stevens et al. (1999, unpublished), data obtained by analysing the result of Stevens et al. cannot be used as a baseline for the conditioning experiments since the perception of Shepard tritones are not only affected by the physical properties of the tones but also the context in which they are presented (Giangrande et al., 2003). In the experiment conducted by Stevens et al. Shepard tritones were interspersed with Shepard tone-pairs that consisted of pitch intervals other than tritone. On the other hand, in the conditioning experiments the perception of Shepard tritones were tested in a context where only Shepard tritones were presented to participants.

6.3 Experiment 1

6.3.1 Aim, Design, and Hypotheses

The aim of Experiment 1 was to modify pitch perception in Shepard tritones using colour-tone associations (i.e., classical conditioning) and corrective feedback (operant conditioning). Training was conducted on all possible pairs of Shepard tones excluding tritones and unisons. After training, pitch contour judgments were tested using tritones constructed of Shepard and sinusoidal tones. Sinusoidal tones were included to investigate whether training on Shepard tones would generalize to pitch perception in tones with unambiguous pitch.

The experiment was based on a between-subjects factorial design with the independent variable being the task-relevant auditory dimension (i.e., auditory dimension that reliably predicted the colour-tone associations). Half of the participants underwent training in which the auditory dimension that reliably predicted the colour-tone associations was the F0-contour (i.e., F0-relevant training). The remaining half of the sample underwent training in which the auditory dimension that reliably predicted the colour-tone associations was the circular dimension of pitch chroma (i.e., F0-irrelevant training). After training, the perception of Shepard tritones
was tested using six ascending tritones and six descending tritones. The dependent variable was the percentage of ascending pitch contour judgements in Shepard tritones.

According to the ecological-conditioning theory of perceptual learning, colour-tone conditioning would direct participants’ attention to the invariant auditory dimension that reliably predicts the colour categories. Thus the perception of Shepard tritones is expected to be more influenced by the F0 contour after F0-relevant conditioning than after F0-irrelevant conditioning. The following related hypotheses were proposed.

1. For Shepard tritones with ascending F0 contour (i.e., 1-7, 2-8, 3-9, 4-10, 5-11, 6-12), it was hypothesised that after training the mean percentage of ascending pitch contour judgments is significantly higher in the F0-relevant training condition than in the F0-irrelevant training condition.

2. For Shepard tritones with descending F0 contour (i.e., 7-1, 8-2, 9-3, 10-4, 11-5, 12-6), it was hypothesised that after training the mean percentage of ascending pitch contour judgments is significantly lower in the F0-relevant training condition than in the F0-irrelevant training condition.

Figure 6.2 below illustrates an idealized set of data, where the F0-contour of each tritone is perceived correctly in the F0-relevant condition and ambiguously in the F0-irrelevant condition.
6.3.2 Method

6.3.2.1 Participants

Participants were 32 Psychology 1 students from the University of Western Sydney, who received course credit for their participation. Participants had no formal musical training in the past and they reported normal hearing and vision. There were 16 participants (1 male and 15 females) in the F0-relevant training condition with a mean age of 19.8 years (SD=3.3 years). These participants were either Australian English monolinguals (5 participants) or they spoke Australian English with the combination of one or more other languages (10 participants). There were 16 participants (3 males and 13 females) in the F0-irrelevant training condition with a mean age of 19.8 years (SD=2.6 years). Six of these participants were Australian English monolinguals and 10 of them spoke Australian English with the combination of one or more languages.

Although data were not analysed as a function of language or languages spoken by participant, linguistic backgrounds were recorded as some researchers (e.g., Deutsch et al., 1990) suggest that it influences the perception of octave-related tritones.
6.3.2.2 *Stimuli*

Auditory stimuli included 132 Shepard tone pairs that were constructed from the 12 Shepard tones used in the experiment conducted by Stevens et al. (1999, unpublished). Frequencies and amplitudes for Shepard tones are shown in Appendix A. For convenience both Shepard and sinusoidal tones were numbered from 1 to 12 according to their order based on their F0 (with tone 1 having the lowest F0 and tone 12 having the highest F0). The duration of each tone was 120 msec, and the two tones within a pair were separated by 120 msec. The F0 contour of the 132 tone-pairs that were used in the experiment is illustrated in Table 6.2.

<table>
<thead>
<tr>
<th>First tone</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>11</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>10</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>9</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>8</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>7</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>6</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>5</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>3</td>
<td>D</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>1</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
</tbody>
</table>

Note: Letter ‘A’ indicates an ascending pair, and letter ‘D’ indicates a descending pair.

Each cell in Table 6.2 represents a pair of Shepard tones. The numbers on the vertical and horizontal axis represent the first tone and the second tone of the pair, respectively. The letters in the cells indicate the F0 contour in the pair (i.e., A= ascending, D= descending).

Shepard tone-pairs, in which the F0s of the constituent tones were separated by 1, 2, 3, 4, 5, 7, 8, 9, 10, and 11 semitones, were used in the training phase. Tone-pairs, in which the F0s were separated by 6 semitones (i.e., tritones), were used in the testing phase. In addition to the Shepard tritones, sinusoidal tritones were also used in the
testing phase to investigate whether training on Shepard tones would generalize to pitch perception in tones with unambiguous pitch.

Visual stimuli used in the experiment included red and green coloured circles that were displayed on the computer screen during the presentation of the Shepard tone-pairs. Other visual stimuli included in the experiment were tick and cross symbols, which were used to indicate correct and incorrect responses, respectively.

6.3.2.3 Equipment

The experiment was conducted using a Macintosh IBook computer running SuperLab 1.74 experimental software. Instructions were given verbally by the experimenter and also appeared on the computer screen outlining what participants were required to do at each stage. Participants listened to tone-pairs through headphones at a comfortable listening level (approximately 65 dB SPL). Pitch contour judgements were indicated by pressing marked keys on the keyboard. Ascending judgments were indicated by pressing the ‘/’ symbol on the keyboard, which was marked with a letter ‘H’ (i.e., second tone is higher in pitch than the first). Descending judgements were indicated by pressing the ‘z’ letter on the keyboard, which was marked with a letter ‘L’ (i.e., second tone is lower in pitch than the first).

6.3.2.4 Procedure

Before commencing the experiment participants read an information sheet about the experiment, they signed a consent form, and filled out a background questionnaire. The information sheet, consent form, and questionnaire are shown in Appendix B, C, D, respectively.

Training Phase

On each training trial, participants listened to a pair of Shepard tones while a red and a green circle appeared on the computer screen. Coloured circles were presented for 125 msec simultaneously with tones and their order (i.e., red-green, green-red) was determined by either the direction of F0 change in the tone pairs or the direction of the smallest frequency change (i.e., circular dimension of pitch chroma).
In the F0-relevant training condition, the order of the coloured circles was determined by direction of frequency change in the F0 (Refer to Table 6.2). For tone pairs with ascending F0 contour, the first tone of the pair was presented simultaneously with the red-coloured circle and the second tone was presented simultaneously with the green-coloured circle. On the other hand, for descending pairs, the first tone of the pair was presented simultaneously with the green-coloured circle and the second tone was presented simultaneously with the red-coloured circle.

In the F0-irrelevant training condition the order of the coloured circles was determined by direction of smallest frequency change in partials (i.e., circular dimension of pitch chroma). The direction of the smallest frequency change in partials for all possible pairs of Shepard tones is illustrated in Table 6.3 below.

<table>
<thead>
<tr>
<th>First tone</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A/D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>11</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A/D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>10</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A/D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>9</td>
<td>A</td>
<td>A</td>
<td>A/D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>8</td>
<td>A</td>
<td>A/D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>7</td>
<td>A/D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>6</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A/D</td>
</tr>
<tr>
<td>5</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A/D</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>4</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A/D</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>3</td>
<td>D</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A/D</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>2</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A/D</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>1</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A/D</td>
<td>D</td>
<td>D</td>
</tr>
</tbody>
</table>

Note: Letter ‘A’ indicates an ascending pair, and letter ‘D’ indicates a descending pair. The frequency contour of the ambiguous tritones is indicated as A/D.

When comparing Table 6.3 with Table 6.2, it appears that the smallest frequency change in the partials corresponds to the F0 contour for small intervals (i.e., intervals smaller than six semitones), however, it does not correspond to the F0 contour for larger intervals (i.e., intervals larger than six semitones). This illustrates that, overall, the frequency change in the smallest partial is a F0-irrelevant dimension.
During training, participants were asked to indicate the colour of the second circle appearing on the computer screen. Participants entered their responses by pressing marked keys on the keyboard (i.e., letter ‘q’ was marked with red colour and letter ‘p’ was marked green colour). After their response, participants received immediate feedback. A correct response was indicated by a tick symbol, and an incorrect response was indicated by a cross symbol. Duration of the feedback was 125 msec and there was a 125 msec inter stimulus interval before the presentation of the next Shepard tone-pair.

Training consisted of eight blocks and each training pair was presented three times during the training. In the first five blocks, training pairs were systematically presented with regard to their pitch interval size (i.e., size of frequency change in F0 in semitones). In the first block, training pairs with pitch interval size 1 and 11 were presented twice in random order. In Block 2, training pairs with pitch interval size 2 and 10 were presented twice in random order. In Block 3, training pairs with pitch interval size 3 and 9 were presented twice in random order. In Block 4, training pairs with pitch interval size 4 and 8 were presented twice in random order. In Block 5, training pairs with pitch interval size 5 and 7 were presented twice in random order. In Blocks 6, 7, and 8 all training pairs were presented again in random order with each block including 40 pairs. Each participant was presented with the same random orders in order to avoid context effects (i.e., previous percept affecting pitch contour judgments).

*Test Phase*

In the second part of the experiment, participants’ pitch contour judgements were tested using Shepard and sinusoidal tritones. The testing phase consisted of two blocks. In the first block, participants listened to all Shepard tritones twice in random order and they were asked to indicate whether the second tone of the pair was higher or lower than the first tone. In the second block, participants listened to all sinusoidal tritones twice in random order, and they were asked to indicate whether the second tone of the pair was higher or lower in pitch than the first tone. Although the random orders used in block 1 and 2 were different from each other, they were constant across participants. Participants received no feedback during testing.
The experiment took approximately 40 minutes and participants were debriefed at the end of the experiment.

6.3.3 Results

The percentages of ascending pitch contour judgments were calculated for each Shepard tritone. Since during testing each tritone were presented twice, for each tritone participants attained one of the following percentage scores: 100% (i.e., 2 ascending judgments), 50% (one ascending and one descending judgment), 0% (i.e., 2 descending judgments). The mean percentages of ascending pitch contour judgements for each Shepard tritone are shown in Figure 6.3.

![Figure 6.3](image.png)

**Figure 6.3.** Mean percentage of ascending pitch contour judgments in Shepard tritones after F0-relevant and F0-irrelevant training.

The mean percentage of ascending judgments for Shepard tritones with ascending F0-contour (i.e., 1-7, 2-8, 3-9, 4-10, 5-11, 6-12) is 62.5% ($SD= 23.96$) in the F0-relevant condition and 63.02% ($SD= 22.56$) in the F0-irrelevant training conditions. An independent $t$ test$^{11}$ indicated no statistically significant difference in the mean

---

$^{11}$ No outliers were identified in either conditions, and the assumption of normality was accepted for both conditions.
percentage scores between F0-relevant and F0-irrelevant conditions, \( t(30) = -0.06, p = 0.95 \).

The mean percentages of ascending judgments for Shepard tritones with descending F0-contour (i.e., 7-1, 8-2, 9-3, 10-4, 11-5, 12-6) is 25% \((SD= 24.72)\) in the F0-relevant condition and 28.13% \((SD= 18.48)\) in F0-irrelevant training condition. Independent \(t\) test\(^{12}\) indicated no statistically significant difference in the mean percentage scores between F0-relevant and F0 irrelevant conditions, \( t(27.8) = -0.41, p = 0.69 \).

Although the mean percentage scores collapsed across the six ascending and six descending Shepard tritones were not significantly different between the two training conditions, Figure 6.3 illustrates that the difference in the percentage scores between the F0-relevant and F0-irrelevant conditions appear to be more pronounced in some tritones (i.e., 2-8, 3-9, 11-5) than in other tritones. Thus it is warranted to investigate the difference between the F0-relevant and the F0-irrelevant percentage scores for each tritone separately. For each tritone, Mann-Whitney nonparametric tests\(^{13}\) were conducted to determine whether the differences in the percentage of ascending judgments between the two training conditions are significant. Statistics for each Shepard tritones are presented in Table 6.4.

Table 6.4
*Mann-Whitney Statistics for the Percentage of Ascending Judgments in Shepard Tritones*

<table>
<thead>
<tr>
<th>Shepard Tritones</th>
<th>U</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-7</td>
<td>93</td>
<td>-1.45</td>
<td>0.146</td>
</tr>
<tr>
<td>2-8</td>
<td>90.5</td>
<td>-1.50</td>
<td>0.134</td>
</tr>
<tr>
<td>3-9</td>
<td>80</td>
<td>-1.97</td>
<td>0.049</td>
</tr>
<tr>
<td>4-10</td>
<td>110.5</td>
<td>-0.77</td>
<td>0.441</td>
</tr>
<tr>
<td>5-11</td>
<td>109.5</td>
<td>-0.82</td>
<td>0.410</td>
</tr>
<tr>
<td>6-12</td>
<td>93</td>
<td>-1.48</td>
<td>0.138</td>
</tr>
<tr>
<td>7-1</td>
<td>120.5</td>
<td>-0.36</td>
<td>0.719</td>
</tr>
<tr>
<td>8-2</td>
<td>121</td>
<td>-0.29</td>
<td>0.768</td>
</tr>
<tr>
<td>9-3</td>
<td>124</td>
<td>-0.17</td>
<td>0.869</td>
</tr>
<tr>
<td>10-4</td>
<td>124</td>
<td>-0.02</td>
<td>0.983</td>
</tr>
<tr>
<td>11-5</td>
<td>127.5</td>
<td>-1.58</td>
<td>0.113</td>
</tr>
<tr>
<td>12-6</td>
<td>90.5</td>
<td>-0.93</td>
<td>0.351</td>
</tr>
</tbody>
</table>

\(^{12}\) No outliers were identified in either condition; however the assumption of normality was only met for the F0-irrelevant condition.

\(^{13}\) Mann-Whitney test was chosen because the percentage scores are not normally distributed, which is presumably due to the fact that the percentage score for each tritone was calculated based on only two observations.
As can be seen in Table 6.4, the Mann-Whitney tests revealed a significant difference between the two training conditions for Shepard tritone 3-9. For Shepard tritone 3-9, the percentage of ascending judgments was significantly greater after F0-relevant training than after F0-irrelevant training.

The perception of sinusoidal tritones was also analysed to examine the extent that training on Shepard tones generalized to pitch perception in a non-ambiguous context. The mean percentages of ascending pitch contour judgements for each sinusoidal tritone are shown in Figure 6.4.

![Figure 6.4. Percentages of ascending pitch contour judgments in sinusoidal tritones after F0-relevant and F0-irrelevant training](image)

The mean percentage of ascending judgments for ascending sinusoidal tritones is 84.38% (SD= 24.51%) in the F0-relevant training condition and 77.08% (SD= 19.6%) in the F0-irrelevant condition. Independent *t*-test\(^{14}\) yielded no significant difference between the two training conditions, *t*(30) = .93, *p* = .36.

The mean percentage of ascending judgments for descending sinusoidal tritones is 6.25 % (SD= 10.76 %) in the F0-relevant training condition and 17.19 % (SD= 17.86 %) in the F0-irrelevant training condition.  

\(^{14}\) No outliers were identified in either condition; however the assumption of normality is violated in both conditions.
% in the F0-irrelevant condition. Independent $t$-test\textsuperscript{15} revealed that the percentage of ascending judgments is significantly lower in the F0-relevant training condition than in the F0-irrelevant condition, $t(30) = -2.1, p = .044$.

Mann-Whitney tests\textsuperscript{16} were used to analyse the difference between the F0-relevant and the F0-irrelevant percentage scores for each tritone separately. Statistics for each sinusoidal tritone are presented in Table 6.5.

Table 6.5

\begin{tabular}{lcccccccc}
& 1-7 & 2-8 & 3-9 & 4-10 & 5-11 & 6-12 & 7-1 & 8-2 & 9-3 & 10-4 & 11-5 & 12-6 \\
\hline
$U$ & 78.5 & 97 & 118 & 109 & 127 & 120 & 120 & 103.5 & 120 & 87.5 & 72 & 100 \\
$z$ & -2.15 & -1.28 & -0.46 & -0.95 & -0.07 & -1.00 & -1.00 & -1.46 & -0.48 & -2.12 & -2.95 & -1.19 \\
$p$ & 0.032 & 0.2 & 0.644 & 0.344 & 0.948 & 0.317 & 0.317 & 0.143 & 0.632 & 0.034 & 0.003 & 0.235 \\
\end{tabular}

As can be seen in Table 6.5, for 3 of the 12 tritones (i.e., 1-7, 10-4, 11-5) the Mann-Whitney tests yielded significant differences between the two training conditions. For tritone 1-7 (i.e., ascending pair), the percentage of ascending judgment is significantly higher in the F0-relevant condition than in the F0-irrelevant condition. For tritones 10-4 and 11-5 (i.e., descending pairs), the percentages of ascending judgment are significantly lower in the F0-relevant condition than in the chroma-relevant condition.

### 6.3.4 Discussion

The aim of Experiment 1 was to investigate the effect of F0-relevant and F0-irrelevant associative training on pitch contour judgments in Shepard tritones. For Shepard tritones with ascending F0 contour, it was hypothesised that after training the mean percentage of ascending pitch contour judgments would be significantly higher in the F0-relevant training condition than in the F0-irrelevant training condition. For Shepard tritones with descending F0 contour, it was hypothesised that after training

\textsuperscript{15} No outliers were identified in either condition; however the assumption of normality is violated in both conditions.

\textsuperscript{16} Mann-Whitney test was chosen because the percentage scores are not normally distributed, which is presumably due to the fact that the percentage score for each tritone was calculated based on only two observations.
the mean percentage of ascending pitch contour judgments would be significantly lower in the F0-relevant training condition than in the F0-irrelevant training condition. None of these hypotheses was supported by the results. Although significant difference between the two training conditions was found for one of the 12 tritones (i.e., 3-9), the difference is interpreted as a random effect rather than a genuine training effect since the hypothesized trend do not systematically occur across the six ascending and six descending tritones.

The analysis conducted on the sinusoidal data revealed a significant difference between the two training conditions. The percentage of ascending judgments for tritones with descending F0-contour was found to be lower in the F0-relevant training condition than in the F0-irrelevant condition. These results suggest that participants in the F0-relevant training condition were more accurate in perceiving the F0-contour of sinusoidal tritones than participants in the F0-irrelevant condition. Although sinusoidal tones were used in the experiment to examine the extent that training on Shepard tone-pairs generalize to the perception of tones with unambiguous pitch, one should be cautious about interpreting the observed results as a genuine effect of training. It is possible that the difference between the two conditions is not due to the training but simply due to a general ability of pitch perception. Although participants in both conditions were randomly sampled from a musically non-trained population, one cannot exclude the possibility that participants in the F0-relevant condition were better at perceiving F0-contour regardless of the associative training they were given. The two possibilities cannot experimentally be teased out since pitch perception was not measured before training - that is no baseline was established.

The finding that the perception of Shepard tritones did not differ as a function of training suggests that training was ineffective to influence pitch perception. The ineffectiveness of the training may have been due the fact that the colour-tone associations used in the F0-irrelevant training, to some extent, overlapped with the colour-tone associations used in the F0-relevant training. The direction of smallest frequency change in partials (i.e., circular dimension of chroma) was used in the F0-irrelevant training, because the direction the smallest frequency change in the partials only corresponds to the F0 contour for small intervals (i.e., intervals smaller than six semitones), but not for large intervals (i.e., intervals larger than six semitones) (See
Table 6.2 and 6.3). However, it should be noted that the combinations of the 12 Shepard tones results in more small intervals (i.e., 90) than large intervals (i.e., 40). Since during training each interval was used the same number of times, the F0-irrelevant training may have, to some extent, directed participants' attention to the F0 contour. Thus, it is possible that the difference between the F0-relevant and F0-irrelevant training was reduced by the overlap between them.

In addition to the overlap between the two training conditions, the actual task that participants were required to do during training may have also contributed to the lack of training effect. During training, each Shepard tone-pair was presented simultaneously with two consecutively occurring coloured circles (red-green or green-red) and participants were required to identify the colour of the second circle. Since participants could perform this task by simply attending to the visual stimuli it is possible that they were not motivated enough to attend to the auditory invariant that predicted the colour of the circles. In addition, if participants did attend to the auditory invariants that predicted the order of coloured circles, it is possible that the use of two circles directed attention to the absolute rather than the relational properties of the tone-pairs. One could argue that associating a single coloured circle with tone-pairs would be more effective to direct participants' attention to the F0 contour of the tone-pairs than associating two coloured circles with the tone-pairs.

The above mentioned limitations were addressed by Experiments 2 and 3, which were designed to examine the effect of F0-relevant and F0-irrelevant training (respectively) on pitch perception by implementing some methodological changes to Experiment 1.

### 6.4 Experiment 2

#### 6.4.1 Aim, Design, and Hypotheses

The aim of the Experiment 2 was to investigate the effect of F0-relevant conditioning on pitch contour judgements in Shepard and sinusoidal tritones, using a within-subjects design with time (i.e., pre-test and post-test) being the independent variable.

The methodology of Experiment 2 is similar to that of Experiment 1; however some aspects of the experiment were changed in order to enhance the effectiveness of
associative training. Firstly, instead of presenting two circles simultaneously with the tone-pairs, in Experiment 2 only one coloured circle was presented simultaneously with the tone-pairs. This change was introduced because it could be argued that a single colour associated with the pair directs attention to the relational properties of tone pairs (i.e., F0-contour) rather than the absolute properties of individual tones (See Chapter 4). Secondly, in Experiment 2, participants were explicitly instructed to pay attention to the auditory properties of Shepard tones during training by being told that some pairs are associated with the colour red and others are associated with the colour green and their task is to learn which pair is associated with which colour. In addition, during training participants were also presented with the Shepard tone-pairs without the coloured circles and they were asked to recall and indicate the colours that were associated with the pairs previously. Responding to Shepard tone-pairs without the presence of coloured circles ensures that corrective feedback (operant conditioning) is better utilized to direct attention to F0-contour. Corrective feedback given when participants are responding to the colour of the circle may not direct attention to the F0-contour as much since correct responses can be made by attending solely to the visual stimuli.

The experiment was based on a within-subjects design. The independent variable was the time of testing with respect to training (i.e., before training and after training). The dependent variable was the percentage of ascending pitch contour judgements in Shepard tritones (i.e., 1-7, 2-8, 3-9, 4-10, 5-11, 6-12, 7-1, 8-2, 9-3, 10-4, 11-5, 12-6). The following related hypotheses were proposed.

1. For Shepard tritones with ascending F0 contour (i.e., 1-7, 2-8, 3-9, 4-10, 5-11, 6-12), it is hypothesized that the mean percentage of ascending pitch contour judgments would be significantly higher after training than before training.

2. For Shepard tritones with descending F0 contour (i.e., 7-1, 8-2, 9-3, 10-4, 11-5, 12-6), it is hypothesized that the mean percentage of ascending pitch contour judgments would be significantly lower after training than before training.
6.4.2 Method

6.4.2.1 Participants

Participants were 19 (2 males and 17 females) Psychology 1 students from the University of Western Sydney, who received course credit for their participation. Participants had no formal musical training in the past and they reported normal hearing and vision. The mean age of participants was 19.9 years (SD=3.1 years). Participants were either Australian English monolinguals (7 participants) or they spoke Australian English with the combination of one or more other languages (12 participants).  

