

**It's not just what you say,
but also how you say it :**

**Exploring the Auditory and Visual
Properties of Speech Prosody**

Erin Cvejic

B.A. (Hons.) Psych.

B.SSc. (Psych.)

A thesis submitted for the degree of Doctor of Philosophy

MARCS Auditory Laboratories

University of Western Sydney

December, 2011

Acknowledgements

First and foremost, my sincerest thanks go to my supervisors, Assoc. Prof. Jeusun Kim and Prof. Chris Davis, for their expert guidance, persistence, patience and constant encouragement over the last five years. You have both taken so much time (regardless of the geographical distance or time difference) to help me become a better cognitive scientist, and while I still have much to learn from both of you I am so very grateful for all that you have given me.

I have also been fortunate to have a diverse range of skills available to me from the rest of my supervisory panel, Prof. Gerard Bailly, Dr. Christian Kroos, and Dr. Takaaki Kuratate. I thank you all for your various comments, suggestions, feedback and technical assistance at various stages throughout my candidature. I also thank Dr. Guillaume Gibert and Dr. Virginie Attina for their generous contribution of Matlab code and technical assistance with the guided principal component analysis of the visual data.

I have had so many amazing opportunities as a consequence of completing my PhD at MARCS, all of which have been made possible by the hard work and dedication of Prof. Denis Burnham and Prof. Kate Stevens, the administration team (Karen McConachie, Gail Charlton, Darlene Williams, Sonya O'Shanna, and Kym Buckley) and the tech team (Colin Schoknecht, Steven Fazio, and Johnson Chen and Donovan Govan). Thank you all for your assistance and support throughout the project, and for making MARCS such a great place to do research.

To the other MARCS students (past and present) who over the years have become like family; Tonya Agostini, Rachel Bennetts, Laura Bishop, Janise Farrell, Ming-Wen Kuo, Jacques Launay, Karen Mulak, Dr. Kirk Olsen, Staci Parlato-Harris, Dr. Nathan Perry, Richard Salmon, Ben Schultz, Dr. Damien Smith and Josephine Terry, thank you all for the thought-provoking discussions and friendship over the last few years.

To my office mates Tim Paris and Michael Fitzpatrick, and office neighbours Amanda Reid and Dr. Bronson Harry, thank you for being a constant source of motivation, inspiration, discussion, and distraction, and for the numerous teas, coffees, and Tiny Teddies that we have shared together. You have all made the journey bearable in your own way.

I am eternally grateful for the support (both emotional and financial) of my parents, Durica and Sylvia Cvejic, and to my sister, Erika Cvejic, for her encouragement and constant offers to make my thesis look “less boring”.

And finally, to my partner, Karolina Ratusznik, thank you for the ongoing support, patience, encouragement and tolerance over the last few years while I finish the thesis.

Statement of Authentication

The work presented in this thesis is, to the best of my knowledge and belief, original except as acknowledged in the text. I hereby declare that I have not submitted this material, either in full or in part, for a degree at the University of Western Sydney, or any other educational institution.

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

.....

Erin Cvejic

Table of Contents

Volume 1

Table of Contents	i
Contents of Appendices on DVD-ROM	xi
List of Tables	xii
List of Figures.....	xviii
List of Publications.....	xxix
Abstract.....	xxxii
Chapter 1. Introduction.....	1
1.1. Prosody.....	2
1.2. Linguistic Prosodic Contrasts	3
1.2.1. Prosodic Focus	3
1.2.2. Utterance Phrasing	4
1.3. Visible Aspects of Speech Communication.....	5
1.4. Research Questions and Thesis Overview	6
Chapter 2. Prosody off the Top of the Head: Prosodic Contrasts can be Discriminated from Head Motion	11
2.1. Experiment 1: Visual-Visual (VV) Prosody Matching.....	17
2.1.1. Method	18
2.1.1.1. Materials.....	18

2.1.1.2. Preliminary Acoustic Analysis.....	21
2.1.1.3. Participants.....	23
2.1.1.4. Procedure.....	24
2.1.2. Results and Discussion.....	26
2.2. Experiment 2: Auditory-Visual (AV) Prosody Matching.....	29
2.2.1. Method	29
2.2.1.1. Participants.....	29
2.2.1.2. Materials and Procedure.....	29
2.2.2. Results and Discussion.....	31
2.3. General Discussion.....	33
Chapter 3. Recognising Prosody across Modalities, Face Areas and Talkers: Examining Perceivers' Sensitivity to Variable Realisations of Visual Prosody .37	
3.1. Experiment 3: Perceiving Prosody from the Lower Face	42
3.1.1. Method	42
3.1.1.1. Participants.....	42
3.1.1.2. Materials.....	42
3.1.1.3. Procedure.....	43
3.1.2. Results and Discussion.....	44
3.2. Experiment 4: Matching Prosody across Talkers	47
3.2.1. Method	48
3.2.1.1. Participants.....	48

3.2.1.2. Materials and Procedure.....	48
3.2.2. Results and Discussion.....	49
3.3. Experiments 5 and 6: Matching Prosody across Face Areas	52
3.3.1. Method	53
3.3.1.1. Participants.....	53
3.3.1.2. Stimuli and Procedure	53
3.3.2. Results and Discussion.....	54
3.4. General Discussion.....	57
Chapter 4. Recording of an Auditory-Visual Speech Prosody Corpus.....	62
4.1. Method	64
4.1.1. Equipment	64
4.1.2. Marker Configuration	66
4.1.3. Materials.....	72
4.1.4. Participants.....	73
4.1.5. Procedure.....	73
4.1.6. Interactive Settings.....	74
4.1.6.1. Face –to-Face (FTF).....	74
4.1.6.2. Auditory-Only (AO)	75
4.2. Preliminary Data Processing	77
4.3. Summary	77
Chapter 5. Auditory Analysis of the Speech Prosody Corpus.....	78

5.1. Data Preparation.....	81
5.1.1. Classification of Utterance Types	81
5.1.2. Identification of Utterance Phases	82
5.1.3. Acoustic Feature Extraction.....	82
5.1.4. Data Normalisation	83
5.2. Acoustic Analysis	83
5.2.1. Realisation of Prosodic Contrasts	83
5.2.1.1. Realisation of Focus Contrasts.....	84
5.2.1.2. Realisation of Phrasing Contrasts	85
5.2.2. Effect of Utterance Type.....	88
5.2.2.1. Focus as a function of Utterance Type.....	88
5.2.2.2. Phrasing as a function of Utterance Type	92
5.2.3. Effects of Interactive Setting.....	94
5.2.3.1. Focus Contrasts across Interactive Settings	95
5.2.3.2. Phrasing Contrasts across Interactive Settings.....	96
5.2.4. Idiosyncratic Talker Strategies.....	96
5.2.4.1. Idiosyncrasies in Focus Realisation	97
5.2.4.2. Idiosyncrasies in Phrasing Realisation.....	98
5.2.4.3. Idiosyncrasies across Interactive Settings.....	99
5.3. Spectral Analysis.....	106
5.3.1. Data Selection	106

5.3.2. Spectral Feature Extraction and Processing	107
5.3.3. Analysis Results	109
5.4. Summary	112
5.4.1. Realisation of Prosodic Contrasts	112
5.4.2. Differences as a function of Interactive Setting	115
5.4.3. Talker Idiosyncrasies	116
Chapter 6. Perceptual Rating of Auditory Prosody	117
6.1. Experiment 7: Effects of Seeing the Interlocutor on the Production of Prosodic Contrasts.....	118
6.1.1. Method	120
6.1.1.1. Participants.....	120
6.1.1.2. Materials.....	120
6.1.1.3. Procedure.....	123
6.1.2. Results.....	124
6.1.2.1. Perceptual Rating Scores for Focus Contrasts	124
6.1.2.1. Perceptual Rating Scores for Phrasing Contrasts.....	127
6.1.2.2. Regression Analyses between Acoustic and Perceptual Measures	130
6.1.3. Discussion	134

Volume 2

Chapter 7. Visual Analysis of the Speech Prosody Corpus.....	136
7.1. Data Preparation.....	141
7.1.1. Face Shape Normalisation.....	141
7.1.2. Dimensionality Reduction.....	142
7.2. Visual Analysis	145
7.2.1. Guided Principal Component Analysis (gPCA)	145
7.2.2. Correlation amongst Extracted Components	150
7.2.3. Area under PC Amplitude Curves.....	151
7.2.3.1. Time Normalisation	151
7.2.3.2. Area Calculation and Normalisation.....	152
7.2.3.3. Visual Realisation of Prosodic Contrasts.....	155
7.2.3.4. Effects of Utterance Type	158
7.2.3.4.1. Effect of utterance type on focus contrasts	158
7.2.3.4.2. Effect of utterance type on phrasing contrasts	161
7.2.3.5. Differences across Interactive Settings	164
7.2.3.6. Talker Idiosyncrasies	167
7.2.3.6.1. Variable visual realisation of prosodic focus.....	168
7.2.3.6.2. Variable visual realisation of prosodic phrasing..	172
7.2.3.6.3. Idiosyncratic prosody production across interactive conditions	176

7.2.3.6.4. Use of rigid head motion across interactive settings	186
7.2.4. Discussion	188
Chapter 8. The Relationship between Auditory and Visual Prosody	192
8.1. Correlation between Auditory and Visual Properties	194
8.1.1. Method	195
8.1.2. Results	197
8.1.2.1. Global Correlations	197
8.1.2.2. Correlations as a function of Prosodic and Interactive Settings	198
8.1.2.2.1. Correlations across focus contrasts	204
8.1.2.2.2. Correlations across phrasing contrasts	204
8.1.2.2.3. Correlations across interactive settings	205
8.1.2.2.4. Correlations during critical utterance phases	205
8.1.3. Discussion	207
8.2. Temporal Alignment between Auditory and Visual Prosody	210
8.2.1. Method	213
8.2.2. Results	215
8.2.2.1. Eyebrow Raising (PC 7)	215
8.2.2.2. Rigid Pitch Rotations (R 1)	219
8.2.2.3. Co-occurrence of eyebrow and rigid head movements	222

8.2.3. Discussion	225
8.3. <i>F0</i> Rises and Visual Prosodic Markers	226
8.3.1. Method	227
8.3.2. Results and Discussion.....	228
8.4. Summary	237
Chapter 9. Perceptual Rating of Auditory-Visual Prosody	239
9.1. Experiment 8: Perceptual Rating of Prosody across Modalities.....	242
9.1.1. Method	242
9.1.1.1. Participants.....	242
9.1.1.2. Materials.....	242
9.1.1.3. Procedure.....	245
9.1.2. Results.....	247
9.1.2.1. Ratings of Focus across Modalities	247
9.1.2.2. Ratings of Phrasing across Modalities	250
9.1.2.3. Regression of VO and AV Rating Scores	253
9.1.3. Discussion	257
Chapter 10. Perceiving Prosody From Augmented Point Light Displays	261
10.1. Experiment 9: Cross-Modal Prosody Matching using Point-Light Displays..	263
10.1.1. Method	265
10.1.1.1. Participants.....	265
10.1.1.2. Materials.....	266

10.1.1.3. Procedure.....	267
10.1.1.3.1. Cross-modal matching task	267
10.1.1.3.2. Auditory prosody rating task.....	269
10.1.2. Results and Discussion.....	270
10.1.2.1. Matching Point-Light Displays of Prosody across Modalities	270
10.1.2.2. Auditory Ratings of Prosodic Contrasts.....	274
10.1.2.3. Relationship between Auditory Ratings and Item Accuracy	277
10.2. Experiment 10: Perceiving Prosody from Manipulated Point-Light Displays	280
10.2.1. Method	281
10.2.1.1. Participants.....	281
10.2.1.2. Materials and Procedure.....	281
10.2.2. Results and Discussion.....	282
10.3. Summary	289
Chapter 11. Summary and Conclusions	291
11.1. Perceiving Prosody	292
11.1.1. Perceptual Sensitivity to Visual Prosody	293
11.1.2. Beneficial Face Areas for Specific Prosodic Contrasts.....	295
11.1.3. Tolerating Variability in Prosodic Realisation.....	295
11.2. Producing Prosody	297

11.2.1. Auditory Correlates of Prosodic Focus and Phrasing Contrasts	297
11.2.2. Visual Correlates of Prosodic Focus and Phrasing Contrasts	298
11.2.3. Variability in the Production of Prosodic Contrasts	299
11.2.4. Relationship between Auditory and Visual Prosodic Signals.....	301
11.3. Linking Production and Perception.....	303
11.3.1. Perceptual Effects of the Talker Seeing the Interlocutor	303
11.3.2. Perceptual Effects of Seeing the Talker	304
11.3.3. Movement Requirements for Perceiving Prosody	305
11.4. Potential Communicative Functions of Visual Prosody	306
11.5. Limitations and Future Directions	309
11.6. Summary	312
References	314
Appendix A. IEEE Stimulus Sentence Properties.....	339
Appendix B. Auditory Visual Speech Prosody Corpus	341
B.1. Corpus Instructions and File Naming Convention.....	341

Contents of Appendices on DVD-ROM

Appendices Disk 1

Appendix B. Auditory Visual Speech Prosody Corpus

- B.2. Auditory Corpus Files
- B.3. Phonemic Transcription Files
- B.4. Motion Capture Files (.n3d Format)
- B.5. Motion Capture Files (.avi Format)
- B.6. Visualisation of Principal Components
- B.7. Parameterised Motion Capture Files
- B.8. Software for Corpus Playback

Appendices Disk 2

Appendix C. Statistical Analyses Tables

- C.1 Auditory AnalysisC-1
- C.2 Visual Analysis.....C-11

Appendix D. Pearsons Correlation Analyses

Appendix E. Spearman Rank Correlation Analyses

Appendix F. Example Stimuli from Experiment 10

- F.1. All Motion
- F.2. Non-Rigid Movement Only
- F.3. Articulatory Movement Only
- F.4. Rigid Movement Only
- F.5. Eyebrow Movement Only

List of Tables

Table 2.1. Sentence material used for the audio-visual recordings of Experiment 1. The prosodically marked word is italicised.	19
Table 2.2. Mean percent of correct responses in the VV matching task as a function of the stimulus presentation condition for each prosodic speech condition. Degrees of freedom (<i>df</i>) are indicated in brackets and ** indicates $p <$ 0.001.....	27
Table 2.3. Mean percent of correct matching responses in the 2AFC auditory-visual matching, as a function of stimulus presentation condition and prosodic speech condition. Degrees of freedom are indicated in brackets and ** indicates $p < 0.001$	32
Table 3.1. Mean percent of correct responses in the within-talker VV and AV matching tasks as a function of visible face area in each prosodic speech condition. Data in italics are from Experiments 1 and 2. Degrees of freedom (<i>df</i>) are indicated in brackets and ** indicates $p < 0.001$	45
Table 3.2. Mean percent of correct responses in the VV and AV matching tasks as a function of visible face area in each prosodic speech condition, when items within pairs were produced by different talkers. $df = 15$ and ** indicates $p < 0.001$	50
Table 3.3. Mean percent of correct responses in the 2AFC visual-visual matching task across face areas using within- and cross-talker stimuli, presented as a function of presentation order (upper to lower; lower to upper face), separated by prosodic speech condition. $df = 39$ and ** indicates $p <$ 0.001.....	55

Table 4.1. Summary of marker configurations used in previous studies utilising optical tracking to measure visual speech.	67
Table 4.2. The location of the 39 IR emitting OPTOTRAK markers on the head and face of the talkers. Due to high rates of marker occlusions and dropouts, the larynx marker was not included in the analysis.	70
Table 5.1. Sub-set of sentences within the recorded corpus containing the corner vowels within the critical item. Critical words appear in brackets. IPA glossing is of standard Australian English.	107
Table 5.2. Spectral properties of critical vowels (with standard deviations) for the corner vowel subset as a function of interactive condition and prosodic context, collapsed across talkers, $n = 6$	111
Table 6.1. Stimuli sentences used in the prosody rating tasks. The critical constituent is italicised.	121
Table 6.2. Standard multiple regression of acoustic features on ratings of narrow focus.	132
Table 6.3. Standard multiple regression of acoustic features on echoic question ratings.	133
Table 7.1. A priori non-rigid components used to drive the gPCA, and the rigid movement parameters that were extracted based on rotation and translation around the centre of rotation.	145
Table 7.2. Visualisation of non-rigid principle components, derived from guided principal components analysis.	147
Table 7.3. Pearson's r correlation values within the principal components extracted using gPCA and rigid parameters for the unique movement database. ..	150

Table 7.4. Significant main effects of prosody for the critical constituent during the production of prosodic focus contrasts.	156
Table 7.5. Significant effects of prosody for the phrasing contrasts for the critical constituents.....	157
Table 7.6. Significant interactions between prosody and utterance type for the post-critical utterance phase of narrow focus tokens. Analyses were interpreted with 3 between, and 26 error degrees of freedom.	160
Table 7.7. Significant interactions between prosody and utterance type for the post-critical utterance phase of echoic question tokens. Analyses were interpreted with 3 between, and 26 error degrees of freedom.....	163
Table 7.8. Main effects of interactive setting on the production of narrow focus. ...	165
Table 7.9. Main effects of interactive setting on the production of echoic questions.	166
Table 7.10. Idiosyncratic talker realisations of prosodic focus. Analyses were interpreted with 1 between and 29 error degrees of freedom.....	169
Table 7.11. Idiosyncratic talker realisations of prosodic phrasing. Analyses were interpreted with 1 between and 29 error degrees of freedom.....	173
Table 7.12. Idiosyncratic talker realisations of narrow focus and echoic questions across interactive settings. Analyses were interpreted with 1 between and 29 error degrees of freedom.	177
Table 7.13. Idiosyncratic talker uses of rigid head motion during the production of narrow focus and echoic questions across interactive settings. Analyses were interpreted with 1 between and 29 error degrees of freedom.....	187

Table 8.1. Global correlation properties collapsed across utterances, prosodic contrasts and talkers, $n = 2160$	198
Table 8.2. Median correlation (r) values between acoustic properties and visual components, as a function of the prosodic condition and interactive setting. $n = 360$ per cell. Median correlation values above 0.30 are presented in bold.	200
Table 8.3. Median correlation (r) values between acoustic properties and visual components for the critical constituent of each utterance, as a function of the prosodic condition and interactive setting. $n = 360$ per cell.	206
Table 8.4. Distribution of non-articulatory gestures accompanying the production of the critical constituent, dependant on the occurrence of an $F0$ rise, as a function of the interactive setting and prosodic condition.	230
Table 8.5. Temporal distribution of non-articulatory gestures in relation to the onset of an $F0$ rise, as a function of the interactive setting and prosodic condition. Values within brackets indicate the number of movements expressed as a percentage value.	233
Table 9.1. Stimuli sentences used in each task version of Experiment 8. The critical constituent of each utterance is italicised.....	243
Table 9.2. Standard multiple regression of the magnitude of visible movements on ratings of focus and questions in the VO presentation modality.	255
Table 9.3. Standard multiple regression of the relationship between auditory prosodic markers and non-articulatory movements on ratings of focus and questions in the AV presentation modality.	257

Table 10.1. Stimuli sentences used in Experiment 9. The critical constituent of each utterance is italicised.	266
Table 10.2. Matching performance against chance for the cross-modal matching task of Experiment 10, as a function of talker and prosodic condition for each of the stimulus conditions.	285
Table A.1. IEEE stimulus sentences and associated properties. The critical constituent is indicated in italics.....	339
Table B.1. File naming convention for the files in the Auditory Visual Speech Prosody Corpus.....	343
Table C.1. Statistical values for the sentence (F_I) ANOVA comparing the acoustic properties of broad and narrow focus renditions in the AO interactive setting.	C-1
Table C.2. Statistical values for the talker (F_S) ANOVA comparing the acoustic properties of broad and narrow focus renditions in the AO interactive setting.	C-2
Table C.3. Statistical values for the sentence (F_I) ANOVA comparing the acoustic properties of statements and echoic questions renditions in the AO interactive setting.	C-3
Table C.4. Statistical values for the talker (F_S) ANOVA comparing the acoustic properties of statements and echoic questions renditions in the AO interactive setting.	C-4
Table C.5. Statistical values for the sentence (F_I) ANOVA comparing the acoustic properties of narrow focus utterances as a function of utterance type....	C-5

Table C.6. Pairwise comparisons for the effect of utterance type on the production of narrow focus in the AO interactive setting.	C-6
Table C.7. Statistical values for the sentence (F_I) ANOVA comparing the acoustic properties of echoic questions as a function of utterance type.....	C-6
Table C.8. Pairwise comparisons for the effect of utterance type on the production of echoic questions in the AO interactive setting.....	C-7
Table C.9. Statistical values for the sentence (F_I) ANOVA comparing acoustic properties of narrow focus across AO and FTF interactive settings.....	C-7
Table C.10. Statistical values for the talker (F_S) ANOVA comparing acoustic properties of narrow focus across AO and FTF interactive settings.....	C-8
Table C.11. Statistical values for the sentence (F_I) ANOVA comparing acoustic properties of echoic questions across AO and FTF interactive settings.	C-9
Table C.12. Statistical values for the talker (F_S) ANOVA comparing acoustic properties of echoic questions across AO and FTF interactive settings.	C-10
Table C.13. Statistical values for the ANOVA comparing broad and narrow focus renditions in the AO interactive setting	C-11
Table C.14. Statistical values for the ANOVA comparing statements and echoic question renditions in the AO interactive setting	C-13
Table C.15. Statistical values for the ANOVA comparing narrow focus across AO and FTF interactive settings	C-15
Table C.16. Statistical values for the ANOVA comparing echoic questions across AO and FTF interactive settings	C-16

List of Figures

- Figure 2.1. Originally recorded tokens (left) were cropped at the nose tip, generating textured upper face stimuli (right upper panel). A thresholding filter was then applied to copies of these stimuli to create the outline only videos (right lower panel).....21
- Figure 2.2. Acoustic properties of the critical constituent expressed as a proportion of the mean broad focus rendition, collapsed across utterances and repetitions, for Talker 1 (upper panel) and Talker 2 (lower panel). Error bars indicate the standard error of the mean.23
- Figure 2.3. Schematic representation of the 2AFC task used in the visual-visual matching task of Experiment 1. The same item appeared first for both pairs, and was the standard that the matching judgment was to be made on.....26
- Figure 2.4. Schematic representation of the 2AFC task used in the auditory-visual matching task used in Experiment 2. The matching pair was always taken from a different recorded token, and the distracter stimuli were always an alternate prosodic type. Stimuli within both audio-video pairs were always produced by the same talker.....31
- Figure 3.1. The original audiovisual recordings made in Chapter 2 were cropped at the nose tip to generate lower face stimuli.....43
- Figure 3.2. Schematic representation of the 2AFC visual-visual (left) and auditory-visual (right) matching tasks used in Experiment 3. The same item appeared first for both pairs, and was the standard that the matching judgment was to be made on. The matching item within pairs was always

taken from a different recorded token, and non-matching items were the same sentence produced as a different prosodic type.44

Figure 3.3. Schematic representation of the 2AFC visual-visual (left) and auditory-visual (right) matching tasks used in Experiment 4. The same item appeared first for both pairs produced by one talker, and was the standard from which the matching judgment was to be made on. The second item within pairs was produced by a different talker, with the non-matching item being the same sentence produced with a different prosody.

Participants completed the task with either upper or lower face stimuli. .49

Figure 3.4. Schematic representation of the 2AFC tasks used in Experiment 5 (left) and Experiment 6 (right). In each item either the upper face or lower face stimuli were displayed first, followed by the opposite face area within each pair. In Experiment 5, items within-pairs were produced by the same talker, with the matching video always taken from a different recorded token. In Experiment 6, items within-pairs were produced by different talkers. The non-matching video in both tasks were always the broad focused rendition of the same sentence.....54

Figure 4.1. The OPTOTRAK infrared camera unit (left) tracks the three-dimensional position over time of infrared emitting markers, measuring 7mm in diameter (right) with high spatial and temporal resolutions.65

Figure 4.2. A polystyrene foam head was used as a visual guide to ensure that placement of the optical markers was consistent across talkers.72

Figure 4.3. Location of the 39 optical markers (with size exaggerated for clarity) on the head and face of the talker, reflecting articulatory and non-articulatory

gestures. Four markers were placed on a head rig and used to estimate rigid movements around the centre of rotation.	74
Figure 4.4. The experimental setup used in the face-to-face (FTF) interactive setting. The talker and interlocutor communicated over a distance of approximately 2.5 meters, and were able to both see and hear each other clearly. To minimise extraneous noise, the OPTOTRAK SCU was located outside of the testing booth.	75
Figure 4.5. The experimental setup used in the auditory-only (AO) interactive setting. The talker and interlocutor communicated over a double microphone and headphone system and were only able to hear each other.	76
Figure 4.6. Diagrammatic representation of the double microphone and headphone system used in the AO interactive condition.....	77
Figure 5.1. Acoustic properties of narrow focus and echoic question renditions (collapsed across utterances and talkers) in the AO and FTF interactive conditions, represented as proportion values of the mean broad focused rendition in their respective interactive condition. Thus, a value greater than 1 indicates an increase on the parameter compared to the broad focused rendition.....	87
Figure 5.2. Acoustic properties of narrow focused renditions recorded in the AO interactive condition (expressed as a proportion of the broad focused AO rendition), as a function of utterance type, collapsed across talkers and sentences.	91

Figure 5.3. Acoustic properties of echoic question renditions recorded in the AO interactive condition (expressed as a proportion of the broad focused AO rendition), as a function of utterance type, collapsed across talkers and sentences.	94
Figure 5.4. Mean syllable duration (represented as a proportion of the broad focused rendition) produced by each talker.....	101
Figure 5.5. Mean fundamental frequency (represented as a proportion of the broad focused rendition) produced by each talker.	102
Figure 5.6. Fundamental frequency range (represented as a proportion of the broad focused rendition) produced by each talker.	103
Figure 5.7. Mean intensity (represented as a proportion of the broad focused rendition) produced by each talker.....	104
Figure 5.8. Intensity range (represented as a proportion of the broad focused rendition) produced by each talker.....	105
Figure 5.9. Vowel triangles for broad, narrow focus and echoic question renditions in the AO (left) and FTF interaction settings (right) collapsed across talkers.	112
Figure 5.10. Between-category displacements for broad focus, narrow focus and echoic question utterance renditions in the AO (left) and FTF interaction settings (right), collapsed across talkers.	112
Figure 6.1. Mean ratings of focus (collapsed across sentences and raters) as a function of talker for the AO and FTF conditions. Error bars indicate the standard error of the mean, $n = 90$ observations per column.....	125

Figure 6.2. Mean ratings of focus (collapsed across talkers and raters) as a function of sentence for the AO and FTF conditions. Error bars indicate the standard error of the mean, $n = 60$ observations per column.....	126
Figure 6.3. Mean ratings of focus (collapsed across talkers and sentences) as a function of rater for the AO and FTF conditions. Error bars indicate the standard error of the mean, $n = 54$ observations per column.....	127
Figure 6.4. Mean ratings of phrasing (collapsed across sentences and raters) as a function of talker for the AO and FTF conditions. Error bars indicate the standard error of the mean, $n = 90$ observations per column.....	128
Figure 6.5. Mean ratings of focus (collapsed across talkers and raters) as a function of sentence for the AO and FTF conditions. Error bars indicate the standard error of the mean, $n = 60$ observations per column.....	129
Figure 6.6. Mean ratings of phrasing (collapsed across talkers and sentences) as a function of rater for the AO and FTF conditions. Error bars indicate the standard error of the mean, $n = 54$ observations per column.....	130
Figure 7.1. The mean talker face shapes (left) were used to generate the normalised average face model (right). Bones (lines between markers) have been added to assist in interpretation.....	142
Figure 7.2. A database of unique motion was created (upper panel) to avoid over-representation of particular marker configurations. The rigid movements of the head were then calculated and removed from the database (lower panel). Also shown are the directions of the X, Y, and Z axes.....	144
Figure 7.3. Cumulative percent of variance explained by each component from the guided principal components analysis.	146

Figure 7.4. Linear spline interpolation was used to normalisation the time of each recorded token, allowing for comparisons across talkers and repetitions to be made.	151
Figure 7.5. Mean area under curve (collapsed across talkers and sentences) of principal components 1 to 5, as a function of prosodic condition represented as a proportion of the broad focused rendition.	153
Figure 7.6. Mean area under curve (collapsed across talkers and sentences) of principal components 6 to 8, and R1 and R2, as a function of prosodic condition represented as a proportion of the broad focused rendition.	154
Figure 7.7. Proportion values (relative to broad focus renditions) for post-focal utterance phases in narrow focus renditions, as a function of utterance type.	160
Figure 7.8. Proportion values (relative to broad focus renditions) for pre-critical utterance phases for echoic questions, as a function of utterance type.	161
Figure 7.9. Proportion values (relative to broad focus renditions) for post-focal utterance phases in echoic question renditions, as a function of utterance type.	163
Figure 7.10. Visual realisation of prosodic contrasts (represented as a proportion of the broad focused rendition) of utterances produced by Talker 1.	180
Figure 7.11. Visual realisation of prosodic contrasts (represented as a proportion of the broad focused rendition) of utterances produced by Talker 2.	181
Figure 7.12. Visual realisation of prosodic contrasts (represented as a proportion of the broad focused rendition) of utterances produced by Talker 3.	182

Figure 7.13. Visual realisation of prosodic contrasts (represented as a proportion of the broad focused rendition) of utterances produced by Talker 4.....	183
Figure 7.14. Visual realisation of prosodic contrasts (represented as a proportion of the broad focused rendition) of utterances produced by Talker 5.....	184
Figure 7.15. Visual realisation of prosodic contrasts (represented as a proportion of the broad focused rendition) of utterances produced by Talker 6.....	185
Figure 8.1. Histograms of the distribution of correlation coefficients (calculated between auditory and visual parameters) for broad focus tokens recorded in the AO interactive setting (blue shows the distribution of the movement and intensity correlations; red shows the F0 correlations).....	201
Figure 8.2. Histograms of the distribution of correlation coefficients (calculated between auditory and visual parameters) for broad focus tokens recorded in the FTF interactive setting.	201
Figure 8.3. Histograms of the distribution of correlation coefficients (calculated between auditory and visual parameters) for narrow focus tokens recorded in the AO interactive setting.	202
Figure 8.4. Histograms of the distribution of correlation coefficients (calculated between auditory and visual parameters) for narrow focus tokens recorded in the FTF interactive setting.	202
Figure 8.5. Histograms of the distribution of correlation coefficients (calculated between auditory and visual parameters) for echoic question tokens recorded in the AO interactive setting.	203

Figure 8.6. Histograms of the distribution of correlation coefficients (calculated between auditory and visual parameters) for echoic question tokens recorded in the FTF interactive setting.	203
Figure 8.7. An example of temporal location of the onset (blue line) and peak (red line) of eyebrow movements in the vicinity of the critical constituent (grey line) of an utterance. A value of “0” indicates the average face position.	214
Figure 8.8. Distribution of brow raises as a function of temporal onset of movement preceding the start of the critical constituent for each prosodic condition and interactive setting ($n = 360$ in each condition).	218
Figure 8.9. Distribution of pitch rotation peaks around the start of the critical constituent for each prosodic condition and interactive setting ($n = 360$ in each condition).	221
Figure 8.10. Occurrence of non-articulatory features accompanying the production of the critical constituent within utterances across prosodic conditions and interactive settings ($n = 360$ in each condition).	223
Figure 8.11. Temporal distribution of co-occurring brow movements (blue) and pitch rotation peaks (red) around the start of the critical constituent, as a function of prosodic condition and interactive setting.	224
Figure 8.12. F_0 rises were detected automatically by examining the F_0 contour around the temporal onset of the critical constituent (grey line). A pitch rise was considered to have occurred if the difference between the F_0 minimum (blue line) and peak (red line) was at least 15 Hz.	228

Figure 8.13. Temporal distribution of brow movement onsets co-occurring with an <i>F0</i> rise, presented as a function of the prosodic condition and interactive setting.	234
Figure 8.14. Temporal distribution of pitch rotation peaks co-occurring with an <i>F0</i> rise, presented as a function of the prosodic condition and interactive setting.	235
Figure 8.15. Temporal distribution of brow movement onsets (blue) and rigid rotation peaks (red) when both non-articulatory features accompany an <i>F0</i> rise, presented as a function of the prosodic condition and interactive setting.	236
Figure 9.1. Example frame of a point-light talker used as stimuli in the VO and AV presentation modalities. An animated rendition is included as Appendix F.1.	245
Figure 9.2. Mean rating of focus (collapsed across sentences and talkers) as a function of prosodic condition, interactive setting and presentation modality. Error bars indicate the standard error of the mean.....	249
Figure 9.3. Mean rating of phrasing (collapsed across sentences and talkers) as a function of prosodic condition, interactive setting and presentation modality. Error bars indicate the standard error of the mean.....	252
Figure 10.1. Schematic representation of the 2AFC cross-modal matching task used in Experiment 9. The same auditory token appeared first for both pairs, and was the standard that the matching judgment was to be made on. The matching item within pairs was always taken from a different recorded	

token, and non-matching items were the same sentence produced as a different prosodic type.	269
Figure 10.2. Mean percent of correct responses (with standard error) for the cross- modal prosody matching tasks as a function of prosodic contrast and talker (collapsed sentences) for the “all movement” task version.	271
Figure 10.3. Mean ratings of focus (collapsed across sentences and raters) as a function of talker for broad and narrow focused utterances. Error bars indicate the standard error of the mean.	276
Figure 10.4. Mean ratings of phrasing (collapsed across sentences and raters) as a function of talker for the broad focused statement and echoic question utterances. Error bars indicate the standard error of the mean.	277
Figure 10.5. Scatter plot indicating the relationship between subjective rating of auditory tokens and item accuracy in the cross-modal matching task, for narrow focus (upper panel) and echoic question (lower panel) items. ...	278
Figure 10.6. Mean percent of correct responses (with standard error) for the cross- modal prosody matching tasks as a function of prosodic contrast and talker (collapsed sentences) for the “non-rigid movement only” stimuli.	283
Figure 10.7. Mean percent of correct responses (with standard error) for the cross- modal prosody matching tasks as a function of prosodic contrast and talker (collapsed sentences) for the “articulator movement only” stimuli.	284

Figure 10.8. Mean percent of correct responses (with standard error) for the cross-modal prosody matching tasks as a function of prosodic contrast and talker (collapsed sentences) for the “rigid movement only” stimuli.....284

List of Publications

Several of the experiments and analyses reported in the chapters of this thesis have previously appeared as peer-reviewed conference proceedings, published papers and submitted manuscripts:

Cvejic, E., Kim, J., & Davis, C. (in press). Recognizing prosody across modalities, face areas and speakers: Examining perceivers' sensitivity to variable realizations of visual prosody. *Cognition*.
doi:10.1016/j.cognition.2011.11.013.

Cvejic, E., Kim, J., & Davis, C. (Revision under review). Effects of seeing the interlocutor on the production of prosodic contrasts. *Journal of the Acoustical Society of America*.

Cvejic, E., Kim, J., & Davis, C. (2011a). Temporal relationship between auditory and visual prosodic cues. *Interspeech 2011*, pp. 981-984.

Cvejic, E., Kim, J., & Davis, C. (2011b). Perceiving visual prosody from point-light displays. *Proceedings of the 10th International Conference on Audio-Visual Speech Processing (AVSP 2011)*, pp. 15-20.

Cvejic, E., Kim, J., & Davis, C. (2010a). Prosody off the top of the head: Prosodic contrasts can be discriminated from head motion. *Speech Communication*, 52, 555-564. doi:10.1016/j.specom.2010.02.006.

Cvejic, E., Kim, J., & Davis, C. (2010b). It's all the same to me: Discriminating prosody across face areas and speakers. *Speech Prosody 2010*, 100893, pp. 1-4.

- Cvejic, E., Kim, J., & Davis, C. (2010c). Abstracting visual prosody across speakers and face areas. *Proceedings of the 9th International Conference on Audio-Visual Speech Processing (AVSP 2010)*, pp. 38-43.
- Cvejic, E., Kim, J., & Davis, C. (2010d). Modification of prosodic cues when an interlocutor cannot be seen: The effect of visual feedback on acoustic prosody production. *Proceedings of the 20th International Congress on Acoustics, Sydney, Australia*, Paper ID: 521, pp. 1-7.
- Cvejic, E., Kim, J., Davis, C., & Gibert, G. (2010). Prosody for the eyes: Quantifying visual prosody using Guided Principal Component Analysis. *Interspeech 2010*, pp. 1433-1436.

Abstract

This thesis investigated the production and perception of prosodic cues for focus and phrasing contrasts from auditory and visual speech (i.e., visible face and head movements). This was done by examining the form, perceptibility, and potential functions of the visual correlates of spoken prosody using auditory and motion analysis and perception-based measures. The first part of the investigation (Chapters 2 to 3) consisted of a series of perception experiments conducted to determine the degree to which perceivers were sensitive to the visual realisation of prosody across face areas. Here, participants were presented with a visual cue (either from the upper or lower half of the face) to match (based on prosody) with another visual or auditory cue. Performance was much better than chance even when the task involved matching cues produced by different talkers. The results indicate that perceivers were sensitive to visual prosodic cues, that considerable variability in the form of these could be tolerated, and that different cues conveying information about the same prosodic type could be matched. The second part of the thesis (Chapters 4 to 8) reported on the construction of a multi-talker speech prosody corpus and the analysis and perceptibility of this production data. The corpus consisted of auditory and visual speech recording of six talkers producing 30 sentences across three prosodic conditions in two interactive settings (face-to-face and auditory-only), with face movements captured using a 3D motion tracking system and characterised using a guided principal components analysis. The analysis consisted of quantifying auditory and visual characteristics of prosodic contrasts separately as well as the relationship between these. Acoustically, the properties of the contrasts corresponded to those typically described in the literature (however, some properties varied systematically as a function of the interactive setting), and were also perceived as conveying the

intended contrasts in subsequent perceptual tasks (reported in Chapter 6). Overall, the types of movements used to contrast narrow from broad focused utterances, and echoic questions from statements, involved the use of both articulatory (e.g., jaw and lip movement) and non-articulatory (e.g., eyebrow and rigid head movement) cues. Both the visual and the acoustic properties varied across talkers and interactive settings. The spatial and temporal relationship between auditory and visual signal modalities was highly variable, differing substantially across utterances. The final part of the thesis (Chapters 9 to 10) reported the results of a series of perception experiments using perceptual rating and cross-modal matching tasks on stimuli resynthesised from the motion capture data. These stimuli showed various combinations of visual cues, and when presented in isolation or combined with the auditory signal, these were perceived as conveying the intended prosodic contrast. However there was no auditory-visual (AV) benefit observed (in the perceptual rating) and the presentation of more cues did not result in better cross-modal matching performance (suggesting there may be limitations in perceivers' ability to process multiple cues). In sum, the thesis showed that perceivers were sensitive to visual prosodic cues despite variability in production, and were able to match different types of cue. The construction of an AV prosody corpus permitted the characteristics of the auditory and visual prosodic correlates (and their relationship) to be quantified, and allowed for the synthesis of visual cues that perceivers subsequently used to successfully extract prosodic information. In all, the experiments reported in this thesis provide a strong case for the development of well-controlled and measured manipulations of prosody and warrants further examination of the visual cues to prosody.

CHAPTER 1.
INTRODUCTION

Chapter 1. Introduction

The experiments reported in this thesis examine the production and perception of auditory-visual speech prosody. Before outlining these experiments and detailing how they are divided across the chapters of the thesis, it is useful to first consider the basic phenomenon of “prosody”.

1.1. Prosody

When two people are conversing with each other in a face to face setting, information is transmitted between the talkers¹ using many different modes. Most prominent is the spoken exchange of messages encoded by propositional symbols, i.e., words and sentences (it is this mode that has been the domain of structural linguistic analysis). However, in addition to the symbolic information that words convey there is an extensive array of other information about how meaning is to be ultimately interpreted. For example, if a talker wishes to modify the communicative message, they can change the segmental content that is used, or can change *how* this segmental content is produced.

Prosody is the broad term used to describe the variations to speech signal properties that are always present, supporting and adapting the meaning of an utterance through the addition of linguistic and non-linguistic information by modulating and manipulating speech features such as duration, pitch contours and loudness (Wagner & Watson, 2010; Wells, 2006). Although a talker’s selection of words is important, the prosody of an utterance (i.e., how these words are actually

¹ Note that the although “speaker” is the more traditional term (and that one speaks a language), the term “talker” has been adopted in this thesis as it avoids confusion with the electronic variety of “speaker”.

produced) greatly impacts on how the utterance is perceived. For example, listeners experience difficulties with comprehension when presented computer-generated synthetic speech that lacks the rhythm and pitch variations apparent in natural human speech (Allan, 1976).

Of its many functions, prosody can assist the listener with the segmentation of a continuous incoming speech signal into individual meaningful units (Cutler, Dahan, von Donselaar, 1997; Shriberg, Ferrer, Kajarekar, Venkataraman & Stolke, 2005). Prosody can also convey demographic information about the talker such as age, gender, emotional and physiological states (Shriberg, 1993; Vaissiere, 2004), as well as serving a linguistic function by conveying information beyond that provided by sentence syntax, grammar, and the symbolic content of speech sounds (Nooteboom, 1997). In this regard, prosody is said to be *suprasegmental*, as it extends beyond the boundaries of individual segmental constituents, affecting the utterance at a sentential level. In this thesis, the linguistic functions of prosody are examined, with particular interest in two prosodic contrast types: prosodic focus and utterance phrasing. These contrasts were selected as they are the most easily defined cases of prosody (Crystal, 1991; Gussenhoven, 2007; Selkirk, 1995) and are less subject to individual interpretation by talkers and listeners than affective (i.e., attitudinal or emotional) prosody (Aubergé & Cathiard, 2003; Drahota, Costall & Reddy, 2008; Linnankoski, Leinonen, Vihla, Laakso & Carlson, 2005).

1.2. Linguistic Prosodic Contrasts

1.2.1. Prosodic Focus

Prosodic focus describes the situation where an individual constituent within an utterance is made perceptually more salient (i.e., more prominent) than the remaining

segmental content in the utterance. This can be used to emphasise the newness or importance of the constituent, or when providing feedback to an interlocutor (e.g., in error correction). The item that is emphasised is deemed to have “narrow focus”, as the point of informational importance has been drawn down to that particular constituent (Bolinger, 1972; Ladd, 1980), making it stand out from the remaining utterance content. In contrast, “broad focused” utterances contain no explicit point of informational focus, with all constituents equivalent in their informational status.

The acoustic properties associated with prosodic focus have been intensively studied and are fairly well described in the literature. In summary, narrow focused renditions (relative to the same word produced within a broad focused context) are articulated with higher mean F_0 (Eady, Cooper, Klouda, Mueller & Lotts, 1986), greater F_0 range (Ladd & Morton, 1997; Xu & Xu, 2005), higher intensity levels (Kochanski, Grabe, Coleman & Rosner, 2005) and consists of syllables of lengthened durations (Krahmer & Swerts, 2001). Similarly, vowels produced within a narrowly focused context are produced with higher first formant (F_1) values (Summers, 1987), with vowel categories in the F_1 - F_2 space moving a greater distance away from the vowel space midpoint, making the categories more distinct from each other (Hay, Sato, Coren, Moran & Diehl, 2006).

1.2.2. Utterance Phrasing

Utterance phrasing refers to the manner in which a sentence is produced, for example, as a statement or as a question. By mimicking the segmental content of a declarative statement, an “echoic question” can be phrased without the use of an interrogative pronoun. That is, echoic questions contain identical words in the same

order as a statement, yet imply a level of uncertainty through the manipulation of suprasegmental acoustic features (Bolinger, 1989).

Acoustically, the different phrasing types typically vary in the following ways: statements can be characterised as having a steadily falling *F0* contour ending with a sharp and definite fall signalling finality (Wells, 2006), whereas the converse pattern is observed for echoic questions in which a gradually rising *F0* contour is observed throughout the time course of the utterance, with a sharp final rise indicating that a response may be required from an interlocutor (Eady & Cooper, 1986). Statements also tend to have slightly shorter final syllable durations and steeper final falls in intensity compared to the same utterances phrased as questions. In addition to affecting the utterance at a global level, echoic questions can also possess a narrowed point of informational focus (i.e., one particular constituent within the utterance is questioned). In this case, the questioned word also differs from broad focused renditions of the same constituent, produced with a gradually rising pitch contour over the duration of the constituent (Pell, 2001).

1.3. Visible Aspects of Speech Communication

Prosody has primarily been studied in terms of how it affects spoken word and sentence recognition in the auditory modality (Xu, 2011), however the auditory speech signal is accompanied by a wealth of additional visible movements that can compliment (and in some cases supplement) the verbal aspects of communication (Goldin-Meadow, 1999). For example, talkers produce a range of different gestures when they speak (i.e., co-speech gestures) such as movements of the arms, hands and head that are believed to serve a communicative function (Bernardis & Gentilucci, 2006; Kendon, 1994; Kraemer & Swerts, 2007; Streeck, 1993).

Similarly, the visual information generated by the process of speech production (i.e., visual speech) has been shown to play an important role in speech perception (Benoît & Le Goff, 1998). Although visual speech alone is a relatively poor source of information for speech intelligibility (Dodd, 1977), listeners can use congruent visual information accompanying speech in degraded conditions (such as noise) to enhance aspects of the auditory signal that may be difficult to hear, but are quite easy to see, such as place of articulation (Sumbly & Pollack, 1954; Summerfield, 1992; Walden, Proesk, Montgomery, Scherr & Jones, 1977). Furthermore, visual speech information is not confined to the lower half of a talker's face. Movements in areas distant to the oral aperture, such as cheek and brow movements, and nods and tilts of the head, although not linked in a direct way to the production of speech acoustics, have also been demonstrated to assist listeners in language processing tasks (Davis & Kim, 2006; Munhall, Jones, Callan, Kuratate & Vatikiotis-Bateson, 2004; Thomas & Jordan, 2004). This suggests that visual information in areas other than those directly involved in articulation may share a relationship with the auditory signal at some higher communicative level than the disambiguation of speech sounds (Granström & House, 2007), potentially providing linguistically relevant suprasegmental content to perceivers. It is these types of visual speech movements that are of interest for this thesis.

1.4. Research Questions and Thesis Overview

This thesis investigated the production and perception of the auditory and visual correlates of linguistic spoken prosody. Although a number of studies have identified the visual correlates of prosody (a detailed review of these studies will be provided when they are linked to the relevant topics in the subsequent chapters), their

outcomes have prompted many questions that remain to be answered. For instance, questions remain about the specific location, range and temporal properties of these prosody related movements (visual cues); about the precise relationship that may exist between the auditory and visual signals; about the consistency of these cues across talkers; about the perceptual relevance of visual prosodic cues, etc.

Furthermore, what has been lacking in previous studies is a well-constrained and systematic examination of the above questions that uses multiple participants with a diverse stimulus set. The current investigation aimed to provide some answers to (or at least ways to go about answering) these questions, by collecting, analysing, and using (for perceptual studies) three-dimensional data of head and face movements that accompany speech (prosody) production of multiple talkers for many sentences.

The first series of perceptual experiments explored whether perceivers are sensitive to visual cues to prosody that are available from the head and face of talkers, whether these visual cues are related by perceivers to auditory prosodic information, and what type of visual movements are of most benefit for conveying prosodic contrasts to perceivers. To answer these questions, Chapter 2 reports two experiments that employed within-modal (visual to visual) and cross-modal (auditory to visual) prosody matching tasks utilising visual stimuli showing the upper head and face of the talkers. Furthermore, these video stimuli were manipulated so that only rigid movements of the whole head were visually available in one condition. Performance in the within-modal task will reflect perceiver's sensitivity to the visible differences between prosodic contrasts (i.e., that they perceptually salient), while the cross-modal matching task performance indicates whether the prosodic content of the auditory and visual tokens can be related to each

other. These questions are further explored in the first experiment of Chapter 3, where the within-modal and cross-modal matching tasks were once again utilised but with visual stimuli showing only the lower half of the talkers face. These results were compared with those obtained in Chapter 2 to determine whether a particular area of the face (i.e., upper half or lower half) is more effective for conveying prosodic focus or phrasing contrasts.

The remaining three experiments of Chapter 3 examined how sensitive perceivers are to variable realizations of prosody, by requiring perceivers to match prosodic tokens across talkers and differing face areas (i.e., matching the lower half of the face to the upper half, and vice versa, both within and across talkers). This task explored whether perceivers are able to tolerate signal-level differences between matching items (i.e., whether accurate prosody matching can be achieved despite any individual variation in how the prosodic cues were realised across talkers or face areas).

After confirming that perceivers are sensitive to the visual correlates of prosody, Chapter 4 outlines the recording of an auditory-visual speech prosody corpus, containing three-dimensional motion capture data for 2160 sentences. These recordings comprised of two repetitions of 30 mundane sentences produced across three prosodic conditions (i.e., broad focus, narrow focus, echoic question) in two interactive settings (face-to-face with the interlocutor or auditory only communication) by six native male talkers of standard Australian English, elicited in a dialogue exchange task which controlled the occurrence (and location) of the prosodically marked constituent. The data obtained from the corpus formed the basis of the subsequent analyses and perceptual experiments.

In Chapter 5, the auditory properties of the recorded corpus were quantified. Although the auditory characteristics of both focus and phrasing contrasts have been well described previously, examination of these properties for the current corpus ascertained whether the produced tokens showed the typical characteristics of these prosodic contrasts. In addition, it was determined whether the auditory properties of prosodic contrasts changed as a function of whether or not the talker was able to see their conversational partner, as well as exploring the degree of consistency across talkers in the realisation of prosodic contrasts.

The perceptual relevance of the auditory prosodic cues, and whether the changes as a function of the interactive setting had any impact on the perceptual salience of the contrasts, was evaluated in Chapter 6. This was done by using subjective rating tasks of the degree of focus received on the prosodically marked constituent, and the clarity of the statement-question contrast.