6.4.2.2 Stimuli

Auditory stimuli included 132 Shepard tone pairs that were constructed from the 12 Shepard tones used in the experiment conducted by Stevens et al. (1999, unpublished). Frequencies and amplitudes for Shepard tones are shown in Appendix A. For convenience both Shepard and sinusoidal tones were numbered from 1 to 12 according to their order based on their F0 (with tone 1 having the lowest F0 and tone 12 having the highest F0). The duration of each tone was 120 msec, and the two tones within a pair were separated by 120 msec.

Shepard tone pairs, in which the F0s were separated by 1, 2, 3, 4, 5, 7, 8, 9, 10, and 11 semitones, were used in the training phase. Shepard tone pairs, in which the F0s were separated by six semitones (i.e., tritones), were used in the pre- and post-test phase. In addition to the Shepard tritones, sinusoidal tritones were also used in the pre- and post-test phase to investigate whether training on Shepard tones would generalize to pitch perception in tones with unambiguous pitch.

Visual stimuli used in the experiment included red and green coloured circles to define tone-colour associations, and tick and cross symbols to indicate correct and

---

17 Although data were not analysed as a function of languages spoken by participants, linguistic backgrounds were recorded as some researchers (e.g., Deutsch et al., 1990) argue that it influences the perception of octave-related tritones.
incorrect responses, respectively. Visual stimuli were displayed on the computer screen.

6.4.2.3 Equipment

The experiment was conducted using a Macintosh IBook computer running SuperLab 1.74 experimental software. Instructions were given verbally by the experimenter and also appeared on the computer screen outlining what participants were required to do at each stage. Participants listened to tone-pairs through headphones at a comfortable listening level (approximately 65 dB SPL). Pitch contour judgements were indicated by pressing marked keys on the keyboard. Ascending judgments were indicated by pressing the ‘/’ symbol on the keyboard, which was marked with a letter ‘H’ (i.e., second tone is higher in pitch than the first). Descending judgements were indicated by pressing the ‘z’ letter on the keyboard, which was marked with a letter ‘L’ (i.e., second tone is lower in pitch than the first).

6.4.2.4 Procedure

Before commencing the experiment participants read an information sheet about the experiment, they signed a consent form, and filled out a background questionnaire. The information sheet, consent form, and questionnaire are shown in Appendix E, F, G, respectively.

Pre-Test Phase

In the first part of the experiment, participants’ pitch contour perception was tested using Shepard and sinusoidal tritones. The testing phase consisted of two blocks. In the first block, participants listened to all Shepard tritones twice in random order and they were asked to indicate whether the second tone of the pair was higher or lower in pitch than the first tone. In the second block, participants listened to all sinusoidal tritones twice in random order, and they were asked to indicate whether the second tone of the pair was higher or lower in pitch than the first tone. Each tritone was preceded with a ‘get ready’ sign displayed on the computer screen and participants received no feedback after their response. All participants were presented with the same random order of tritones.
Training Phase

In the second part of the experiment participants underwent F0-relevant associative training. The presentation of either a single red or a single green circle was used to define the direction of F0 change in Shepard tritones. Ascending pairs were presented simultaneously with the green-coloured circle, and descending pairs were presented simultaneously with the red-coloured circle. Coloured circles were presented for 375 msec simultaneously with each tone pair. Similar to Experiment 1, training was conducted on all possible pairs of Shepard tones excluding unisons and tritones. However, in Experiments 2 each training pair was presented 4 times during the training (Note: in Experiment 1, each pair was presented 3 times).

Training consisted of 10 Blocks, and each block consisted of two parts. In the first half of each block, 24 Shepard tone pairs were presented simultaneously with the coloured circles and participants were asked to indicate the colour of the circle. Similarly to Experiment 1, participants’ responses were followed by immediate feedback. A correct response was indicated by the display of a tick, and an incorrect response was indicated by the display of a cross. Duration of feedback was .125 sec and there was a .25 sec inter stimulus interval before the presentation of the next Shepard tone pair. In the second half of each block, participants listened to the same 24 Shepard tone pairs again, but this time the coloured circles did not appear on the computer screen. Rather, participants were asked to recall and indicate the colour that was associated with the pair previously. A correct response was followed by the display of a tick symbol, and an incorrect response was followed by the display of a cross symbol.

In the first 5 Blocks, training pairs were systematically presented with regard to their pitch interval size. In Block 1, training pairs separated by 1 and 11 semitones were presented. In Block 2, training pairs separated by 2 and 10 semitones were presented. In Block 3, training pairs separated by 3 and 9 semitones were presented. In Block 4, training pairs separated by 4 and 8 semitones were presented. In Block 5, training pairs separated by 5 and 7 semitones were presented. In Block 6, 7, 8, 9, 10 all training pairs were presented again irrespective of pitch interval size. In the first half of Block 1, 2, 3, 4, 5 the presentation order of the Shepard tone pairs was randomised, and all participants were presented with the same random order. In the first half of Block 6, 7,
8, 9, 10 and in the second half of each block Shepard tones were presented in various random orders across participants.

Post-Test Phase

In the third part of the experiment, participants’ pitch contour perception was re-tested using exactly the same procedure as in the pre-test. Pitch contour judgments were tested in two blocks. In the first block, participants listened to all Shepard tritones twice in random order and they were asked to indicate whether the second tone of the pair was higher or lower in pitch than the first tone. In the second block, participants listened to all sinusoidal tritones twice in random order, and they were asked to indicate whether the second tone of the pair was higher or lower in pitch than the first tone. Each tritone was preceded with a ‘get ready’ sign displayed on the computer screen and participants received no feedback after their response. All participants were presented with the same random order of tritones.

The experiment took approximately 50 minutes including instructions and breaks. Participants were debriefed at the end of the experiment.

6.4.3 Results

The percentages of ascending pitch contour judgments were calculated for each Shepard tritone. Since during testing each tritone were presented twice, for each tritone participants attained one of the following percentage scores: 100% (i.e., 2 ascending judgments), 50% (one ascending and one descending judgment), 0% (i.e., 2 descending judgments). The percentage of ascending pitch contour judgments before and after training for all Shepard tritones are illustrated in Figure 6.5.
For Shepard tritones with ascending F0 contour (i.e., 1-7, 2-8, 3-9, 4-10, 5-11, 6-12), it is hypothesized that the mean percentage of ascending pitch contour judgments is significantly higher after training than before training. For Shepard tritones with descending F0 contour (i.e., 7-1, 8-2, 9-3, 10-4, 11-5, 12-6), it is hypothesized that the mean percentage of ascending pitch contour judgments is significantly lower after training than before training.

For ascending pairs, the mean percentages of ascending pitch contour judgments before and after training were 70.18 % (SD= 19.51) and 81.58 % (SD=18.55), respectively. A dependent t test indicated that the difference between the before and after training scores is significant, $t(18)=-2.66, p=.016^{18}$.

For descending pairs, the mean percentages of ascending pitch contour judgments before and after training were 27.19 % (SD= 26.48) and 14.91 % (SD=14.85). A

---

18 The normality assumption is met for both conditions and no outliers were identified in either condition.
dependent $t$ test indicated that the difference between the before and after training scores is significant $t(18)=2.59, p=.018^{19}$.

Wilcoxon tests$^{20}$ were used to analyse the difference between the before and after training percentage scores for each tritone separately. Statistics for each Shepard tritone are presented in Table 6.6.

Table 6.6
\textit{Wilcoxon Statistics for the Percentage of Ascending Judgments in Shepard Tritones}

<table>
<thead>
<tr>
<th>Shepard Tritones</th>
<th>1-7</th>
<th>2-8</th>
<th>3-9</th>
<th>4-10</th>
<th>5-11</th>
<th>6-12</th>
<th>7-1</th>
<th>8-2</th>
<th>9-3</th>
<th>10-4</th>
<th>11-5</th>
<th>12-6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$z$</td>
<td>-0.165</td>
<td>-2.126</td>
<td>-1.508</td>
<td>-0.828</td>
<td>-1.897</td>
<td>-1.342</td>
<td>-1.826</td>
<td>-1.134</td>
<td>-1.633</td>
<td>-1.941</td>
<td>-1.081</td>
<td>-0.25</td>
</tr>
<tr>
<td>$p$</td>
<td>0.869</td>
<td>0.033</td>
<td>0.132</td>
<td>0.408</td>
<td>0.058</td>
<td>0.18</td>
<td>0.068</td>
<td>0.257</td>
<td>0.102</td>
<td>0.052</td>
<td>0.279</td>
<td>0.803</td>
</tr>
</tbody>
</table>

As can be seen in Table 6.6, significant difference between the before and after training percentage scores was only found for one of the 12 Shepard tritones (i.e., 2-8).

The perception of sinusoidal tritones were also analysed to examine whether training generalized to pitch perception in non-ambiguous context. The before and after training percentages of ascending pitch contour judgements for each sinusoidal tritone are shown in Figure 6.6.

---

$^{19}$ The normality assumption is only met for the after training condition and no outliers were identified in either condition.

$^{20}$ Wilcoxon test was chosen because the percentage scores are not normally distributed, which is presumably due to the fact that the percentage score for each tritone was calculated based on only two observations.
For ascending pairs, the mean percentages of ascending pitch contour judgments before and after training were 75.88% (SD=21.85) and 84.65% (SD=17.84), respectively. Dependent t test indicated no significant difference between the before and after training scores, t(18)=−2.02, p=.059\textsuperscript{21}.

For descending pairs, the mean percentages of ascending pitch contour judgments before and after training were 25% (SD=25) and 14.47% (SD=19.41). Dependent t test indicated no significant difference between the before and after training scores, t (18)=1.95, p=.065\textsuperscript{22}.

\textsuperscript{21} The normality assumption is only met for the before training condition and no outliers were identified in either condition.

\textsuperscript{22} The normality assumption is only met for the before training condition and no outliers were identified in either condition.
Wilcoxon tests\textsuperscript{23} were used to analyse the difference between the before and after training percentage scores for each tritone separately. Statistics for each sinusoidal tritone are presented in Table 6.7.

Table 6.7

<table>
<thead>
<tr>
<th>Sinusoidal Tritones</th>
<th>1-7</th>
<th>2-8</th>
<th>3-9</th>
<th>4-10</th>
<th>5-11</th>
<th>6-12</th>
<th>7-1</th>
<th>8-2</th>
<th>9-3</th>
<th>10-4</th>
<th>11-5</th>
<th>12-6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$z$</td>
<td>-0.491</td>
<td>-3.051</td>
<td>-1.342</td>
<td>-0.707</td>
<td>-0.378</td>
<td>-0.276</td>
<td>-0.378</td>
<td>-0.513</td>
<td>-2.271</td>
<td>0</td>
<td>-1.613</td>
<td>-1.645</td>
</tr>
<tr>
<td>$p$</td>
<td>0.623</td>
<td>0.002</td>
<td>0.18</td>
<td>0.48</td>
<td>0.705</td>
<td>0.783</td>
<td>0.705</td>
<td>0.608</td>
<td>0.023</td>
<td>1</td>
<td>0.107</td>
<td>0.1</td>
</tr>
</tbody>
</table>

As indicated in Table 6.7, significant differences were found in two of tritones. The percentage of ascending pitch contour judgments in tritone 2-8 is higher after training than before training. The percentage of ascending pitch contour judgments in tritone 9-3 is significantly lower after training than before training.

6.4.4 Discussion

As hypothesized, the percentages of ascending judgments in ascending Shepard tritones increased and the percentages of ascending judgments in descending Shepard tritones decreased after F0-relevant training. Improvement in pitch perception was also observed in two of the sinusoidal tritones, and trends in that direction are evident in most of the sinusoidal tritones. These results suggest that F0-relevant conditioning training improved pitch perception in both Shepard and sinusoidal tones. However, one could argue that the observed improvement in pitch perception may simply be due to mere exposure to Shepard tones rather than conditioning. It is possible that as a result of mere exposure to Shepard tones during training, participants’ sensitivity to the invariant structural property of Shepard tones (i.e., octave-related harmonic structure) increased, which increased their ability to perceive the F0 (i.e., F0 can be perceptually deduced from the octave-related harmonics).

\textsuperscript{23} Wilcoxon test was chosen because the percentage scores are not normally distributed, which is presumably due to the fact that the percentage score for each tritone was calculated based on only two observations.
6.5 Experiment 3

6.5.1 Aim, Design, Hypotheses

The aim of Experiment 3 was to investigate the effect of F0-irrelevant conditioning on the perception of Shepard tritones, using an F0-irrelevant dimension that is maximally inconsistent with the F0-contour dimension. Although Experiment 3 was methodologically the same as Experiment 2, the colour-tone associations in Experiment 3 were determined by the frequency change in the loudest partial. The frequency change in the loudest partial (i.e., partial from 110 to 208 Hz) was chosen as the dimension that determined the colour-tone associations because it is maximally inconsistent with the frequency change in the F0 (See Table 6.8).

Table 6.8
Direction of the frequency change in the loudest partial for all possible pairs of Shepard tones excluding unisons

<table>
<thead>
<tr>
<th>First tone</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>11</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>D</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>A</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>4</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>3</td>
<td>D</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>2</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>1</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
</tr>
</tbody>
</table>

Note: Letter ‘A’ indicates an ascending pair, and letter ‘D’ indicates a descending pair.

Each cell in Table 6.8 represents a pair of Shepard tones. The numbers on the vertical and horizontal axis represent the first tone and the second tone of the pair, respectively. The letters in the cells indicate the frequency change in the loudest partial (i.e., A= ascending, D= descending). When comparing Table 6.8 with Table 6.2, which illustrates the F0 contour for Shepard tone-pairs, it appears that the overlap between the two dimensions is at chance level (i.e., frequency contours only correspond for half of the tone pairs). Thus using the frequency change in the loudest partial to determine colour-tone associations (unlike the chroma dimension used in
Experiment 1) can be considered as a F0-irrelevant training, which is *maximally* inconsistent with the F0-contour dimension.

Similar to Experiment 2, Experiment 3 was based on a within-subjects design. The independent variable was the time of testing with respect to training (i.e., before training and after training). The dependent variable was the percentages of ascending pitch contour judgements in Shepard tritones (i.e., 1-7, 2-8, 3-9, 4-10, 5-11, 6-12, 7-1, 8-2, 9-3, 10-4, 11-5, 12-6).

Since F0 contour does not specify colour categories during training, it is expected that conditioning training would decrease participants’ attention to the F0 contour. Thus the following related hypotheses were proposed.

1. For Shepard tritones with ascending F0 contour (i.e., 1-7, 2-8, 3-9, 4-10, 5-11, 6-12), it is hypothesized that the mean percentage of ascending pitch contour judgments would be significantly lower after training than before training.

2. For Shepard tritones with descending F0 contour (i.e., 7-1, 8-2, 9-3, 10-4, 11-5, 12-6), it is hypothesized that the mean percentage of ascending pitch contour judgments would be significantly higher after training than before training.

6.5.2 Method

6.5.2.1 Participants

Participants were 18 (6 males and 12 females) Psychology 1 students from the University of Western Sydney, who received course credit for their participation. Participants had no formal musical training in the past and they reported normal hearing and vision. The mean age of participants was 20.22 years (SD=3.56 years). Participants were either Australian English monolinguals (8 participants) or they
spoke Australian English with the combination of one or more other languages (10 participants).24

6.5.2.2 Stimuli

Auditory stimuli included 132 Shepard tone pairs that were constructed from the 12 Shepard tones used in the experiment conducted by Stevens et al. (1999, unpublished). Frequencies and amplitudes for Shepard tones are shown in Appendix A. For convenience both Shepard and sinusoidal tones were numbered from 1 to 12 according to their order based on their F0 (with tone 1 having the lowest F0 and tone 12 having the highest F0). The duration of each tone was 120 msec, and the two tones within a pair were separated by 120 msec.

Shepard tone pairs, in which the F0s were separated by 1, 2, 3, 4, 5, 7, 8, 9, 10, and 11 semitones, were used in the training phase. Shepard tone pairs, in which the F0s were separated by six semitones (i.e., tritones), were used in the pre- and post-test phase. In addition to the Shepard tritones, sinusoidal tritones were also used in the pre- and post-test phase to investigate whether training on Shepard tones would generalize to pitch perception in tones with unambiguous pitch.

Visual stimuli used in the experiment included red and green coloured circles to define tone-colour associations, and tick and cross symbols to indicate correct and incorrect responses, respectively. Visual stimuli were displayed on the computer screen.

6.5.2.3 Equipment

The experiment was conducted using a Macintosh IBook computer running SuperLab 1.74 experimental software. Instructions were given verbally by the experimenter and also appeared on the computer screen outlining what participants were required to do at each stage. Participants listened to tone-pairs through headphones at a comfortable

24 Although data were not analysed as a function of languages spoken by participants, linguistic backgrounds were recorded as some researchers (e.g., Deutsch et al., 1990) argue that it influences the perception of octave-related tritones
listening level (approximately 65 dB SPL). Pitch contour judgements were indicated by pressing marked keys on the keyboard. Ascending judgments were indicated by pressing the ‘/’ symbol on the keyboard, which was marked with a letter ‘H’ (i.e., second tone is higher in pitch than the first). Descending judgements were indicated by pressing the ‘z’ letter on the keyboard, which was marked with a letter ‘L’ (i.e., second tone is lower in pitch than the first).

6.5.2.4 Procedure

Before commencing the experiment participants read an information sheet about the experiment, they signed a consent form, and filled out a background questionnaire. The information sheet, consent form, and questionnaire were the same as the ones used in Experiment 2 (See Appendix E, F, G, respectively).

Pre-Test Phase
In the first part of the experiment participants’ pitch contour perception was tested using Shepard and sinusoidal tritones. The testing phase consisted of two blocks. In the first block, participants listened to all Shepard tritones twice in random order and they were asked to indicate whether the second tone of the pair was higher or lower in pitch than the first tone. In the second block, participants listened to all sinusoidal tritones twice in random order, and they were asked to indicate whether the second tone of the pair was higher or lower in pitch than the first tone. Each tritone was preceded with a ‘get ready’ sign displayed on the computer screen and participants received no feedback after their response. All participants were presented with the same random order of tritones.

Training Phase
In the second part of the experiment participants were conditioned to educate their attention to the direction of frequency change in the loudest frequency partial. The presentation of either a single red or a single green circle was used to define the direction of frequency change in the loudest frequency partial. Ascending pairs were presented simultaneously with the green-coloured circle, and descending pairs were presented simultaneously with the red-coloured circle. Coloured circles were presented for 375 msec simultaneously with each tone-pair.
Training was conducted on all possible pairs of Shepard tones excluding unisons and tritones and each training pair was presented 4 times during the training. Training consisted of 10 Blocks, and each block consisted of two parts. In the first half of each block, 24 Shepard tone pairs were presented simultaneously with the coloured circles and participants were asked to indicate the colour of the circle. Participants’ responses were followed by immediate feedback. A correct response was indicated by the display of a tick, and an incorrect response was indicated by the display of a cross. Duration of feedback was 125 msec and there was a 25 msec inter stimulus interval before the presentation of the next Shepard tone pair. In the second half of each block, participants listened to the same 24 Shepard tone pairs again, but this time the coloured circles did not appear on the computer screen. Rather, participants were asked to recall and indicate the colour that was associated with the pair previously. A correct response was followed by the display of a tick symbol, and an incorrect response was followed by the display of a cross symbol.

In the first 5 Blocks, training pairs were systematically presented with regard to their pitch interval size. In Block 1, training pairs separated by 1 and 11 semitones were presented. In Block 2, training pairs separated by 2 and 10 semitones were presented. In Block 3, training pairs separated by 3 and 9 semitones were presented. In Block 4, training pairs separated by 4 and 8 semitones were presented. In Block 5, training pairs separated by 5 and 7 semitones were presented. In Block 6, 7, 8, 9,10 all training pairs were presented again irrespective of pitch interval size. In the first half of Block 1, 2, 3, 4, 5 the presentation order of the Shepard tone pairs was randomised, and all participants were presented with the same random order. In the first half of Block 6, 7, 8, 9, 10 and in the second half of each block Shepard tones were presented in various random orders across participants.

Post-Test Phase
In the third part of the experiment, participants’ pitch contour perception was re-tested using exactly the same procedure as in the pre-test. Pitch contour judgments were tested in two blocks. In the first block, participants listened to all Shepard tritones twice in random order and they were asked to indicate whether the second tone of the pair was higher or lower in pitch than the first tone. In the second block, participants
listened to all sinusoidal tritones twice in random order, and they were asked to indicate whether the second tone of the pair was higher or lower in pitch than the first tone. Each tritone was preceded with a ‘get ready’ sign displayed on the computer screen and participants received no feedback after their response. All participants were presented with the same random order of tritones.

The experiment took approximately 50 minutes including instructions and breaks. Participants were debriefed at the end of the experiment.

6.5.3 Results

The percentages of ascending pitch contour judgments were calculated for each Shepard tritone. Since during testing each tritone was presented twice, for each tritone participants attained one of the following percentage scores: 100% (i.e., 2 ascending judgments), 50% (one ascending and one descending judgment), 0% (i.e., 2 descending judgments). The percentage of ascending pitch contour judgments before and after training for all Shepard tritones are illustrated in Figure 6.7.

![Figure 6.7: Percentage of ascending pitch contour judgments for all Shepard tritones before and after training.](image)

For Shepard tritones with ascending F0 contour (i.e., 1-7, 2-8, 3-9, 4-10, 5-11, 6-12), it is hypothesized that the mean percentage of ascending pitch contour judgments is significantly lower after training than before training. For Shepard tritones with
descending F0 contour (i.e., 7-1, 8-2, 9-3, 10-4, 11-5, 12-6), it is hypothesized that the mean percentage of ascending pitch contour judgments is significantly higher after training than before training.

For ascending pairs, the mean percentages of ascending pitch contour judgments before and after training were 72.69 % (SD= 14.52) and 75.93% (SD=15.36), respectively. Dependent \( t \) test indicated that the difference between the before and after training scores is not significant, \( t(17)=-.65, p=.53 \).

For descending pairs, the mean percentages of ascending pitch contour judgments before and after training were 23.15 % (SD= 24.01) and 18.89 % (SD=22.29). Dependent \( t \) test indicated that the difference between the before and after training scores is not significant \( t(17)=1.41, p=.18 \).

Although the mean percentage scores collapsed across the six ascending and six descending Shepard tritones were not significantly different between the before and after training conditions, it appears (See Figure 6.7) that the difference in the percentage scores between the two conditions is more pronounced in some tritones (i.e., 2-8, 4-10, 8-2) than in other tritones. Thus the difference between the before and after training percentage scores was investigated for each tritone separately. For each tritone, Wilcoxon nonparametric tests\(^{27}\) were conducted to determine whether the differences in the percentage of ascending judgments between the two conditions are significant. Statistics for each Shepard tritones are presented in Table 6.9.

\(^{25}\) The normality assumption is met for both conditions and no outliers were identified in either condition.

\(^{26}\) The normality assumption is not met for either condition and no outliers were identified in either condition.