Chapter 7 examined the visual correlates of prosody in the recorded corpus. As with the analysis of the auditory properties, the visual analysis explored 1) what are the visual correlates of prosodic contrasts; 2) does the production of these visual cues differ as a function of whether or not they will be seen by an interlocutor; and 3) do talkers utilise idiosyncratic movement features to contrast the different prosodic types?

Once the auditory and visual correlates of the prosodic contrasts were established, the spatial and temporal relationship between the signal modalities was explored in Chapter 8 to determine the potential role of visual prosody. Three possible systematic relationships are considered between auditory and visual signals; that both signals occur simultaneously (suggesting an automatic coupling between

modalities); that visual prosodic cues precede the auditory prosodic markers (with visual cues acting as a signaling device to perceivers that important information is about to occur in the auditory modality); or that the visual markers occur after the auditory signal has been produced (in which case, the gestures may be produced to reinforce the auditory content).

In Chapter 9, the perceptual relevance of the visual correlates to prosody was investigated by presenting items to perceivers for subjective ratings in auditory only, visual only and auditory-visual conditions. By comparing the rating data between auditory only and auditory-visual presentation modalities, the perceptual benefits of visual prosodic cues can be explored.

An alternate perceptual task was used in the final experimental series presented in Chapter 10. Point-light representations of talkers' visual speech movements were presented to perceivers in a cross-modal matching task (as used in the initial experimental series) to investigate which visual motion cues (i.e., non-rigid, articulatory, or rigid head movements only) may be responsible for conveying prosodic content to perceivers, and to evaluate whether the different auditory and visual strategies used to prosodically mark constituents across talkers are as equally effective.

Finally, Chapter 11 draws together the findings of the experiments and highlights the limitations and suggests future directions.

CHAPTER 2.

PROSODY OFF THE TOP OF THE HEAD: PROSODIC CONTRASTS CAN BE DISCRIMINATED FROM HEAD MOTION

Chapter 2. Prosody off the top of the head: Prosodic contrasts can be discriminated from head motion²

The current chapter examined people's sensitivity to prosodic contrasts (focus: broad vs. narrow, and phrasing: statements vs. echoic questions) likely signalled by visual speech. Here, the visual speech information was restricted to the upper half of the talker's head and face. The motivation for concentrating on signals from the upper face and head motion was to determine whether visual signals not directly related to speech articulation could support prosody-related judgments. The proposition that visual prosodic cues can be obtained from peri-oral regions has received tentative support in the finding that people presented with extended monologues spent a considerable amount of time (65%) looking at the eyes and upper half of the talker's face (Vatikiotis-Bateson, Eigsti, Yano & Munhall, 1998; see also Buchan, Paré & Munhall, 2008). Furthermore, even when noise was added to the auditory signal, perceivers still looked at the upper face approximately half the time even though it might be expected that a person's gaze would shift to the mouth and jaw regions (as these would provide more beneficial cues for determining segmental content, Summerfield, 1979, 1992). The maintenance of gaze towards the upper face suggests that other speech-related information, such as prosody, is obtained from these regions.

More direct evidence for the role of the upper face in providing prosodic cues comes from the study by Lansing and McConkie (1999). These authors ascertained

² The contents of this chapter appear in published form as: Cvejic, E., Kim, J., & Davis, C. (2010). Prosody off the top of the head: Prosodic contrasts can be discriminated from head motion. *Speech Communication*, 52, 555-564. doi:10.1016/j.specom.2010.02.006.

where people looked when explicitly trying to decide on lexical content, prosodic focus, or utterance phrasing. To do this, participant's gaze direction was tracked when viewing visual only presentations of two-word sentences (e.g., "We won", "Ron ran") while identifying what was said, which word was narrow focused, or whether the sentence was a statement or question. It was found that the pattern of how long people looked at the upper, middle and lower parts of the face changed depending on the type of judgment being made. For judgements of prosodic focus and utterance phrasing, people looked longer at the middle and upper areas of the face, whereas they looked longer at the lower face when deciding upon what was said. However, eye-gaze patterns do not necessarily provide a complete picture of all the visual information that can be processed, i.e., such behavioural observations do not index information processed in peripheral vision. Indeed, it appears that motion information can be accurately processed in the periphery (Lappin, Tadin, Nyquist & Corn, 2009; McKee & Nakayama, 1984), thus movements associated with speech production (such as jaw, lip and mouth motion) do not need to be visually fixated in order to be accurately processed (Kim & Davis, 2011; Paré, Richler, ten Hove, & Munhall, 2003). To be certain as to what signals were in fact available to perceivers, Lansing and McConkie (1999) restricted face motion signals to particular facial regions (either full face or only lower face motion). When presented with full face or lower face motion, the identification of word content and sentence focus exceeded 95% correct, indicating that motion information in the lower half of the face was sufficient to perform such tasks. However, when identifying utterance phrasing, performance markedly declined when upper face motion was unavailable in comparison to full facial motion. This latter result suggests that visual information

from the upper face may be important for the accurate perception of phrasing, and raises the question of what face or head signals convey this information.

Since many auditory and visual speech properties originate from the same temporal process (i.e., speech production), it is clear why visual speech linked to the motion of the articulators is closely related to acoustics. That is, typically, articulatory movements (i.e., lip and mouth opening) or closely related movements (e.g., chin and cheek motion) are strongly correlated with aspects of the produced acoustics such as intensity variation over time that are used to signal prosody (Yehia, Rubin & Vatikiotis-Bateson, 1998). Indeed, in order to produce a speech sound over an extended duration (a common acoustic property of narrowly focused syllables), the speaker must maintain the configuration of the articulators for this time (de Jong, 1995). Similarly, increases in amplitude are likely to be accompanied by more dynamic jaw movements that end in a lower jaw position (Edwards, Beckman & Fletcher, 1991; Summers, 1987). However, it is less clear why the visible regions of the face *beyond* the mouth and jaw need be linked to speech acoustics. It is intriguing therefore that correlations have been found between different types of head and face motion and the change in acoustics as a sentence is uttered. The auditory property most studied is F_0 , with changes in this measure related to movements of the eyebrows (Cavé, Guaitella, Bertrand, Santi, Harley & Espesser, 1996; Granström & House, 2005; Guaitella, Santi, Lagrue & Cavé, 2009) and rigid head motion (Burnham, Reynolds, Vignali, Bollwerk & Jones, 2007; Ishi, Haas, Wilbers, Ishiguro & Hagita, 2007; Munhall et al., 2004; Yehia, Kuratate & Vatikiotis-Bateson, 2002). In general, it has been found that a significant positive correlation exists between F_0 and face and head motion. For example, Cavé et al. (1996) observed the non-rigid

eyebrow movements of ten talkers across various conversational settings, and found that rising F_0 patterned with eyebrow movements. However, these movements did not occur for every change in F_0 , suggesting that the coupling was functional, rather than an automatic uncontrolled consequence of articulation. Similarly, a functional relationship was suggested between variation in F_0 and rigid head motion in Yehia et al. (2002). These results indicate that prosody may be signalled both by non-rigid eyebrow movements and rigid head motion.

More recent studies have attempted to identify the nature of the visual speech signals that co-occur with prosodic focus, and to examine whether visual prosody can be identified in perception experiments. They include naturalistic production studies that have examined the motion information produced in conjunction with the auditory signal (with particular focus on either oral or peri-oral areas), and manipulation studies that have examined how changes in auditory signals affect the way prosody is perceived. The review of these studies will concentrate on results pertinent to peri-oral signals related to prosody.

Building on the work of Dohen and Lœvenbruck (2005) that showed the production of the focal syllables involved significantly larger lip areas, Dohen, Lœvenbruck and Hill (2006) investigated whether movements beyond the oral area (eyebrow and head movements) might also be associated with the production of prosodic focus. This study used motion capture to measure lower face movements (lip opening, lip spreading and jaw motion) as well as head and eyebrow movements in five French talkers. A relationship was found between eyebrow motion (rising) and the production of prosodic focus for three out of the five talkers, and a relationship between head nods and focus production for one talker. Scarborough,

Keating, Mattys, Cho and Alwan (2009), who also used motion capture, examined the visual correlates of lexical (using reiterant syllable-based versions of words) and phrasal stress in three male talkers of Southern Californian English. For lexical stress, it was found that there was greater head motion for the stressed syllables, but no differences in eyebrow movement. For phrasal stress, every measure (including eyebrow measures) distinguished stressed from unstressed words.

Dohen and Løevenbruck (2005) and Scarborough et al. (2009) conducted additional perception studies to determine if observers were sensitive to visual prosody. Dohen and Løevenbruck showed that when participants were presented with soundless videos of talkers uttering a sentence that had narrow focus on the subject, verb or object phrase (or broad focus), they could successfully identify the focused constituent at better than chance levels. Scarborough et al. also showed that when participants were presented with three stimuli to decide which received stress (with an additional “no stress” option), lexical and phrasal stress in visual only presentation could both be perceived at better than chance levels. Likewise, a similar study by Srinivasan and Massaro (2003) showed that participants could identify whether a sentence presented in a silent video was a statement or an echoic question, indicating that visual speech alone is capable of conveying information relating to utterance phrasing. It should be noted that all of the above perceptual studies presented the talker’s whole head, so it is not possible to separate the effect of visual cues directly related to speech production (mouth and jaw) from those signalled by such things as eyebrows and head motion.

In a study similar in concept to that of Lansing and McConkie (1999), Swerts and Krahmer (2008) used monotonic renditions of broad focused auditory statements

paired with narrow focused visual speech tokens and asked perceivers to identify which word within the utterance received prosodic focus. In the critical conditions for current concerns, Swerts and Kraemer presented participants with either full face videos, videos that restricted visibility to the upper face, or to only the lower face. It was found that performance varied across viewing condition: performance on the videos showing only the upper face was equal to the full face condition (77.3% correct) and significantly better than the lower face presentation condition (51.4% correct, which was itself better than chance performance of 25%). Swerts and Kraemer interpreted their findings as showing that the upper face has more cue value for phrasal prominence (i.e., narrow focus).

In sum, it has been demonstrated that information from the upper face can convey visual cues for prosodic focus and phrasing. However, what remains to be determined is the type of information from this area (e.g., rigid or non-rigid movement) that provides these cues. Given that Scarborough et al. (2009) showed that lexical stress was associated with greater head motion but not with movements of the eyebrows, it would seem that rigid motion might be the principle cue. Thus, the experiments presented in this chapter investigate whether rigid head motion when separated from other face cues (e.g., eyebrow motion, brow shape and textural information) is capable of providing prosodic information.

2.1. Experiment 1: Visual-Visual (VV) Prosody Matching

Experiment 1 gauged perceivers' sensitivity to prosody related visual cues from the talkers' upper face by using a visual-visual discrimination task (adopting the procedure used in Davis & Kim, 2006). The aim of this experiment was to determine whether visual signals related to prosody can be used to drive reliable perceptual

discrimination. That is, if talkers consistently produce visual prosodic cues in upper face regions, then participants should be capable of discriminating between pairs of stimuli that differ only in prosody. Video stimuli are presented in two conditions: fully textured, providing a combination of rigid and non-rigid movements, and outline-only, providing predominantly rigid motions of the head. A comparison of the results across these two stimuli presentation conditions will indicate the contribution of particular types of visual cues to discrimination performance (e.g., a drop in performance when only rigid information is available would suggest that non-rigid cues carry more perceptually relevant information).

2.1.1. Method

2.1.1.1. Materials

The materials consisted of 10 non-expressive sentences drawn from the IEEE (1969) Harvard Sentence list that describe fairly mundane events of minimal emotive content (Table 2.1). Audio-visual recordings were made of two age-matched, native male talkers of standard Australian English ($M_{\text{Age}} = 23$ years) in a well-lit, sound attenuated room against a neutral coloured background using a Sony TRV19E digital video camera (25 fps). Audio was synchronously recorded at 44.1 kHz, 16-bit mono with an externally connected Senheiser e840 lapel microphone.

Table 2.1. Sentence material used for the audio-visual recordings of Experiment 1. The prosodically marked word is italicised.

Sentence	Segmental Content
1	It is a band of <i>steel</i> three inches wide
2	The pipe ran almost the <i>length</i> of the ditch.
3	It was hidden from sight by a <i>mass</i> of leaves and shrubs.
4	The weight of the <i>package</i> was seen on the high scale.
5	Wake and rise, and <i>step</i> into the green outdoors.
6	The green light in the <i>brown</i> box flickered.
7	The brass <i>tube</i> circled the high wall.
8	The lobes of her ears were <i>pierced</i> to hold rings.
9	Hold the <i>hammer</i> near the end to drive the nail.
10	Next <i>Sunday</i> is the twelfth of the month.

Each sentence was recorded as a *broad focused* statement, a *narrow focused* statement and as an *echoic question*. A dialogue exchange task was used to elicit these conditions in which the talker interacted with an interlocutor by either repeating what the interlocutor said (broad focused statement), making a correction to an error made by the interlocutor (narrow focused statement, Example 1), or questioning an emphasised item that was produced by the interlocutor (echoic question, Example 2). An example of this dialogue is provided below:

Example 1.

(a) The pipe ran almost the [width]_{Error} of the ditch.

(b) The pipe ran almost the [**length**]_{Correction} of the ditch.

Example 2.

(a) The pipe ran almost the [**length**]_{Emphasised} of the ditch.

(b) The pipe ran almost the [*length*]_{Questioned} of the ditch?

Chapter 2: Prosody off the Top of the Head

The critical item within each utterance (i.e., the word erroneously produced or emphasised by the interlocutor that in turn received narrow focus or question intonation when produced by the talker) was a content word, began with a consonant, and was not located in phrase-initial or phrase-final position, with the position within the utterance varying across the ten sentences. The same critical item was maintained across prosodic speech conditions and talkers. Two repetitions of each utterance were recorded several minutes apart. This recording procedure resulted in 120 auditory and 120 visual speech tokens for use as stimuli.

The visual tokens were then processed using custom designed scripts in VirtualDub (Lee, 2008) to generate two versions of visual stimuli. The videos were cropped at the tip of the talkers' nose, generating stimuli showing only the upper half of the talkers face (i.e., 'textured'). A thresholding filter that converts image sequences into black and white based on colour values was then applied to copies of the upper face videos. This process removed most of the non-rigid movements of the talker's face (eyebrow and skin deformations), leaving only a basic outline of the face, hair and eyes (referred to as 'outline only' stimuli). An example of these stimuli is displayed in Figure 2.1.



Figure 2.1. Originally recorded tokens (left) were cropped at the nose tip, generating textured upper face stimuli (right upper panel). A thresholding filter was then applied to copies of these stimuli to create the outline only videos (right lower panel).

2.1.1.2. Preliminary Acoustic Analysis

To ascertain that the recorded auditory stimuli showed the expected differences across the varying prosodic speech conditions, the acoustic properties (i.e., duration, mean $F0$, mean intensity, $F0$ range and intensity range) of the critical constituent within each utterance was extracted using custom-designed scripts in Praat (Boersma, 2001). These values were then normalised to a proportion value of the average broad focused rendition (per sentence and talker, collapsed across repetitions, as in Dohen et al., 2009). These proportion values for each acoustic feature were then subjected to a 2×3 repeated measures analysis of variance (ANOVA), with talker (Talker 1; Talker 2) and prosodic speech condition (broad focus; narrow focus; echoic question) treated as within-items independent factors.

With α set to 0.05, the main effect of prosodic speech condition was significant for duration, $F(2,18) = 135.66, p < 0.001, \eta_p^2 = 0.938$, mean $F0$, $F(2,18) = 36.85, p < 0.001, \eta_p^2 = 0.804$, mean intensity, $F(2,18) = 5.97, p = 0.010, \eta_p^2 =$

0.399, F_0 range, $F(2,18) = 6.42$, $p = 0.008$, $\eta_p^2 = 0.416$, and intensity range, $F(2,18) = 14.43$, $p < 0.001$, $\eta_p^2 = 0.616$. Sidak post-hoc comparisons indicated that the critical items recorded in the narrow focused condition differed from broad focused renditions with longer durations [$M_{\text{Diff}} = 0.96$, Sidak 95% CI: 0.80 – 1.12], higher mean F_0 [$M_{\text{Diff}} = 0.05$, Sidak 95% CI: 0.03 – 0.08] and greater intensity range [$M_{\text{Diff}} = 0.52$, Sidak 95% CI: 0.26 – 0.78]. Echoic questions also differed relative to broad focused statement renditions on the critical constituent in terms of duration [$M_{\text{Diff}} = 0.49$, Sidak 95% CI: 0.29 – 0.69], with questioned content being lengthened, higher mean F_0 [$M_{\text{Diff}} = 0.08$, Sidak 95% CI: 0.05 – 0.10], and larger F_0 range [$M_{\text{Diff}} = 2.47$, Sidak 95% CI: 0.44 – 5.00]. The results of the acoustic analyses confirm that the auditory content produced by the talkers differed between the prosodic conditions on the critical constituent as expected (see Figure 2.2).

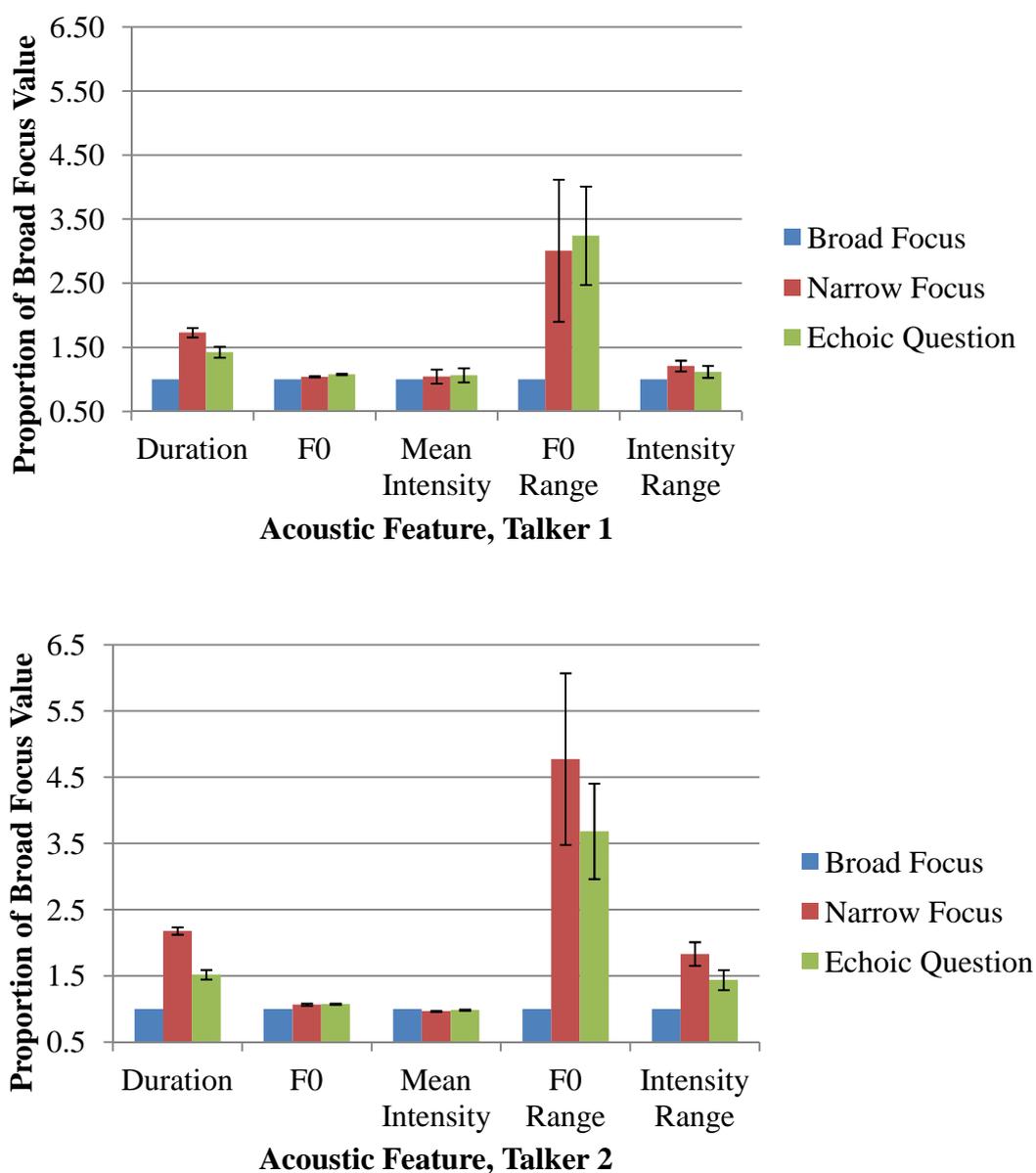


Figure 2.2. Acoustic properties of the critical constituent expressed as a proportion of the mean broad focus rendition, collapsed across utterances and repetitions, for Talker 1 (upper panel) and Talker 2 (lower panel). Error bars indicate the standard error of the mean.

2.1.1.3. Participants

Twenty undergraduate psychology students ($M_{\text{Age}} = 20.3$ years) from the University of Western Sydney (UWS) participated in the experimental tasks for course credit.

All were fluent talkers of English, and self-reported normal or corrected-to-normal

vision, normal hearing, with no known communicative deficits. Participants were randomly allocated to a visual stimuli condition (i.e., textured upper face or outline only). All participants were treated in accordance with the ethical protocols outlined by the UWS Human Research Ethics Committee.

2.1.1.4. Procedure

Participants were tested individually in a sound-attenuated booth. The experiment was run in DMDX (Forster & Forster, 2003), with video stimuli displayed on a 17" LCD display monitor in a two-interval alternate forced choice (2AFC) matching task (see Figure 2.3), in which each interval included a pair of stimuli to be compared, with the participant's task to select the pair produced as the same prosodic type. The 2AFC procedure was chosen (opposed to an AXB, ABX or single interval identification task) as it provides participants with a constant reference that is always one item back from the to-be judged token, keeping the circumstances of comparison consistent across items.

Participants were informed of the three prosodic conditions used (in straightforward language to ensure that the distinctions were clear), and that the sentences they would be judging differed only in prosody, not segmental content. To rule out instance-specific matching strategies, the matching items within pairs were always taken from a different recorded token. All items within pairs were produced by the same talker. Participants indicated their response as to which pair was produced with the same prosody via a selective button press. No feedback was given as to the correctness of their response. Video display, item randomisation and collection of response data was controlled by the stimulus presentation software.

Chapter 2: Prosody off the Top of the Head

A total of 40 matching responses were involved across two prosodic speech conditions (i.e., narrow focus and echoic questions), with the broad focused renditions always acting as the non-matching item within pairs. Participants were first presented with “PAIR 1” displayed on the screen for 1000 ms, followed by the first pair of silent videos. “PAIR 2” was then displayed on the screen for 1000ms, before the second silent video pair was presented. This was followed by a prompt to respond, with a response required within 10 second. Failure to respond within this time limit was counted as an incorrect response. The first silent video within each pair was identical and the standard with which the other stimulus item within pairs were to be judged against. The order of correct response pair was counter-balanced, so the correct option appeared equally in the first and second interval. Videos were presented for the entire experiment in one of two presentation conditions; as textured videos (providing both rigid and non-rigid cues), and outline only (showing only rigid motion cues).

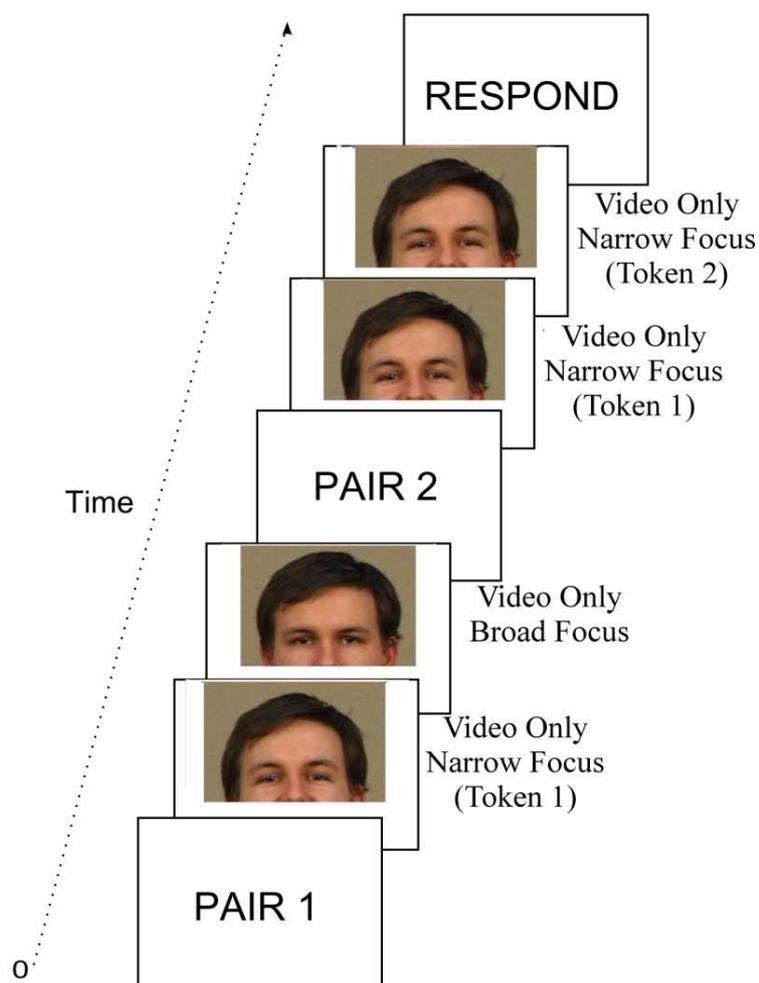


Figure 2.3. Schematic representation of the 2AFC task used in the visual-visual matching task of Experiment 1. The same item appeared first for both pairs, and was the standard that the matching judgment was to be made on.

2.1.2. Results and Discussion

The mean percent of correct responses for the VV matching task of Experiment 1 are displayed in Table 2.2. Performance was substantially better than that expected by chance alone (i.e., 50%) for both stimulus presentation conditions, as confirmed by a series of significant one-sample *t*-tests.

Table 2.2. Mean percent of correct responses in the VV matching task as a function of the stimulus presentation condition for each prosodic speech condition. Degrees of freedom (*df*) are indicated in brackets and ** indicates $p < 0.001$.

Stimulus Presentation Condition	Prosodic Speech Condition	Mean Correct (%)	Standard Error of Mean	<i>t</i>-test vs. chance (50%)
Textured (<i>df</i> = 10)	Narrow Focus	82.7	2.97	11.03**
	Echoic Question	87.7	3.26	11.58**
Outline Only (<i>df</i> = 8)	Narrow Focus	93.3	2.04	21.23**
	Echoic Question	91.7	2.50	16.67**

Further analyses were conducted comparing the results across presentation conditions. A 2×2 mixed repeated measures ANOVA (interpreted with an α of 0.05) was used to determine if task performance (percent correct responses) varied as a function of the presentation condition (upper face vs. outline only), with prosodic condition (narrow focus; echoic question) as a within-subjects factor, and presentation condition as a between-subjects factor. A significant main effect of stimulus presentation condition was found, $F(1,18) = 7.05$, $p = 0.016$, $\eta_p^2 = 0.282$, with superior performance observed for the outline only stimuli (92.5% correct) in comparison to textured upper face stimuli (85.2% correct). The main effect of prosodic condition, $F(1,18) = 0.32$, $p = 0.579$, and the prosody by presentation condition interaction, $F(1,18) = 1.28$, $p = 0.273$, failed to reach significance.

Post-hoc comparisons (interpreted with a Bonferroni adjusted α of 0.025 for multiple comparisons) showed that the presentation condition main effect was driven primarily by the significant difference observed between textured upper face and outline only displays for narrow focused items, $F(1,18) = 7.90$, $p = 0.012$, $\eta_p^2 =$

0.305, since the difference across presentation conditions for echoic questions was not significant, $F(1,18) = 0.85, p = 0.368$.

The results showed that people could use visual displays from the talker's upper face to distinguish narrow focus and echoic question utterances from broad focused renditions. As different recorded tokens were used within stimulus pairs, it is expected that there would have been some minor movement and temporal variations across the recordings of the same segmental content. As such, participants could not simply base their matching judgments on identifying *any* difference between serially presented videos, but rather needed to identify particular patterns of motion that were consistent across the video pairs.

The result that performance was maintained when non-rigid information was removed from the visual signal (i.e., in the outline only stimulus presentation condition) suggests that rigid head movements and iris position provide sufficient cues to perform the task. Performance was indeed better for the outline only than the textured video condition in the VV task. One potential explanation for this occurrence may be that subtle differences between the prosodic contrasts were made more apparent in the outline videos, removing potentially redundant visual information thus making it easier to select the correct pair.

However, it should be noted that successful performance in the within-modal (VV) task indicates only that there are differences between the visual cues used by talkers to contrast prosodic types (i.e., broad vs. narrow focus; statements vs. questions) and that they are perceptually salient. It does not necessarily indicate that perceivers base their performance on having recognised prosody (although the matching task instructions were phrased in terms of selecting the pair that has the

same prosody). Thus, the following experiment used an auditory-visual discrimination task to test whether people can relate prosodic contrasts to specific upper head and face movements.

2.2. Experiment 2: Auditory-Visual (AV) Prosody Matching

Experiment 2 followed the same basic design of Experiment 1 (i.e., using a 2AFC discrimination task), except that this experiment uses auditory-visual stimulus pairs, requiring the perceiver to select the pair in which the auditory and visual stimuli matched.

2.2.1. Method

2.2.1.1. Participants

The same participants that completed Experiment 1 took part in Experiment 2. The order in which participants completed the experimental tasks was counter-balanced (i.e., some took part in the auditory-visual task first before completing the visual-visual matching task, and vice versa). There was a break of several minutes between sessions in order to minimise potential order, exposure, or fatigue effects. Stimuli were presented to participants in the same visual condition as in Experiment 1 (i.e., textured or outline only).

2.2.1.2. Materials and Procedure

The stimuli used for Experiment 2 were the same as those outlined for Experiment 1. Participants were presented with two pairs of stimuli, each consisting of an auditory-only stimulus followed by a silent video showing the upper head of the talker uttering the same sentence as in the auditory-only stimulus. The participants' task

was to select the auditory and visual pair in which the prosody of the utterance matched (Figure 2.4), with the set of instructions issued as for Experiment 1. The matching auditory and visual items were always taken from a different recorded token to rule out instance-specific matching strategies (e.g., absolute duration). The mismatching items within pairs were the same segmental content produced as one of the alternate prosodic types (i.e., the non-matching items for half of the narrow focus trials were broad focused renditions and echoic question renditions for the remaining half). The initial auditory token that appeared at the start of each pair was the same and the standard against which the silent videos were to be matched, with each of the 120 recorded auditory tokens appearing as the target once.

Stimuli were randomly presented in four blocks of 30 items, with each block containing one of each sentence in all three prosodic speech conditions produced by an individual talker. Within-block randomisation was controlled by the stimulus presentation software. Auditory stimuli were presented binaurally via Senheiser HD650 stereo headphones. Participants completed both tasks with either textured upper face or outline only stimuli. All other procedural details are the same as outlined for Experiment 1.

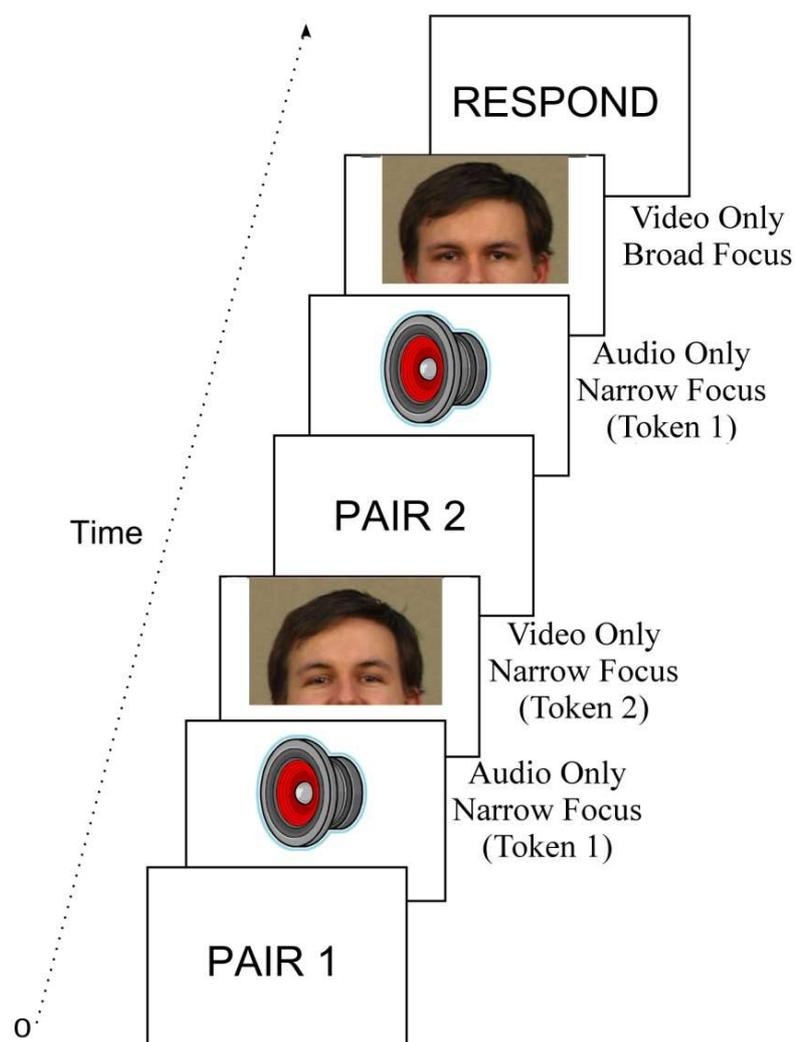


Figure 2.4. Schematic representation of the 2AFC task used in the auditory-visual matching task used in Experiment 2. The matching pair was always taken from a different recorded token, and the distracter stimuli were always an alternate prosodic type. Stimuli within both audio-video pairs were always produced by the same talker.

2.2.2. Results and Discussion

The mean percent of correct responses for both textured and outline presentation conditions are shown in Table 2.3. Performance was substantially better than what would be expected by chance alone. This was confirmed with a series of one-sample *t*-tests, the values of which are also reported in Table 2.3.

Table 2.3. Mean percent of correct matching responses in the 2AFC auditory-visual matching, as a function of stimulus presentation condition and prosodic speech condition. Degrees of freedom are indicated in brackets and ** indicates $p < 0.001$.

Stimulus Presentation Condition	Prosodic Speech Condition	Mean Correct (%)	Standard Error of Mean	<i>t</i>-test vs. chance (50%)
Textured (<i>df</i> = 10)	Broad Focus	88.9	2.14	18.15**
	Narrow Focus	94.8	1.61	27.79**
	Echoic Question	88.9	2.39	16.25**
Outline Only (<i>df</i> = 8)	Broad Focus	82.5	2.89	11.26**
	Narrow Focus	92.2	2.37	17.81**
	Echoic Question	91.9	1.58	26.56**

A 2×3 mixed repeated measures ANOVA (with an α level of 0.05) was conducted to determine whether performance in the AV matching task differed across stimulus presentation conditions, with prosodic speech condition (broad focus; narrow focus; echoic question) as the within-subjects factor, and presentation condition as the between-subjects factor. The main effect of prosodic condition was significant, $F(2,36) = 7.93$, $p = 0.001$, $\eta_p^2 = 0.306$, however the presentation condition main effect, $F(1,18) = 0.85$, $p = 0.369$, and the prosody by presentation condition interaction, $F(2,36) = 2.85$, $p = 0.071$, did not reach significance. Post-hoc comparisons (interpreted with a Bonferroni adjusted α of 0.025 for multiple comparisons) revealed no significant differences for any of the prosodic speech conditions between textured and outline presentation conditions (broad focus, $F(1,18) = 3.27$, $p = 0.087$; narrow focus, $F(1,18) = 0.83$, $p = 0.374$; echoic questions, $F(1,18) = 0.99$, $p = 0.332$).

The results obtained in Experiment 2 (using auditory-visual stimuli) were highly similar to those obtained for Experiment 1 (using visual-visual item pairs). The similarity of the findings across the two experiments suggests that people were aware of and used visual prosody in both matching tasks. Furthermore, in both experiments it was found that perceivers could perform just as well using the outline videos as the full textured ones. Given that in outline videos visual cues such as eyebrow motion and skin deformations are eliminated, but the rigid head motion cue remains unaffected, it would appear that rigid motion provides sufficient cues to reliably match prosodic contrasts.

2.3. General Discussion

The experiments presented in the current chapter examined perceivers' ability to use restricted visual displays showing only the talker's upper face to match prosody within and across modalities. Cues from the upper face were used not because they are likely to provide stronger or more salient cues than the lower face (indeed, previous studies have reported that visual cues from the lips, mouth and jaw are capable of signalling prosody, see Dohen et al., 2009; Erickson, Fujimura & Pardo, 1998; Swerts & Kraemer, 2008), but because they are not so directly tied to articulation, and as such may represent a reinforced behaviour that achieves better communication. The current experiments investigated whether people could use the visual cues from the talker's upper head and face to discriminate sentences that differed only in prosody, and whether such cues could be matched to the prosody of auditory stimuli. Additionally, the amount of visual information available was manipulated by using textured videos that provided a combination of rigid and non-

rigid movement cues, and videos showing only an outline of the head and iris position.

The results of the VV matching task indicated that people are able to discriminate echoic question and narrow focused sentences from broad focused renditions even though all had the same segmental content. This was the case for both fully textured (containing a combination of rigid and non-rigid head motion) and outline only (rigid motion only) presentations, indicating that rigid head motion provides reliable cues about the prosodic contrasts. A parsimonious explanation of the overall pattern of results is that perceivers use only the rigid motion cue (this would account for the similar performance levels across stimuli presentation conditions). However, the ability of participants to perform both VV and AV matching tasks despite the removal of non-rigid information does not necessarily mean that *only* rigid motion was used to make matching judgements, but rather that perceivers may be flexible in the cues that they use to discriminate between contrast types. That is, when a combination of rigid and non-rigid cues are presented, either of these cues can be exploited; however, when rigid motion is the only cue available it provides sufficient contrastive detail to still allow for accurate prosodic discrimination.

Indeed, the results of previous studies suggest that not only are eyebrows movements associated with phrasal stress but that people use this cue in prosody perception. For example, Scarborough et al. (2009) reported that their talkers raised an eyebrow on almost all focused words, and that eyebrow displacement accounted for significant variance in subsequent perception tasks. Furthermore, in an experiment that required Dutch participants to produce nonsense three-word

sentences using reiterant speech with focus on one word, Kraemer and Swerts (2004) found that more talkers (nine out of twenty) raised their eyebrows to indicate narrow focus (compared to only four who used head movements). Of course, it may be that the perceivers in the current study were less inclined to use such eyebrow movement cues because such a cue might have had a differential value for the type of stimuli they were judging (i.e., narrow focussed sentences and echoic questions). That is, Flecha-Garcia (2006, 2010) found that the frequency of produced eyebrow raises did not distinguish questions from other types of utterances. If this were true for the current stimuli, then using eyebrow motion would not be a useful cue for half of the stimulus trials (i.e., the echoic question renditions).

The results of the auditory-visual matching task showed that people could reliably match auditory prosody to the corresponding visual signal, despite the visual tokens available differing only in prosody, not segmental content (c.f., Davis & Kim, 2006). This indicates that participants knew how specific prosodic contrasts played out in both the auditory and visual signals. Furthermore, when non-rigid motion such as eyebrow movements were removed from the visual signal, task performance was maintained. Once again, this suggests that the rigid head motion (nods and tilts) produced by a talker were sufficient to convey information about the prosodic nature of an utterance.

It is important to consider the scope of the study and what the results are capable of saying about visual prosody. The aim of the study was to determine whether the visual cues for prosody could be reliably extracted from the talker's upper head and face. For this purpose, the task of discriminating paired utterances that differed only on visual cues related to prosody (Experiment 1) and task of

selecting auditory-visual pairs that matched on prosody (Experiment 2) were used when the visual stimuli only presented a talker's upper face. In this regard, human perceivers were employed as sensitive measurement devices, with their performance reflecting that the visual cues for prosody were available from the talker's upper head, and that these signals manifested consistently enough over utterance repetitions to drive high levels of correct performance. Having said this, it should be clear that this type of inquiry does not reveal whether people actually *use* these cues when not specifically directed to do so.

Furthermore, it is likely that performance on the matching tasks represents a generous estimate of the information that is available, since the sequential display of minimal prosodic pairs may highlight key differences and possibly shape correct responding. Indeed, performance on the matching tasks was very good; in fact, scores were higher than other perception studies in which participants were required to identify which stimulus had been prosodically marked. For instance, overall correct performance on deciding which of three stimulus words (names) received focus (with an additional "no focus" option) reported by Scarborough et al. (2009) was 54% (with chance being 25%); in Bernstein, Eberhardt and Demorest (1989) correct performance was 76% (with chance equal to 33%), and stimuli produced by one talker in Dohen, Løevenbruck, Cathiard and Schwartz.(2004) was identified with 71% accuracy (with chance at 25%).

CHAPTER 3.
RECOGNISING PROSODY ACROSS MODALITIES,
FACE AREAS AND TALKERS:
EXAMINING PERCEIVERS' SENSITIVITY TO
VARIABLE REALISATIONS OF VISUAL PROSODY

Chapter 3. Recognising prosody across modalities, face areas and talkers: Examining perceivers' sensitivity to variable realisations of visual prosody³

In Chapter 2, it was demonstrated that perceivers were capable of matching prosodic contrasts within and across modalities when presented with restricted visual displays that showed only the upper face of the talker. This ability was maintained even when non-rigid movements of the face were removed from the signal. It was proposed that participants' ability to perform this task stemmed from there being multiple cues to prosody contained within the visual speech signal, and when one of these cues is no longer available, those that remain allow the underlying prosodic category to be determined (i.e., perceivers are flexible in the cues that they can use to ascertain prosodic information from). In the current chapter, this hypothesis is further explored by examining perceivers' sensitivity to variable realisations of visual prosody.

Production studies examining visual cues to prosody indicate that the manner in which these are realised is quite variable across talkers. For example, Dohen, Løevenbruck and Hill (2009) examined five native French talkers' utterances that had narrow focus (on the subject, verb or object of the base sentences) compared to broad focus and found that in general, narrow focused syllables attracted

³ The contents of this chapter appear in the following peer-reviewed published works:
Cvejjic, E., Kim, J., & Davis, C. (in press). Recognizing prosody across modalities, face areas and speakers: Examining perceivers' sensitivity to variable realizations of visual prosody. *Cognition*. doi:10.1016/j.cognition.2011.11.013.

Cvejjic, E., Kim, J., & Davis, C. (2010a). It's all the same to me: Prosodic discrimination across speakers and face areas. *Proceedings of 5th International Conference on Speech Prosody 2010*, 100893, pp. 1-4.

Cvejjic, E., Kim, J., & Davis, C. (2010b). Abstracting visual prosody across speakers and face areas. *Proceedings of the 9th International Conference on Auditory Visual Speech Processing (AVSP2010)*, pp.38-43.

hyperarticulation (larger mouth/jaw opening gestures). However, it appeared that talkers used different strategies to signal focus, with one talker consistently showing larger mouth area and significantly longer gesture duration for focused versus non-focused syllables, whereas another showed considerable variation in whether a focused syllable was marked by duration and/or enhanced mouth opening. Similarly, while all talkers hyperarticulated the prosodically marked constituent, some talkers did so to a lesser degree yet complimented this by hypoarticulating post-focal utterance content.

This variability also extends to prosodic cues occurring outside of the oral region that are not closely linked to speech articulation. Dohen et al. (2006) found that although all five of their recorded talkers moved their head to some degree, only one talker showed a significant correlation between rigid head tilts and the production of focus, a link that appeared to be non-systematic and highly variable. Moreover, only three of the five talkers raised their eyebrows on focused syllables (and these movements did not always accompany the production of a focused constituent). Consistent with this, the results of a study by Cavé et al. (1996) that examined whether changes in *F0* were accompanied by eyebrow movements showed considerable variability both within and between talkers in whether eyebrow raises were accompanied by a rise in *F0* (see also Guaitella et al., 2009). The above two studies examined French talkers, however similar patterns of movement and inter-talker variability have been shown for talkers of American English in producing words with lexical stress and phrasal focus (Scarborough et al., 2009).

In general, the above production studies have found both within- and between-talker variation in when and indeed whether particular visual cues for

Chapter 3: Recognising Prosody across Modalities, Face Areas and Talkers

prosody are used. Further it appears that there are differences between the upper and lower face cues both in the strength of the prosodic information signaled, and in the regularity that such signals are emitted. Despite this, a range of perception studies (as reviewed in Chapters 1 and 2) have shown that people readily perceive and use such visual cues, affecting how auditory prosody is perceived or directly affecting the interpretation of a spoken utterance (e.g., Dohen & Løevenbruck, 2009; Foxton, Riviere & Barone, 2010; Lansing & McConkie, 1999; Swerts & Krahmer, 2008).

This apparent mismatch between signal variability and constant perception highlights a basic issue in human pattern recognition (i.e., how variable form is mapped onto perception). In speech perception, this has often been characterized in terms of the problem of a lack of invariant cues to support categorical distinctions. Although there have been various proposals to account for such an ability, it remains a fundamental concern (e.g., see McMurray & Jongman, 2011). In the domain of visual prosody research the issue has yet to be considered, but the question of how perceivers cope with variability in the realisation of prosodic cues seems equally important. To begin to answer this question, what is needed is a better understanding of how the visual cues themselves are perceived and how they relate to auditory prosodic cues.

The results obtained in Chapter 2 demonstrated that perceivers were capable of matching prosodic contrasts within and across modalities when presented with restricted visual displays showing only the talker's upper face. The ability of participants to accurately match the type of prosody across different visual tokens and different modalities showed that more than a simple feature-to-feature matching strategy was involved. Indeed, it is proposed that participants were able to achieve

such high levels of correct across-token matching performance because they could classify the type of prosody from the visual cues (e.g., narrow or broad focus; statement or question) and then use the result of this classification to decide which stimulus pair matched.

This idea that good performance in the matching task is based upon the categorization of visual prosody cues suggests that this task may be useful in probing the extent to which a prosodic category can be determined from *different* inputs. Thus, the matching paradigm used in Experiments 1 and 2 provides a well controlled situation to assess the extent to which perceivers can tolerate variation in the production of visual prosody. For instance, the within-modal matching task can be used to investigate whether people can determine prosodic counterparts across different face regions by testing whether people can reliably match cues to prosody that are signaled by the upper and lower face (an important division as it picks out those cues that stem more directly from articulation from those that do not). Furthermore, the cross-modal matching task can also reveal whether perceivers can ascertain the underlying prosody type regardless of who is speaking by testing matching ability across different talkers.

To probe the extent to which prosodic categories can be determined from different inputs, the current experimental series tests people's ability to match visual prosody when the cues manifest in different ways; such ability would provide a basis for using visual cues to prosody even though they are variable across talkers and face areas. Specifically, Experiment 3 examined whether perceivers were able to match within-modal (visual to visual) and cross-modal (auditory to visual) cues from the lower face region (as this was not assessed in Experiment 1 or 2, and provides an

opportunity to assess whether cues to visual prosody that are linked to articulation show a different pattern across prosodic types). Experiment 4 examined people's ability to match cues across different talkers for the same face region. Experiment 5 tested if perceivers could successfully match across face areas within a talker, and Experiment 6 examined matching both across face areas and across different talkers.

3.1. Experiment 3: Perceiving Prosody from the Lower Face

The aim of Experiment 3 was to ascertain the extent to which perceivers were able to use visual cues from the lower face region for matching within (i.e., visual to visual) and across modalities (i.e., auditory to visual). These results, taken together with the results of Experiments 1 and 2, will provide a baseline from which to evaluate cross-talker matching (Experiment 4).

3.1.1. Method

3.1.1.1. Participants

Twenty undergraduate students from UWS ($M_{\text{Age}} = 21.3$ years) participated in the experiment in return for course credit. All were fluent talkers of English. None had previously taken part in Experiment 1 or 2, and all reported normal or corrected-to-normal vision with no history of hearing loss, and no known communicative deficits. All participants were treated in accordance with the policies of the UWS Human Research Ethics Committee.

3.1.1.2. Materials

The materials used in this chapter were the same 120 audiovisual tokens as those used in Experiments 1 and 2, comprising of two talkers producing two repetitions of

ten sentences across three prosodic conditions. The visual tokens were processed in VirtualDub (Lee, 2008) to create an additional version of visual stimuli by cropping the video displays at the nose tip, generating lower face videos that showed only the lips, cheeks, chin and jaw of the talker (see Figure 3.1).



Figure 3.1. The original audiovisual recordings made in Chapter 2 were cropped at the nose tip to generate lower face stimuli.

3.1.1.3. Procedure

The experiments were run in DMDX (Forster & Forster, 2003) on a desktop computer connected to a 17" LCD Monitor. Participants were tested individually in a double-walled, sound attenuated booth. Each participant completed two experimental tasks: a visual-visual (VV) matching task (as in Experiment 1) and an auditory-visual (AV) matching task (as in Experiment 2) in a counter-balanced order (i.e., half of the participants completed the VV task first, while the remaining half completed the AV task first). These tasks were procedurally identical to 2AFC tasks used in Experiments 1 and 2, except that the video stimuli presented showed only the talker's lower face (Figure 3.2).