\(^{27}\) Wilcoxon test was chosen because the percentage scores are not normally distributed, which is presumably due to the fact that the percentage score for each tritone was calculated based on only two observations.
As can be seen in Table 6.9, Wicoxon tests revealed significant differences between the before and the after training scores for tritone 2-8 and 8-2. For tritone 2-8 (i.e., ascending pair) the percentage of ascending judgments significantly increased after training, and for tritone 8-2 (i.e., descending pair), the percentage of ascending judgments significantly decreased after training.

The perception of sinusoidal tritones were also analysed to examine pitch perception in unambiguous context. The before and after training percentages of ascending pitch contour judgements for each sinusoidal tritone are shown in Figure 6.8.

\[\begin{array}{cccccccccccc}
1-7 & 2-8 & 3-9 & 4-10 & 5-11 & 6-12 & 7-1 & 8-2 & 9-3 & 10-4 & 11-5 & 12-6 \\
z & -0.108 & -2.14 & 0 & -1.73 & -0.447 & -1.406 & 0 & -2.271 & -1.732 & -1.732 & -0.447 & -0.812 \\
p & 0.914 & 0.032 & 1 & 0.084 & 0.655 & 0.16 & 1 & 0.023 & 0.083 & 0.083 & 0.655 & 0.417 \\
\end{array}\]

Figure 6.8: Percentage of ascending pitch contour judgments for all sinusoidal tritones before and after training.
For ascending pairs, the mean percentages of ascending pitch contour judgments before and after training were 82.41% (SD=16.14) and 87.5% (SD=14.08), respectively. Dependent t test indicated no significant difference between the before and after training scores, t(17)=-1.64, p=.12.  

For descending pairs, the mean percentages of ascending pitch contour judgments before and after training were 17.13% (SD=21.29) and 11.57% (SD=13.14). Dependent t test indicated no significant difference between the before and after training scores, t(17)=1.51, p=.15.

Wilcoxon tests were used to analyse the difference between the before and after training percentage scores for each tritone separately. Statistics for each sinusoidal tritone are presented in Table 6.10.

<table>
<thead>
<tr>
<th>Sinusoidal Tritones</th>
<th>1-7</th>
<th>2-8</th>
<th>3-9</th>
<th>4-10</th>
<th>5-11</th>
<th>6-12</th>
<th>7-1</th>
<th>8-2</th>
<th>9-3</th>
<th>10-4</th>
<th>11-5</th>
<th>12-6</th>
</tr>
</thead>
<tbody>
<tr>
<td>z</td>
<td>-1.26</td>
<td>-1.00</td>
<td>-0.33</td>
<td>-1.63</td>
<td>-1.34</td>
<td>-1.41</td>
<td>-1.63</td>
<td>-1.26</td>
<td>0.00</td>
<td>-0.33</td>
<td>-1.13</td>
<td>-1.00</td>
</tr>
<tr>
<td>p</td>
<td>0.21</td>
<td>0.32</td>
<td>0.74</td>
<td>0.10</td>
<td>0.18</td>
<td>0.16</td>
<td>0.10</td>
<td>0.21</td>
<td>1.00</td>
<td>0.74</td>
<td>0.26</td>
<td>0.32</td>
</tr>
</tbody>
</table>

As can be seen in Table 6.10, no significant differences were found between the before and after training scores for any of the tritones.

6.5.4 Discussion

It was found that F0-irrelevant training did not change the overall perception of Shepard and sinusoidal tritones. Although significant difference between the two training conditions was found for two of the 12 tritones (i.e., 2-8, 8-2), erring on the

---

28 The normality assumption is not met in either condition and no outliers were identified in either condition.
29 The normality assumption is not met in either condition and no outliers were identified in either condition.
30 Wilcoxon test was chosen because the percentage scores are not normally distributed, which is presumably due to the fact that the percentage score for each tritone was calculated based on only two observations.
side of caution, the difference is interpreted as a random effect since no systematic trend can be observed across the 12 tritones. The results do not support the hypotheses which predicted that F0-irrelevant conditioning would decrease the accuracy of F0 contour perception in Shepard tritones.

When interpreting the results of this experiment, similar to that of Experiment 2, one should consider the possibility that mere exposure effect interacted with the conditioning training. Just as it is possible that mere exposure facilitated the hypothesised trend in Experiment 2, it is possible that mere exposure inhibited the hypothesised trend in the present experiment.

6.6 General Discussion – Experimental Set 1

6.6.1 Summary and Interpretation of Results

The present experimental set investigated the effect of F0-relevant and F0-irrelevant conditioning on the perception of Shepard tritones. Hypotheses were on the assumption that F0-relevant conditioning would improve participants’ perception of F0 contour and F0-irrelevant training would deteriorate participants’ perception of F0 contour in Shepard tritones.

Experiment 1 was based on a between-subjects design in which the independent variable was the relevance of F0 contour during conditioning (i.e., F0-relevant vs. F0-irrelevant conditioning). The results revealed no significant difference between the two training conditions. It was suggested that the lack of difference might have been due to some of the methodological limitations of the study, which were addressed in Experiment 2 and 3. Experiment 2 and 3 investigated that effect of F0-relevant and F0-irrelevant conditioning (respectively) using a within-subjects design with time (pre-test and post-test) being the independent variable. Experiment 2 found that the perception of F0-contour in Shepard tritones improved after F0-relevant condition. Experiment 3 found that F0-irrelevant training did not change the perception of Shepard tritones.

The results of Experiment 2 in comparison with Experiment 3 suggest that conditioning played some role in affecting pitch perception, however the extent that
conditioning influenced pitch perception cannot be concluded since the experiments did not control for mere exposure effect. It is possible that as result of simply being exposed to Shepard tones, participants’ sensitivity to the invariant structural property of Shepard tones (i.e., octave-related harmonic structure) increased, which increased their ability to perceive the F0 (i.e., F0 can be perceptually deduced from the octave-related harmonics). If so, then the improvement in pitch perception after F0-relevant training (Experiment 2) cannot be attributed solely to conditioning, and the lack of change in pitch perception after F0-irrelevant training (Experiment 3) might have been due to the possibility that the negative effect of conditioning was cancelled out by the positive effect of mere exposure. Thus, future experiments are needed to further explore that extent that conditioning can be used to increase or decrease participants’ attention to the F0 dimension. Such experiments should be explicitly designed to include a control condition, in which participants receive no conditioning training, but rather they are simply presented with the tone-pairs the same number of times as participants in the experimental conditions (i.e., F0-relevant and F0-irrelevant conditioning).

6.6.2 Explanation of Results in Terms of the Ecological Theory of Perceptual Learning

As proposed in Chapter 3 the effect of conditioning on perception can be explained in terms of the ecological theory of perceptual learning. It was argued in Chapter 3 that conditioning changes the perceived meaning of stimuli (i.e., learning new affordances) during which attention is directed to invariant stimulus properties that relate to the affordances of stimuli. In conditioning experiments, such as the present experiments, the affordance of a stimulus is determined by the experimenter by pairing the stimulus with another stimulus (i.e., classical conditioning) or by making reinforcers contingent on responding to the stimulus (i.e., operant conditioning). In the present experiments, classical conditioning was established by pairing colours with Shepard tones and operant conditioning was established by giving corrective feedback after participants responded to the colour-tone associations. It is suggested that, such conditioning may increased participants’ attention to the invariant property of Shepard tone-pairs that predicted the colour-tone associations and decreased their attention to auditory invariants that did not specify the colour categories.
6.6.3 Implications of the Results

6.6.3.1 Implications for Shepard tone research

The results of the present experimental set suggest that research in the area of Shepard tone perception would benefit by considering the possibility that F0 plays a significant role in the perception of Shepard tritones. Although perceptual cues to the F0 (i.e., harmonic content) are reduced in Shepard tones, the present experiments suggest that the accuracy of perceiving F0-contour in Shepard tritones is above chance even before participants are conditioned to increase their attention to it. The pilot experiment together with the baseline data of Experiments 2 and 3 indicate that pitch contour judgment in Shepard tritones are mostly in accordance with F0 contour. That is, the percentages of ascending judgments in Shepard tritones with ascending F0 contour tend to be above 50% and the percentages of ascending judgments in Shepard tritones with descending F0 contour tend to be below 50%.

The perception of F0 in Shepard tritones can be explained in terms of holistic models of pitch perception, which propose that F0 can be perceptually deduced from the higher harmonics. Although some of the harmonics are removed in Shepard tones, the octave-related harmonics are retained that seem to be sufficient to allow participants to perceptually infer the F0 in tritones. This is not to say that other variables\textsuperscript{31} do not influence the perception of Shepard tritones, but rather suggests that one should consider the possibility that the F0 of Shepard tones is not as ambiguous as previously thought. Of course, sceptically we might argue that the observed sensitivity to F0 contour is not a genuine effect but due to a confound with some other variable (e.g., frequency range of listener’s speech, or preceding percept). Future research is encouraged to further examine the role of F0 in Shepard tritone perception using a design that controls for confounding variables, where possible, that may also influence the perception of Shepard tones.

\textsuperscript{31} Variables that were previously reported to influence the perception of octave-related tritones include the frequency range of the listener’s speech (Deutsch, 1987; Deutsch & Henthorn, 2004; Deutsch et al., 1990), the spectral envelope of the stimuli (Repp, 1994, 1997), and the preceding percept (Giangrande et al., 2003).
6.6.3.2 *Implications for Pitch Perception Research*

Although the experiments reported in this chapter used complex tones that do not occur naturally in the environment, the observed results do have implications for musical pitch perception in natural settings. As described in Chapter 5, musical education, to some extent, already utilizes conditioning principles to enhance relative pitch perception (e.g., associating relative pitch with musical intervals labels). The results of Experiment 2, in combination with Experiment 3, support the idea that pitch perception can be improved by simple conditioning training, and encourages researchers in music education to further investigate the role of conditioning in musical training. For instance, in addition to training children to label pitch stimuli with musical interval labels, musical training could utilize pairing tones with visual stimuli (i.e., classical conditioning) which could further direct attention to relative pitch. The principle of operant conditioning emphasizes the role of feedback in improving musical skills. Teachers are encouraged to use a variety of feedback that involves not only verbal feedback but feedback provided by visual presentation of reinforcing stimuli.

Future research could investigate the effect of conditioning on pitch perception in real life settings, and with the use of harmonic tones (i.e., tones with unambiguous pitch). Music education research could benefit form studies investigating perceptual learning induced by pairing tones with concrete stimuli (e.g., visual stimuli) in addition to associating tones with abstract verbal labels (as commonly used in music education). It is possible that the use of visual stimuli during conditioning would benefit younger children whose understanding of abstract relations is limited\(^\text{32}\). Similarly, it is possible that young children would benefit more from feedback that is based on the presentation of concrete visual stimuli rather than verbal instruction that involves abstract concepts.

---

\(^{32}\) It has been suggested that children before the age of seven do not completely understand the logic of relations (Piaget, 1928)
6.6.3.3 Implications for Perceptual Training Studies

As noted in Chapter 3, the effect of conditioning on perception is not widely acknowledged and has not been extensively researched. The present experiments expand the existing literature on conditioned precepts by investigating the effect of conditioning on perception using stimuli (i.e., Shepard tones) that have not previously been used in conditioning studies. The finding that conditioning principles play some role in changing pitch perception encourages researchers in the area of perceptual learning to further investigate the effect of conditioning principles on perception using various stimuli. Most importantly, researchers investigating the effect of conditioning on perception are advised to control for or further investigate potential effects of mere exposure.

The next chapter describes a set of experiments that were conducted to further investigate the effect of conditioning in a different domain, namely speech perception.
CHAPTER 7

EXPERIMENTAL SET 2: THE EFFECT OF CONDITIONING ON THE PERCEPTION OF McGURK SYLLABLES
7.1 Introduction

The motivation behind the experimental work reported in this chapter is similar to that behind the previously reported experimental set (i.e., Shepard tone experiments). That is, the aim of the present experimental set is to investigate the effect of conditioning on the perception of ambiguous perceptual stimuli. As noted earlier, ambiguous stimuli are chosen as stimuli for the experiments in this thesis as it is assumed that due to their ambiguous nature they require fewer training trials for their perception to be modified. The ambiguous stimuli used in the present experiments consist of incongruent audio-visual speech stimuli that are commonly referred to as McGurk stimuli. Before moving on to the experiments, we will review previous research that has been conducted on the perception of McGurk stimuli.

7.2 Previous Research Investigating the Perception of McGurk Syllables

McGurk stimuli are ambiguous speech stimuli that are constructed by an incongruent pairing of auditory and visual speech syllables. The effect of audio-visual incongruency on perception was first reported by McGurk and MacDonald (1976), who created an incongruent audio-visual stimulus by dubbing an audio signal of /ba/ onto the visual articulation of /ga/. Audio /ba/ paired with visual /ga/ (i.e., McGurk stimulus) is ambiguous in nature because the auditory and the visual cues to the phonetic message are contradictory (people in general are unaware that the audio and visual speech syllables are mismatched).

People often resolve the ambiguity of McGurk stimuli by combining or integrating the auditory and visual cues, which often results in perceiving a phoneme, which corresponds neither to the audio signal nor to the visual articulation cues. For example, the audio signal of /ba/ dubbed on the visual articulation of /ga/ is often perceived as /da/. Such a percept has been referred to as a ‘fused’ response because it results in blending or fusing the perceptual cues from the two modalities (McGurk & MacDonald, 1976). As noted earlier, the /b/, /d/, /g/ consonants differ along a physical continuum (i.e., place of articulation). Since the auditory signal of /ba/ is similar to the auditory signal of /da/, and the visual articulation of /ga/ is similar to the
visual articulation of /da/, the /da/ or fused percept appears to be the ‘common denominator’ of the audio /ba/ and the visual /ga/.

It is important to note that McGurk stimuli do not always result in a clear and distinct /da/ percept. Although the original McGurk experiment reported 98% of ‘da’ responses in adult participants, other studies reported much lower values (Green & Norrix, 1997). In addition, auditory and the visual cues are sometimes perceptually combined in a way that results in a response other than /da/. (e.g., /tha/, /va/, /bga/). In addition, the perceptual integration of auditory and visual cues does not occur for all participants. Some participants respond to what they heard (/ba/ response), others to what they saw (/ga/ response).

Often, the proportion of /ba/ responses elicited to McGurk stimuli is found to be higher than the proportion of /ga/ responses (e.g., Green & Norrix, 1997), which is probably due to the tendency to rely more on auditory than on visual perceptual cues in speech. If so, then reducing the auditory perceptual cues in the McGurk stimuli would increase the effect of visual articulation cues. In fact, a number of experimental studies found that degrading the acoustic speech signal increases the degree of McGurk effect (Green & Norrix, 1997; Massaro, Thomson & Laren, 1986; Sekiyama & Tohkura, 1991). The size of the McGurk effect was found to increase by embedding the auditory signal in noise (Sekiyama & Tohkura, 1991), and by filtering higher formants of the audio signal (Green & Norrix, 1997). Besides adding noise and filtering, auditory perceptual cues to speech are often manipulated by synthesising speech. Massaro, Thomson and Laren (1986) created synthesized speech sounds ranging from /ba/ to /da/ and paired them with either /ba/ or /da/ visual articulations. Visual articulation had a greater effect on perception when paired with an ambiguous speech signal as opposed to being paired with an unambiguous auditory syllable.

Besides the quality of auditory perceptual speech cues, audio-visual integration of stop consonants is also affected by vowel context. It has been found that the proportion of /d/ responses is higher in /i/ than in /a/ vowel context, and lower in /u/ than in /a/ vowel context (Burnham, 1988; Green & Norrix, 1997; Green, Kuhl, & Meltzoff, 1988; Shigeno, 2000). Shigeno (2000) found that McGurk stimuli in /i/ vowel context resulted in 59% of /d/ responses, while the proportion of /d/ responses
in /a/ and /u/ vowel contexts were only 30% and 3%, respectively. In general, experimental research suggests that the McGurk effect is the greatest in /i/ context, moderate in /a/ context, and almost non-existent in /u/ context (Green, Kuhl, & Meltzoff, 1988; Shigeno, 2000). It has been suggested that the proportion of /d/ responses is higher in /i/ than in /a/ vowel context because the similarity between the auditory signal of /b/ and /d/ consonants is greater in /i/ than in /a/ vowel context (Green & Norrix, 1997). In the /i/ vowel context, the second formant transition is rising in both /b/ and /d/. On the other hand, the slope of the second formant transitions in /ba/ and /da/ are inconsistent (i.e., rising in /ba/ and slightly falling in /da/). Therefore, /b/ and /d/ are more likely to be confused when presented in /i/ than in /a/ vowel context.

It has been found that the degree of the McGurk effect differs across cultures, presumably because different cultures do not pay attention equally to the visual cues of speech perception. For example, it has been found that the McGurk effect is weaker in Japanese and Chinese participants than in Americans (Sekiyama & Tohkura, 1991, 1993; Sekiyama, 1996, 1997). There are three reasons that have been proposed to account for these cultural differences. Firstly, it is suggested that Japanese and Chinese people are less sensitive to visual articulation cues than Americans because of their cultural habit of avoiding looking at a speaker’s face (Sekiyama, 1997). Secondly, it is possible that visual articulation cues have a reduced effect on speech perception in Japanese people because articulation in Japanese is less dynamic or pronounced than articulation in English (Shigeno, 2000). Thirdly, Japanese and Chinese people may pay less attention to the visual articulation cues because they have additional acoustic cues (i.e., pitch-accent and tones, respectively) that aid speech perception (Burnham & Dodd, 2004).

In addition to cultural differences, the McGurk effect is also influenced by development. Experimental evidence suggests that the McGurk effect is less pronounced in children than in adults (Boliek, Green, Fohr & Obrzut, 1996; Hockley & Polka, 1994; Massaro, Thompson, & Laren, 1986; McGurk & McDonald, 1976). For instance, McGurk and McDonald (1976) found that the percentages of visually influenced responses to incongruent audio-visual stimuli were higher in adults (i.e., 92%) than in pre- and primary school children (59% and 52 %, respectively). Similar
developmental trends were observed in studies using synthesized auditory speech signals (Massaro, 1984; Massaro et al., 1986). Massaro et al. tested audio-visual speech perception using five synthetic auditory syllables ranging from /ba/ to /da/ paired with either /ba/ or /da/ visual articulation. The average size of the visual effect (measured by the difference in the proportion of /da/ response between /ba/ and /ga/ articulation conditions) for adults was 66% and 40% for children (Massaro, 1984). Thus, developmental studies suggest that speech perception in adults is more affected by articulation cues than speech perception in children.

In conclusion, McGurk stimuli are generally used to investigate the nature of audio-visual speech perception. Previous studies have used McGurk stimuli to investigate the nature of audio-visual speech perception and developmental and cross-cultural differences in audio-visual speech perception. In the present experiments, McGurk stimuli are used as a tool to investigate the effect of conditioning on perceptual learning.

7.3 General Introduction to Experiments

The experiments reported in this chapter were designed to modify the perception of ambiguous McGurk speech stimuli using classical and operant conditioning. As argued in Chapter 3, both classical and operant conditioning result in perceptual learning, during which the learner educates his/her attention to meaningful invariants (i.e., affordances) in the environment. In the present experiments participants were trained educate their attention to pitch contour, which is normally an irrelevant auditory dimension in stop consonant perception (i.e., it does not afford the perception of stop consonants).

During training, participants were repeatedly presented with /b/, /d/, and /g/ stop consonant in falling, flat, and rising pitch contour contexts, respectively (i.e., classical conditioning). Pitch contour was paired with either the auditory signal (i.e., Experiments 4 and 5) or the visual displays of articulations (i.e., Experiment 6). During training, participants were required to identify the stop consonants and they received corrective feedback after their response (i.e., operant conditioning). Thus conditioning principles were used to train participants to learn that falling pitch
contour affords consonant /b/, flat pitch contour affords consonant /d/, and rising pitch contour affords consonant /g/.

The experiments were based on a within-subjects design, in which McGurk stimuli were presented in three different pitch contours (i.e., falling, flat, rising) and their perception was tested before and after training. In general, hypotheses were based on the assumption that the change in the perception of McGurk stimuli as a result of training would be influenced by pitch contour.

### 7.4 Experiment 4

#### 7.4.1 Aim, Design, and Hypotheses

The aim of Experiment 4 was to change the perception of the McGurk stimuli using classical and operant conditioning principles. McGurk stimuli used in the experiment were constructed by pairing the visual articulation of the /ga/ syllable with the auditory /ba/ syllables presented in three different pitch contours (i.e., falling, flat, and rising).

During training, participants were repeatedly presented with auditory /ba/, /da/, and /ga/ syllables in falling, flat, and rising pitch contours, respectively (i.e., classical conditioning) and they received corrective feedback after they identified the consonants (i.e., operant conditioning).

The experiment was based on a within-subjects design, in which the perception of McGurk stimuli was tested before and after training (i.e., pre-test and post-test) and in three different pitch contour conditions (i.e., falling, flat, rising). The perception of McGurk stimuli was tested by asking participants to identify each stimulus as /ba/ or /da/ or /ga/ syllables. Thus the independent variables were time (pre-test and post-test) and pitch contour (falling, mid, rising) and the dependent variables were the number of /ba/, /da/, and /ga/ identifications in response to McGurk stimuli. The reaction time of responses was also analysed. It should also be noted that there were two types of McGurk stimuli used in the experiments, which were implemented as a between-subject factor. The two types of McGurk stimuli differed in terms of the speaker used in the construction of the stimuli (i.e., Asian male and Caucasian
female). Speaker type was added to the design simply to increase the variability of responses, however, the training effect was not tested separately for the two speaker types.

The following hypotheses were proposed:

1. The number of /ba/ responses for the McGurk stimuli increases with training more in the falling than in the rising and flat contour conditions.
2. The number of /da/ responses for the McGurk stimuli increases with training more in the flat than in the falling and rising contour conditions.
3. The number of /ga/ responses for the McGurk stimuli increases with training more in the rising than in the flat and falling contour conditions.

7.4.2 Method

7.4.2.1 Participants

Participants were 37 (6 males, 31 females) Psychology 1 students from the University of Western Sydney, who received course credit for their participation. All participants reported normal hearing and vision. The mean age of participants was 20.8 years ($SD=5$). Seventeen participants were presented with syllables uttered by a Caucasian female speaker and 20 participants were presented with syllables uttered by an Asian male speaker.

7.4.2.2 Stimuli

Audio-visual recordings used for stimulus construction consisted of an Asian male and a Caucasian female speaker uttering /ba/, /da/, and /ga/ syllables. These recordings have previously been used to investigate the McGurk effect in developmental and cross-cultural settings (Sekiyama & Burnham, 2004).

33 It is possible that the proportion of /ba/, /da/, and /ga/ responses for McGurk stimuli differ as a function of speakers used in the construction of stimuli. For instance, it is likely that the proportion of visual (i.e., /ga/) responses is higher for speakers who clearly articulate the syllable than for speakers who do not.
To construct stimuli for the present experiment, audio signals were first separated from the video files using the Virtual Dub audio-video editing program. Then, the Praat speech analysis software program was used to extract the duration and the mean pitch of the auditory syllables (See Appendix H) and to modify their pitch contour. There were three different pitch contour modified stimuli (i.e., falling, flat, rising) created for each original auditory syllable. The pitch values in the flat contour modifications corresponded to the mean pitch values of the original syllables. The pitch change in the rising and falling pitch contour modification was set to be 1.7 semitones per 100 msec, retaining the mean pitch value of the original signal.