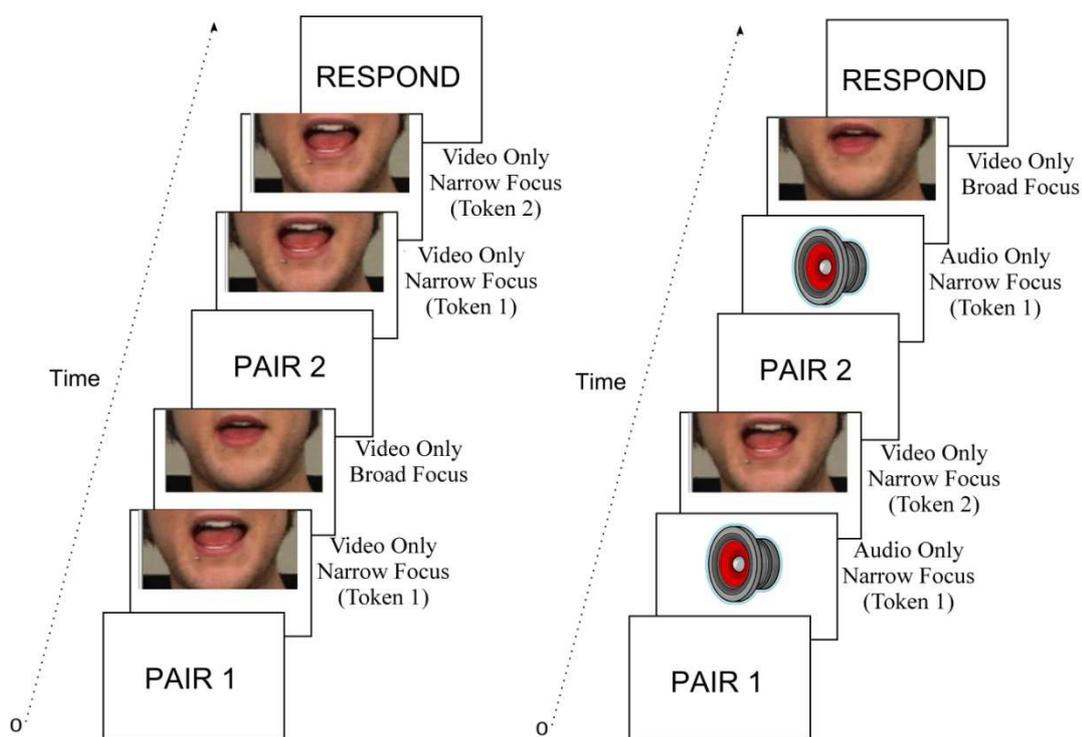


Figure 3.2. Schematic representation of the 2AFC visual-visual (left) and auditory-visual (right) matching tasks used in Experiment 3. The same item appeared first for both pairs, and was the standard that the matching judgment was to be made on. The matching item within pairs was always taken from a different recorded token, and non-matching items were the same sentence produced as a different prosodic type.

3.1.2. Results and Discussion

The mean percent of correct responses for the VV and AV matching tasks are shown in Table 3.1. The results are presented together with the textured upper face data from Experiments 1 and 2 to allow for full comparisons between upper and lower face stimuli presentation. As can be seen, performance was considerably greater than chance (i.e., 50%) across all prosodic speech conditions in both tasks, as confirmed by a series of significant one-sample *t*-tests.

Table 3.1. Mean percent of correct responses in the within-talker VV and AV matching tasks as a function of visible face area in each prosodic speech condition. Data in italics are from Experiments 1 and 2. Degrees of freedom (*df*) are indicated in brackets and ** indicates $p < 0.001$.

Visible Face Area	Prosodic Speech Condition	Mean Correct (%)	Standard Error of Mean	<i>t</i>-test vs. Chance (50%)
<u>Visual-Visual (VV) Matching Task</u>				
<i>Upper Half</i> <i>(df = 10)</i>	<i>Narrow Focus</i>	<i>82.7</i>	<i>2.97</i>	<i>11.03**</i>
	<i>Echoic Question</i>	<i>87.7</i>	<i>3.26</i>	<i>11.58**</i>
<i>Lower Half</i> <i>(df = 19)</i>	<i>Narrow Focus</i>	<i>91.7</i>	<i>2.87</i>	<i>20.92**</i>
	<i>Echoic Question</i>	<i>80.5</i>	<i>2.84</i>	<i>10.28**</i>
<u>Auditory-Visual (AV) Matching Task</u>				
<i>Upper Half</i> <i>(df = 10)</i>	<i>Broad Focus</i>	<i>88.9</i>	<i>2.14</i>	<i>18.15**</i>
	<i>Narrow Focus</i>	<i>94.8</i>	<i>1.61</i>	<i>27.79**</i>
	<i>Echoic Question</i>	<i>88.9</i>	<i>2.39</i>	<i>16.25**</i>
<i>Lower Half</i> <i>(df = 16)[#]</i>	<i>Broad Focus</i>	<i>87.4</i>	<i>3.12</i>	<i>14.70**</i>
	<i>Narrow Focus</i>	<i>91.4</i>	<i>2.69</i>	<i>17.71**</i>
	<i>Echoic Question</i>	<i>84.4</i>	<i>2.85</i>	<i>12.34**</i>

[#]For the AV task, three participants were not included in the analysis as they recorded matching accuracy of 0% due to a technical error.

Further analyses were conducted to compare the results from the upper and the lower face conditions. For VV matching performance, a 2×2 mixed repeated measures ANOVA was conducted to determine if task performance (percent correct responses) varied as a function of the visible face area, with prosodic speech condition (narrow focus; echoic question) as the within-subjects factor, and face area (upper vs. lower half) as a between-subjects factor. No significant main effect was found for prosody condition, $F(1,29) = 1.58$, $p = 0.219$, or visible face area, $F(1,29) = 0.08$, $p = 0.787$. However, the interaction was significant, $F(1,29) = 10.67$, $p =$

0.003, $\eta_p^2 = 0.269$, reflecting the pattern that displays of the lower face produced better discrimination of focus whereas phrasing was better discriminated from the upper face displays (consistent with Lansing & McConkie, 1999).

A series of post-hoc between-subjects ANOVAs (with a Bonferroni adjusted α of 0.025 for multiple comparisons) revealed that narrow focused items were discriminated with significantly greater accuracy from the lower face compared to upper face presentation condition, $F(1,29) = 6.75$, $p = 0.015$, $\eta_p^2 = 0.189$, but the difference in discriminating echoic questions was not statistically significant, $F(1,29) = 2.38$, $p = 0.134$.

For AV matching performance, a 2 (upper vs. lower face) \times 3 (broad focus; narrow focus; echoic question) mixed repeated measures ANOVA was conducted, with visible face area as the between-subjects factor, and prosodic condition as a within-subjects factor. The main effect of prosodic speech condition was significant, $F(2,52) = 9.62$, $p < 0.001$, $\eta_p^2 = 0.270$, this difference appears to be driven by participants being better able to discriminate narrow focus renditions across both upper and lower face presentations. The main effect of visible face area, $F(1,26) = 0.97$, $p = 0.333$, and the interaction, $F(2,52) = 0.46$, $p = 0.632$, did not reach statistical significance.

Unlike VV matching results, no significant interaction between face area and prosodic conditions was found. This result could be due to differences in how well prosody was specified by the initial item of a pair. For visual presentation, it seems the lower face provides a better cue to narrow focus, so matching performance is very good for narrow focus items when this information has been clearly presented by the initial item of the lower face VV trials (compared to a less clear specification

in the upper face VV displays). A similar argument applies for the echoic question items (only here, it is the upper face that provides the clearest information to the relevant prosodic condition). This interaction between face area and prosodic condition was not found in the AV trials because the auditory specification of prosodic type is the same regardless of whether it is followed by a lower or upper face item.

In sum, reliable matching of visual speech to other visual or auditory speech tokens based on prosodic differences alone was observed regardless of whether the upper or lower face area was presented. This result is consistent with the proposal that perceivers are able to resolve the type of prosody from any of multiple visual cues, and when a particular cue is not available, the underlying prosody can still be determined from those cues that remain. The current within face region and within talker results provide a baseline measure of performance from which to further examine this proposal by investigating people's ability to match prosodic counterparts across face region and across different talkers.

3.2. Experiment 4: Matching Prosody across Talkers

Experiment 4 examined perceivers' ability to match visual speech tokens within and across modalities when the signals were produced by different talkers (with the same face area shown across talkers). If perceivers can categorise the type of prosody (e.g., narrow or broad focus) from the visual cues regardless of who produced the token, then it is expected that they should be able to successfully perform the VV and AV matching tasks. That is, if task performance is based on matching information at the level of abstract form (category type), then accurate prosody

matching can be achieved despite any individual variation in how the prosodic cues were realized.

3.2.1. Method

3.2.1.1. Participants

Thirty-two undergraduate students ($M_{Age}=22$ years) from UWS participated in the experiment in return for course credit. All were fluent talkers of English, and self-reported normal or corrected-to-normal vision with no history of hearing loss. None of these participants had previously taken part in any other experiment reported in the current chapter.

3.2.1.2. Materials and Procedure

The stimuli used and task procedures were the same as described in Experiment 3; however in this experiment, the paired stimuli for matching consisted of two tokens produced by *different* talkers. Participants were randomly allocated to a visible face area condition (upper or lower face, 16 in each condition), and completed both a within-modal (VV) and cross-modal (AV) 2AFC matching task in counter-balanced order. Figure 3.3 outlines the composition of the experimental tasks used. The initially presented item was the same for both pairs produced by one talker, and was the standard against which the matching judgment was to be made on. The second item within each pair was produced by a different talker. In the VV task, the non-matching item was always a broad focused rendition; in the AV task, the non-matching item was one of the alternate prosodic types. All other material and procedural details are the same as Experiment 3.

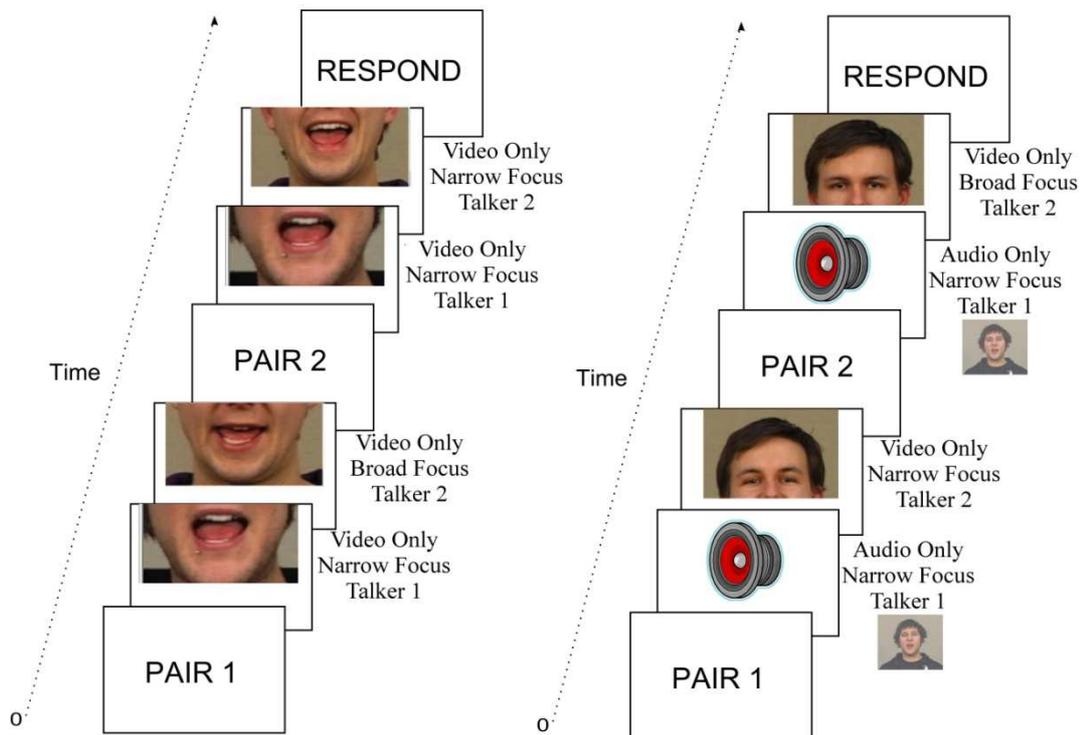


Figure 3.3. Schematic representation of the 2AFC visual-visual (left) and auditory-visual (right) matching tasks used in Experiment 4. The same item appeared first for both pairs produced by one talker, and was the standard from which the matching judgment was to be made on. The second item within pairs was produced by a different talker, with the non-matching item being the same sentence produced with a different prosody. Participants completed the task with either upper or lower face stimuli.

3.2.2. Results and Discussion

The mean percent of correct responses for both VV and AV tasks are shown in Table 3.2. A series of significant one-sample *t*-tests of the results showed that despite the within-pair signals originating from different talkers, participants were able to perform the task at levels well above chance for all conditions.

Chapter 3: Recognising Prosody across Modalities, Face Areas and Talkers

Table 3.2. Mean percent of correct responses in the VV and AV matching tasks as a function of visible face area in each prosodic speech condition, when items within pairs were produced by different talkers. $df = 15$ and ** indicates $p < 0.001$.

Visible Face Area	Prosodic Speech Condition	Mean Correct (%)	Standard Error of Mean	<i>t</i> -test vs. Chance (50%)
<u>Cross-Talker Visual-Visual (VV) Matching</u>				
Upper Half	Narrow Focus	70.9	4.33	4.83**
	Echoic Question	78.4	3.38	8.42**
Lower Half	Narrow Focus	85.9	2.89	12.42**
	Echoic Question	70.0	3.03	6.61**
<u>Cross-Talker Auditory-Visual (AV) Matching</u>				
Upper Half	Broad Focus	79.8	4.85	6.15**
	Narrow Focus	85.9	4.49	8.01**
	Echoic Question	81.4	3.93	8.00**
Lower Half	Broad Focus	81.6	2.39	13.19**
	Narrow Focus	94.2	1.22	36.16**
	Echoic Question	84.8	2.30	15.17**

The results for cross-talker prosody matching (Experiment 4) were compared to the results obtained for within-talker matching (Experiment 1-3). A $2 \times 2 \times 2$ ANOVA was conducted for VV task performance, and a $2 \times 2 \times 3$ ANOVA for AV performance, each with talker congruency (within-; cross-talker) and visible face area (upper face; lower face) as between-subjects factors, and prosodic speech condition as the within-subjects factor. The main effect of talker congruency was significant for the VV task, $F_{VV}(1,59) = 11.00$, $p = 0.002$, $\eta_p^2 = 0.157$, but not for the AV task, $F_{AV}(1,56) = 3.52$, $p = 0.066$. Overall, performance across both tasks was greater when the matching speech tokens were produced by the same talker,

suggesting that although non-talker specific suprasegmental content can be extracted from visible movements, there is also a talker-specific component.

The main effect of visible face area was not significant for either task, $F_{VV}(1,59) = 0.55, p = 0.461, F_{AV}(1,56) = 0.08, p = 0.782$, suggesting that both the upper and lower face provide equally effective prosodic cues. For both tasks, the main effect of prosody was significant, $F_{VV}(1,59) = 5.50, p = 0.022, \eta_p^2 = 0.085; F_{AV}(2,112) = 11.67, p < 0.001, \eta_p^2 = 0.172$, which appears to be driven by narrow focus being easier to visually discriminate than broad focused statements and echoic questions. The prosody by visible face area interaction for the VV task was significant, $F_{VV}(1,59) = 40.15, p < 0.001, \eta_p^2 = 0.405$, however this interaction for the AV task was not significant, $F_{AV}(2,112) = 0.44, p = 0.643$. Likewise, the prosody by talker congruence interactions, $F_{VV}(1,59) = 0.12, p = 0.728; F_{AV}(2,112) = 1.06, p = 0.351$, and the three-way interactions, $F_{VV}(1,59) = 1.32, p = 0.256; F_{AV}(2,112) = 0.81, p = 0.449$, failed to reach significance for either task.

The results showed that despite visual cues originating from two different talkers, perceivers were able to visually match prosodic content when information was restricted to either the upper or lower face. This finding is consistent with previous evidence that visual speech cues can be processed at an abstract level. For example, observing a silently spoken word facilitates lexical decisions on the same word subsequently presented in either written or spoken form (Kim, Davis & Krins, 2004). This priming effect occurs even when auditory and visual speech tokens are produced by different talkers (Buchwald, Winters & Pisoni, 2009). Similarly, McGurk effects (i.e., the integration of incongruent auditory and visual information resulting in a “fused” percept, McGurk & McDonald, 1976) are observed when

auditory and visual signals originate from different talkers (Green, Kuhl, Meltzoff & Stevens, 1991), even when a male face is paired with a female voice. These studies demonstrate that equivalent phonemic information is extracted from the visual speech signal regardless of the talker that produced the signal.

Indeed, in the current task, perceivers performed just as well when matching the auditory prosody of one talker to the visual prosody of another. These results suggest that information for visual prosody, much like phonemic information, is processed in terms of abstract visual speech events, allowing for generalisation across tokens, modalities and talkers. Furthermore, participants can match visual cues to prosody regardless of whether they were from the upper or lower face, supporting the notion that multiple (and potentially redundant) visual cues to prosody are distributed across face areas, and that perceivers must be sensitive to prosodic counterparts across different face regions. This suggestion concerning the flexibility of perceivers' use of visual prosody is further tested in Experiments 5 and 6.

3.3. Experiments 5 and 6: Matching Prosody across Face Areas

In these experiments, perceivers were shown visual-only tokens of the upper face, with the task to match these to the prosody displayed from the lower face (and vice versa). In Experiment 5, the upper and lower face stimuli were of the same talker (but from different tokens), whereas in Experiment 6, the tokens were from different talkers.

3.3.1. Method

3.3.1.1. Participants

Forty undergraduate students from UWS ($M_{Age} = 21.5$ years) participated in the experiment for course credit. All were fluent talkers of English, and had self-reported normal or corrected-to-normal vision and no history of hearing loss. None had taken part in the experiments previously reported. Participants took part in both Experiment 5 and 6 in a counter-balanced order (i.e., 20 completed Experiment 5 first; the remaining 20 completed Experiment 6 first).

3.3.1.2. Stimuli and Procedure

Experiments 5 and 6 used the same materials as outlined in the VV task of Experiment 4. Figure 3.4 provides an overview of the sequence of displays used in the task of each experiment. In Experiment 5, items within-pairs were produced by the same talker, but displayed one half of the face in the first item, then the opposite half of the face in the second video. Experiment 6 was identical to Experiment 5, except that items within-pairs were produced by different talkers.

For each experiment, two different versions of the experimental task were created, so that the upper and lower face stimulus of each item appeared as the target only once across both versions. Participants completed only one version of the task for each experiment ($n = 20$ in each version), and were never exposed to the same face area producing the same sentence by either talker more than once. In total, each version required 40 matching responses across two prosodic speech conditions (i.e., narrow focus and echoic questions), with the broad focused rendition always acting as the non-matching item within pairs. Half of the trials displayed the upper face followed by the lower face within pairs, while the remaining half displayed the lower

face before the upper face. All other material and procedural details are the same as the VV task of Experiment 4.

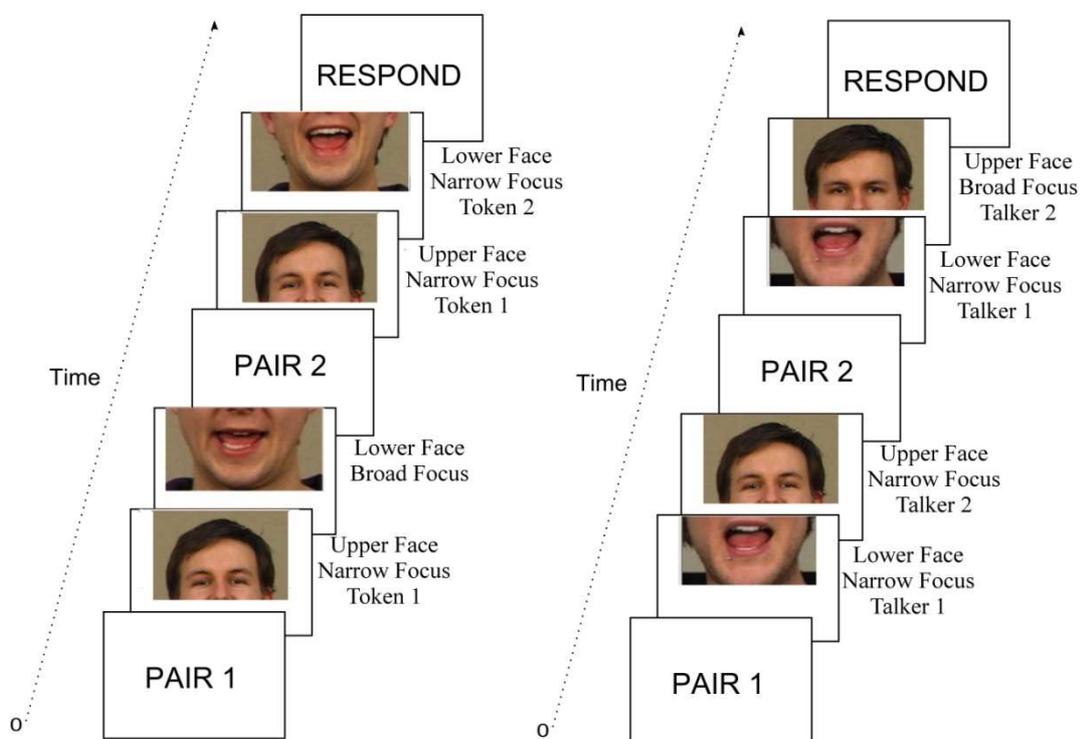


Figure 3.4. Schematic representation of the 2AFC tasks used in Experiment 5 (left) and Experiment 6 (right). In each item either the upper face or lower face stimuli were displayed first, followed by the opposite face area within each pair. In Experiment 5, items within-pairs were produced by the same talker, with the matching video always taken from a different recorded token. In Experiment 6, items within-pairs were produced by different talkers. The non-matching video in both tasks were always the broad focused rendition of the same sentence.

3.3.2. Results and Discussion

Table 3.3 shows the mean percent of correct responses for the 2AFC visual-visual matching task across face areas of the same talker (Experiment 5) or different talkers

(Experiment 6). A series of one-sample *t*-tests indicated that performance was significantly greater than that expected by chance for both experiments.

Table 3.3. Mean percent of correct responses in the 2AFC visual-visual matching task across face areas using within- and cross-talker stimuli, presented as a function of presentation order (upper to lower; lower to upper face), separated by prosodic speech condition. *df* = 39 and ** indicates $p < 0.001$.

Presentation Order	Prosodic Speech Condition	Mean Correct (%)	Standard Error of Mean	<i>t</i>-test vs. Chance (50%)
<u>Within-Talker Stimuli (Experiment 5)</u>				
Upper to Lower	Narrow Focus	78.0	2.56	10.93**
	Echoic Question	86.3	2.42	14.98**
Lower to Upper	Narrow Focus	88.0	2.21	17.17**
	Echoic Question	70.8	2.22	9.35**
<u>Cross-Talker Stimuli (Experiment 6)</u>				
Upper to Lower	Narrow Focus	79.3	3.07	9.54**
	Echoic Question	82.3	2.44	13.21**
Lower to Upper	Narrow Focus	86.5	2.25	16.21**
	Echoic Question	61.0	3.05	3.60**

The results of Experiment 5 and 6 were compared using a $2 \times 2 \times 2$ repeated measures ANOVA. Talker congruency within-pairs (congruent; incongruent), presentation order (upper first; lower first) and prosodic condition (narrow focus; echoic question) were all treated as within-subjects factors. In general, performance was better when items within-pairs were produced by the same talker, however no main effect was observed for talker congruency, $F(1,39) = 1.64$, $p = 0.208$. The main effect of prosody was significant, $F(1,39) = 44.85$, $p < 0.001$, $\eta_p^2 = 0.535$, an effect driven by the participants superior performance in discriminating prosodic focus in

comparison to prosodic phrasing. The main effect of presentation order was also significant, $F(1,39) = 14.76, p < 0.001, \eta_p^2 = 0.275$. As expected, the prosody by presentation order interaction was significant, $F(1,39) = 91.05, p < 0.001, \eta_p^2 = 0.700$, as was the talker congruence by prosody interaction, $F(1,39) = 5.67, p = 0.022, \eta_p^2 = 0.127$. The talker congruence by presentation order interaction, $F(1,39) = 2.84, p = 0.100$ and three-way interaction, $F(1,39) = 0.26, p = 0.616$, did not reach significance.

In the above analyses, it is clear that presentation of the lower face before the upper face resulted in better matching accuracy for focus, whereas the opposite presentation order (i.e., the upper face followed by the lower face) yielded better results for judgments of phrasing, regardless of talker congruency. In explaining this significant interaction between presentation order and prosodic contrast, it is useful to consider the relative effectiveness of prosodic cues from the lower and upper face regions. The results of Experiment 3 and 4 indicated that, compared to the upper face, the lower face provides more effective cues for determining whether a constituent has been focused or not, whereas the upper face appears to provide more effective cues concerning phrasing. Given this, it may be that when the face area containing more robust, salient visual cues was initially presented, matching performance was facilitated because the prosodic type (i.e., category) resolved from the salient cues guides the perceiver as to the type of cues they should seek when viewing the second item within the pair, increasing their sensitivity to subtle, non-salient cues. In contrast, when the initially presented face area included less salient cues, the perceiver (without any guide) might be relatively less sensitive to subsequently presented cues in the second interval. Effects of presentation order

have been observed in a variety of psychophysical studies with it being a common idea that participants employ categorical coding to compare stimuli (Repp & Crowder, 1990).