The audio-visual stimuli used in the pre- and post-test phases of the experiment included nine congruent and three incongruent audio-visual stimuli. The congruent stimuli were created by matching the nine pitch contour modified auditory syllables (i.e., /ba/-falling, /ba/-flat, and /ba/-rising, /da/-falling, /da/-flat, /da/-rising, /ga/-falling, /ga/-flat, and /ga/-rising) with the corresponding visual articulation. Incongruent stimuli were created by matching the /ba/-rising, /ba/-flat, and /ba/-falling audio syllables with the visual articulation of /ga/. The auditory signal was paired with the visual signal so that the onset of the consonant in the auditory signal was synchronised with the onset of the consonant in the visual signal. Auditory stimuli used in training consisted of the three syllables (i.e., /ba/, /da/, /ga/) in three different pitch contours (i.e., rising, flat, and falling).

Additional stimuli included the original (i.e., not pitch-modified) /ba/, /da/, and /ga/ audio-visual syllables that were used in practice trials and visual signals that were displayed on the computer screen. Visual signals included a ‘Get Ready’ sign (printed in blue) and a ‘Respond Now’ sign (printed in blue) that were used in pre- and post-test. Visual stimuli used during training consisted of the words ‘YES’ (printed in black) and ‘NO’ (printed in black) that were used as corrective feedback. If participants failed to enter their response within a certain time limit\(^{34}\) the word ‘LATE’ (printed in red) appeared on the computer screen.

\(^{34}\) The ‘late’ sign was set to appear at 3000 ms (in Block 1), 2500 ms (in Block 2), and 2000 ms (in Block 3) after the onset of the audio files.
7.4.2.3 Equipment

The experiment was conducted using a laptop computer running DMDX experimental software. Instructions were given verbally by the experimenter and also appeared on the computer screen outlining what participants were required to do at each stage. Participants listened to the syllables through headphones at a comfortable listening level (approximately 65 dB SPL). Identification responses were indicated by pressing marked keys on the keyboard; /ba/ responses were indicated by pressing the left shift button (marked ‘BA’), /da/ responses were indicated by pressing the shift bar (marked ‘DA’), and /ga/ responses were indicated by pressing the right shift button (marked ‘GA’).

7.4.2.4 Procedure

Before commencing the experiment participants read an information sheet about the experiment, they signed a consent form, and filled out a background questionnaire. The information sheet, consent form, and questionnaire are shown in Appendices E, F, G, respectively.

Pre-test Phase

Prior to the pre-test participants were given three practice trials with feedback. The stimuli used in the practice trials included the original (i.e., not pitch-modified) /ba/, /da/, and /ga/ audio-visual syllables. In the pre-test phase, participants’ audio-visual stop consonant perception was tested using nine congruent and three McGurk-type incongruent audio-visual stimuli. Participants were presented with the 12 stimuli in random order and they were asked to indicate the syllable they perceived by pressing marked buttons on the keyboard. The presentation of each stimulus was preceded by a ‘Get Ready’ sign and followed by a ‘Respond Now’ sign. No feedback was given after the responses.

Training Phase

Training was divided into three blocks. In each block, participants were presented with 38 /ba/-falling, 38 /da/-flat, and 38 /ga/-rising auditory syllables. Syllables were presented in a different random order to each participant. In addition to the training stimuli, in each block participants were also presented with six catch-trials that
consisted of one /ba/-flat, one /ba/-rising, one /da/-rising, one /da/-falling, one /ga/-falling, and one /ga/-flat stimuli. Catch-trials were randomly interspersed with the training trials. On each trial, participants were asked to indicate the syllable they heard as fast as they could. Correct responses were followed by display of the word ‘YES’, and incorrect responses were followed by display of the word ‘NO’. In order to keep participants motivated during training, they were told that they have a limited time to respond to stimuli and that the time limit for entering the response would gradually decrease during the training. Participants were told that in Block 1 the time limit is 2 seconds, in Block 2 the time limit is 1.5 seconds and in Block 3 the time limit is 1 second. If participants failed to respond within a certain time limit, the word ‘LATE’ was displayed on the computer screen. Participants were instructed to try to avoid giving late and incorrect responses. Before training, participants were given three practice-trials with feedback that included /ba/-falling, /da/-flat, /ga/-rising auditory syllables.

**Post-test Phase**

Training was followed by the post-test, which was methodologically the same as the pre-test. Participants’ audio-visual stop consonant perception was tested using nine congruent and three McGurk-type incongruent audio-visual stimuli. Participants were presented with the 12 stimuli in random order and they were asked to indicate the syllable they perceived by pressing marked buttons on the keyboard. The presentation of each stimulus was preceded by a ‘Get Ready’ sign and followed by a ‘Respond Now’ sign. No feedback was given after the responses.

The experiment took approximately 25 minutes including instructions and breaks. Participants were debriefed at the end of the experiment.

---

35 Catch-trials were used as they allow tracking learning during training by comparing the reaction times for training stimuli with the reaction times for catch trials.

36 The time limits told to participants were only approximates of the actual appearance of the ‘late’ sign. The ‘late’ sign in Block 1, 2, and 3 was set to appear at 3000, 2500, and 2000 millisecond after the onset of the audio files, respectively.
7.4.3 Results

7.4.3.1 Pre-Test Analysis

A preliminary analysis was conducted on the pre-test data to examine the proportions of /ba/, /da/, and /ga/ responses for McGurk stimuli and to investigate whether the proportions differ as a function of pitch contours. The number of /ba/, /da/, and /ga/ responses for the McGurk stimuli as a function of pitch contour is shown in Figure 7.1.

The Friedman two-way analysis of variance by ranks revealed no significant difference among the number of /ba/, /da/, and /ga/ responses (i.e., 25, 36, 50, respectively), $X_r^2=4.47$, $N=37$, $C=3$, $p>.1$.

The Friedman two-way analysis of variance by ranks revealed no significant relationship between response category and pitch contour, $X_r^2=3.5$, $N=3$, $C=3$, $p>.1$. That is, the proportions of /ba/, /da/, and /ga/ responses were not significantly different for the three different pitch contours. This suggests that falling, flat and rising pitch contours have no pre-existing association with /b/, /d/ and /g/ consonants.
7.4.3.2 Training Effect Analysis

Hypotheses were based on the assumption that, as a result of contour-consonant associative training, the perception of McGurk stimuli would be influenced by pitch contour.

The first hypothesis stated that the number of /ba/ responses for the McGurk stimuli would increase with training more in the falling than in the rising and flat contour conditions. The number of /ba/ responses as a function of time and pitch contour is shown in Figure 7.2.

![Figure 7.2. The number of /ba/ responses as a function of time and pitch contour.](image)

Friedman two-way (time x contour) analysis of variance by ranks was conducted to investigate whether the change in the number of /ba/ responses as a result of training is different for the three pitch contour conditions. The Friedman test revealed no significant relationship between time and pitch contour, $X_r^2 = 3.25$, $N=2$, $C=3$, $p>.1$. In other words, the change in the number of /ba/ responses as a result of training was not significantly influenced by pitch contour.

The second hypothesis stated that the number of /da/ responses for the McGurk stimuli increases with training more in the flat than in the falling and rising contour conditions. The number of /da/ responses as a function of time and pitch contour is shown in Figure 7.3.
Friedman two-way (time x contour) analysis of variance by ranks was conducted to investigate whether the change in the number of /da/ responses as a result of training is different for the three pitch contour conditions. The Friedman test revealed no significant relationship between time and pitch contour, \( X^2 = 3.25, N=2, C=3, p > .1 \) suggesting that the change in the number of /da/ responses is not affected by pitch contour.

The third hypothesis stated that the number of /ga/ responses for the McGurk stimuli increases with training more in the rising than in the flat and falling contour conditions. The number of /ga/ responses as a function of time and pitch contour is shown in Figure 4.
Figure 7.4. The number of /ga/ responses as a function of time and pitch contour.

Friedman two-way (time x contour) analysis of variance by ranks was conducted to investigate whether the change in the number of /ga/ responses as a result of training is different for the three pitch contour conditions. The Friedman test revealed no significant relationship between time and pitch contour, \( X^2 = 3, N = 2, C = 3, p > .1 \). Thus, the change in the number of /ga/ responses (similar to the change in the number of /ba/ and /da/ responses) was not significantly influenced by pitch contour.

7.4.3.3 Reaction Time Analysis

Analyses were also conducted to investigate whether the contour-consonant associative learning is reflected in the reaction times. The effect of contour on reaction times was examined in (1) the training data, (2) McGurk stimuli used in pre- and post-test, and (3) congruent audio-visual stimuli used in pre- and post-test.

Dealing with outliers in reaction time data requires special considerations (Ratcliff, 1993). Outliers in reaction time data reflect responses that are generated by cognitive processes that are not the ones being studied (e.g., guesses based on participant’s inattention or failure to reach decision). Reaction time data are usually skewed to the right and often contain some long spurious reaction times that increase the mean and inflate the standard deviation. After carefully reviewing different methods of dealing with reaction time outliers, Ratcliff concluded that the greatest power is obtained by eliminating outliers longer than an absolute cutoff value. He suggested that the cutoff
value should be selected by considering the proportion of responses eliminated, and recommended choosing a cutoff value that eliminates 10-15% of the data.

Following the recommendations of Ratcliff (1993), outliers in the pre- and post-test data were dealt with by eliminating reaction times longer than a specific cutoff value. The cutoff values used for eliminated reaction time outliers for McGurk stimuli and for congruent audio-visual stimuli were 2000 and 1500 msec, respectively. Eliminating outliers in the training data was not necessary because, due to the nature of the task, training data already contained reaction times lower than a specific cutoff (i.e., during training participants had limited time to enter their response and failing to do so automatically resulted in a missing value).

Prior to conducting statistical analyses, the assumption of normality was tested in each data set. However, it was taken into consideration that relatively simple parametric tests (e.g., t-test and ANOVA) are robust with respect to the violation of normality in the case of large sample sizes (i.e., $N>30$) (Tabachnick & Fidell, 2001). Since the sample used in the present experiment was relatively large (i.e., $N>30$) data were analysed using either t-test or ANOVA even when the assumption of normality was violated.

**Reaction Times Analyses for Training Stimuli**

Assuming that the learning of contour-consonant associations would be most pronounced towards the end of the training, reaction times in training data were only analysed in the last block of training (i.e., Block 3). It is expected that if participants learned the contour-consonant associations then they would identify the actual training stimuli (i.e., /ba/-falling, /da/-flat, and /ga/-rising) faster than the catch-trial stimuli (i.e., /ba/-flat, /ba/-rising, /da/-rising, /da/-falling, /ga/-falling, /ga/-flat). The mean reaction times for correctly identified training and catch-trial stimuli are shown in Figure 7.5.

---

37 The use of longer cutoff values for McGurk stimuli is due to the fact that reaction times for McGurk stimuli were generally longer than reaction times for congruent stimuli.
A paired sample t-test revealed that the mean reaction times for the /ba/ syllables were not significantly faster in the falling than in the flat/rising condition, \( t(36)=1.6, p=.12, d=.37 \). On the other hand, paired sample t-tests conducted on the reaction times for /da/ and /ga/ syllables revealed significant differences between training and catch-trial stimuli. Firstly, reaction times for /da/ syllables were significantly lower in the flat contour condition than in the falling/rising contour condition, \( t(35)=2.4, p=.02, d=.56 \). Secondly, the reaction times for the /ga/ syllables were significantly lower in the rising contour condition than in the falling/flat contour condition, \( t(36)=2.2, p=.03, d=.51 \). These results suggest that participants learned to associate the flat pitch contour with the /da/ auditory syllable and the rising pitch contour with the /ga/ auditory syllable.

**Reaction Times Analyses for McGurk Stimuli**

Reaction times for McGurk stimuli were analysed separately for /ba/, /da/, and /ga/ response categories using a 2x3 (time x contour) repeated measures ANOVAs. It is important to note that, statistical analysis on reactions times for McGurk stimuli is limited by the fact that /ba/, /da/, and /ga/ responses are unequally distributed as a function of time and pitch contour. The unequal distribution of response categories poses a problem for data analysis because it generates missing values for reaction times. Conducting ANOVAs on a data set with a considerable number of missing values is limited in detecting differences due to the listwise deletion of data points.
(Schafer & Graham, 2002). To overcome this problem, missing reaction time data for McGurk stimuli in each cell was replaced by the average of the observed values (i.e., mean substitution). Since mean substitution retains the cell means, it is an often used practice to deal with missing data. However, it should be noted that the mean substitution reduces the standard deviations which may result in overestimating the significance of the observed effects (Schafer & Graham, 2002). To reduce the distortion of standard deviations in the McGurk data set, mean substitution in each response category was only applied to missing values that would otherwise result in listwise deletion of reaction time data. Mean reaction times for McGurk stimuli in the /ba/ response category are shown in Figure 7.6.

![Figure 7.6. Mean reaction times for McGurk stimuli in the /ba/ response category.](image)

A 2x3 (time x contour) repeated measures ANOVA revealed a significant main effect of time, $F(1,20)=25.12$, $p<.001$, partial $\eta^2=.68$. However, the main effect of contour and the time by contour interaction were not significant $F(1.4,16.6)=3.58$, $p=.07$, partial $\eta^2=.23^{39}$, $F(2,24)=.87$, $p=.43$, partial $\eta^2=.08$, respectively. Thus although participants responded faster to McGurk stimuli after training than before training, the decrease in their tendency to respond faster was not significantly affected by pitch contour.

---

38 Error bars in the figure correspond to the standard errors prior to mean substitution.

39 Greenhouse-Geisser adjustment was applied to the degrees of freedom since the homogeneity of covariance assumption was violated.
Mean reaction times for McGurk stimuli in the /da/ response category are shown in Figure 7.7\(^{40}\).

![Figure 7.7](image-url)

**Figure 7.7.** Mean reaction times for McGurk stimuli in the /da/ response category.

A 2x3 (time x contour) repeated measures ANOVA revealed that not only the main effect of time but also the time x contour interaction was significant, \(F(1,20)=147.06, p<.001, \text{partial } \eta^2=.88; F(2,40)=11.54, p<.001, \text{partial } \eta^2=.37\), respectively. There were no significant differences observed in the main effect of contour, \(F(2,40)=1.17, p=.32, \text{partial } \eta^2=.06\).

Since the time x contour interaction was significant, Post hoc pairwise comparisons with Bonferroni adjustment were conducted on the decrease in reaction times\(^{41}\) across the three pitch contour conditions. It was found that the reaction time of /da/ responses decreased more in the flat contour condition than in the rising and falling contour conditions (\(p=.002, p=.002\), respectively). No significant difference was found between the rising and falling contour conditions (\(p=1\)). These results suggest that the learned associations between the auditory /da/ syllable and the flat pitch contour increased participants’ confidence to identify McGurk stimuli as /da/ syllable.

---

\(^{40}\) Error bars in the figure correspond to the standard errors prior to mean substitution.

\(^{41}\) The decrease in reaction times were calculated by subtracting the post-test reaction times from the pre-test reaction times.
Reaction times for McGurk stimuli in the /ga/ response category are shown in Figure 7.8.

![Graph showing reaction times for McGurk stimuli in the /ga/ response category.](image)

**Figure 7.8.** Mean reaction times for McGurk stimuli in the /ga/ response category.

A 2x3 (time x contour) repeated measures ANOVA revealed that main effect of both time and contour was significant, $F(1,24)=97.02, p<.001$, partial $\eta^2=.8$; $F(2,48)=3.97, p=.03$, partial $\eta^2=.14$, respectively. In addition, the time x contour interaction was also found to be significant, $F(2,48)=4.65, p=.01$, partial $\eta^2=.16$. Since the time x contour interaction was significant, post hoc pairwise comparisons with Bonferroni adjustment were conducted on the decrease in reaction times across the three pitch contour conditions. It was found that the reaction time of /ga/ responses decreased more in the rising contour condition than in the flat contour condition ($p=.035$). No significant differences were found between the rising and falling conditions ($p=1$) and between the falling and flat conditions ($p=.057$). These results suggest that participants’ confidence to identify McGurk stimuli as /ga/ syllable increased most in the rising contour condition.

---

42 Error bars in the figure correspond to the standard errors prior to mean substitution.
43 The decrease in reaction times were calculated by subtracting the post-test reaction times from the pre-test reaction times.
**Reaction Times Analyses for Congruent Audio-Visual Stimuli**

Reaction times for congruent stimuli were analysed separately for /ba/, /da/, and /ga/ syllables using 2x3 (time x contour) repeated measures ANOVAs. Prior to analysis, missing values (which were primarily due to eliminating reaction times that were longer than 1500 msec) in each cell were replaced by the mean of the observed values.

Mean reaction times for congruent /ba/ audio-visual syllables are shown in Figure 7.9.

![Figure 7.9](image_url)

**Figure 7.9. Mean reaction times for congruent /ba/ audio-visual syllables.**

Repeated measures ANOVA revealed that after training the reaction times for /ba/ audio-visual syllable decreased significantly; main effect of time: $F(1,36)=62.92$, $p<.001$, partial $\eta^2 =.64$. However, no significant main effect of contour and time by contour interaction was found, $F(2,72)=.08$, $p=.93$, partial $\eta^2 =.002$; $F(2,72)=.13$, $p=.88$, partial $\eta^2 =.003$, respectively.

Mean reaction times for congruent /da/ audio-visual syllables are shown in Figure 7.10.
A 2x3 (time x contour) repeated measures ANOVA revealed that after training the reaction times for /da/ audio-visual syllable decreased significantly; main effect of time: $F(1,36)=39.35$, $p<.001$, partial $\eta^2 = .52$. However, no significant main effect of contour and time x contour interaction was found, $F(2,72)=1.12$, $p=.33$, partial $\eta^2 = .03$; $F(2,72)=.5$, $p=.61$, partial $\eta^2 = .01$, respectively.

Mean reaction times for congruent /ga/ audio-visual syllables are shown in Figure 7.11.
A 2x3 (time x contour) repeated measures ANOVA revealed that after training the reaction times for /ga/ audio-visual syllable decreased significantly; main effect of time: $F(1,36)=61.97, p<.001$, partial $\eta^2=.36$. However, no significant main effect of contour and time by contour interaction was found, $F(2,72)=.05, p=.95$, partial $\eta^2=.001$; $F(2,72)=1.19, p=.31$, partial $\eta^2=.03$, respectively.

7.4.4 Discussion

Experiment 4 investigated whether the perception of pitch contour modified McGurk stimuli would change as result of contour-consonant associative training. During training, participants were repeatedly presented with auditory /ba/, /da/, and /ga/ syllables in falling, flat, and rising pitch contours, respectively. It was hypothesized that such training would change the proportion of /ba/, /da/, and /ga/ responses for McGurk stimuli as a function of pitch contour. Specifically, it was hypothesized that as a result of training (1) the number of /ba/ responses for the McGurk stimuli increases more in the falling than in the rising and flat contour conditions, (2) the number of /da/ responses for the McGurk stimuli would increase more in the flat than in the falling and rising contour conditions, and (3) the number of /ga/ responses for the McGurk stimuli would increase more in the rising than in the flat and falling contour conditions. None of these hypotheses were supported by the results. This suggests that contour-consonant associative training was not effective in changing participants’ perception of pitch contour modified McGurk stimuli or the dependent measure was not sensitive enough to pick up differences.

Although no change was observed in participants’ responses to McGurk stimuli, it should not be concluded that participants failed to learn the contour-consonant associations. Reaction times for stimuli used in training and for McGurk stimuli suggest that participants did learn that certain pitch contour affords a certain consonant. Specifically, reaction time results suggest that participants learned the association between consonant /d/ and flat pitch contour and the association between consonant /g/ and rising pitch contour. During the last block of training participants identified the /da/ and /ga/ syllables the fastest when they were presented in flat and rising pitch contours, respectively. In addition, reaction times of /da/ responses for
McGurk stimuli decreased the most in the flat contour condition, and reaction times of /ga/ responses for McGurk stimuli decreased the most in the rising contour condition.

Since reaction time data suggest that contour-consonant associations were learned (at least implicitly), the question to be asked is why such learning is not evident in the identification responses to McGurk stimuli. It is possible that the relatively small number of training trials resulted in a subtle learning of associations, which is only reflected in reaction time. In addition, it is likely that the measure of McGurk stimulus perception was not powerful enough to reflect the change in the perception. That is, during pre- and post-test, the perception of McGurk stimuli was only measured once in each contour condition. It is plausible that more training trials and more test trials of McGurk stimuli would result in the hypothesised change in McGurk stimulus perception. This possibility was examined in Experiment 5.

7.5 Experiment 5

7.5.1 Aim, Design, and Hypotheses

The aim of Experiment 5, similar to that of Experiment 4, was to modify the perception of ambiguous audio-visual speech stimuli (i.e., McGurk stimuli) using classical and operant conditioning. The method of Experiment 5 is similar to that of Experiment 4, however, there were some changes introduced to Experiment 5 that concern the number and the type of training and test stimuli.

Firstly, instead of using syllables in a single vowel context, Experiment 5 used vowel-consonant-vowel /vcv/ disyllables in three different vowel contexts (i.e., /a/, /i/, /u/). Thus, McGurk stimuli used in Experiment 5 consisted of audio signals of /aba/, /ibi/, and /ubu/ disyllables paired with visual displays of speakers articulating /aga/, /igi/, and /ugu/, respectively. The reason for using /vcv/ disyllables instead of /cv/ (consonant-vowel) syllables was to increase the prominence of pitch contour. Introducing a vowel before the consonants ensures that pitch contour information precedes the consonants, which may enhance contour-consonant associative learning. In addition to vowel context, stimuli used in Experiment 5 also varied as a function of speaker. There were eight different speakers used in the construction of stimuli. According to Iverson et al. (2003) the effectiveness of speech-related perceptual
training can be increased by using “multi-talker high-variability stimulus sets…, because the variability provides information about which cues are most robust and trains individuals to ignore irrelevant variation” (p. 54). Both vowel context and speaker were included as a within-subjects factor thus each participant was presented with 24 (i.e., 3 vowel context x 8 speaker) different McGurk stimuli in each pitch contour condition.

Similar to Experiment 4, training in Experiment 5 consisted of pairing the auditory signals of consonant /b/, /d/, and /g/ with falling, flat and rising pitch contours, respectively. Participants were required to identify the stop consonant in the audio signals (as /b/, /d/, or /g/) and they received corrective feedback after each response.

The experiment was based on a within-subjects design, where the perception of McGurk stimuli was tested before and after training in three different pitch contours. Dependent variables were the change in the percentages of /b/, /d/, and /g/ responses as a result of training (measured by the difference between the after and before training percentage scores). It was hypothesised that the change in the perception of McGurk stimuli (i.e., change in the percentages of /b/, /d/, and /g/ responses) would be significantly influenced by pitch contour.

The following related hypotheses were proposed:

1. The percentages of /b/ responses would increase with training more in the falling contour condition than in the flat and rising contour conditions.

2. The percentages of /d/ responses would increase with training more in the flat contour condition that in the falling and rising contour conditions.

3. The percentages of /g/ responses would increase with training more in the rising contour condition that in the falling and flat contour conditions.
7.5.2 Method

7.5.2.1 Participants

Participants were 25 (5 males and 20 females) Psychology 1 students from the University of Western Sydney, who received course credit for their participation. All participants reported normal hearing and vision. The mean age of participants was 20 years (SD=9.6).

7.5.2.2 Stimuli

Stimulus Recording

Eight speakers (4 males, 4 females) were video-recorded uttering /aba/, /ada/, /aga/, /ibi/, /idi/, /igi/, /ubu/, /udu/, and /ugu/ disyllables. The production of disyllables was recorded using a digital video camera and an external microphone attached to the video camera. Recordings were done in a recording studio with black background at the University of Western Sydney. Speakers were sitting in front of the video camera with the microphone located at approximately 200 cm from their mouth. For the recording, speakers were wearing a black T-shirt and they removed glasses and jewellery. Lighting conditions were set so that the speakers’ articulations could be clearly recorded. Speakers were asked to utter each disyllable three times in following order: /aba/x3, /ada/x3, /aga/x3, /ibi/x3, /idi/x3, /igi/x3, /ubu/x3, /udu/x3, and /ugu/x3. Speakers were instructed to utter the disyllables naturally (i.e., not over or under articulate) and with approximately equal length and flat pitch contour. To reduce variations in the lengths of produced disyllables across speaker, prior to recording the speakers listened to a recording of a female speaker producing each disyllable three times. The nine disyllables were also written on an A4 paper which was shown to the speakers to remind them about the order of the to-be-produced disyllables. Although speakers were allowed to look at the written disyllables, they were instructed to look towards the camera during the production of disyllables. Speakers were also told to close their lips before starting and after finishing the production of each disyllable.

Stimulus Editing

The recorded disyllables were cut (using a video editing program ULead) into 1800 msec long video segments with the criterion that the utterance of the disyllable
(measured from the onset of mouth opening) started at 600 msec. River Past Audio Converter program was used to separate the audio signals from the video files and to normalize loudness. Praat speech analysis software program was used to extract the duration and the mean pitch of the auditory disyllables (See Appendix L) and to create pitch contour modified disyllables. The pitch value in the flat contour modification corresponded to the mean pitch value of the original disyllable. The pitch change in the rising and falling pitch contour modification was set to be six semitones per 650 msec duration (i.e., 0.92 semitones per 100 msec), retaining the mean pitch value of the original signal. For disyllables with consonant /b/ (i.e., /vbv/ disyllables), three pitch contour modified (falling, mid, rising) stimuli were created. Disyllables with /d/ consonant (i.e., /vdv/ disyllables) were modified to have flat pitch contour, and disyllables with /g/ consonant (i.e., /vgv/ disyllables) were modified to have rising pitch contour.

McGurk stimuli used in pre- and post-test were constructed by pairing /vbv/ audio signals in each pitch contour (falling, flat, rising) with /vgv/ visual articulations resulting in 72 (8 speakers x 3 pitch contour x 3 vowel context) audio-video files using Virtual Dub audio-video editing program. The auditory signal was paired with the visual signal so that onset of consonant in the auditory signal was synchronised with the onset of consonant in the visual signal.

Auditory stimuli used in training consisted of 24 (8 speaker x 3 vowel context) audio /vbv/ disyllables with falling pitch contour, 24 (8 speaker x 3 vowel context) audio /vdv/ disyllables with flat pitch contour, and 24 (8 speaker x 3 vowel context) audio /vgv/ disyllables with rising pitch contour.

Additional stimuli included nine unedited congruent audio-visual disyllables (i.e., /aba/, /ada/, /aga/, /ibi/, /idi/, /igi/, /ubu/, /udu/, /ugu/), which were used in the practice trials. Visual stimuli included the words ‘YES’ (printed in black) and ‘NO’ (printed in black) that were displayed on the computer screen as corrective feedback. If
participants failed to enter their response within a certain time limit\textsuperscript{44} the word ‘LATE’ (printed in red) appeared on the computer screen.

7.5.2.3 Equipment

The experiment was conducted using a laptop computer running DMDX experimental software. Instructions were given verbally by the experimenter and also appeared on the computer screen outlining what participants were required to do at each stage. Participants listened to the syllables through headphones at a comfortable listening level (approximately 65 dB SPL). Identification responses were indicated by pressing marked keys on the keyboard; \textasciitilde b/ responses were indicated by pressing the left shift button (marked ‘B’), \textasciitilde d/ responses were indicated by pressing the shift bar (marked ‘D’), and \textasciitilde g/ responses were indicated by pressing the right shift button (marked ‘G’).

7.5.2.4 Procedure

Before commencing the experiment participants read an information sheet about the experiment, they signed a consent form, and filled out a background questionnaire. The information sheet, consent form, and questionnaire are shown in Appendix L, M, N, respectively.

\textit{Pre-Test Phase}

Before pre-test, participants were familiarized with the task by receiving nine practice trials with feedback. In the practice trials participants were presented with congruent audio-visual stimuli (i.e., /aba/, /ada/, /aga/, /ibi/, /idi/, /igi/, /ubu/, /udu/, /ugu/) and they were told to indicate the consonant they perceived by pressing marked keys on the keyboard (left shift=\textasciitilde b/, space=\textasciitilde d/, right shift=\textasciitilde g/). McGurk stimuli were not used in the practice trials because, due to their ambiguity, no objective feedback could have been given to participants. During testing participants were presented with 72 McGurk stimuli (3 contour x 3 vowel context x 8 speaker) in random order. Participants were asked to indicate the consonant they perceived by pressing marked keys on the keyboard (left shift=\textasciitilde b/, space=\textasciitilde d/, right shift=\textasciitilde g/). Participants were also

\textsuperscript{44} The ‘late’ sign was set to appear at 2300 msec (in Block 1), 2100 msec (in Block 2), and 1900 msec (in Block 3) after the onset of the audio files.
told that some of the stimuli used in the test phase may be harder to identify than stimuli used in the practice trial. During pre-test participants received no feedback after their response.

**Training Phase**

During training, participants were presented with the /vbv/, /vdv/, and /vgv/ auditory disyllables with falling, flat, and rising pitch contours, respectively. Training was divided into three blocks. In each blocks, participants were presented with the 72 training stimuli (i.e., 3 consonant x 3 vowel context x 8 speaker) twice in random order. At each trial, participants were asked to indicate the consonant (as /b/, /d/, or /g/) they heard in the disyllables as fast as they could. Correct responses were followed by the display of the word ‘YES’, and incorrect responses were followed by the display of the word ‘NO’. If participants failed to respond within a certain time limit, the word ‘LATE’ was displayed on the computer screen. In order to keep participants motivated during training, they were told that the time limit for entering the response would gradually decrease during the training. Participants were told that in Block 1 the time limit is 2 seconds, in Block 2 the time limit is 1 second and in Block 3 the time limit is .5 second\(^45\). Participants were instructed to avoid late and incorrect responses. Training was preceded with nine practice trials that contained some of the items used during training.

**Post-Test Phase**

The post-test was methodologically the same as the pre-test. Participants were presented with 72 McGurk stimuli (3 contour x 3 vowel context x 8 speaker) in random order and they were asked to indicate the consonant they perceived by pressing marked keys on the keyboard (left shift=/b/, space=/d/, right shift=/g/). Participants received no feedback after their response.

The experiment took approximately 45 minutes including instructions and breaks. Participants were debriefed at the end of the experiment.

\(^45\) The time limits told to participants were only approximates of the actual appearance of the ‘late’ sign. The ‘late’ sign in Block 1, 2, and 3 was set to appear at 2300, 2100, and 1900 millisecond after the onset of the audio files, respectively.
7.5.3 Results

Prior to conducting statistical analyses, the assumption of normality was tested in each data set. However, it was taken into consideration that relatively simple parametric tests (e.g., t-test and ANOVA) are robust with respect to the violation of normality in the case of large sample sizes (i.e., \(N>30\)) (Tabachnick & Fidell, 2001). Since the sample used in the present experiment was relatively large (i.e., \(N>30\)) data were analysed using either t-test or ANOVA even when the assumption of normality was violated.

7.5.3.1 Pre-Test Analysis

Before analysing the effect of training on the perception of McGurk stimuli, the baseline perception of McGurk stimuli was analysed to examine the proportion of /b/, /d/ and /g/ responses as a function of pitch contour and vowel context. The before training percentages of /b/ responses are shown in Figure 7.12.

![Figure 7.12](image)

*Figure 7.12. Percentages of /b/ responses before training as a function of pitch contour and vowel context.*
Data were analysed by a two way (contour x vowel context) repeated measures ANOVA\(^{46}\). The main effect of pitch contour was not significant, \(F(2,48) = .05, p = .95\), partial \(\eta^2 = .002\). However, a significant main effect of vowel context was found, \(F(2,48) = 4.99, p = .01\), partial \(\eta^2 = .17\). Post hoc pair wise comparisons with Bonferroni adjustment revealed that the mean percentage of /b/ responses is significantly higher in the /u/ vowel context than in the /a/ vowel context \((p=.02)\). There was no significant difference found in the percentages of /b/ responses between the /i/ and /u/ vowel context \((p=.76)\) and the /a/ and /i/ vowel context \((p=.11)\).

The vowel context x pitch contour interaction was also significant, \(F(4, 96) = 2.99, p = .02\), partial \(\eta^2 = .11\). Since there was an interaction between vowel context and pitch contour, the pitch contour effect was examined separately in each vowel context using one-way repeated measures ANOVAs. A significant main effect of pitch contour was only found in the /u/ vowel context, \(F(2, 48)=4.83, p=.01\), partial \(\eta^2 = .17\). Post hoc pair wise comparisons with Bonferroni adjustment revealed that the mean percentage of /b/ responses is significantly higher in flat than in the rising conditions \((p=.015)\); however, no significant differences were found between the falling and flat and the falling and rising conditions \((p=.115, p=1\), respectively).

The before training percentages of /d/ responses are shown in Figure 7.13.

\(^{46}\) The assumption of normality was violated in the /i/-falling, /i/-flat, /i/-rising, /u/-falling, /u/-flat conditions; however no outliers were identified in any of the conditions.
Figure 7.13. Percentages of /d/ responses before training as a function of pitch contour and vowel context.

Data were analysed by a two-way (contour x vowel context) repeated measures ANOVA\(^{47}\). The main effect of pitch contour, was not significant $F(2,48)=.35$, $p=.71$, partial $\eta^2 = .01$, however the main effect of vowel context was significant $F(2,48)=12.95$, $p=.000$, partial $\eta^2 = .35$. Post hoc pair wise comparisons with Bonferroni adjustment revealed that the percentage of /d/ responses is significantly lower in the /u/ vowel context than in the /a/ and /i/ vowel contexts ($p=.004$, $p=.001$, respectively). No significant differences were found between the /a/ and /i/ vowel contexts ($p=.17$). The vowel context by pitch contour interaction was not significant, $F(2.5,59.1)=2.34$, $p=.09$, partial $\eta^2 = .09^{48}$.

The before training percentages of /g/ responses are shown in Figure 7.14.

\(^{47}\) The assumption of normality was violated in all conditions and three outliers were detected (one in the /u/-falling condition, one in the /u/-flat condition, and one in the /u/-rising condition). According to the recommendations of Tabachnick and Fidell (2001, p.71) the outlier values were changed so they are only one unit (i.e., one percent) more extreme than the next most extreme value.

\(^{48}\) Greenhouse-Geisser adjustment was applied to the degrees of freedom since the homogeneity of covariance assumption was violated.
Figure 7.14. Percentages of /g/ responses before training as a function of pitch contour and vowel context.

Data were analysed by a two way (contour x vowel context) repeated measures ANOVA. The main effect of pitch contour was not significant, F(2,48) = .04, p = .97, partial $\eta^2 = .001$. However, there was a significant main effect of vowel context, $F(2,48) = 10.6, p = .000$, partial $\eta^2 = .31$. Post hoc pair wise comparisons with Bonferroni adjustment revealed that the percentage of /g/ responses is significantly lower in the /i/ vowel context than in the /a/ and /u/ vowel contexts ($p=.000$, $p=.007$, respectively). No significant differences were found between the /a/ and /u/ vowel contexts ($p=1$).

There was a significant interaction found between vowel context and pitch contour interaction, $F(4,96) = 2.74, p = .03$, partial $\eta^2 = .1$. Due to the interaction, pitch contour effect was examined separately in each vowel context using one-way repeated measure ANOVAs. A significant main effect of pitch contour was only found in the /u/ vowel context, $F(2, 48) = 4.26, p = .02$, partial $\eta^2 = .15$. Post hoc pair wise comparisons with Bonferroni adjustment revealed that the mean percentage of /g/ responses is significantly higher in rising than in the flat conditions ($p=.03$); however, no significant differences were found between the falling and flat and the falling and rising conditions ($p=.12$, $p=1$, respectively).

$^{49}$ The assumption of normality was only met in the /u/-rising condition; however no outliers were identified in any of the conditions.
7.5.3.2 Training Effect Analysis

The training effect was measured by calculating the difference between the post-test and the pre-test percentage values. Positive difference scores indicate an increase in the percentages of responses as an effect of training; and negative values indicate a decrease in the percentages of responses as an effect of training.

The first hypothesis stated that the percentage of /b/ responses increases with training more in the falling contour condition than in the flat and rising contour conditions. The hypothesis was tested using four orthogonal planned comparisons, which included comparisons between falling and flat/rising contours (1) across vowel contexts, (2) in /a/ vowel context, (3) in /i/ vowel context, and (4) in /u/ vowel context. Difference scores for the percentages of /b/ responses are illustrated in Figure 7.15 and the results of planned comparisons are presented in Table 7.1.

![Figure 7.15](image)

Figure 7.15. Difference scores for the percentages of /b/ responses as a function of pitch contour and vowel context.

Table 7.1
Planned comparison statistics for /b/ responses

50 Analysing the difference between the post- and pre-test values is statistically more powerful to investigate training effect than simply analysing the post-test values because the prior removes concomitant variations in responses (Steel & Torrie, 1980, p. 401).
As can be seen in Table 7.1, significant differences between the falling and the flat/rising conditions were found in the /a/ vowel context. A paired-sample t-test revealed that in the /a/ vowel context the increase of /b/ responses in the falling conditions is significantly higher than the mean increase in the flat and rising conditions.\(^{51}\)

The second hypothesis stated that the percentage of /d/ responses increases with training more in the flat contour condition that in the falling and rising contour conditions. The hypothesis was tested by four orthogonal planned comparisons, which included comparisons between flat and falling/rising contours (1) across vowel contexts, (2) in /a/ vowel context, (3) in /i/ vowel context, and (4) in /u/ vowel context. Difference scores for the percentages of /d/ responses are illustrated in Figure 7.16 and the results of planned comparisons are presented in Table 7.2.

---

\(^{51}\) Alpha level was set at .05, following the argument that orthogonal comparisons driven by the hypothesis require no adjustment to the alpha level (Catham, 1999; O’Keefe, 2003).
Figure 7.16. Difference scores for the percentages of /d/ responses as a function of pitch contour and vowel context.

Table 7.2
Planned comparison statistics for /d/ responses

<table>
<thead>
<tr>
<th>Flat vs. Falling/Rising</th>
<th>t</th>
<th>df</th>
<th>p</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Across Vowel Context</td>
<td>0.65</td>
<td>24</td>
<td>0.52</td>
<td>0.18</td>
</tr>
<tr>
<td>/a/ Vowel Context</td>
<td>1.13</td>
<td>24</td>
<td>0.27</td>
<td>0.32</td>
</tr>
<tr>
<td>/i/ Vowel Context</td>
<td>-0.65</td>
<td>24</td>
<td>0.52</td>
<td>-0.18</td>
</tr>
<tr>
<td>/u/ Vowel Context</td>
<td>1.45</td>
<td>24</td>
<td>0.16</td>
<td>0.41</td>
</tr>
</tbody>
</table>

Note: Following the recommendations of Volker (2006), the value of $d$ was calculated by dividing the difference between the paired scores by the SD of the differences between the paired scores and then multiplying the result by the square root of 2.

As can be seen in Table 7.2, no significant differences were found between the flat and the falling/rising contour conditions.

The third hypothesis stated that the percentage of /g/ responses increases with training more in the rising contour condition than in the falling and flat contour conditions. The hypothesis was tested by four orthogonal planned comparisons, which included comparisons between rising and falling/flat contours (1) across vowel contexts, (2) in /a/ vowel context, (3) in /i/ vowel context, and (4) in /u/ vowel context. Difference scores for the percentages of /g/ responses are illustrated in Figure 7.17 and the planned comparisons statistics are presented in Table 7.3.
Figure 7.17. Difference scores for the percentages of /g/ responses as a function of pitch contour and vowel context.

Table 7.3

<table>
<thead>
<tr>
<th>Planned comparison statistics for /g/ responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rising vs. Falling/Flat</td>
</tr>
<tr>
<td>t</td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>Across Vowel Context</td>
</tr>
<tr>
<td>/a/ Vowel Context</td>
</tr>
<tr>
<td>/i/ Vowel Context</td>
</tr>
<tr>
<td>/u/ Vowel Context</td>
</tr>
</tbody>
</table>

Note: Following the recommendations of Volker (2006), the value of $d$ was calculated by dividing the difference between the paired scores by the $SD$ of the differences between the paired scores and then multiplying the result by the square root of 2.

As can be seen in Table 7.3, no significant differences were found between the rising and the falling/flat contour conditions.

It is worth noting that the trends in Figures 7.15, 7.16, and 7.17 suggest that the percentages of /b/ responses increased (i.e., difference scores are above 0) and the percentages of /d/ and /g/ responses appear to have decreased (i.e., the difference scores are predominantly below 0) as a result of training. An additional analysis was conducted to examine the overall change in the percentages of /b/, /d/, and /g/ responses as result of training. One sample t-tests were conducted to investigate whether the overall increase/decrease in the percentages of /b/, /d/, and /g/ responses was significant (i.e., different from 0). It was found that the increase in the mean percentages of /b/ responses ($M=7.21$, $SD=7.47$) was significant (i.e., significantly different from 0), $t(24)=4.82$, $p<.001$, $d=.96$; and the decrease in the mean percentages
of /g/ responses ($M=-5.77, SD=5.05$) was significant, $t(24)=-5.71, p<.001, d=-1.14$. However, the overall change in percentages of /d/ responses ($M=-1.49, SD=5.43$) was not significant, $t(24)=-1.38, p=.18, d=-.28$. Thus, exposing participant to the auditory disyllables during training increased the auditory responses (i.e., /b/ responses) and decreased the visual responses (i.e., /g/ responses) to McGurk stimuli.

### 7.5.4 Discussion

Experiment 5 investigated whether the perception of pitch contour modified McGurk stimuli would change as result of contour-consonant associative training. During training, participants were repeatedly presented with auditory /vbv/, /vdv/, and /vgv/ disyllables in falling, flat, and rising pitch contours, respectively. It was hypothesized that such training would change the percentages of /b/, /d/, and /g/ responses for McGurk stimuli as a function of pitch contour. Specifically, it was hypothesized that (1) the percentages of /b/ responses for the McGurk stimuli increases more in the falling than in the rising and flat contour conditions, (2) the percentages of /d/ responses for the McGurk stimuli increases more in the flat than the in falling and rising contour conditions, and (3) the percentages of /g/ responses for the McGurk stimuli increase more in the rising than in the flat and falling contour conditions.

Support for the hypotheses was only found in the /a/ vowel context condition. It was found that the increase in percentages of /b/ responses was higher in the falling contour condition than in the flat and rising contour conditions. There are two questions that need to be discussed with regard to this finding. Firstly, the question of why the learning of contour-consonant association was only reflected by the change in the /b/ responses. Secondly, the question of why training effect was only observed in the /a/ vowel context conditions.

In addressing the first question, it should be emphasized that training was conducted on the auditory syllables and the auditory syllable in the McGurk stimuli involves the consonant /b/. Since participants were exposed to the contour-consonant associations in the auditory dimension, it is possible that the learned contour-consonant associations are more likely to manifest in the change of the auditory responses.
In addressing the second question, one should consider that the proportion of /b/, /d/ and /g/ responses differ as a function of vowel context. The preliminary analysis conducted on the pre-test scores revealed that the percentage of auditory responses to McGurk stimuli (i.e., /b/ responses) is the lowest in the /a/ vowel context condition. It is possible that the training effect only manifested in the /a/ vowel context because the high pre-test percentages of /b/ responses in the /i/ and /u/ vowel context resulted in a ceiling effect.

In conclusion, the present experiment found some support that the perception of pitch contour modified McGurk stimuli can be changed by contour-consonant associative training. It was found that the /b/ responses to the McGurk stimuli in the /a/ vowel context increased the most in the falling contour condition, which was associated with the consonant /b/ during training. It has been suggested that the fact that significant changes were only observed in the change of /b/ responses is due to the auditory nature of perceptual training. If that reasoning is valid then one could argue that associative training in which pitch contour is paired with the visual signal of consonants would be more likely to result in the hypothesized trend in the /g/ (i.e., visual) responses to the McGurk stimuli than the /b/ (i.e. auditory) responses. This question was investigated in Experiment 6.

7.6 Experiment 6

7.6.1 Aim, Design, and Hypotheses

The aim of Experiment 6, similar to that of Experiment 5, was to modify the perception of McGurk stimuli using contour-consonant associative training. However, in Experiment 6 pitch contour was paired with the visual signal (i.e., video display of articulation) of disyllables. Visual signals of disyllables with consonant /b/, /d/, and /g/ were paired with falling, flat, and rising pitch contours, respectively. Participants were required to identify the stop consonant in the visual signals (as /b/, /d/, or /g/) and they received corrective feedback after their response.
The experiment was based on a within-subjects design, where the perception of McGurk stimuli was tested before and after training in three different pitch contours. Dependent variables were the change in the percentages of /b/, /d/, and /g/ responses as a result of training (measured by the difference between the after and before training percentage scores). It was hypothesised that the change in the perception of McGurk stimuli (i.e., change in the percentages of /b/, /d/, and /g/ responses) would be significantly influenced by pitch contour.

The following related hypotheses were proposed.

1. The percentages of /b/ responses would increase with training more in the falling contour condition than in the flat and rising contour conditions.

2. The percentages of /d/ responses would increase with training more in the flat contour condition that in the falling and rising contour conditions.

3. The percentages of /g/ responses would increase with training more in the rising contour condition that in the falling and flat contour conditions.

7.6.2 Method

7.6.2.1 Participants

Participants were 26 (4 males and 22 females) Psychology 1 students from the University of Western Sydney, who received course credit for their participation. All participants reported normal hearing and vision. The mean age of participants was 19.6 years (SD=2.9).

7.6.2.2 Stimuli

Stimulus Recording

The stimuli used in Experiment 6 were constructed based on the same recorded material that was used in Experiment 5. Eight speakers (4 males, 4 females) were video-recorded uttering /aba/, /ada/, /aga/, /ibi/, /idi/, /igi/, /ubu/, /udu/, and /ugu/ disyllables. The production of disyllables was recorded using a digital video camera and an external microphone attached to the video camera. Recordings were done in a
recording studio with black background at the University of Western Sydney. Speakers were sitting in front of the video camera with the microphone located at approximately 200 cm from their mouth. For the recording, speakers were wearing a black T-shirt and they removed glasses and jewellery. Lighting conditions were set so that the speakers’ articulations could be clearly recorded. Speakers were asked to utter each disyllable three times in following order: /aba/x3, /ada/x3, /aga/x3, /ibi/x3, /idi/x3, /igi/x3, /ubu/x3, /udu/x3, and /ugu/x3. Speakers were instructed to utter the disyllables naturally (i.e., not over or under articulate) and with approximately equal length and flat pitch contour. To reduce variations in the lengths of produced disyllables across speaker, prior to recording the speakers listened to a recording of a female speaker producing each disyllable three times. The nine disyllables were also written on an A4 paper which was shown to the speakers to remind them about the order of the to-be-produced disyllables. Although speakers were allowed to look at the written disyllables, they were instructed to look towards the camera during the production of disyllables. Speakers were also told to close their lips before starting and after finishing the production of each disyllable.