3.4. General Discussion

Motivated by the apparent mismatch between behavioural studies indicating that perceivers effectively use visual prosodic cues, and the observed variability in the production of such cues across talkers, the experiments presented in this chapter examined perceivers' sensitivity to visual cues to prosody from upper and lower face regions. To account for the effectiveness of visual cues, it was proposed that perceivers are able to resolve the type of prosody from any of multiple visual cues, so that if a particular cue is not available, other cues will still permit the underlying prosody to be determined. To test whether this was the case, a series of experiments examined whether people could match upper and lower face cues as well as auditory to visual cues (not only from the same talker but also from different talkers). The results showed that despite differences in the form and temporal structure of prosodic cues across modalities and face areas, perceivers could reliably match both auditory and visual items to visual tokens of the same prosodic type across talkers regardless of the face area presented, confirming that perceivers can use visual prosody effectively (Lansing & McConkie, 1999). More importantly, the results show that perceivers can match prosody from visual cues provided by the upper face to the lower face (and vice versa) across different talkers. This ability to match very different cues suggests that matching was performed at an abstract level (i.e., the perceiver used the available cues in the visual display to determine the prosodic category and performed matching at this level).

These results, and the interpretation that the ability to perform prosody matching (despite very different visual cues) is based on matching at an abstract level, raises a set of issues regarding the different sources of visual prosodic cues, and precisely what is meant by matching at an abstract level. Given that the description of visual prosody and theories concerning how variable realisations might map onto prosodic categories are still in their infancy, the discussion is developed around two recent theoretical accounts proposed within the auditory domain. The first issue considered concerns the nature of different types of prosodic cues, while the second explores how variable signals might be mapped onto categories (and how these categories are specified).

The outcome that perceivers can reliably match prosody using visual cues based on very different signals indicates that visual prosody may be derived from more than one source. Recently, Watson (2010) has developed a multisource account of acoustic prominence that appears relevant to this notion. This view proposes that prosody (at least in terms of prominence) is the product of a number of different cognitive processes that give rise to different realisations. This proposal allows a distinction to be drawn between sources. For example, Watson suggested that although acoustic changes in intensity, *F0* and duration are all linked in some way to important or focused information, duration may occasionally be a less reliable cue as it is related to talker-centric production processes. That is, it was argued that different acoustic factors will influence prominence only in as much as they mark relevant information for the listener.

It is this latter suggestion that appears relevant to the current results, as it seems that the perception of what is relevant information determines the extent to

which a visual cue was influential. More specifically, consider the interaction between face region and prosody type in visual matching performance of Experiment 3. The level of AV matching showed that both the lower and upper face stimuli could be matched to narrow focus equally well (both above 90% correct). However, this was only the case when the participant knew to look for narrow focus (when this was indicated by the auditory signal). That is, when presented with visual cues from the upper face (in the VV task), performance was worse (around 80%) whereas performance remained above 90% when visual cues from the lower face were presented. This may be because the lower face provides distinct cues related to prosodic focus (e.g., cues for change in amplitude or duration, Kochanski et al., 2005) whereas the upper face provides more reliable visual cues (head and eyebrow movements) to distinguish echoic questions from statements (as these cues are linked with variations in F_0 , Yehia et al., 1998; Cavé et al., 1996, an acoustic feature that can differentiate statements from echoic questions, see Eady & Cooper, 1986).

Having considered that there are multiple visual cues to prosody and that the perception of what is relevant information may determine how a visual cue is weighted, the related issue of how such diverse source cues might be mapped to prosodic categories is now considered. Once again, the discussion is based on ideas derived from studies of auditory speech processing as these have a long history of development and refinement. It should also be noted that the models considered have been proposed with regard to speech recognition (i.e., distinguishing phonemes) and not *prosodic* recognition. Given this, the discussion will focus on the setting out of alternative models and whether they are suited for describing the current results rather than attempting to specify particular mechanisms.

Chapter 3: Recognising Prosody across Modalities, Face Areas and Talkers

The first approach to the issue of variability across sources is to regard it as a problem that needs to be overcome. Under such an “invariance” approach, sources of variability are removed by processes of normalization or compensation (with the latter being a more general term typically including processes that deal with coarticulation). The basic assumption of this approach is that a few invariant underlying signal properties can be revealed by recoding the signal by grouping overlapping cues or gestures, or by exploring mechanisms of contrast (using either other signal events, or long-term expectations about cues). However, a problem with applying this approach to visual prosody is that it is not clear that visual prosodic cues can be defined in terms of a small number of invariant properties; not only is the relationship between auditory and visual signals variable (in terms of how and when they occur, if at all), but visual cues also vary with respect to each other. That is, the visible movements of the upper and lower face do not occur simultaneously or systematically (e.g., see Cavé et al., 1996; McClave, 1998).

The exemplar approach provides an alternate scheme for the mapping of variable form onto categories. Here, it is assumed that the input is encoded in detail by using all available cues, with context dependency overcome because perceivers store multiple exemplars and make categorization decisions by comparing the incoming input to these collections of stored exemplars. Such an approach provides a natural way of dealing with the effects of context and talker variability (without compensation *per se*), but it is unclear how it can deal with completely novel input, such as having to match cues derived from one face region to another.

Intermediate between the above two approaches are models referred to as cue-integration approaches (McMurray & Jongman, 2011). These models propose

that if a sufficient number of cues are encoded, then in combination, the variability of any one cue can be overcome (possibly without the need for compensation). Examples of this type of model include the fuzzy-logic model of perception (FLMP; Oden & Massaro, 1978) and TRACE (Elman & McClelland, 1986), with the former having been used to successfully model how prosodic cues (duration and pitch) are integrated to influence syntactic identification (Beach, 1991). The most recent model of this type, called the “computing cues relative to expectations” (C-CuRE) model proposed by McMurray and colleagues, combines aspects of the invariance approach (compensation) with aspects of exemplar approaches (retaining every cue, e.g., McMurray & Jongman, 2011; McMurray, Cole & Munson, 2011). That is, like exemplar models, C-CuRE maintains a continuous representation of cue values that include variations due to the talker, context and coarticulation. However, unlike exemplar models, it uses this variation to build categories relating to such variables as context and talker and these in turn are used to interpret the informational content of the speech signal. Such a model offers a more principled way of taking into account instance-based variation by partitioning out of sources of variance prior to cue-integration. It is this feature of C-CuRE that seems an attractive framework for the integration of auditory and visual cues to prosody while taking into account their variability.

CHAPTER 4.
RECORDING OF AN AUDITORY-VISUAL SPEECH
PROSODY CORPUS

Chapter 4. Recording of an Auditory-Visual Speech Prosody Corpus

In Chapters 2 and 3, it was established that prosody could be perceived from the visible movements of the talker's head and face. A follow-up issue concerns the quantification of these movements. That is, what are the specific visual correlates of prosodic contrasts? There have been several studies (e.g., Beskow, Granström & House, 2006; Dohen et al., 2006, 2009; Graf et al., 2002; Munhall et al., 2004; Scarborough et al., 2009; Srinivasan & Massaro, 2003; Swerts & Kraemer, 2010) that have tried to quantitatively determine the visual correlates of prosody. However, these studies have been limited in several ways. For example, the size of the recorded corpus has typically been small (making generalization of the results potentially unreliable). The analysis has often been based on single token productions examining only the initially stressed syllable (rather than examining properties of the entire prosodically marked constituent and its sentential context), or has been based on perceiver-driven annotations of video data. Moreover, local movement features rather than whole head and face movements are measured, missing the potential relationships between gestures occurring across face areas.

Given the above, the current study examined visual prosodic cues by measuring the overall head and face movements produced by six talkers for 30 different sentences across a range of prosodic contrasts elicited using a dialogue exchange task. Additionally, the setting in which the interactions took place was experimentally manipulated so that the talker could both see and hear the interlocutor in one condition (i.e., face-to-face; FTF); or only hear the interlocutor (i.e., auditory-

only; AO) while engaging in the dialogue exchange task. This variable was included as it has been proposed that talkers modify both their auditory and their visual speech as a function of whether they can see or only hear each other (see Fitzpatrick, Kim & Davis, 2011; Garnier, Henrich & Dubois, 2010).

In the current chapter, the recording of the multi-talker auditory-visual speech prosody corpus is outlined. The data from this corpus will allow for a more detailed analysis of the spatiotemporal properties of linguistic auditory and visual spoken prosody, including the type of movements and acoustic features used, the distribution of gestures across the face and their temporal characteristics, and the nature of the relationship between the auditory and visual signals across the two interactive settings over the six talkers (these analyses are detailed in Chapters 5, 7 and 8).

4.1. Method

4.1.1. Equipment

To synchronously record both auditory and visual speech data, an OPTOTRAK 3020 (Northern Digital Inc.) was utilised, a system capable of capturing the three-dimensional position of infrared (IR) light emitting diode (LED) markers at higher sampling rates than a typical digital video camera along with temporally synchronised analogue data at a different sampling rate. The system is composed of an infrared camera unit (Figure 4.1, left), a collection of wired IR-LED markers (Figure 4.1, right) connected to strober units (that generate strobing patterns in the IR-LED markers), a system control unit, an OPTOTRAK data acquisition unit (ODAU II; for acoustic recording) and a control PC with an OPTOTRAK communication board. The OPTOTRAK system is an active one in which the

markers repetitively strobe IR light in a cyclical pattern allowing for their individual positions to be determined by the system. The location of the markers over time and space is recorded by the NDI First Principles software package. The three-dimensional resolution of the tracked movements from such a system is highly accurate, within the range of 0.005 to 0.01mm (Kroos, Kuratate & Vatikiotis-Bateson, 2002; Maletssky, Sun & Morton, 2007; Schmidt, Berg, Ploeg & Ploeg, 2009; States & Pappas, 2006).



Figure 4.1. The OPTOTRAK infrared camera unit (left) tracks the three-dimensional position over time of infrared emitting markers, measuring 7mm in diameter (right) with high spatial and temporal resolutions.

Although active marker systems such as OPTORAK require the talker to be “wired” to the system, there are corresponding benefits: minimum amounts of pre-calibration and data post-processing are necessary. As a physical connection exists between the strobing markers and the camera unit, any markers that drop out (i.e., are no longer visible by at least two of the three cameras due to orientation or occlusion) are easily recovered once they re-emerge inside the camera’s field of vision. In comparison, passive marker systems such as QUALISYS (Qualisys

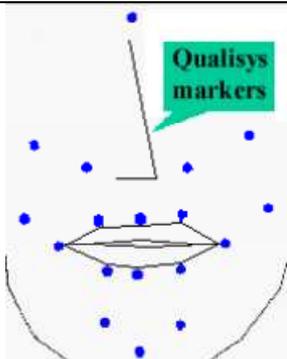
Medical) and VICON (Vicon Motion Systems) require no physical connection with the subject, relying instead on the detection of IR light reflection from non-wired reflective markers by multiple IR cameras. However, these systems require substantially more calibration before data can be obtained (e.g., the position of every IR camera relative to each other needs to be determined), are prone to “phantom” markers caused by IR reflections off environmental objects, and involve a more arduous post-processing procedure (e.g., if a marker drops out or is occluded during a recording trial, its reappearance in the visual scene is treated as a “new” marker). As both the active and passive marker systems are capable of generating data of equivalent sampling rates (~100 Hz), the use of the active marker system was chosen on the basis of the amount of processing required.

4.1.2. Marker Configuration

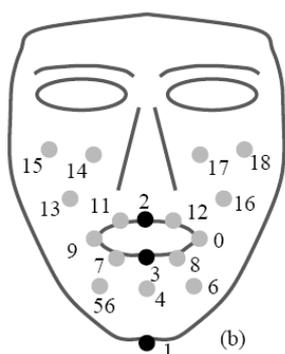
Previous studies utilising optical tracking for the purpose of auditory-visual speech analysis and synthesis have typically used between 18 and 38 markers distributed across the talker’s face to measure temporary non-rigid deformations, as well as rigid rotations and translations of the whole head (e.g., Jiang, Alwan, Bernstein, Keating, & Auer, 2000). These marker configurations (summarised in Table 4.1 with a brief description of the research focus of these studies) appear adequate to capture prosody-related head and face movements; for example, Jiang, Alwan, Auer and Bernstein (2001) showed that there was a strong correlation (of around 0.8) between visual confusions (measured by lip-reading of silent videos of consonant vowel syllables) and facial movement measurements when only 17 optical markers were utilised.

For the current corpus, marker locations were chosen to reflect a combination of articulatory gestures from the lips, jaw, chin and cheeks, non-articulatory gestures such as eyebrow movements, along with the rigid movements (i.e., rotations and translations around the centre of rotation) of the whole head. The locations chosen were akin to those used by Lucero, Maciel, Johns and Munhall (2005), with the exception of markers on the eyelids and lip surface. (Lucero and colleagues used a passive marker system, so it was possible to place smaller markers in more awkward positions such the nostrils, eyelids or lip surface as no wires attached to the markers. In contrast, the OPTOTRAK markers are slightly larger and wired, making them impractical to place on some facial surfaces). In total, 39 markers⁴ were placed on the talker's head and face (four of which were attached to a head rig and used to determine rigid movements), the positions of which are detailed in Table 4.2.

Table 4.1. Summary of marker configurations used in previous studies utilising optical tracking to measure visual speech.

Marker Configuration	Research Focus	Marker <i>N</i>
 <p data-bbox="432 1787 655 1816">Jiang et al., 2000</p>	<p data-bbox="810 1406 1254 1664">Relationship between articulatory movements, produced acoustics and tongue movements (used in combination with electromagnetic articulography; EMA).</p>	<p data-bbox="1326 1406 1361 1440">18</p>

⁴ The total number of markers used was also constrained by the equipment available at the time of conducting the motion capture sessions. The set-up contained two serially connected strober units (1× 16 Channel and 1× 24 Channel), allowing for a total of 40 markers to be tracked. However, one of the channels was faulty, leaving only 39 operational channels/markers.



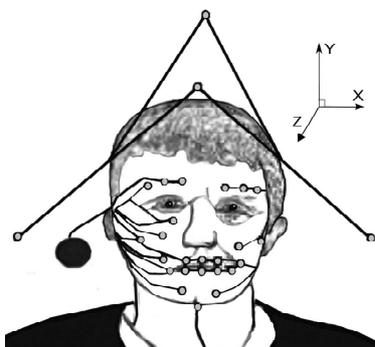
Vatikiotis-Bateson & Yehia, 2000

Modelling the relationship between articulatory movements and produced acoustics. 18



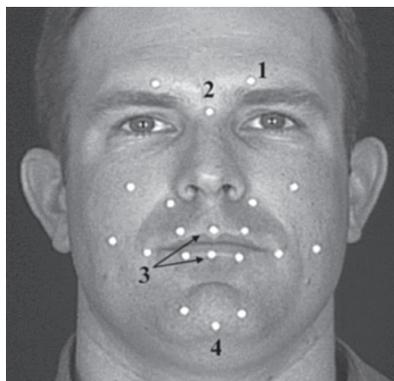
Engwall & Beskow, 2003

Optical tracking used in conjunction with EMA to resynthesise tongue movements from face motion data. 28



Kim, Sironic & Davis, 2011

Characterising the visual properties of speech produced in noise that functions to increase speech intelligibility. 28



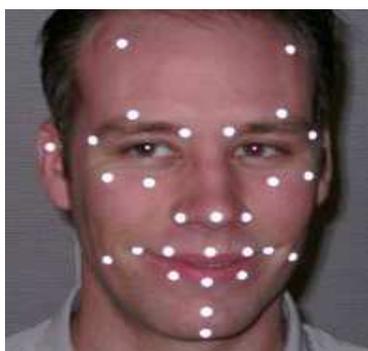
Scarborough et al., 2009

Determining the optical correlates of lexical and phrasal stress for three American-English talkers. 22



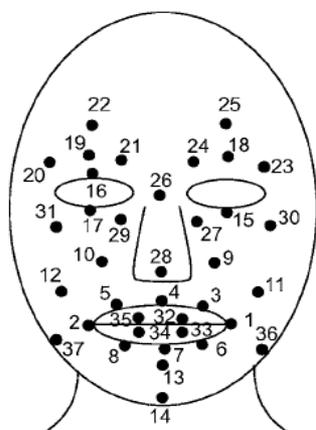
Dohen et al., 2009

Identification of the visual correlates of producing narrow focus, and the identification of idiosyncratic talker strategies. 28



Granström & House, 2005

Driving animations of embodied conversational agents (ECAs) to examine the various expressive functions of visual prosody (e.g., emotion, focus). 30



Lucero et al., 2005

Modelling of human face kinematics using clustering (i.e., estimating the movement of secondary marker positions based on a weighted relationship to primary markers). 37

Table 4.2. The location of the 39 IR emitting OPTOTRAK markers on the head and face of the talkers. Due to high rates of marker occlusions and dropouts, the larynx marker was not included in the analysis.

Marker Number	Marker Placement	Marker Number	Marker Placement	Marker Number	Marker Placement	Marker Number	Marker Placement
1	Right Rigid Body	11	Left Inner Orbital	21	Right Mid Cheek	31	Right Lower Lip
2	Centre Top Rigid Body	12	Left Mid Cheek	22	Right Puffer	32	Right Forehead
3	Left Rigid Body	13	Left Puffer	23	Right Outer Chin	33	Left Forehead
4	Centre Bottom Rigid Body	14	Left Outer Chin	24	Right Lip Corner	34	Left Outer Cheek
5	Left Outer Brow	15	Middle Lower Chin	25	Right Upper Lip	35	Left Lower Cheek
6	Left Mid Brow	16	Right Outer Brow	26	Middle Upper Lip	36	Right Lower Cheek
7	Left Inner Brow	17	Right Mid Brow	27	Left Upper Lip	37	Right Outer Cheek
8	Nose Bridge	18	Right Inner Brow	28	Left Lip Corner	38	Middle Upper Chin
9	Nose Tip	19	Right Outer Orbital	29	Left Lower Lip	39	<i>Larynx</i>
10	Left Outer Orbital	20	Right Inner Orbital	30	Middle Lower Lip		

Of the 39 markers, one was placed on the approximated location of the thyroid cartilage (which surrounds the larynx) slightly below the laryngeal prominence (i.e., “Adams Apple”). This position was chosen based on the proposal of Honda, Hirai, Masaki and Shimada (1999) that fundamental frequency modulation and control may be a product of vertical larynx movements. As such, the movements recorded may share a strong relationship with the produced acoustic signal, or may provide a visual signal that perceivers are able to exploit to assist with the interpretation of prosodic content. However, the markers position was occluded for a substantial proportion of the recording sessions (e.g., by the opening and closing gestures of the jaw), and as a result, was not included in the analysis.

Markers were attached to the talker’s face using double-sided medical tape. To ensure that markers were placed in the same location across talkers, a polystyrene foam display head with drawing pins in the desired marker locations (Figure 4.2) was used as a visual guide when attaching the markers. The positions of the markers in three-dimensional space were sampled at 60Hz (i.e., every 17ms). Videos of the tracking sessions were also recorded using a Sony TRV19E digital video camera (25 fps). Auditory data was synchronously captured using a Behringer C-2 condenser microphone connected to the ODAU II through a Eurorack MX602A mixer. The microphone was placed approximately 30cm away from the talker’s mouth, held in place by a boom-arm microphone stand. Auditory data was sampled at 44.1 kHz, digitized mono.



Figure 4.2. A polystyrene foam head was used as a visual guide to ensure that placement of the optical markers was consistent across talkers.

4.1.3. Materials

The corpus consisted of three randomly selected lists of ten non-expressive sentences drawn from the IEEE (1969) Harvard sentence list, describing mundane events with minimal emotive content (Appendix A). The sentences ranged in length between six and twelve words ($M = 8.33$, $SD = 1.47$), and contained between seven and eleven syllables ($M = 9.50$, $SD = 1.50$).

Each sentence was recorded across three prosodic conditions: as a *broad focused statement*, a *narrow focused statement*, and as an *echoic question*. To elicit these conditions, a dialogue exchange task was used requiring talkers to interact with an interlocutor and either repeat what they heard the interlocutor say (broad focus), make a correction to an error made by the interlocutor (narrow focus), or question an emphasised item within the sentence that the interlocutor produced (echoic question). The “critical” word (i.e., the word within the sentence that was erroneously produced or produced with emphasis by the interlocutor and

subsequently focused or questioned by the talker) was selected before recording and was kept consistent across talkers, prosodic conditions and repetitions, allowing for comparisons to be made across these factors. The location of the critical word within each sentence varied, but never appeared in phrase-final position, and was always a content word. The interlocutor was always a male confederate who was aware of the purpose of the data recording.

4.1.4. Participants

Six university educated male native talkers of standard Australian English ($M_{Age} = 23.2$ years) participated in the data capture sessions, recruited via convenience sampling. All talkers self-reported normal vision and hearing, with no known communicative deficits. Participants were financially compensated for their time, and treated in accordance with the ethical protocols outlined by the UWS Human Research Ethics Committee.

4.1.5. Procedure

All recordings were conducted in the Face and Voice Lab of MARCS Auditory Laboratories. Recording sessions began by placing the movement sensors on the face of the talker in the configuration shown in Figure 4.3. Each talker was recorded individually while seated in a height-adjustable dentist's chair within a double-walled, sound insulated recording booth. Participants were recorded producing the prosodic contrasts with the interlocutor across two interactive settings (outlined below). Two repetitions of each sentence were recorded in each of the three prosodic conditions. Motion capture sessions lasted approximately 180 minutes resulting in 360 recorded tokens ($30 \text{ sentences} \times 2 \text{ repetitions} \times 3 \text{ prosodic conditions} \times 2 \text{ interactive settings}$) per talker.

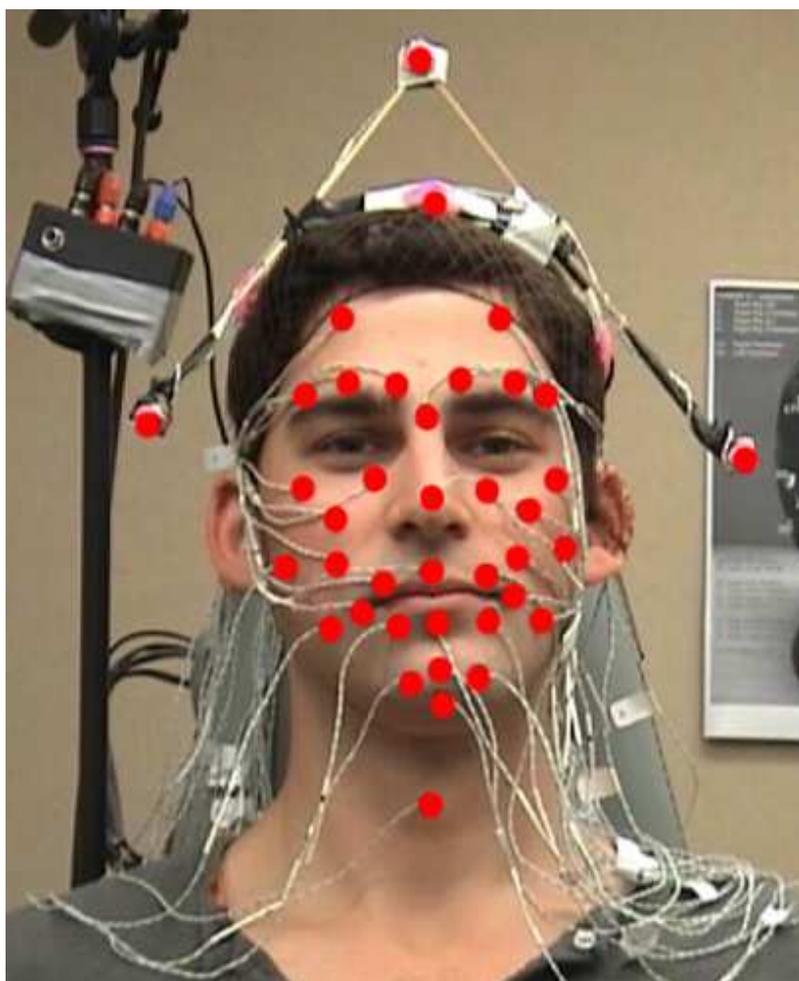


Figure 4.3. Location of the 39 optical markers (with size exaggerated for clarity) on the head and face of the talker, reflecting articulatory and non-articulatory gestures. Four markers were placed on a head rig and used to estimate rigid movements around the centre of rotation.

4.1.6. Interactive Settings

4.1.6.1. Face-to-Face (FTF)

In the FTF setting, the talker and the interlocutor were facing each other, and were able to both see and hear each other. The talkers were instructed to direct their speech towards the interlocutor who was located approximately 2.5 meters away from them (standing behind the OPTOTRAK unit, see Figure 4.4).