Stimulus Editing
The recorded disyllables were cut (using a video editing program ULead) into 1800 msec long video segments with the criterion that the utterance of the disyllable (measured from the onset of mouth opening) started at 600 msec. River Past Audio Converter program was used to separate the audio signals from the video files and to normalize loudness. Praat speech analysis software program was used to extract the duration and the mean pitch of the auditory disyllables (See Appendix L) and to create pitch contour modified disyllables. The pitch value in the flat contour modification corresponded to the mean pitch value of the original disyllable. The pitch change in the rising and falling pitch contour modification was set to be six semitones per 650 msec duration (i.e., 0.92 semitones per 100 msec), retaining the mean pitch value of the original signal. For disyllables with consonant /b/, three pitch contour modified (falling, mid, rising) stimuli were created. Disyllables with /d/ consonant were modified to have flat pitch contour, and disyllables with /g/ consonant were modified to have rising pitch contour.
McGurk stimuli used in the Experiment 6 were the same as those used in Experiment 5. That is, there were 72 (8 speakers x 3 pitch contour x 3 vowel context) McGurk stimuli that were constructed by pairing /vbv/ audio signals in each pitch contour (falling, flat, rising) with /vgv/ visual articulations. The auditory signal was paired with the visual signal so that onset of consonant in the auditory signal was synchronised with the onset of consonant in the visual signal.

Training stimuli were constructed by pairing 24 (8 speaker x 3 vowel context) video signals of /vbv/, /vdv/, and /vgv/ disyllables with falling, flat and rising sinusoidal tones, respectively. Falling sinusoidal tones were constructed (using Praat speech analysis software program) by extracting the duration and pitch contour of the falling-/vbv/ disyllables. Flat sinusoidal tones were constructed by extracting the duration and pitch contour of the flat-/vdv/ disyllables. Rising sinusoidal tones were constructed by extracting the duration and pitch contour of the rising-/vdv/ disyllables.

Additional stimuli included nine unedited congruent audio-visual disyllables (i.e., /aba/, /ada/, /aga/, /ibi/, /idi/, /igi/, /ibu/, /udu/, /ugu/) which were used in the practice trials. Visual stimuli included the words ‘YES’ (printed in black) and ‘NO’ (printed in black) that were displayed on the computer screen as corrective feedback. If participants failed to enter their response within a certain time limit the word ‘LATE’ (printed in red) appeared on the computer screen.

7.6.2.3 Equipment

The experiment was conducted using a laptop computer running DMDX experimental software. Instructions were given verbally by the experimenter and also appeared on the computer screen outlining what participants were required to do at each stage. Participants listened to the syllables through headphones at a comfortable listening level (approximately 65 dB SPL). Identification responses were indicated by pressing marked keys on the keyboard; /b/ responses were indicated by pressing the left shift key.

---

52 The ‘late’ sign was set to appear at 2300 msec (in Block 1), 2100 msec (in Block 2), and 1900 msec (in Block 3) after the onset of the audio files.
button (marked ‘B’), /d/ responses were indicated by pressing the shift bar (marked ‘D’), and /g/ responses were indicated by pressing the right shift button (marked ‘G’).

7.6.2.4 Procedure

Before commencing the experiment participants read an information sheet about the experiment, they signed a consent form, and filled out a background questionnaire. The information sheet, consent form, and questionnaire are shown in Appendix O, P, Q, respectively.

Pre-test Phase

Before pre-test, participants were familiarized with the task by receiving nine practice trials with feedback. In the practice trials participants were presented with congruent audio-visual stimuli (i.e., /aba/, /ada/, /aga/, /ibi/, /idi/, /igi/, /ubu/, /udu/, /ugu/) and they were told to indicate the consonant they perceived by pressing marked keys on the keyboard (left shift=/b/, space=/d/, right shift=/g/). McGurk stimuli were not used in the practice trials because, due to their ambiguity, no objective feedback could have been given to participants. During pre-test participants were presented with 72 McGurk stimuli (3 contour x 3 vowel context x 8 speaker) in random order. Participants were asked to indicate the consonant they perceived by pressing marked keys on the keyboard (left shift=/b/, space=/d/, right shift=/g/). Participants were also told that some of the stimuli used in the test phase may be harder to identify than stimuli used in the practice trial. During pre-test participants received no feedback after their response.

Training Phase

During training participant were presented with the /vbv/, /vdv/, and /vgv/ visual signals paired with falling, flat, and rising sinusoidal tones, respectively. Training was divided into 3 blocks. In each block, participants were presented with the 72 training stimuli twice in random order. At each trial, participants were asked to indicate the consonant (as /b/, /d/, or /g/) they perceived in the visual disyllables as fast as they could by pressing marked keys on the keyboard (left shift=/b/, space=/d/, right shift=/g/). Correct responses were followed by the display of the word ‘YES’, and incorrect responses were followed by the display of the word ‘NO’. In order to
keep participants motivated during training, they were told that they have limited time for entering their response and the time limit would gradually decrease during the training. Participants were told that in Block 1 the time limit is 2 seconds, in Block 2 the time limit is 1 second and in Block 3 the time limit is .5 second\textsuperscript{53}. If participants failed to respond within a certain time limit, the word ‘LATE’ was displayed on the computer screen. Participants were instructed to avoid late and incorrect responses. Training was preceded with nine practice trials that contained some of the items used during training.

\textit{Post-Test Phase}
Post-test was methodologically the same as pre-test. Participants were presented with 72 McGurk stimuli (3 contour x 3 vowel context x 8 speaker) in random order and they were asked to indicate the consonant they perceived by pressing marked keys on the keyboard (left shift=/b/, space=/d/, right shift=/g/). Participants received no feedback after their response.

The experiment took approximately 45 minutes including instructions and breaks. Participants were debriefed at the end of the experiment.

\textbf{7.6.3 Results}
Prior to conducting statistical analyses, the assumption of normality was tested in each data set. However, it was taken into consideration that relatively simple parametric tests (e.g., t-test and ANOVA) are robust with respect to the violation of normality in the case of large sample sizes (i.e., $N>30$) (Tabachnick & Fidell, 2001). Since the sample used in the present experiment was relatively large (i.e., $N>30$) data were analysed using either t-test or ANOVA even when the assumption of normality was violated.

\textsuperscript{53} The time limits told to participants were only approximates of the actual appearance of the ‘late’ sign. The ‘late’ sign in Block 1, 2, and 3 was set to appear at 2300, 2100, and 1900 millisecond after the onset of the audio files, respectively.
7.6.3.1 Pre-Test Analysis

Before analysing the effect of training on the perception of McGurk stimuli, the baseline perception of McGurk stimuli was analysed to examine the proportion of /b/, /d/ and /g/ responses as a function of pitch contour and vowel context. The before training percentages of /b/ responses are shown in Figure 7.18.

![Figure 7.18. Percentages of /b/ responses before training as a function of pitch contour and vowel context.](image)

Data were analysed by a two-way (contour x vowel context) repeated measures ANOVA. No significant main effect of pitch contour and vowel context by pitch contour interaction was found, $F(2,50) = 2.27, p = .11$, partial $\eta^2 = .08$, $F(4, 100) = .62$, $p = .65$, partial $\eta^2 = .02$, respectively. However the main effect of vowel context was significant, $F(1.6,39.5) = 7.33, p = .004$, partial $\eta^2 = .23$. Post hoc pairwise comparisons with Bonferroni adjustment revealed that the mean percentage of /b/ responses is significantly lower in the /a/ vowel context than in the /i/ and /u/ vowel contexts ($p=.05$ and $p=.004$, respectively). There was no significant difference in the percentages of /b/ responses between the /i/ and /u/ vowel context ($p=.29$).

---

54 The normality of assumption was only met in the /a/-flat condition; however no outliers were identified in any of the conditions.

55 Greenhouse-Geisser adjustment was applied to the degrees of freedom since the homogeneity of covariance assumption was violated.
The before training percentages of /d/ responses are shown in Figure 7.19.

![Figure 7.19. Percentages of /d/ responses before training as a function of pitch contour and vowel context.](image)

A two-way (contour x vowel context) repeated measures ANOVA revealed no significant main effect of pitch contour, $F(2,50) = 2.11, p = .13, \text{partial } \eta^2 = .08$ and no significant contour by vowel context interaction, $F(4, 100) = 1.24, p = .3, \text{partial } \eta^2 = .05$. On the other hand, the main effect of vowel context was significant, $F(1.4,35.3) = 9.52, p = .002, \text{partial } \eta^2 = .28^{56}$. Post hoc pairwise comparisons with Bonferroni adjustment revealed that the mean percentage of /d/ responses is significantly lower in the /u/ vowel context than in the /a/ and /i/ vowel contexts ($p=.006$ and $p=.006$, respectively). There was no significant difference in the percentages of /d/ responses between the /i/ and /u/ vowel context ($p=.18$).

The before training percentages of /g/ responses are shown in Figure 7.20.

---

56 Greenhouse-Geisser adjustment was applied to the degrees of freedom since the homogeneity of covariance assumption was violated.
A two-way (contour x vowel context) repeated measures ANOVA revealed no significant main effect of pitch contour, $F(2,50) = .87$, $p = .42$, partial $\eta^2 = .03$, and no significant contour by vowel context interaction, $F(4, 100) = .17$, $p = .95$, partial $\eta^2 = .01$. On the other hand, the main effect of vowel context was significant, $F(2,50) = 6.81$, $p = .002$, partial $\eta^2 = .21$. Post hoc pairwise comparisons with Bonferroni adjustment revealed that the mean percentage of /g/ responses is significantly lower in the /i/ vowel context than in the /a/ vowel context ($p<.001$). There was no significant difference in the percentages of /g/ responses between the /a/ and /u/ vowel contexts ($p=.38$) and the /i/ and /u/ vowel contexts ($p=.23$).

### 7.6.3.2 Training Effect Analysis

The training effect was measured by calculating the difference between the post-test and the pre-test percentage values. Positive difference scores indicate an increase in the percentages of responses as an effect of training; and negative values indicate a decrease in the percentages of responses as an effect of training.

---

57 Analysing the difference between the post- and pre-test values is statistically more powerful to investigate training effect than simply analysing the post-test values because the prior removes concomitant variations in responses (Steel & Torrie, 1980, p. 401).
The first hypothesis stated that the percentage of /b/ responses increases with training more in the falling contour condition than in the flat and rising contour conditions. The hypothesis was tested by four orthogonal planned comparisons, which included comparisons between falling and flat/rising contours (1) across vowel contexts, (2) in /a/ vowel context, (3) in /i/ vowel context, and (4) in /u/ vowel context. Difference scores for the percentages of /b/ responses are illustrated in Figure 7.21, and the results of planned comparisons are presented in Table 7.4.

![Figure 7.21](image-url)  
*Figure 7.21*. Difference scores for the percentages of /b/ responses as a function of pitch contour and vowel context.

<table>
<thead>
<tr>
<th>Planned comparison statistics for /b/ responses</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Falling vs. Flat/Rising Contrasts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Across Vowel Context</td>
<td>-0.29</td>
<td>25</td>
<td>0.78</td>
<td>-0.08</td>
</tr>
<tr>
<td>/a/ Vowel Context</td>
<td>-0.65</td>
<td>25</td>
<td>0.52</td>
<td>-0.18</td>
</tr>
<tr>
<td>/i/ Vowel Context</td>
<td>0.33</td>
<td>25</td>
<td>0.75</td>
<td>0.09</td>
</tr>
<tr>
<td>/u/ Vowel Context</td>
<td>-0.01</td>
<td>25</td>
<td>0.99</td>
<td>0.00</td>
</tr>
</tbody>
</table>

*Note:* Following the recommendations of Volker (2006), the value of $d$ was calculated by dividing the difference between the paired scores by the $SD$ of the differences between the paired scores and then multiplying the result by the square root of 2.

As can be seen in Table 7.4, no significant differences were found between the falling and the flat/rising contour conditions.

The second hypothesis stated that the percentage of /d/ responses increases with training more in the flat contour condition that in the falling and rising contour...
conditions. The hypothesis was tested by four orthogonal planned comparisons, which included comparisons between flat and falling/rising contours (1) across vowel contexts, (2) in /a/ vowel context, (3) in /i/ vowel context, and (4) in /u/ vowel context. Difference scores for the percentages of /d/ responses are illustrated in Figure 7.22, and the results of planned comparisons are presented in Table 7.5.

![Figure 7.22. Difference scores for the percentages of /d/ responses as a function of pitch contour and vowel context.](image)

<table>
<thead>
<tr>
<th>Vowel Context</th>
<th>Increase of 'D' Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>/a/</td>
<td>-15</td>
</tr>
<tr>
<td>/i/</td>
<td>-10</td>
</tr>
<tr>
<td>/u/</td>
<td>-5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Vowel Context</th>
<th>Increase of 'D' Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>/a/</td>
<td>0</td>
</tr>
<tr>
<td>/i/</td>
<td>5</td>
</tr>
<tr>
<td>/u/</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Vowel Context</th>
<th>Increase of 'D' Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>/a/</td>
<td>15</td>
</tr>
<tr>
<td>/i/</td>
<td>10</td>
</tr>
<tr>
<td>/u/</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 7.5

Planned comparison statistics for /d/ responses

<table>
<thead>
<tr>
<th>Flat vs. Falling/Rising</th>
<th>t</th>
<th>df</th>
<th>p</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Across Vowel Context</td>
<td>-0.19</td>
<td>25</td>
<td>0.85</td>
<td>-0.05</td>
</tr>
<tr>
<td>/a/ Vowel Context</td>
<td>-1.14</td>
<td>25</td>
<td>0.26</td>
<td>-0.32</td>
</tr>
<tr>
<td>/i/ Vowel Context</td>
<td>0.92</td>
<td>25</td>
<td>0.36</td>
<td>0.26</td>
</tr>
<tr>
<td>/u/ Vowel Context</td>
<td>-0.12</td>
<td>25</td>
<td>0.90</td>
<td>-0.03</td>
</tr>
</tbody>
</table>

Note: Following the recommendations of Volker (2006), the value of \(d\) was calculated by dividing the difference between the paired scores by the SD of the differences between the paired scores and then multiplying the result by the square root of 2.

As can be seen in Table 7.5, no significant differences were found between the flat and the falling/rising contour conditions.

The third hypothesis stated that the percentage of /g/ responses increases with training more in the rising contour condition than in the falling and flat contour conditions. The hypothesis was tested by four orthogonal planned comparisons, which included comparisons between rising and falling/flat contours (1) across vowel contexts, (2) in
/a/ vowel context, (3) in /i/ vowel context, and (4) in /u/ vowel context. Difference scores for the percentages of /g/ responses are illustrated in Figure 7.23, and the results of planned comparisons are presented in Table 7.6.

![Figure 7.23](image)

**Figure 7.23.** Difference scores for the percentages of /g/ responses as a function of pitch contour and vowel context.

<table>
<thead>
<tr>
<th>Table 7.6</th>
<th>Planned comparison statistics for /g/ responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rising vs. Falling/Flat</td>
<td>t</td>
</tr>
<tr>
<td>Across Vowel Context</td>
<td>2.87 *</td>
</tr>
<tr>
<td>/a/ Vowel Context</td>
<td>0.88</td>
</tr>
<tr>
<td>/i/ Vowel Context</td>
<td>1.94</td>
</tr>
<tr>
<td>/u/ Vowel Context</td>
<td>1.39</td>
</tr>
</tbody>
</table>

Note: Following the recommendations of Volker (2006), the value of d was calculated by dividing the difference between the paired scores by the SD of the differences between the paired scores and then multiplying the result by the square root of 2.

As indicated in Table 7.6, the overall increase of /g/ responses was significantly higher in the rising contour condition than in the falling/flat contour conditions. The effect size values in the three different vowel contexts suggest that the training effect was the largest in the /i/ vowel context (i.e., d=.54).

---

58 Alpha level was set at .05, following the argument that orthogonal comparisons driven by the hypothesis require no adjustment to the alpha level (Catham, 1999; O’Keefe, 2003).
It is worth noting that the trends in the difference scores (See Figure 7.21, 7.22, and 7.23) suggest that the percentages of /b/ response decreased (i.e., difference scores are below 0) and the percentages of /d/ and /g/ responses increased (i.e., the difference scores are above 0) as a result of training. One sample t-tests were conducted to investigate whether the overall increase/decrease in the percentages of responses was significant (i.e., different from 0). It was found that the decrease in the mean percentages of /b/ responses was significant, \( t(25)=-3.06, \ p=0.005, \ d=-0.6 \); and the increase in the mean percentages of /d/ and /g/ responses was also significant, \( t(25)=2.42, \ p=0.02, \ d=0.47, \ t(25)=3.26, \ p=0.003, \ d=0.64 \), respectively. Thus exposing participants to the visual signals of the disyllables during training increased the percentages of visual and fused responses (i.e., /g/ and /d/ responses) to McGurk stimuli and decreased the percentages of auditory responses (i.e., /b/ responses) to McGurk stimuli.

### 7.6.4 Discussion

Experiment 6 investigated whether the perception of pitch contour modified McGurk stimuli would change as result of contour-consonant associative training. During training, participants were repeatedly presented with visual /v/bv/, /v/dv/, and /v/gv/ disyllables paired with falling, flat, and rising sinusoidal pitch contours, respectively. It was hypothesized that such training would change the percentages of /b/, /d/, and /g/ responses for pitch contour modified McGurk stimuli. Specifically, it was hypothesized that (1) the percentages of /b/ responses for the McGurk stimuli would increase more in the falling than in the rising and flat contour conditions, (2) the percentages of /d/ responses for the McGurk stimuli would increase more in the flat than in the falling and rising contour conditions, and (3) the percentages of /g/ responses for the McGurk stimuli would increase more in the rising than in the flat and falling contour conditions.

Support for the hypotheses was only found in the /g/ response category; the increase in percentages of /g/ responses was higher in the rising contour condition than in the falling and flat contour conditions. It is possible that the expected trend was only observed in the visual responses (i.e., /g/ responses) to McGurk stimuli because training was conducted on the visual disyllables (i.e., during training pitch contour
was paired with the visual disyllables). These findings complement the findings of Experiment 5, which revealed a change in the auditory responses (i.e., /b/ responses) to McGurk stimuli as a result of auditory contour-consonant training.

7.7 General Discussion – Experimental Set 2

7.7.1 Summary and Interpretation of Results

Experiments 4, 5, 6 investigated the effect of conditioning on the perception of pitch contour modified McGurk stimuli. The McGurk stimuli were constructed by pairing the visual articulation of /g/ consonant with the auditory signal of /b/ consonants presented in three different pitch contours (i.e., falling, flat, and rising). During training, participants were repeatedly presented with /b/, /d/, and /g/ consonants in falling, flat, and rising pitch contour contexts, respectively (i.e., classical conditioning). In Experiment 4 and 5, pitch contour was paired with the auditory signal of the consonants, and in Experiment 6, it was paired with the visual displays of articulations. During training, participants were asked to identify the stop consonants and they received corrective feedback (i.e., operant conditioning). The perception of McGurk stimuli was tested both before and after training and it was hypothesized that, (1) the percentages of /b/ responses would increase with training more in the falling contour condition than in the flat and rising contour conditions, (2) the percentages of /d/ responses would increase with training more in the flat contour condition that in the falling and rising contour conditions, and (3) the percentages of /g/ responses would increase with training more in the rising contour condition that in the falling and flat contour conditions.

McGurk stimuli used in Experiment 4 consisted of the pairing of visual /ga/ and auditory /ba/ syllables uttered by either a male or a female speaker (speaker type included as a between subject design). It was found that the change in the identification of McGurk stimuli (i.e., proportion of /b/, /d/, and /g/ responses) as a result of training was not affected by pitch contour. Although no change was observed in participants’ responses to McGurk stimuli, analysis conducted on the reaction times suggest that participants, to some extent, learned the contour-consonant. Reaction times of /da/ responses for McGurk stimuli decreased the most in the flat contour condition, and reaction times of /ga/ responses for McGurk stimuli
decreased the most in the rising contour condition. It was argued that the lack of change in the identification responses might have been due to the fact that in the pre- and post-tests participants were presented with only one McGurk stimulus in each pitch contour. In other words, the training effect was analysed using categorical data, which might have not been powerful enough to detect the subtle changes in perception. This issue was addressed in Experiments 5 and 6.

The method of Experiments 5 and 6 was similar to that of Experiment 4, however, the number and the variability of stimuli were increased in order to improve the effectiveness of training and the sensitivity of test measures. Stimuli in Experiments 5 and 6 were constructed using eight different speakers and three different vowel contexts (/a/, /i/, /u/). In addition, the prominence of pitch contour was increased by presenting consonants in /vcv/ (vowel-consonant-vowel) disyllables context. In Experiment 5, participants were trained to associated pitch contours with the auditory signals of disyllables, and in Experiment 6 participants were trained to associate pitch contours with the visual articulations of disyllables. Support for the hypotheses in Experiment 5 was found in the /b/ response category (i.e., auditory response for McGurk stimuli) and support for the hypotheses in Experiment 6 was found in the /g/ response category (i.e., visual response for the McGurk stimuli). These results are considered complementary and have been explained by the fact that disyllables during conditioning were presented aurally in Experiment 5 and visually in Experiment 6.

In addition to the conditioning effect, in Experiment 5 the overall percentage of /b/ responses to McGurk stimuli increased with training. On the contrary, training in Experiment 6 increased the overall percentage of /g/ responses to McGurk stimuli. These trends are, again, complementary and are due presumably to the fact that training was based on the presentation of auditory speech signals in Experiment 5 and the presentation of visual speech signals in Experiment 6. In other words, the perception of McGurk stimuli was not only influenced by conditioning but also by mere exposure.

The finding that the hypotheses were only supported in one response category in both Experiment 5 and 6 indicates that the observed effect of conditioning was subtle. The fragility of the effect may be due to the nature of the conditioning training. Although
contours and consonants were paired during training, the training task itself (i.e.,
identification of the consonants within a certain time limit) did not require
participants’ to overtly attend to pitch contour. It is thus suggested that modifying the
task in a way that would increase participants’ attention to pitch contour would
increase conditioning. For instance, pairing pitch contour with auditory syllables
embedded noise or with hypo-articulated visual syllable signals may increase
participants’ tendency to utilize pitch contour as an invariant that affords the
identification of consonants. Similar effects could be obtained by increasing the
salience of pitch contour. The salience of pitch contour could be amplified by
increasing the duration of vowels in the /vcV/ syllables and/or by increasing the slope
of pitch change. These possibilities could be further investigated in future
experiments.

Another aspect of the experiments that presumably contributed to the fragility of the
results is the fact that perceptual training in the present experiments was relatively
short (i.e., 20-30 minutes). Previous perceptual training experiments in the speech
perception domain (See Chapter 6) often used multiple training sessions that added up
to many hours of training (e.g., 80.5 hours in the study conducted by Brooks et al.,
1986; 45 hours in the study conducted by Barlow et al., 1997). Such experiments
suggest that speech-related conditioning requires extensive training. Thus, it is likely
that the results obtained in the present experiments could be further strengthened by
increasing the number of training trials.

7.7.2 Explanation of Results in Terms of the Ecological-Conditioning Theory of
Perceptual Learning

According to the ecological theory of perception, perceptual learning involves
discovering perceptual information (i.e., invariants) that conveys meaning (i.e.,
affordances) to the perceiver. As Michaels and Carello (1981) noted, perceptual
learning is underlined by “the education of attention to those invariants …that specify
the affordances of the local environment” (p. 81). Within the ecological framework,
the results of the present experiments can be explained by suggesting that participants
educated their attention to pitch contour information because it specified the
consonant that participants were required to identify during training.
Participants’ attention was directed to pitch contour by simple conditioning principles. That is, perceptual training, involved pairing certain pitch contours with certain consonants (i.e., classical conditioning) and by providing corrective feedback after participants identified the consonant (i.e., operant conditioning). As argued in Chapter 3, pairing a conditioned stimulus with an unconditioned stimulus changes the perceived meaning (i.e., affordances) of the conditioned stimulus. In the present experiment, pitch contour is considered as a conditioned stimulus and the auditory and visual invariants of consonant are considered as the unconditioned stimuli. It is argued that classical conditioning changed the perceived meaning of pitch contour which was observed as a change in the perception of pitch contour modified McGurk stimuli.

Operant conditioning (providing corrective feedback) directs attention to invariants that predict correct responses. Such invariants in the present experiment involved both the invariants that normally specify the stop-consonant (i.e., place of articulation) and the pitch contour, which was consistently paired with the consonant. Thus, it is possible that operant conditioning directed attention to both the place of articulation and pitch contour information. In fact, the increased attention to the place of articulation information is demonstrated by the overall increase in the auditory response to McGurk stimuli after auditory training (i.e., Experiment 5) and the overall increase in the visual response to McGurk stimuli after visual training (i.e., Experiment 6). Although, it is theoretically valid to suggest that operant conditioning contributed to the increase in the auditory and visual responses in Experiment 5 and 6 (respectively), the observed effects can also be partly due to mere exposure. It is possible that participants would have also increased their sensitivity to the place of articulation cues even if they were given no corrective feedback.

### 7.7.3 Implications of the Results

#### 7.7.3.1 Implications for McGurk Effect Research

Analyses conducted on the perception of McGurk stimuli in the pre-test phase support the findings of previous studies. Previous studies found that McGurk effect (i.e., /d/ or ‘fusion’ responses to McGurk stimuli) is the greatest in /i/ context, moderate in /a/ context, and almost non-existent in /u/ context (Burnham, 1988; Green & Norrix,
1997; Green, Kuhl, & Meltzoff, 1988; Shigeno, 2000). Such a trend is supported by the findings of both Experiments 5 and 6. In addition, by demonstrating that conditioning and mere exposure influence the perception of McGurk stimuli, the present experimental set adds to the existing research investigating McGurk effect. The experimental demonstration of perceptual learning is particularly relevant to developmental studies on McGurk effect (Boliek et al., 1996; Hockley & Polka, 1994; Massaro et al., 1986; McGurk & McDonald, 1976) as they are based on the assumption that the perception of McGurk stimuli depends on previous exposure to audio-visual speech.

7.7.3.2 Implications for Speech Perception Research

The results of the present study extend the small number of previous studies which have already demonstrated speech-related perceptual learning via conditioning (See Chapter 5). Such studies (Brooks & Frost, 1983; Brooks et al., 1986a, 1986b; Plant, 1998; Stephens & Holt, 2006) showed that conditioning can be used to train participants to use novel (i.e., non-speech related) perceptual variables in speech perception. For instance, Stephens & Holt (2006) repeatedly presented participants with auditory /vcv/ disyllables paired with computer generated animated movements, and demonstrated that after such associative training participants were able to identify consonants based on the animation alone. Another study conducted by Brooks et al. (1986) trained participants to associate the auditory speech signals with tactile stimulation and found that after training participants could identify some words based on tactile stimulation alone. The general finding of these studies is extended by the present experiments because they demonstrate that participants can be conditioned to use a non-speech related invariant (i.e., pitch contour) to identify speech sounds.
CHAPTER 8

GENERAL DISCUSSION
8.1 The Proposed Conditioning-Ecological Theory of Perceptual Learning

This thesis has been concerned with the application of conditioning principles to the ecological theory of perceptual learning. The proposed ecological-conditioning theory of perceptual learning was motivated by the limitations of the conventional (i.e., Gibsonian) ecological theory of perceptual learning. As it is admitted by Gibson and Gibson (1955b) themselves, the ecological theory of perceptual learning mainly focuses on the question of what humans learn to perceive as a result of experience, while the question of how learning occurs is only partially addressed by ecological psychologists. Regarding the question of what humans learn to perceive, ecological psychologists propose that humans learn to perceive invariants that specify affordances (i.e., meaning) in their environment (Gibson, 2000; Gibson & Pick, 2000; Michaels & Carello, 1981). Regarding the question of how humans learn to perceive affordances, ecological psychologists focus mainly on the role of action in perceptual learning (e.g., exploratory activity) (Gibson, 2000; Gibson & Pick, 2000) and they tend to neglect the role of simple observation in perceptual learning. It has been argued in this thesis that principles of conditioning, which have previously been used to explain a wide range of human behaviour (Domjan, 2005; Rescorla, 1988), can be used to strengthen and expand the ecological account of perceptual learning.

The theory of classical conditioning has been used to expand the ecological theory of perceptual learning by describing how humans learn affordances by observing certain regularities in the environment. According to the principles of classical conditioning, humans learn that co-occurring stimuli belong together (i.e., are associated) and such learning changes their behavioural responses to stimuli (Rescorla, 1988). In this thesis, classical conditioning has been interpreted as learning about the affordances of stimuli by observing that a certain stimulus is associated with (i.e., affords) another stimulus. This interpretation expands the ecological theory of perceptual learning since it provides a learning mechanism by which humans can learn affordances without physically interacting with (i.e., operating on) their environment. The theory of operant conditioning has been used to strengthen the ecological theory of perceptual learning by describing how humans learn affordances via acting on their
According to the principles of operant conditioning, humans learn to associate environmental stimuli with the consequences of responding to the stimuli (Skinner, 1976). This process has been interpreted in this thesis as learning about the affordances of stimuli by experiencing the consequences of interacting with the stimuli. Such an interpretation strengthens the ecological argument that action (i.e., exploratory activity) plays an important role in perceptual learning. The physiological mechanisms that underpin conditioning have been related to the increase in neural connections as a result of coordinated activity between neurons (classical conditioning) and as a result of the release of ‘reinforcing’ neurotransmitters (operant conditioning) (Donahoe & Palmer, 1994; Stein, 1997).

Since conditioning principles are considered as general principles of learning, it can be argued that their application to the ecological theory of perceptual learning increases both the explanatory and predictive power of that theory. As will be discussed next, the explanatory power of the proposed ecological-conditioning theory of perceptual learning has been demonstrated in two perceptual domains, namely musical pitch perception and speech perception.

### 8.2 Demonstrating the Explanatory Power of the Theory

#### 8.2.1 The Application of the Theory to Musical Pitch Perception

The ecological-conditioning theory of perceptual learning was applied to pitch perception by describing (1) pitch-related invariants and (2) the role of conditioning in educating attention to meaningful pitch-related invariants. Pitch-related invariants involve frequency ratios that exist at the level of harmonic tone and at the level of musical structures. The role of conditioning in directing attention to meaningful frequency ratios has been investigated in the context of *musical training*. Although mere exposure to music without any musical training also results in educating attention to frequency ratios, musical training further enhances pitch-related perceptual learning (Peretz & Babai, 1992; Trainor et al., 1999). Classical conditioning in musical training involves training students to associate frequency ratios (i.e., musical intervals) with certain musical labels (i.e., musical interval names). For instance, the frequency ratio of 2:1 is associated with the interval label of octave. In this thesis, the learning of such associations is interpreted as a type of
perceptual learning which involves learning that a certain frequency ratio affords a certain musical label. According to the ecological framework, the perceptual learning of pitch-related affordances (i.e., musical intervals) is underlined by the process of education of attention to frequency ratios that specify musical intervals. Thus, learning to associate frequency ratios with musical interval names improves the perception of frequency ratios (i.e., relative pitch perception). The role of operant conditioning in musical training involves providing students with feedback on their production of pitch intervals. The production of pitch is considered an operant behaviour as it involves interaction with the environment (e.g., controlling the vocal folds or playing an instrument). The feedback given by music teachers informs the student whether the produced frequency ratios correspond to (or affords) the musical interval they intended to produce. According to the ecological framework, using feedback in musical training educates students’ attention to frequency ratios that specify musical intervals. Thus both classical and operant conditioning play a role in enhancing pitch perception in the musically trained population.

8.2.2 The Application of the Theory to Speech Perception

The conditioning-ecological theory of perceptual learning was applied to speech perception by describing (1) speech-related invariants and (2) the role of conditioning in educating attention to meaningful speech-related invariants. Speech-related invariants have been described in terms of both articulatory (e.g., place of articulation) and auditory (e.g., formant frequencies) dimensions. Perceptual learning in speech involves educating attention to invariants that specify speech sounds that distinguish meaning (i.e., phonemes) (Best, 1994). The role of conditioning in directing attention to phonemes has been described in early developmental settings, which involves the interaction between infants and their caregivers. The role of classical conditioning has been related to learning the association between an invariant phonetic sequence (e.g., a word) and the object or event to which it relates (i.e., affords) in the environment. It has been argued that such learning is underlined by educating attention to speech sounds that are used to distinguish meaning in a given language. The role of operant conditioning in speech-related perceptual learning has been described in terms of infants’ production of speech and their caregivers’ feedback. Since caregivers reinforce infants’ production of native speech sounds, infants educate
their attention of invariants that specify native speech sounds. It has been argued that the learning of naturally occurring associations between words produced by the caregiver and their referents (i.e., classical conditioning); and associations between speech sounds produced by the infant and the caregivers feedback (i.e., operant conditioning) provide examples of learning mechanisms that describe how human infants educate their attention to meaningful speech sounds in their language.

8.3 Investigating the Predictive Power of the Theory

8.3.1 Experimental Set 1

Experimental Set 1 (i.e., Experiments 1, 2, and 3) was designed to modify the perception of ambiguous pitch stimuli (i.e., Shepard tones) using conditioning. During training, Shepard tone-pairs were paired with coloured circles (classical conditioning) in a way that the colour of the circles could be predicted by a certain invariant property of the Shepard tone-pair. During training participants were required to identify the colour of the circles that was associated with the tone-pair and they received corrective feedback after their response (operant conditioning). According to the ecological-conditioning theory of perceptual learning, participants’ attention would be directed to the invariant property of Shepard tone-pairs that are relevant in the colour-tone associative arrangements.

Experiment 1 was based on a between-subjects factorial design with the independent variable being the task-relevant auditory dimension (i.e., auditory dimension that reliably predicted the colour-tone associations). Half of the participants underwent training in which the auditory dimension that reliably predicted the colour categories was the F0 contour (i.e., F0-relevant training). The other half of the participants underwent training in which the auditory dimension that reliably predicted the colour-tone associations was the circular dimension of pitch chroma (i.e., F0-irrelevant training). After training, the perception of Shepard tritones was tested using six tritones with ascending F0 contour and six tritones with descending F0 contour. For Shepard tritones with ascending F0 contour, it was hypothesised that after training the percentage of ascending pitch contour judgments is higher in the F0-relevant training condition than in the F0-irrelevant training condition. For Shepard tritones with descending F0 contour, it was hypothesised that after training the percentage of
ascending pitch contour judgments is lower in the F0-relevant training condition than in the F0-irrelevant training condition. The results revealed no significant difference between the two training conditions. It was suggested that the lack of difference might have been due to nature of the associative training task. That is, during training participants’ could identify the colour of the circles without overtly directing attention to the auditory invariants that predicted the colours. This limitation was addressed in Experiment 2 and 3 by including training trials in which Shepard tone-pairs were presented without the coloured circles and participants were asked to indicate the colour that was associated with the tone-pair.

Experiment 2 investigated that effect of F0-relevant conditioning using a within-subjects design with time (pre-test and post-test) being the independent variable. For Shepard tritones with ascending F0 contour, it is hypothesized that the percentage of ascending pitch contour judgments would be higher after training than before training. For Shepard tritones with descending F0 contour, it was hypothesized that the percentage of ascending pitch contour judgments would be lower after training than before training. Both of these hypotheses were supported. However, it was emphasized that the possibility that the observed improvement in pitch perception was due to mere exposure to Shepard tones rather than conditioning cannot be excluded. It is possible that as result of simply being exposed to Shepard tones, participants’ sensitivity to octave-related harmonic structure of Shepard tones increased, which increased their ability to perceive the F0 (i.e., F0 can be perceptually deduced from the octave-related harmonics).

Experiment 3 investigated the effect of F0-irrelevant conditioning using a within-subjects design with time (pre-test and post-test) being the independent variable. For Shepard tritones with ascending F0 contour, it is hypothesized that the percentage of ascending pitch contour judgments would be lower after training than before training. For Shepard tritones with descending F0 contour, it was hypothesized that the percentage of ascending pitch contour judgments would be higher after training than before training. None of these hypotheses were supported as it was found that F0-irrelevant training did not change the perception of Shepard tritones. However, similar to Experiment 2, the possibility that mere exposure interacted with the conditioning cannot be excluded. It is possible that F0-irrelevant conditioning, in
itself, did direct participants’ attention away from the F0 dimension, but that negative effect was cancelled out by the positive effect of mere exposure.

Since the experiments did not control for mere exposure effect, conclusions about the effect of conditioning on pitch perception need to be carefully interpreted. The findings obtained in Experiment 2, in combination with that obtained in Experiment 3, suggest that conditioning can play a role in influencing pitch perception. However, the question remains whether the difference between the observed trends in Experiments 2 and 3 were due to the facilitatory effect of F0-relevant conditioning or the inhibitory effect of F0-irrelevant conditioning. Thus, future experiments with control for mere exposure are needed to further explore the extent that conditioning can be used to increase or decrease participants’ attention to the F0 dimension.

8.3.2 Experimental Set 2

Experimental Set 2 (i.e., Experiments 4, 5, 6) was designed to modify the perception of pitch contour modified McGurk stimuli using conditioning. McGurk stimuli used in the experiment were constructed by pairing the visual articulation of /g/ consonant with the auditory signal of /b/ consonants presented in three different pitch contours (i.e., falling, flat, and rising). During training, participants were repeatedly presented with /b/, /d/, and /g/ consonants in falling, flat, and rising pitch contour contexts, respectively (i.e., classical conditioning). Pitch contour was paired with either the auditory signal (i.e., Experiments 4 and 5) or the visual displays of articulations (i.e., Experiment 6). During training participants were required to identify the stop consonants and they received corrective feedback after their response (i.e., operant conditioning). Thus participants were conditioned to learn that falling pitch contour affords consonant /b/, flat pitch contour affords consonant /d/, and rising pitch contour affords consonant /g/. The experiments were based on a within-subjects design, in which McGurk stimuli were presented in three different pitch contours (i.e., falling, flat, rising) and their perception was tested before and after training. The perception of McGurk stimuli was tested by asking participants to identify stop consonant in each stimulus as /b/ or /d/ or /g/ consonants. There were three related hypotheses proposed namely, (1) the percentages of /b/ responses would increase with training more in the falling contour condition than in the flat and rising contour conditions, (2) the
percentages of /d/ responses would increase with training more in the flat contour condition that in the falling and rising contour conditions, and (3) the percentages of /g/ responses would increase with training more in the rising contour condition that in the falling and flat contour conditions.

McGurk stimuli used in Experiment 4 consisted of the pairing of visual /ga/ and auditory /ba/ syllables uttered by either a male or a female speaker (speaker type included as a between subject design). It was found that the change in the identification of McGurk stimuli (i.e., proportion of /b/, /d/, and /g/ responses) as a result of training was not affected by pitch contour. Although no change was observed in participants’ identification responses to McGurk stimuli, analysis of reaction times suggests that responses to McGurk stimuli were, to some extent, influenced by pitch contour. It was found that as result of training, the reaction times of /da/ responses for McGurk stimuli decreased the most in the flat contour condition, and reaction times of /ga/ responses for McGurk stimuli decreased the most in the rising contour condition. It was suggested that the lack of change in the identification responses might have been due to the fact that during the pre- and post-test phases participants were presented with only one McGurk stimulus in each pitch contour. In other words, the effect of training was analysed using categorical data, which might have not been powerful enough to detect the subtle changes in perception. This limitation was addressed in Experiments 5 and 6.

The method of Experiments 5 and 6 was similar to that of Experiment 4, however, the number and the variability of stimuli used in training and in testing (pre- and post-test) were increased in order to improve the effectiveness of training and the sensitivity of test measures. Instead of using stimuli presented in a single speaker and a single vowel context, stimuli in Experiments 5 and 6 were constructed using eight different speakers and three different vowel contexts (/a/, /i/, /u/). Also, to increase the prominence of pitch contour, consonants were presented in /vcv/ (vowel-consonant-vowel) disyllables context. In Experiment 5, participants were trained to associated pitch contours with the auditory signals of disyllables, and in Experiment 6 participants were trained to associate pitch contours with the visual articulations of disyllables. Both experiments found some support for the hypotheses. Support for the hypotheses in Experiment 5 was found in the /b/ response category (i.e., auditory
response for McGurk stimuli) and support for the hypotheses in Experiment 6 was found in the /g/ response category (i.e., visual response for the McGurk stimuli). These results complement each other considering that disyllables during conditioning were presented aurally in Experiment 5 and visually in Experiment 6.

The finding that the hypotheses were only supported in one response category in both Experiment 5 and 6 indicates that the observed effect of contour-consonant conditioning was fairly subtle. It is suggested that the fragility of the conditioning effect was due to the nature of the training task. Although pitch contours and consonants were consistently paired during training, the identification of consonants (i.e., training task) did not require participants to overtly direct their attention to pitch contour. Thus it is suggested that modifying the task in a way that would increase participants’ attention to pitch contour would increase conditioning. For instance, pairing pitch contour with auditory syllables embedded noise or with hypo-articulated visual syllable signals may increase the likelihood that participants’ learn to use pitch contour as an affordance that helps them identify consonants. Alternatively, the attention to pitch contour could also be increased by making the change in pitch more salient. For instance, the salience of pitch contour could be amplified by increasing the duration of vowels in the /vcv/ syllables and/or by increasing the slope of pitch change.

8.4 Implications and Future Directions

Although the training effects obtained in the present experiments are subtle, they are consistent with previous studies demonstrating the role of conditioning in perceptual learning (See Chapter 3, 4, and 5). For instance, in the domain of pitch perception, Jones (1971) demonstrated that learning to associate pitch with spatial position improves the perception of relative pitch. Previous studies in the speech perception domain showed that participants can be conditioned to perceive non-speech related perceptual variables, such as animated movements (Stephens & Holt, 2006) and tactile stimulations (Brooks et al., 1986), as affordances of speech sounds. These findings are extended by the results of the present experiments by demonstrating that pitch perception can be modified by tone-colour associative training (Experiments 2
and 3) and that participants can be trained to use pitch contour as an affordance of consonants (Experiments 5 and 6).

It is important to point out that in both experimental sets, the hypothesized training effects were only observed after certain modifications were made to the initial experiments. In other words, the first experiments in each experimental set can be considered as stepping stones that have led to a training design which, to some extent, resulted in the hypothesized trend in subsequent experiments (i.e., Experiment 2, 3, 5, and 6). This suggests that the Experiments 2, 3, 5 and 6 should also be considered as potential stepping stones that allow further investigation of the sufficient and necessary conditions of associative training. Keeping in mind that the discovery of learning principles requires consistent experimental work, one should consider expanding the present experimental set by addressing the previously mentioned methodological issues (e.g., investigating effects of mere exposure, increasing the salience of pitch contour).

In addition to the methodological limitations specific to each experiments, there is a general issue that requires further investigation. Although classical and operant conditioning have been described as different means of learning about affordances (See Chapter 3) they have not been separately investigated in the present experiments. In each experiment, affordances were experimentally determined by stimulus-stimulus associations (i.e., classical conditioning), and operant conditioning (i.e., giving feedback) was used to further direct attention to invariants that predicted the stimulus-stimulus associations. However, the extent to which operant conditioning contributed to perceptual learning is questionable since there were no control conditions without feedback trials. A previous study (McCandliss et al., 2002, See Chapter 5) investigating the role of feedback in perceptual learning found that giving feedback to participants is the most effective when the training task is difficult. In most of the present experiments (with the exception of Experiment 2s and 3), the training task was relatively easy, because participants could identify colours (Experiment 1) and consonants (Experiment 4, 5, and 6) without overtly attending to the invariants that predicted them. It was only in Experiments 2 and 3, that participants’ attention was explicitly directed to the invariants that predicted colour categories by presenting them with the tones only and asking them to identify the colour that was associated
with the tones. Thus, it is possible that, with the exception of Experiments 2 and 3, feedback given to participants (i.e., operant conditioning) was not fully utilized. Future studies are encouraged to investigate the differential effect of classical and operant conditioning on perceptual learning, by using experimental designs in which the two learning mechanisms are delineated.

Although the present thesis is based on the argument that conditioning changes perception, it does not claim that perception can only be changed by conditioning. Just as learning involves mechanisms other than conditioning, perception can be changed by mechanisms other than conditioning. One such mechanism suggested in the thesis was the mere exposure effect. Mere exposure effect refers to a change in perception as a result of repeated exposure to stimuli. In perception research, mere exposure is regarded as a type of sensory adaptation rather than perceptual learning (Gibson, 1963). As described earlier, the experiments conducted as part of this thesis were limited as they did not control for mere exposure effects. Thus the interpretation of the experimental findings is limited as changes to perception cannot solely be attributed to conditioning.

While the empirical findings are not fully conclusive, the theoretical contribution of the thesis expands the conventional versions of both the ecological theory of perception and theories of conditioning. The conventional (i.e., Gibsonian) view of perception has been expanded by broadening the notion of affordance. Ecological psychologists mainly use the term affordance to describe the meaning of objects in relation to action (i.e., how an animal can interact with an object). It has been argued here that the concept of affordance should encompass not only the action-related functionality of objects but also their affective or symbolic meaning. As described in Chapters 4 and 5, the meaning of culturally constructed stimuli - such as music and speech – are often abstract and cannot be described in terms of action. For instance, the affordances of stimuli used in the present experiments have been described in terms of the associative context in which they were presented (i.e., Shepard tones afforded colours and consonants afforded pitch contours).