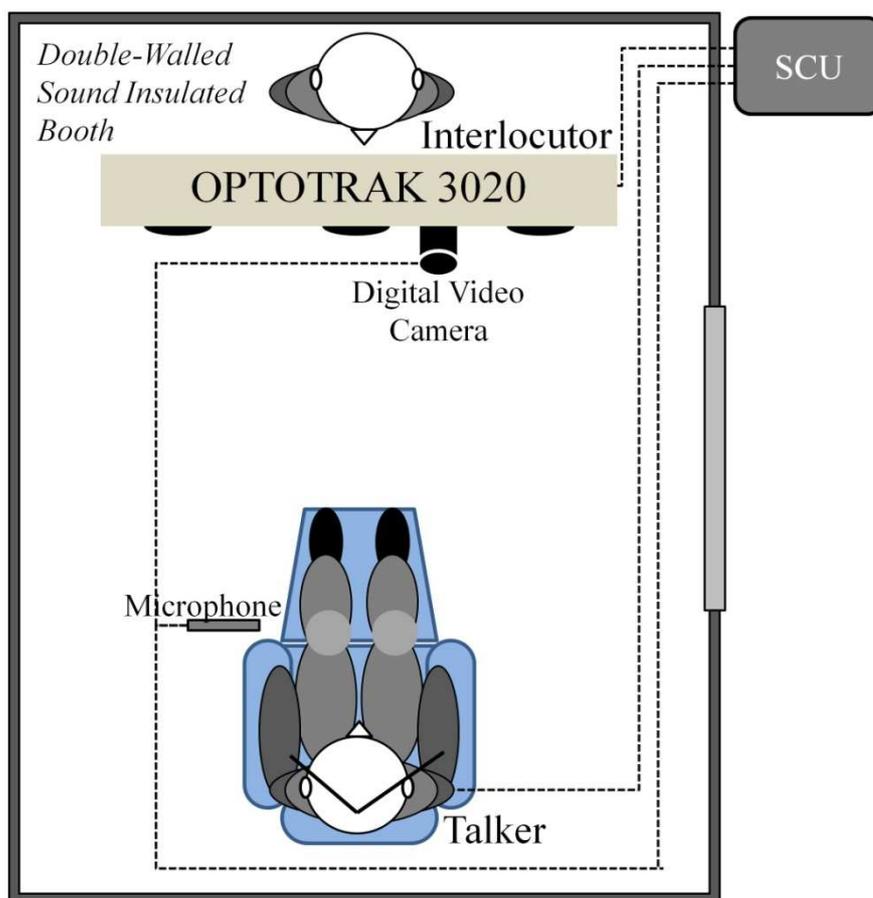


Figure 4.4. The experimental setup used in the face-to-face (FTF) interactive setting. The talker and interlocutor communicated over a distance of approximately 2.5 meters, and were able to both see and hear each other clearly. To minimise extraneous noise, the OPTOTRAK SCU was located outside of the testing booth.

4.1.6.2. Auditory-Only (AO)

In the AO interactive setting, the talker conversed with the interlocutor (who was located outside of the testing booth, Figure 4.5) over a double microphone and headphone system (Figure 4.6). An Edirol UA-25 USB Audio Capture Device was used in streaming mode to mix the auditory signals. Two Behringer C-2 condenser microphones were used as inputs for the left (talker) and right (interlocutor) channels. The opposing output channel was played through to the individual (i.e., the

talker received output from the right channel, and vice versa). Thus, the auditory input from the talker was heard by the interlocutor, and vice versa. The auditory signal was played to the talker through Skullcandy Riot in-ear stereo headphones, and to the interlocutor through Senheiser HD650 stereo headphones. In this interactive condition, the talker and interlocutor could still hear each other clearly, but were not visible to each other. Talkers were given no explicit instructions as to how they should communicate.

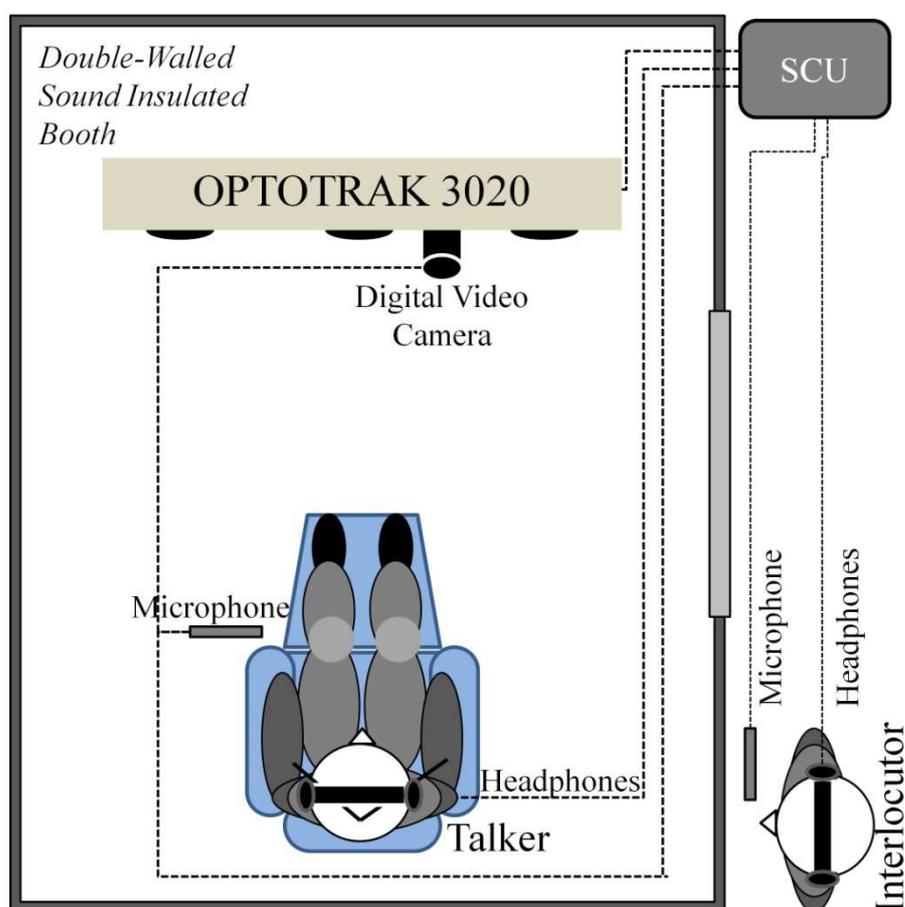


Figure 4.5. The experimental setup used in the auditory-only (AO) interactive setting. The talker and interlocutor communicated over a double microphone and headphone system and were only able to hear each other.

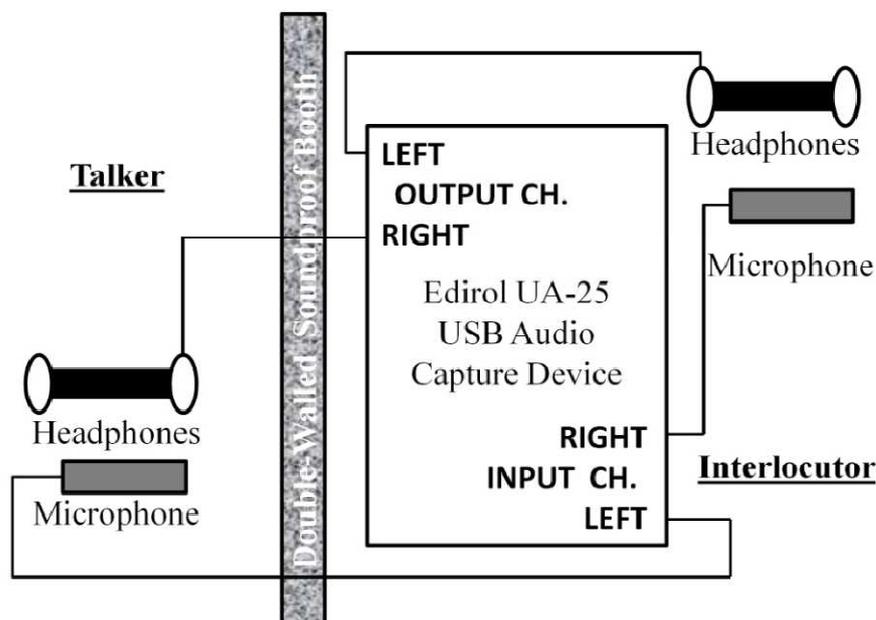


Figure 4.6. Diagrammatic representation of the double microphone and headphone system used in the AO interactive condition.

4.2. Preliminary Data Processing

Captured auditory data were subjected to a semi-automatic forced phonemic alignment using the MARY text-to-speech engine (Schröder & Trouvain, 2003), before manual alignment correction in Praat (Boersma, 2001) where necessary. These transcriptions were used to temporally locate the segmental boundaries of individual constituents within each utterance, and identify the location of the prosodically marked critical constituent.

4.3. Summary

In total, the recorded corpus consists of 2160 auditory-visual tokens, comprising of optically tracked movement data and corresponding phonemically transcribed auditory data. A copy of the corpus is included as Appendix B.

CHAPTER 5.
AUDITORY ANALYSIS OF THE SPEECH PROSODY
CORPUS

Chapter 5. Auditory Analysis of the Speech Prosody Corpus

Chapter 4 detailed the recording of an audiovisual speech prosody corpus that included focus and phrasing contrasts produced across two interactive settings by six talkers. In the current chapter, the recorded utterances were analysed by measuring selected acoustic properties across the elicited prosodic contrasts. The acoustic characteristics of both prosodic focus and phrasing have been extensively studied and are well described in the literature (i.e., in terms of duration, $F0$, $F0$ range, intensity, intensity range and vowel space properties). These properties were thus examined for the current corpus to ascertain whether the produced tokens showed the typical characteristics of these prosodic contrasts.

Further analyses were also conducted on the above acoustic measures to determine whether the production of prosodic contrasts differed across interactive settings (i.e., FTF and AO). The rationale for this contrast is developed in detail in Chapter 6 that perceptually evaluated the focus and phrasing contrasts. In essence, the rationale for investigating FTF versus AO settings stems from an extension of Limblom's Hyper-Hypospeech (H-H) theory (1990, 1996) as applied to the production of prosody. Finally, the acoustic measures were compared across individual talkers in order to address the issue of talker variation in the realisation of these specific prosodic contrasts. Even within homogenous samples, talkers can differ in the way that they exploit particular suprasegmental features to prosodically mark a constituent within an utterance (Eady & Cooper, 1986; Lieberman, 1960). For example, Peppé, Maxim and Wells (2000) found that the production of digit

strings containing a narrowly focused constituent were predominantly produced by talkers with a falling *F0* contour on post-focal syllables, however these markings were variably accompanied by either an increase in intensity on the focused constituent, an insertion of a pre-focal silence, durational manipulation throughout the utterance, or the use of no additional features. Given that there is a potential for variation across talkers, so too may there be differences in how effective these strategies are for conveying linguistic content to perceivers. To explore these questions, it is necessary first to examine if there are indeed signal-level differences across talkers.

In sum, the current chapter aimed to quantify the acoustic and spectral characteristics of the produced prosodic contrasts, and determined whether the production of these contrasts changed as a function of whether or not the talker could see their conversational partner. In addition, the degree of consistency across talkers in the realisation of prosodic contrasts was determined.

To examine the above questions, the 30 sentences were grouped according to their overall length (reflected by number of syllables in the sentence), as this factor can impact on the suprasegmental features used to realise prosodic contrasts. For example, long sentences (i.e., 10 syllables) tend to be produced with anticipatory reductions to pre-focal syllable duration when the critical constituent occurs towards the end of the utterance, however no differences are found for short utterances (i.e., 6 syllables; Pell, 2001). In contrast, utterance length impacts little on *F0* properties for focus contrasts, however short echoic questions are produced with a higher terminal *F0* than longer echoic questions.

Similarly, the location of the critical constituent within the utterance can affect prosodic marking. In terms of duration, prosodically marked words occurring in sentence-initial or sentence-medial positions have been shown to receive a greater syllable lengthening than when in sentence-final position (e.g., 40% increase compared to only a 15% increase respectively, relative to broad focused renditions, Cooper, Eady & Mueller, 1985). This pattern in lengthening was apparent for both narrow focus and echoic question renditions (Eady & Cooper, 1986). The location of the critical constituent also affects *F0* contours. For narrow focus, a post-focal reduction has been seen when focused items occur sentence-initially or medially. In contrast, the *F0* marking of focused words at the end of an utterance appears much less pronounced (Cooper et al., 1985). Furthermore, the differences in *F0* between statements and echoic questions are far less pronounced when the critical word occurs late in the utterance (Pell, 2001). Thus, utterances of the corpus were classified on the basis of both utterance length and location of the critical constituent within the utterance.

5.1. Data Preparation

5.1.1. Classification of Utterance Types

Given that the mean number of syllables in the utterances of the speech prosody corpus was 9.50, a sentence was classified as being “short” if it had fewer than ten syllables, and “long” if it contained ten or more syllables. If the critical constituent occurred in the first half of the utterance, the location was classified as “early”, whereas utterances containing the critical constituent in the second half of the utterance were deemed as being “late”. Thus, the 30 sentences were allocated into

one of four possible classifications⁵: short sentences with an early critical constituent (S/E; $n = 10$), short sentences with a late occurring critical constituent (S/L; $n = 5$), long sentences with an early occurring critical content (L/E; $n = 7$), and long sentences with a late critical item (L/L; $n = 8$). The classification of each sentence is provided in Appendix A.

5.1.2. Identification of Utterance Phases

As detailed in Chapter 4, each of the 2160 sentence tokens was subjected to semi-automatic phonemic transcription with manual alignment correction. These transcriptions were used to locate the prosodically marked constituent (i.e., critical word) within each utterance. Any content that occurred before the critical word was labelled as “pre-critical”, and any content that followed the critical word as “post-critical”. The auditory features were extracted for each of these utterance phases.

5.1.3. Acoustic Feature Extraction

All acoustic features were determined in Praat (Boersma, 2001) using custom-created scripts. The duration of each utterance phase was extracted to derive the mean syllable duration (calculated by dividing the phase duration in milliseconds by the number of syllables in the phase, see Appendix A), which can be used as a relative index of speech rate (Dahan & Bernard, 1996). To determine the F_0 characteristics of the utterances, the F_0 contour was initially extracted for the entire utterance at 1ms time steps, with a pitch floor of 65Hz and ceiling of 300Hz. Octave jumps were removed from the resulting contour, and interpolated over voiceless content before applying a 10Hz smoothing filter. The mean values for each utterance

⁵ Given that the decision to classify the utterances on the basis of length and location of the critical constituent was not an experimental manipulation and were made after the corpus had already been recorded, the group sizes are somewhat uneven.

phase was then determined. Similarly, the F_0 range was calculated by determining the difference between the minimum and maximum F_0 within each utterance phase. Finally, mean relative intensity was calculated for each utterance phase, with the intensity range determined by calculating the difference between the minimum and maximum intensity values for each utterance phase.

5.1.4. Data Normalisation

To allow for comparisons across different sentences, the obtained values were subjected to a normalisation procedure, with the broad focused rendition considered as the baseline version for both narrow focus and echoic question renditions. For each interaction setting, the mean value of the broad focus renditions (per talker and sentence) was calculated, with the remaining prosodic renditions divided by these mean values. Thus, a value of 1 after normalisation corresponds to no variation relative to the broad focus rendition (within the interactive setting), a value below 1 represents a decrease on the measured parameter, and a value greater than 1 indicates an increase on the parameter.

5.2. Acoustic Analysis

5.2.1. Realisation of Prosodic Contrasts

The acoustic properties expressed as proportions of the broad focused renditions in each of the interactive conditions are displayed in Figure 5.1. For each prosodic contrast (i.e., focus and phrasing) recorded in the AO interactive condition, the data for the extracted acoustic features (at each utterance phase) were analysed using two analyses of variance, one for the sentence data (item analyses; F_I , collapsed across talkers and repetitions), and one for talker data (subject analyses; F_S , collapsed

across sentences and repetitions). For the sentence analyses, 4×2 mixed repeated measures ANOVAs were conducted, with utterance type (S/E; S/L; L/E; L/L) as a between-items factor, and prosodic condition (broad focus AO; narrow focus AO / broad focus AO; echoic question AO) as the repeated within-items measure. For the talker analyses, repeated measures ANOVAs were used, with prosodic condition as the within-subjects factor.

The purpose of these analyses was to confirm that the auditory tokens recorded in the AO condition of the corpus conformed to the typical acoustic properties descriptive of prosodic focus and phrasing contrasts (i.e., that the experimental manipulation and dialogue exchange task were effective in eliciting prosodic contrasts). Given this objective and to streamline reporting, the current section highlights only the significant main effects of prosody (with full statistical tables included as Appendix C.1 for both sentence and talker data).

5.2.1.1. *Realisation of Focus Contrasts*

As expected, the mean syllable durations of narrowly focused constituents, $F_1(1,26) = 456.04, p < 0.001, \eta_p^2 = 0.946$; $F_S(1,5) = 45.29, p = 0.001, \eta_p^2 = 0.901$, and post-focal content, $F_1(1,26) = 46.69, p < 0.001, \eta_p^2 = 0.642$; $F_S(1,5) = 6.68, p = 0.049, \eta_p^2 = .572$, were significantly longer in narrow than broad focused renditions.

The mean F_0 of post-critical content was significantly lower in narrow than broad focused renditions in the sentence analysis, $F_1(1,26) = 9.88, p = 0.004, \eta_p^2 = 0.275$, but failed to reach significance in the talker analysis, $F_S(1,5) = 3.94, p = 0.104, \eta_p^2 = 0.441$. The fundamental frequency range covered during the critical constituent was significantly greater for narrow focused than broad focused renditions, $F_1(1,26) = 109.03, p < 0.001, \eta_p^2 = 0.807$; $F_S(1,5) = 22.22, p = 0.005, \eta_p^2 =$

= 0.816. This pattern was also mirrored in the sentence analysis for post-focal content, $F_{I}(1,26) = 35.56, p < 0.001, \eta_p^2 = 0.578$, but was not found in the talker analysis, $F_S(1,5) = 4.76, p = 0.081, \eta_p^2 = 0.487$.

Mean relative intensity was significantly lower in narrow focused renditions (relative to broad focused ones) for both pre-critical, $F_{I}(1,26) = 10.16, p = 0.004, \eta_p^2 = 0.281$; $F_S(1,5) = 7.67, p = 0.039, \eta_p^2 = 0.605$, and post-critical utterance content, $F_{I}(1,26) = 261.95, p < 0.001, \eta_p^2 = 0.910$; $F_S(1,5) = 34.21, p = 0.002, \eta_p^2 = 0.872$, relative to broad focused productions. Although the mean relative intensity was not higher, the intensity range of the critical constituent was significantly greater for narrow focused renditions, $F_{I}(1,26) = 52.70, p < 0.001, \eta_p^2 = 0.670$; $F_S(1,5) = 67.17, p < 0.001, \eta_p^2 = 0.931$.

5.2.1.2. Realisation of Phrasing Contrasts

Relative to content produced as statements, the echoically questioned critical constituents were produced with an expectedly greater mean syllable duration, $F_{I}(1,26) = 279.09, p < 0.001, \eta_p^2 = 0.915$; $F_S(1,5) = 77.60, p < 0.001, \eta_p^2 = 0.939$.

The mean F_0 of the critical constituent was higher for questions, however this difference only occurred in the sentence analysis, $F_{I}(1,26) = 61.22, p < 0.001, \eta_p^2 = 0.702$; $F_S(1,5) = 4.31, p = 0.093, \eta_p^2 = 0.463$. The post-critical content was produced with a significantly higher mean F_0 in echoic questions than for statements, $F_{I}(1,26) = 348.32, p < 0.001, \eta_p^2 = 0.931$; $F_S(1,5) = 32.43, p = 0.002, \eta_p^2 = 0.866$. The F_0 range covered was also significantly greater in the echoic question renditions than in the matched statements for both critical, $F_{I}(1,26) = 104.98, p < 0.001, \eta_p^2 = 0.801$; $F_S(1,5) = 34.35, p = 0.003, \eta_p^2 = 0.862$, and post-critical content, $F_{I}(1,26) = 89.66, p < 0.001, \eta_p^2 = 0.775$; $F_S(1,5) = 14.54, p = 0.012, \eta_p^2 = 0.744$.

The post-critical content in echoic questions were also produced with significantly greater mean intensity than statements, $F_1(1,26) = 388.82, p < 0.001, \eta_p^2 = 0.937$; $F_S(1,5) = 22.68, p = 0.005, \eta_p^2 = 0.819$. The intensity range covered was also significantly greater for echoic question renditions for both critical constituents, $F_1(1,26) = 33.04, p < 0.001, \eta_p^2 = 0.560$; $F_S(1,5) = 23.35, p < 0.001, \eta_p^2 = 0.979$, and post-critical content, $F_1(1,26) = 93.92, p < 0.001, \eta_p^2 = 0.783$; $F_S(1,5) = 18.88, p = 0.007, \eta_p^2 = 0.791$.

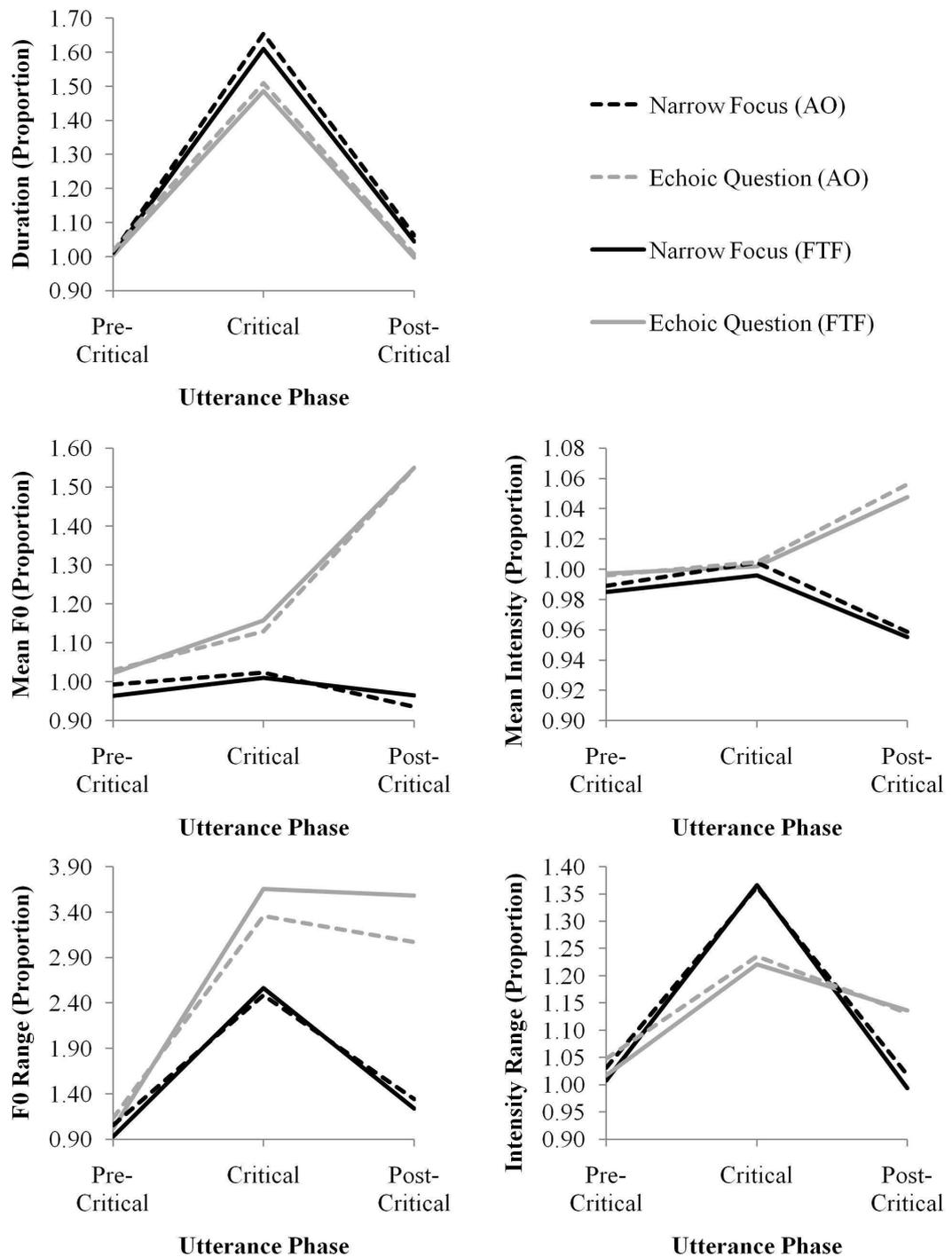


Figure 5.1. Acoustic properties of narrow focus and echoic question renditions (collapsed across utterances and talkers) in the AO and FTF interactive conditions, represented as proportion values of the mean broad focused rendition in their respective interactive condition. Thus, a value greater than 1 indicates an increase on the parameter compared to the broad focused rendition.

5.2.2. Effect of Utterance Type

To determine the effect that utterance length and location of the critical constituent within the utterance had on prosodic realisation, the interactions between prosody and utterance type from the sentence analysis (Appendix C.1) were examined for focus (Figure 5.2) and phrasing contrasts (Figure 5.3) produced in the AO interactive condition.

The basis of the significant interactions were investigated with a series of univariate ANOVAs (conducted individually for focus and phrasing in AO interactive condition) with utterance type as the between-items factor, interpreted with a Bonferroni adjusted α of 0.025 for multiple comparisons⁶. Sidak post-hoc pairwise comparisons (with 97.5% confidence intervals) were used to identify where significant differences occurred between utterance types.