Describing the perceptual learning of affordances in terms of associative principles, not only expands the Gibsonian notion of affordance but also the conventional theory
of conditioning. Conditioning is usually interpreted within the behaviourists’ framework as simply the learning of associations without any implications for cognition. In this thesis conditioning has been described in terms of learning to perceive affordances, which implies that cognition (i.e., perception of meaning) is an essential part of learning. Such an implication diminishes the stigma of behaviourism that is often associated with conditioning principles and ignorance of cognition. Thus the integration of conditioning theory and the ecological theory of perceptual learning enriches both theories and eliminates some of the criticisms that are directed to the conventional versions of these theories.
CHAPTER 9

REFERENCES


© August, 2008 – Bettina Stevanovic


APPENDIX A

FREQUENCIES AND AMPLITUDES OF SHEPARD TONES USED IN EXPERIMENTS 1, 2, AND 3
### Appendix A

<table>
<thead>
<tr>
<th>Shephard tones</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency (Hz)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tone 1</td>
<td>4.86</td>
<td>9.73</td>
<td>19.45</td>
<td>38.90</td>
<td>77.81</td>
<td>155.62</td>
<td>311.23</td>
<td>622.46</td>
<td>1244.93</td>
<td>2489.86</td>
</tr>
<tr>
<td>Tone 2</td>
<td>5.15</td>
<td>10.30</td>
<td>20.61</td>
<td>41.22</td>
<td>82.43</td>
<td>164.87</td>
<td>329.74</td>
<td>659.48</td>
<td>1318.95</td>
<td>2637.91</td>
</tr>
<tr>
<td>Tone 3</td>
<td>5.46</td>
<td>10.92</td>
<td>21.83</td>
<td>43.67</td>
<td>87.34</td>
<td>174.67</td>
<td>349.35</td>
<td>698.69</td>
<td>1397.38</td>
<td>2794.77</td>
</tr>
<tr>
<td>Tone 4</td>
<td>5.78</td>
<td>11.57</td>
<td>23.13</td>
<td>46.26</td>
<td>92.53</td>
<td>185.06</td>
<td>370.12</td>
<td>740.24</td>
<td>1480.48</td>
<td>2960.95</td>
</tr>
<tr>
<td>Tone 5</td>
<td>6.13</td>
<td>12.25</td>
<td>24.51</td>
<td>49.02</td>
<td>98.03</td>
<td>196.06</td>
<td>392.13</td>
<td>784.26</td>
<td>1568.51</td>
<td>3137.02</td>
</tr>
<tr>
<td>Tone 6</td>
<td>6.49</td>
<td>12.98</td>
<td>25.97</td>
<td>51.93</td>
<td>103.86</td>
<td>207.72</td>
<td>415.44</td>
<td>830.89</td>
<td>1661.78</td>
<td>3323.56</td>
</tr>
<tr>
<td>Tone 7</td>
<td>6.88</td>
<td>13.75</td>
<td>27.51</td>
<td>55.02</td>
<td>110.04</td>
<td>220.07</td>
<td>440.15</td>
<td>880.30</td>
<td>1760.59</td>
<td>3521.19</td>
</tr>
<tr>
<td>Tone 8</td>
<td>7.29</td>
<td>14.57</td>
<td>29.15</td>
<td>58.29</td>
<td>116.58</td>
<td>233.16</td>
<td>466.32</td>
<td>932.64</td>
<td>1865.28</td>
<td>3730.57</td>
</tr>
<tr>
<td>Tone 9</td>
<td>7.72</td>
<td>15.44</td>
<td>30.88</td>
<td>61.76</td>
<td>123.51</td>
<td>247.02</td>
<td>494.05</td>
<td>988.10</td>
<td>1976.20</td>
<td>3952.40</td>
</tr>
<tr>
<td>Tone 10</td>
<td>8.18</td>
<td>16.36</td>
<td>32.71</td>
<td>65.43</td>
<td>130.86</td>
<td>261.71</td>
<td>523.43</td>
<td>1046.86</td>
<td>2093.71</td>
<td>4187.42</td>
</tr>
<tr>
<td>Tone 11</td>
<td>8.66</td>
<td>17.33</td>
<td>34.66</td>
<td>69.32</td>
<td>138.64</td>
<td>277.28</td>
<td>554.55</td>
<td>1109.10</td>
<td>2218.21</td>
<td>4436.42</td>
</tr>
<tr>
<td>Tone 12</td>
<td>9.18</td>
<td>18.36</td>
<td>36.72</td>
<td>73.44</td>
<td>146.88</td>
<td>293.76</td>
<td>587.53</td>
<td>1175.06</td>
<td>2350.11</td>
<td>4700.22</td>
</tr>
<tr>
<td>Amplitude (dB)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tone 1</td>
<td>22.00</td>
<td>25.24</td>
<td>33.74</td>
<td>44.24</td>
<td>52.74</td>
<td>56.00</td>
<td>52.77</td>
<td>44.29</td>
<td>33.79</td>
<td>25.28</td>
</tr>
<tr>
<td>Tone 2</td>
<td>22.02</td>
<td>25.78</td>
<td>34.59</td>
<td>45.08</td>
<td>53.25</td>
<td>55.98</td>
<td>52.23</td>
<td>43.44</td>
<td>32.95</td>
<td>24.77</td>
</tr>
<tr>
<td>Tone 3</td>
<td>22.09</td>
<td>26.36</td>
<td>35.45</td>
<td>45.90</td>
<td>53.71</td>
<td>55.91</td>
<td>51.66</td>
<td>42.57</td>
<td>32.13</td>
<td>24.30</td>
</tr>
<tr>
<td>Tone 4</td>
<td>22.21</td>
<td>26.97</td>
<td>36.33</td>
<td>46.70</td>
<td>54.14</td>
<td>55.80</td>
<td>51.04</td>
<td>41.70</td>
<td>31.32</td>
<td>23.88</td>
</tr>
<tr>
<td>Tone 5</td>
<td>22.37</td>
<td>27.62</td>
<td>37.21</td>
<td>47.48</td>
<td>54.52</td>
<td>55.63</td>
<td>50.40</td>
<td>40.82</td>
<td>30.54</td>
<td>23.49</td>
</tr>
<tr>
<td>Tone 6</td>
<td>22.58</td>
<td>28.30</td>
<td>38.10</td>
<td>48.24</td>
<td>54.86</td>
<td>55.43</td>
<td>49.73</td>
<td>39.93</td>
<td>29.78</td>
<td>23.15</td>
</tr>
<tr>
<td>Tone 7</td>
<td>22.83</td>
<td>29.00</td>
<td>38.99</td>
<td>48.98</td>
<td>55.16</td>
<td>55.18</td>
<td>49.02</td>
<td>39.04</td>
<td>29.04</td>
<td>22.85</td>
</tr>
<tr>
<td>Tone 8</td>
<td>23.13</td>
<td>29.73</td>
<td>39.88</td>
<td>49.68</td>
<td>55.41</td>
<td>54.88</td>
<td>49.28</td>
<td>38.15</td>
<td>28.34</td>
<td>22.59</td>
</tr>
<tr>
<td>Tone 9</td>
<td>23.47</td>
<td>30.49</td>
<td>40.76</td>
<td>50.36</td>
<td>55.62</td>
<td>54.54</td>
<td>47.53</td>
<td>37.26</td>
<td>27.66</td>
<td>22.38</td>
</tr>
<tr>
<td>Tone 10</td>
<td>23.85</td>
<td>31.27</td>
<td>41.64</td>
<td>51.01</td>
<td>55.79</td>
<td>54.16</td>
<td>46.75</td>
<td>36.38</td>
<td>27.01</td>
<td>22.22</td>
</tr>
<tr>
<td>Tone 11</td>
<td>24.28</td>
<td>32.08</td>
<td>42.52</td>
<td>51.62</td>
<td>55.90</td>
<td>53.74</td>
<td>45.95</td>
<td>35.51</td>
<td>26.40</td>
<td>22.10</td>
</tr>
<tr>
<td>Tone 12</td>
<td>24.74</td>
<td>32.90</td>
<td>43.38</td>
<td>52.20</td>
<td>55.98</td>
<td>53.27</td>
<td>45.13</td>
<td>34.64</td>
<td>25.82</td>
<td>22.03</td>
</tr>
</tbody>
</table>
APPENDIX B

INFORMATION SHEET USED IN EXPERIMENT 1
Re: Research Project: Pitch Perception in Musical Tones

Intersensory perception is often discussed in psychology. It generally refers to our ability to simultaneously perceive information from different modalities. We investigate this area by examining visual and auditory perception.

Participants in the experiment complete simple visual and auditory perceptual tasks. The experiment takes approximately 45 minutes.

Data collected will not be identified with individuals and will be kept confidential. Results of the study will be reported in a PhD thesis, and will be written up as a journal article and a conference presentation.

Your participation in this experiment is voluntary and you are free to discontinue your participation at any time. Withdrawing from the experiment will not affect your future relationship with the University of Western Sydney.

If you have any additional questions please feel free to ask the investigator.

If you decide to participate please complete the attached consent form and questionnaire.

Thank you for your time.

Bettina Keresztesi
PhD Student
MARCS Auditory Laboratories
Phone: +61 2 9772 6660
Fax: +61 2 9772 6736
Email: b.keresztesi@uws.edu.au
Web: http://www.uws.edu.au/marcs/

Dr Kate Stevens
Supervisor
MARCS Auditory Laboratories
Phone: +61 2 9772 6324
Fax: +61 2 9772 6736
Email: kj.stevens@uws.edu.au
Web: http://www.uws.edu.au/marcs/

NOTE: This study has been approved by the University of Western Sydney Human Research Ethics Committee (HEC 03/167). If you have any complaints or reservations about the ethical conduct of this research, you may contact the Ethics Committee through the Research Ethics Officers (tel: 02 4570 1136). Any issues you raise will be treated in confidence and investigated fully, and you will be informed of the outcome.
APPENDIX C

CONSENT FORM USED IN EXPERIMENTS 1, 2, AND 3
Consent Form

Re: Research Project: Pitch Perception in Musical Tones

I __________________________ (Please print your name) have decided to participate in the experiment. I have read and understood the information provided. Any additional questions I asked have been answered to my satisfaction.

_____________________________________ __________________________
(Please sign your name) Date

NOTE: This study has been approved by the University of Western Sydney Human Research Ethics Committee. If you have any complaints or reservations about the ethical conduct of this research, you may contact the Ethics Committee through the Research Ethics Officers (tel: 02 4570 1136). Any issues you raise will be treated in confidence and investigated fully, and you will be informed of the outcome.
APPENDIX D

QUESTIONNAIRE USED IN EXPERIMENT 1, 2, AND 3
Questionnaire

Re: Research Project: Pitch Perception in Musical Tones

Name:

Age:

Gender: Male/Female (Please circle one)

Do you have any hearing difficulties? Yes/No (Please circle one)
If yes, please comment:

Have you had any musical training? Yes/No (Please circle one)
If yes, please comment (e.g., How many years?; What instrument? Etc):

What language/languages do you speak and how proficient are you in this/these languages?
APPENDIX E

INFORMATION SHEET USED IN EXPERIMENTS 2 AND 3
Re: Research Project: Pitch Perception in Musical Tones

The effect of learning on perception is often discussed in psychology. We investigate this by examining pitch perception in musical tones. It has previously been reported that associative strategies enhance learning and perception. We suggest that associating musical tones with colours improve pitch perception because it directs attention to certain properties of tone.

To investigate this question, we invite you to participate in an experiment. The experiments consist of 3 phases, that are pre-test, training, and post-test. In the pre- and post-test phases, participants’ pitch perception is tested. In the training phase, participants are trained to associate pairs of musical tones with certain colours. The experiment takes approximately 1 hour.

Data collected will not be identified with individuals and will be kept confidential. Results of the study will be reported in a PhD thesis, and will be written up as a journal article and a conference presentation.

Your participation in this experiment is voluntary and you are free to discontinue your participation at any time. Withdrawing from the experiment will not affect your future relationship with the University of Western Sydney.

If you have any additional questions please feel free to ask the investigator.

If you decide to participate please complete the attached consent form and questionnaire.

Thank you for your time.

Bettina Keresztesi
PhD Student
MARCS Auditory Laboratories
Phone: +61 2 9772 6660
Fax: +61 2 9772 6736
Email: b.keresztesi@uws.edu.au
Web: http://www.uws.edu.au/marcs/

Dr Kate Stevens
Supervisor
MARCS Auditory Laboratories
Phone: +61 2 9772 6324
Fax: +61 2 9772 6736
Email: kj.stevens@uws.edu.au
Web: http://www.uws.edu.au/marcs/

NOTE: This study has been approved by the University of Western Sydney Human Research Ethics Committee (HEC 03/167). If you have any complaints or reservations about the ethical conduct of this research, you may contact the Ethics Committee through the Research Ethics Officers (tel: 02 4570 1136). Any issues you raise will be treated in confidence and investigated fully, and you will be informed of the outcome.
APPENDIX F

DURATION AND MEAN PITCH OF SYLLABLES
USED IN EXPERIMENT 4
<table>
<thead>
<tr>
<th>Syllables</th>
<th>Duration (msec)</th>
<th>Mean pitch (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male speaker</td>
<td></td>
<td></td>
</tr>
<tr>
<td>/ba/</td>
<td>230</td>
<td>121</td>
</tr>
<tr>
<td>/da/</td>
<td>210</td>
<td>125</td>
</tr>
<tr>
<td>/ga/</td>
<td>300</td>
<td>120</td>
</tr>
<tr>
<td>Female speaker</td>
<td></td>
<td></td>
</tr>
<tr>
<td>/ba/</td>
<td>310</td>
<td>167</td>
</tr>
<tr>
<td>/da/</td>
<td>350</td>
<td>174</td>
</tr>
<tr>
<td>/ga/</td>
<td>310</td>
<td>166</td>
</tr>
</tbody>
</table>
APPENDIX G

INFORMATION SHEET USED IN EXPERIMENT 4
Previous research has found that watching a person’s lip while he/she is talking helps us understand what the person is saying. It has also been observed that the more often we hear a certain word the faster we recognise it in the future. We investigate these phenomena using audio and video clips consisting of /ba/ /da/ and /ga/ syllables uttered by a male/female speaker.

The experiment consists of three parts. In the first part, participants will listen to and watch the audio-video clips and they will be asked to indicate what they perceive. In the second part, we examine how fast participants recognise the same audio clips without the visual display of articulation, and how much their reaction time improves after they hear the stimuli a number of times. In this part, participants will repeatedly listen to the syllables and they will have to indicate what they hear as fast as they can. The final part of the experiment will be the same as the first part. The experiment will take approximately 25 minutes, including breaks and instructions.

Data collected will not be identified with individuals and will be kept confidential. Results of the study will be reported in a PhD thesis, and will be written up as a journal article and a conference presentation.

Your participation in this experiment is voluntary and you are free to discontinue your participation at any time. Withdrawing from the experiment will not affect your future relationship with the University of Western Sydney. If you have any additional questions please feel free to ask the investigator. If you decide to participate please complete the attached consent form and questionnaire.

Thank you for your time.

Bettina Keresztesi
PhD Student
MARCS Auditory Laboratories
Phone: +61 2 9772 6589
Fax: +61 2 9772 6736
Email: b.keresztesi@uws.edu.au
Web: http://www.uws.edu.au/marcs/

Dr Kate Stevens
Supervisor
MARCS Auditory Laboratories
Phone: +61 2 9772 6324
Fax: +61 2 9772 6736
Email: kj.stevens@uws.edu.au
Web: http://www.uws.edu.au/marcs/

NOTE: This study has been approved by the University of Western Sydney Human Research Ethics Committee (HEC 03/167). If you have any complaints or reservations about the ethical conduct of this research, you may contact the Ethics Committee through the Research Ethics Officers (tel: 02 4570 1136). Any issues you raise will be treated in confidence and investigated fully, and you will be informed of the outcome.
APPENDIX H

CONSENT FORM USED IN EXPERIMENTS 4, 5, AND 6
Consent Form

Re: Audio-Visual Speech Perception

I __________________________ (Please print your name) have decided to participate in the experiment. I have read and understood the information provided. Any additional questions I asked have been answered to my satisfaction.

_________________________________ __________________________
(Please sign your name)      Date

NOTE: This study has been approved by the University of Western Sydney Human Research Ethics Committee. If you have any complaints or reservations about the ethical conduct of this research, you may contact the Ethics Committee through the Research Ethics Officers (tel: 02 47360883). Any issues you raise will be treated in confidence and investigated fully, and you will be informed of the outcome.
APPENDIX I

QUESTIONNAIRE USED IN EXPERIMENTS 4, 5, AND 6
Questionnaire

Re: Audio-Visual Speech Perception

Name:

Age:

Gender: Male/Female (Please circle one)

Do you have any hearing difficulties? Yes/No (Please circle one)
If yes, please comment:

Have you had any musical training? Yes/No (Please circle one)
If yes, please comment (e.g., How many years?; What instrument? Etc):

What language/languages do you speak and how proficient are you in this/these languages?

Are you left or right handed? Left/Right (Please circle one)
APPENDIX J

DURATION AND MEAN PITCH OF DISYLLABLES USED IN EXPERIMENTS 5 AND 6
<table>
<thead>
<tr>
<th>Speakers</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration (msec)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/aba/</td>
<td>473</td>
<td>353</td>
<td>388</td>
<td>401</td>
<td>418</td>
<td>290</td>
<td>381</td>
<td>491</td>
</tr>
<tr>
<td>/ada/</td>
<td>501</td>
<td>360</td>
<td>355</td>
<td>427</td>
<td>456</td>
<td>341</td>
<td>359</td>
<td>458</td>
</tr>
<tr>
<td>/aga/</td>
<td>510</td>
<td>411</td>
<td>392</td>
<td>395</td>
<td>442</td>
<td>320</td>
<td>342</td>
<td>478</td>
</tr>
<tr>
<td>/ibi/</td>
<td>528</td>
<td>433</td>
<td>417</td>
<td>419</td>
<td>407</td>
<td>390</td>
<td>463</td>
<td>471</td>
</tr>
<tr>
<td>/idi/</td>
<td>570</td>
<td>481</td>
<td>384</td>
<td>396</td>
<td>425</td>
<td>368</td>
<td>426</td>
<td>496</td>
</tr>
<tr>
<td>/igi/</td>
<td>535</td>
<td>452</td>
<td>395</td>
<td>409</td>
<td>448</td>
<td>367</td>
<td>491</td>
<td>508</td>
</tr>
<tr>
<td>/ibu/</td>
<td>509</td>
<td>462</td>
<td>455</td>
<td>481</td>
<td>454</td>
<td>430</td>
<td>506</td>
<td>602</td>
</tr>
<tr>
<td>/udu/</td>
<td>504</td>
<td>470</td>
<td>411</td>
<td>472</td>
<td>481</td>
<td>462</td>
<td>505</td>
<td>653</td>
</tr>
<tr>
<td>/ugu/</td>
<td>509</td>
<td>486</td>
<td>394</td>
<td>503</td>
<td>497</td>
<td>484</td>
<td>531</td>
<td>579</td>
</tr>
<tr>
<td>Mean pitch (Hz)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/aba/</td>
<td>129</td>
<td>97</td>
<td>82</td>
<td>99</td>
<td>210</td>
<td>212</td>
<td>210</td>
<td>211</td>
</tr>
<tr>
<td>/ada/</td>
<td>131</td>
<td>94</td>
<td>84</td>
<td>101</td>
<td>208</td>
<td>212</td>
<td>221</td>
<td>203</td>
</tr>
<tr>
<td>/aga/</td>
<td>130</td>
<td>92</td>
<td>82</td>
<td>99</td>
<td>210</td>
<td>212</td>
<td>195</td>
<td>210</td>
</tr>
<tr>
<td>/ibi/</td>
<td>133</td>
<td>106</td>
<td>82</td>
<td>105</td>
<td>216</td>
<td>208</td>
<td>232</td>
<td>210</td>
</tr>
<tr>
<td>/idi/</td>
<td>132</td>
<td>102</td>
<td>83</td>
<td>101</td>
<td>215</td>
<td>232</td>
<td>211</td>
<td>207</td>
</tr>
<tr>
<td>/igi/</td>
<td>134</td>
<td>102</td>
<td>86</td>
<td>100</td>
<td>216</td>
<td>222</td>
<td>232</td>
<td>211</td>
</tr>
<tr>
<td>/ibu/</td>
<td>134</td>
<td>100</td>
<td>84</td>
<td>104</td>
<td>215</td>
<td>228</td>
<td>230</td>
<td>207</td>
</tr>
<tr>
<td>/udu/</td>
<td>139</td>
<td>98</td>
<td>84</td>
<td>106</td>
<td>218</td>
<td>234</td>
<td>229</td>
<td>205</td>
</tr>
<tr>
<td>/ugu/</td>
<td>139</td>
<td>105</td>
<td>84</td>
<td>107</td>
<td>219</td>
<td>231</td>
<td>229</td>
<td>208</td>
</tr>
</tbody>
</table>
APPENDIX K

INFORMATION SHEET USED IN EXPERIMENT 5
Re: Audio-Visual Speech Perception

Previous research has found that watching a person’s lip while he/she is talking helps us understand what the person is saying. It has also been observed that the more often we hear a certain word the faster we recognise it in the future. We investigate these phenomena using audio and video clips consisting of various syllables uttered by different speakers.

The experiment consists of three parts. In the first part, participants will listen to and watch the audio-video clips and they will be asked to indicate what they perceive. In the second part, we examine how fast participants recognise the same audio clips without the visual display of articulation, and how much their reaction time improves after they hear the stimuli a number of times. In this part, participants will repeatedly listen to the syllables and they will have to indicate what they hear as fast as they can. The final part of the experiment will be the same as the first part. The experiment will take approximately 45 minutes, including breaks and instructions.

Data collected will not be identified with individuals and will be kept confidential. Results of the study will be reported in a PhD thesis, and will be written up as a journal article and a conference presentation.

Your participation in this experiment is voluntary and you are free to discontinue your participation at any time. Withdrawing from the experiment will not affect your future relationship with the University of Western Sydney. If you have any additional questions please feel free to ask the investigator. If you decide to participate please complete the attached consent form and questionnaire.

Thank you for your time.

Bettina Keresztesi
PhD Student
MARCS Auditory Laboratories
Phone: +61 2 9772 6589
Fax: +61 2 9772 6736
Email: b.keresztesi@uws.edu.au
Web: http://www.uws.edu.au/marcs/

Dr Kate Stevens
Supervisor
MARCS Auditory Laboratories
Phone: +61 2 9772 6324
Fax: +61 2 9772 6736
Email: kj.stevens@uws.edu.au
Web: http://www.uws.edu.au/marcs/

NOTE: This study has been approved by the University of Western Sydney Human Research Ethics Committee (HEC 03/167). If you have any complaints or reservations about the ethical conduct of this research, you may contact the Ethics Committee through the Research Ethics Officers (tel: 02 4570 1136). Any issues you raise will be treated in confidence and investigated fully, and you will be informed of the outcome.
APPENDIX L

INFORMATION SHEET USED IN EXPERIMENT 6
Re: Audio-Visual Speech Perception

This study investigates the effect of articulation on speech perception. Previous research has found that watching a person’s lip while he/she is talking helps us understand what the person is saying. We investigate this phenomenon using video clips consisting of various syllables uttered by different speakers.

The experiment consists of three parts. In the first part, participants will listen to and watch the audio-video clips and they will be asked to indicate what they perceive. In the second part, we examine how well participants recognise the articulation without auditory information. In this part, participants will repeatedly watch the video clips without hearing the syllables and they will have to indicate what the person articulates. The final part of the experiment will be the same as the first part. The experiment will take approximately 45 minutes, including breaks and instructions.

Data collected will not be identified with individuals and will be kept confidential. Results of the study will be reported in a PhD thesis, and will be written up as a journal article and a conference presentation.

Your participation in this experiment is voluntary and you are free to discontinue your participation at any time. Withdrawing from the experiment will not affect your future relationship with the University of Western Sydney. If you have any additional questions please feel free to ask the investigator. If you decide to participate please complete the attached consent form and questionnaire.

Thank you for your time.

Bettina Keresztesi
PhD Student
MARCS Auditory Laboratories
Phone: +61 2 9772 6589
Fax: +61 2 9772 6736
Email: b.keresztesi@uws.edu.au
Web: http://www.uws.edu.au/marcs/

Dr Kate Stevens
Supervisor
MARCS Auditory Laboratories
Phone: +61 2 9772 6324
Fax: +61 2 9772 6736
Email: kj.stevens@uws.edu.au
Web: http://www.uws.edu.au/marcs/

NOTE: This study has been approved by the University of Western Sydney Human Research Ethics Committee (HEC 03/167). If you have any complaints or reservations about the ethical conduct of this research, you may contact the Ethics Committee through the Research Ethics Officers (tel: 02 4570 1136). Any issues you raise will be treated in confidence and investigated fully, and you will be informed of the outcome.