5.2.2.1. Focus as a function of Utterance Type

The interaction between prosody and utterance type was significant for the mean syllable duration of pre-critical content, $F_1(3,26) = 7.64$, $p = 0.001$, $\eta_p^2 = 0.469$. Sidak post-hoc comparisons showed that S/E utterances had longer pre-critical syllable durations than S/L [$M_{\text{Diff}} = 0.08$, Sidak 97.5% CI: 0.01 – 0.15] and L/L utterances [$M_{\text{Diff}} = 0.07$, Sidak 97.5% CI: 0.01 – 0.13]. L/E pre-critical syllable durations were also longer than in S/L utterances [$M_{\text{Diff}} = 0.08$, Sidak 97.5% CI: 0.00 – 0.15]. When the critical content occurred late in the utterance (i.e., in S/L and L/L utterances), pre-focal reduction in syllable duration occurred compared to broad

⁶ As the data used in the sentence ANOVA were proportion values, the value of broad focused renditions was always “1”, with no variability. Thus, the interactions between prosody and utterance type are driven exclusively by variations in the production of the narrow focus/echoic question renditions. As such, the between-subjects ANOVA used to interpret the interactions generated the same F , p and η_p^2 values as the interaction.

focused productions. By contrast, the S/E and L/E utterances (i.e., when the critical constituent occurred early in the utterance) were produced with pre-focal lengthening.

The post-critical duration also differed across prosodic contrasts as a function of utterance type, $F_1(3,26) = 13.42$, $p < 0.001$, $\eta_p^2 = 0.608$. A significant difference was found between S/E and S/L [$M_{\text{Diff}} = 0.158$, Sidak 97.5% CI: 0.06 - 0.26], and between L/E and L/L [$M_{\text{Diff}} = 0.11$, Sidak 97.5% CI: 0.02 - 0.20] utterances, with the post-critical content produced with longer syllable durations when the critical constituent occurred late in the utterances. Presumably, this difference came about because after the production of a narrowly focused constituent (marked by enhanced syllable durations), the production of the remaining content (i.e., post-critical phase) only gradually returns to something resembling the pre-critical rate. Given this, when the critical constituent occurs early in the utterance, a greater amount of time is available for the return in rate compared to when the critical word occurs in the latter half of the utterance.

A significant interaction was observed between prosody and utterance type for post-focal F_0 , $F_1(3,26) = 3.89$, $p = 0.02$, $\eta_p^2 = 0.310$, however pairwise comparisons revealed no statistically significant differences between any individual utterance class. With the exception of S/L utterances, all utterance types were produced with a lower mean F_0 in the narrow focused renditions (when compared to broad focus productions). As with the above account for differences in duration, it would seem plausible that after an increase in F_0 to prosodically mark the focused constituent, F_0 would gradually fall back to baseline; an occurrence more likely to

occur when the utterance is long, or when the critical constituent occurs early in the utterance, but not in S/L utterances where there may be insufficient time.

The F_0 range of post-critical content also varied across utterance types, $F_{1(3,26)} = 10.19, p < 0.001, \eta_p^2 = 0.540$, with post-hoc comparisons showing that S/L utterances covered a significantly larger post-focal F_0 range than S/E [$M_{\text{Diff}} = 0.99$, Sidak 97.5% CI: 0.35 - 1.64], L/L [$M_{\text{Diff}} = 1.02$, Sidak 97.5% CI: 0.34 - 1.69] and L/E utterances [$M_{\text{Diff}} = 1.06$, Sidak 97.5% CI: 0.37 - 1.75]. When this results is considered in conjunction with the data for mean F_0 across utterance types (where the greatest difference from broad focused renditions on the critical constituent is achieved in S/L utterances, see Figure 5.2), the post-focal F_0 for S/L utterances requires the greatest amount of change to return back to a baseline (although this may not be entirely successful, as the mean F_0 is also higher for S/L utterances in post-focal phases, see above).

The intensity range covered during the production of post-critical content showed a prosody by utterance type interaction, $F_{1(3,26)} = 3.13, p = 0.043, \eta_p^2 = 0.265$. However, no pairwise comparisons were significant, with a similar pattern of intensity range observed across all four utterance types.

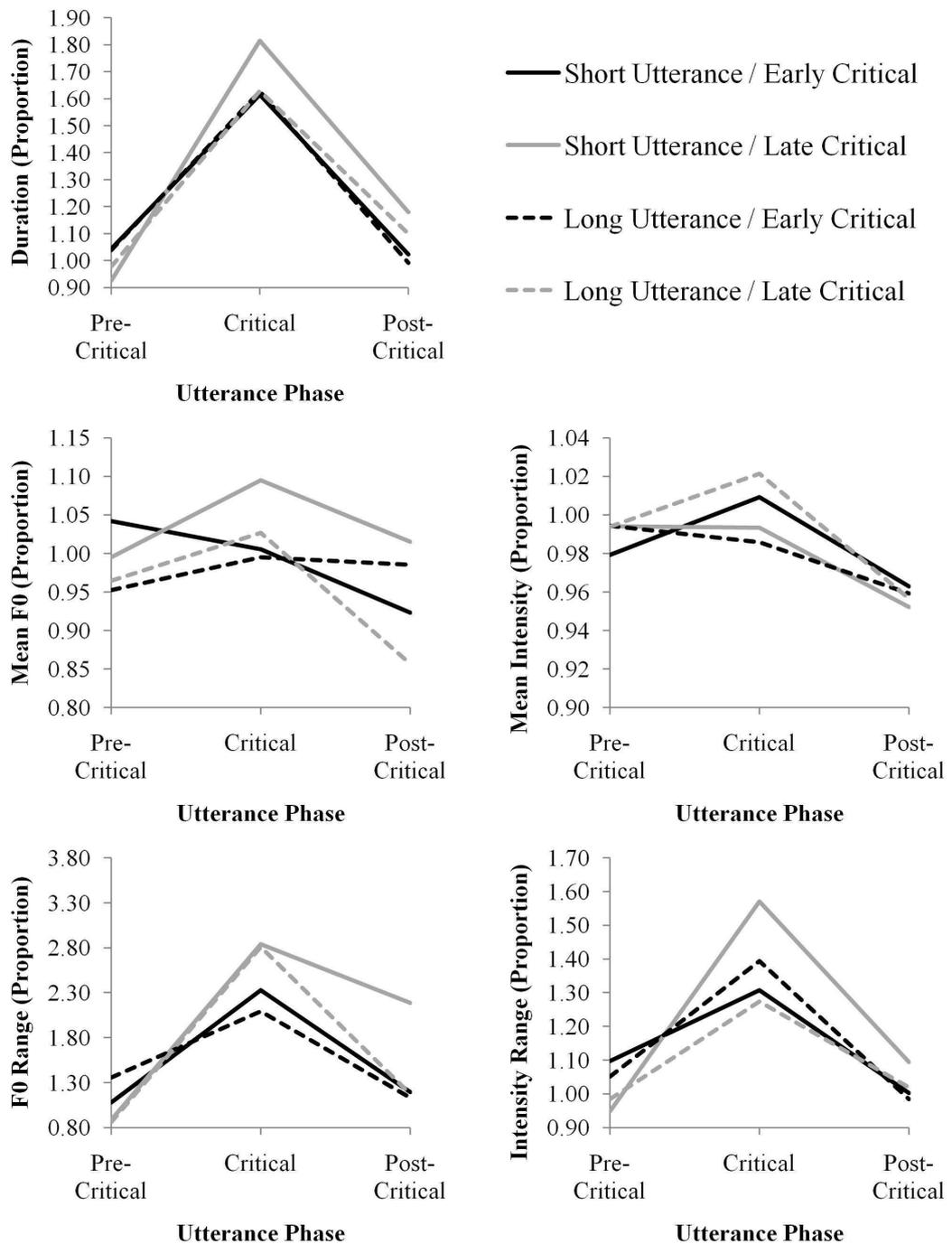


Figure 5.2. Acoustic properties of narrow focused renditions recorded in the AO interactive condition (expressed as a proportion of the broad focused AO rendition), as a function of utterance type, collapsed across talkers and sentences.

5.2.2.2. *Phrasing as a function of Utterance Type*

The prosody by utterance type interaction was significant for pre-critical content duration, $F_1(3,26) = 4.03$, $p = 0.018$, $\eta_p^2 = 0.317$. Although utterances containing a late critical constituent tended to be pre-critically shortened (and pre-critical lengthening when the critical word occurred early in the utterance), none of the pairwise comparisons were statistically significant.

The interaction was also significant for the mean syllable duration of post-critical content, $F_1(3,26) = 8.86$, $p < 0.001$, $\eta_p^2 = 0.506$; however the opposite pattern was observed for pre-critical duration. The post-critical syllable durations of S/L utterances were significantly greater than both S/E [$M_{\text{Diff}} = 0.13$, Sidak 97.5% CI: 0.04 - 0.22] and L/E utterances [$M_{\text{Diff}} = 0.14$, Sidak 97.5% CI: 0.04 - 0.23].

For post-critical F_0 , a significant interaction between utterance type and prosody was found, $F_1(3,26) = 3.12$, $p = 0.043$, $\eta_p^2 = 0.265$. Although the mean F_0 was greater in S/L than other utterance types, no pairwise comparisons were significant.

A significant interaction between prosody and utterance type was found for the F_0 range of the critical constituent, $F_1(3,26) = 4.11$, $p = 0.016$, $\eta_p^2 = 0.322$. Although no pairwise comparisons were significant, the pattern of data indicated that utterances with critical constituents occurring in the latter half were produced with greater F_0 ranges than when the critical constituent occurred in the first half.

Post-critically, the interaction was also significant, $F_1(3,26) = 6.87$, $p = 0.001$, $\eta_p^2 = 0.442$, driven by the S/L utterances covering a much greater F_0 range than all other utterance types [vs. S/E; $M_{\text{Diff}} = 2.96$, Sidak 97.5% CI: 0.73 – 5.18; vs. L/E;

$M_{\text{Diff}} = 2.99$, Sidak 97.5% CI: 0.61 – 5.36; vs. L/L; $M_{\text{Diff}} = 2.58$, Sidak 97.5% CI: 0.27 – 4.90].

The interaction between prosody and utterance type was significant for the mean relative intensity of post-critical content, $F_1(3,26) = 3.60$, $p = 0.027$, $\eta_p^2 = 0.294$. Pairwise comparisons showed that the post-critical phase of S/L utterances were produced with greater intensity than L/E utterances [$M_{\text{Diff}} = 0.03$, Sidak 97.5% CI: 0.00 - 0.06].

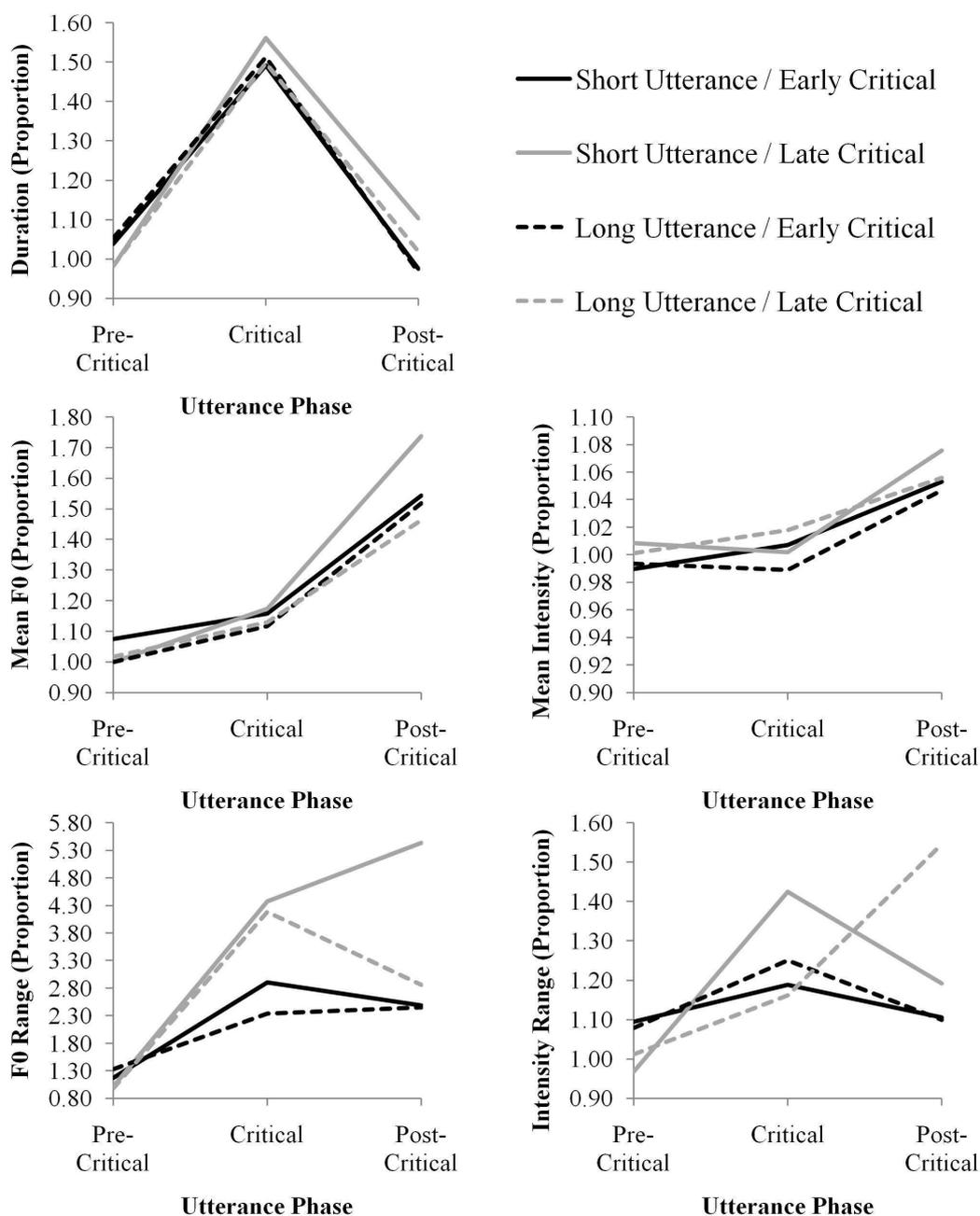


Figure 5.3. Acoustic properties of echoic question renditions recorded in the AO interactive condition (expressed as a proportion of the broad focused AO rendition), as a function of utterance type, collapsed across talkers and sentences.

5.2.3. Effects of Interactive Setting

To determine whether the talker being able to see the interlocutor had any impact on the realisation of prosodic contrasts, the renditions recorded in the AO setting were

compared to those recorded in the FTF setting for both narrow focus and echoic questions (see Figure 5.1). For each of the acoustic properties at each utterance phase, the renditions across interaction settings were compared with two analyses; one for the sentence data (item analyses: F_I , collapsed across talkers and repetitions), and one for talker data (subject analyses: F_S , collapsed across sentences and repetitions). For both analyses, a repeated-measures ANOVA was conducted, with interaction setting (AO; FTF) as the within-subjects item. Full statistical tables are included in Appendix C.1 for sentence and talker data.

5.2.3.1. Focus Contrasts across Interactive Settings

Syllable duration for the critical constituent was greater in the AO than in the FTF setting across both analyses, $F_I(1,29) = 8.26, p = 0.008, \eta_p^2 = 0.222$; $F_S(1,5) = 6.99, p = 0.046, \eta_p^2 = 0.583$. The post-critical utterance content was also produced over longer durations in the AO than FTF setting, $F_I(1,29) = 7.70, p = 0.010, \eta_p^2 = 0.210$, however this difference did not reach significance in the analysis by talker, $F_S(1,5) = 4.36, p = 0.091, \eta_p^2 = 0.466$.

The interactive setting also had an effect in the item analysis for mean intensity of critical constituents, with greater differences in intensity relative to baseline (broad focused renditions) being found for AO than FTF recordings, $F_I(1,29) = 11.60, p = 0.002, \eta_p^2 = 0.286$; $F_S(1,5) = 4.18, p = 0.096, \eta_p^2 = 0.455$. The intensity range of post-critical utterance phases was also larger compared to baseline for AO than FTF renditions, $F_I(1,29) = 7.38, p = 0.011, \eta_p^2 = 0.203$; $F_S(1,5) = 2.15, p = 0.202, \eta_p^2 = 0.301$.

5.2.3.2. *Phrasing Contrasts across Interactive Settings*

Only one difference between interactive settings was observed for phrasing contrasts. This difference was found for the mean intensity of post critical utterance content, $F_{I(1,29)} = 17.68, p < 0.001, \eta_p^2 = 0.379$, with the difference between echoic question renditions relative to baseline (broad focus) being greater in AO than FTF recordings. However, this difference was observed only for the sentence analysis, not the talker analysis, $F_S(1,5) = 2.57, p = 0.170, \eta_p^2 = 0.340$.

5.2.4. **Idiosyncratic Talker Strategies**

It was evident from both the analysis of prosodic realisations and the examination of the effect of differing interactive settings that not all talkers used the same pattern of acoustic features to contrasts focus and phrasing, or in the different interactive settings. This is reflected by the absence of an effect in the talker analysis (with data collapsed across sentences) when the sentence analysis (collapsed across talkers) showed significant differences. Indeed, some features such as lengthening of syllable durations for the critical constituent in the realisation of narrow focus and echoic questions (Figure 5.4), or an increase in post-critical mean F_0 for echoic question realisation (Figure 5.5), are consistently produced across talkers and sentences; however some features are utilised only by a single or small selection of talkers.

To investigate this further, a series of analyses were conducted individually for each talker, comparing the realisation of prosodic focus and phrasing, and differences in the realisation of these contrasts across interactive conditions. The way each contrast was realised by individual talkers is shown in Figure 5.4 to Figure 5.8.

5.2.4.1. *Idiosyncrasies in Focus Realisation*

A series of repeated measures within-subjects ANOVAS were conducted individually for each talker, comparing the broad and narrow focused renditions (in the AO interactive condition) for each acoustic feature at each utterance phase (full statistical tables are included in Appendix C.1).

For the realisation of narrow focus (relative to the broad focused rendition), all six talkers consistently elongated both the critical and post-critical syllable durations, reduced the intensity of the pre- and post-focal utterance content (Figure 5.7), and covered a greater intensity range during the production of the critical constituent (Figure 5.8). Furthermore, all talkers increased the range of $F0$ covered for the critical constituent, but did this at varying degrees hence explaining the absence of a talker effect in the original sentence analysis (see Figure 5.6, particularly Talker 4).

Two other features achieved significance in the original sentence analysis without an effect across talkers. The first of these was a post-focal reduction in $F0$; three of the six talkers produced post-critical utterance content with a lower mean $F0$ than in broad focused renditions [Talker 1: $F(1,29) = 19.20$, $p < 0.001$, $\eta_p^2 = 0.398$; Talker 2: $F(1,29) = 19.87$, $p < 0.001$, $\eta_p^2 = 0.407$; Talker 3: $F(1,29) = 21.18$, $p < 0.001$, $\eta_p^2 = 0.422$]. Interestingly, the remaining three talkers consistently produced post-focal content with an increased $F0$ range relative to broad focused productions [Talker 4: $F(1,29) = 8.19$, $p = 0.008$, $\eta_p^2 = 0.220$; Talker 5: $F(1,29) = 7.59$, $p = 0.01$, $\eta_p^2 = 0.207$; Talker 6: $F(1,29) = 7.26$, $p = 0.012$, $\eta_p^2 = 0.200$, see Figure 5.6].

In addition to using these acoustic features to mark a narrowly focused word, some talker-specific features were also observed. For example, Talker 2 produced

pre-critical content in narrow focused renditions with a lower range of $F0$, $F(1,29) = 8.39$, $p = 0.007$, $\eta_p^2 = 0.224$; and a reduction in the intensity range of post-critical content, $F(1,29) = 23.53$, $p < 0.001$, $\eta_p^2 = 0.448$. By contrasts, two other talkers produced post-critical utterance content with an increased intensity range (Talker 5: $F(1,29) = 11.85$, $p = 0.002$, $\eta_p^2 = 0.290$; and Talker 6: $F(1,29) = 8.93$, $p = 0.006$, $\eta_p^2 = 0.236$). Finally, Talker 4 uniquely realised narrowly focused critical constituents with an increase to both mean $F0$, $F(1,29) = 24.38$, $p < 0.001$, $\eta_p^2 = 0.457$; and mean intensity, $F(1,29) = 31.34$, $p < 0.001$, $\eta_p^2 = 0.519$, in comparison to broad focused tokens.

5.2.4.2. Idiosyncrasies in Phrasing Realisation

With respect to the realisation of focus, the acoustic features (at each phase of the utterance) were analysed per talker in a series of repeated measures within-subjects ANOVAS that compared broad focused and echoic question renditions recorded in the AO interactive condition (full statistical tables are included in Appendix C.1).

As expected, the analyses showed that all six talkers consistently produced echoic questions (relative to broad focused tokens) with greater syllable duration, $F0$ range and intensity range on the critical constituents, and greater $F0$, $F0$ range, intensity, and intensity range for post-critical utterance content. Furthermore, all but one talker (i.e., Talker 1) produced the critical constituent in echoic question renditions with a greater mean $F0$ than in the broad focused renditions.

The acoustic properties of pre-critical utterance content were highly variable across talkers. For example, the pre-critical $F0$ in echoic questions was reduced by two talkers [Talker 1: $F(1,29) = 6.09$, $p = 0.020$, $\eta_p^2 = 0.174$; Talker 4: $F(1,29) = 10.34$, $p = 0.003$, $\eta_p^2 = 0.263$], maintained by two talkers (Talker 2 and 3), and

enhanced by the remaining two talkers relative to broad focus renditions [Talker 5: $F(1,29) = 6.47, p = 0.017, \eta_p^2 = 0.182$; Talker 6: $F(1,29) = 10.60, p = 0.003, \eta_p^2 = 0.268$; see Figure 5.5]. Talkers 5 and 6 also elongated the syllable duration of pre-critical content in echoic questions [Talker 5: $F(1,29) = 6.82, p = 0.014, \eta_p^2 = 0.190$; Talker 6: $F(1,29) = 8.16, p = 0.008, \eta_p^2 = 0.219$].

5.2.4.3. *Idiosyncrasies across Interactive Settings*

Whereas the pattern of acoustic features used to realise prosodic contrasts were similar across talkers, the same was not the case across the interactive settings. The acoustic features at each utterance phase were subjected to a series of post-hoc repeated measures ANOVAS for each talker, comparing the narrow focus and echoic question renditions across interactive settings.

The outcome of these analyses showed no consistent strategies across talkers when visual information about their interlocutor was no longer available (i.e., in the AO condition) for either focus or phrasing contrasts. However, talkers produced patterns of change in certain properties across the AO and FTF settings. For example, with respect to narrow focus, Talker 1 and 6 increased mean syllable duration across all three utterance phases more so in the AO than FTF setting, Talker 1 [Pre-Critical: $F(1,29) = 4.61, p = 0.040, \eta_p^2 = 0.137$; Critical: $F(1,29) = 8.27, p = 0.004, \eta_p^2 = 0.222$; Post-Critical: $F(1,29) = 10.42, p = 0.003, \eta_p^2 = 0.264$] and Talker 6 [Pre-Critical: $F(1,29) = 4.86, p = 0.036, \eta_p^2 = 0.143$; Critical: $F(1,29) = 6.17, p = 0.019, \eta_p^2 = 0.175$; Post-Critical: $F(1,29) = 6.31, p = 0.018, \eta_p^2 = 0.179$]. In contrast, Talkers 2 and 3 increased the mean intensity of the critical constituent more (compared to broad focused renditions) in the AO than FTF interactive condition [Talker 2: $F(1,29) = 5.18, p = 0.03, \eta_p^2 = 0.152$; Talker 3: $F(1,29) = 4.99, p = 0.033,$

$\eta_p^2 = 0.147$]; Talker 5 appeared to use a combination of greater fundamental frequency when producing narrow focused tokens in AO situations compared to FTF productions, $F(1,29) = 5.42, p = 0.027, \eta_p^2 = 0.158$; and intensity, $F(1,29) = 17.82, p < 0.001, \eta_p^2 = 0.381$.

Compared to the narrow focus, there were far fewer differences between interactive settings for the echoic question renditions. As it turned out, three talkers produced post-critical content with a greater intensity in the AO than FTF setting [Talker 2: $F(1,29) = 6.32, p = 0.018, \eta_p^2 = 0.179$; Talker 4: $F(1,29) = 15.47, p < 0.001, \eta_p^2 = 0.348$; Talker 5: $F(1,29) = 12.17, p = 0.002, \eta_p^2 = 0.296$]. Talker 5 also produced the critical constituent at a greater intensity, $F(1,29) = 9.54, p = 0.004, \eta_p^2 = 0.247$, while Talker 6 elongated the syllable duration of all utterance phases more in the AO than FTF settings [Pre-Critical: $F(1,29) = 4.45, p = 0.044, \eta_p^2 = 0.133$; Critical: $F(1,29) = 7.27, p = 0.012, \eta_p^2 = 0.200$; Post-Critical: $F(1,29) = 10.12, p = 0.003, \eta_p^2 = 0.259$].

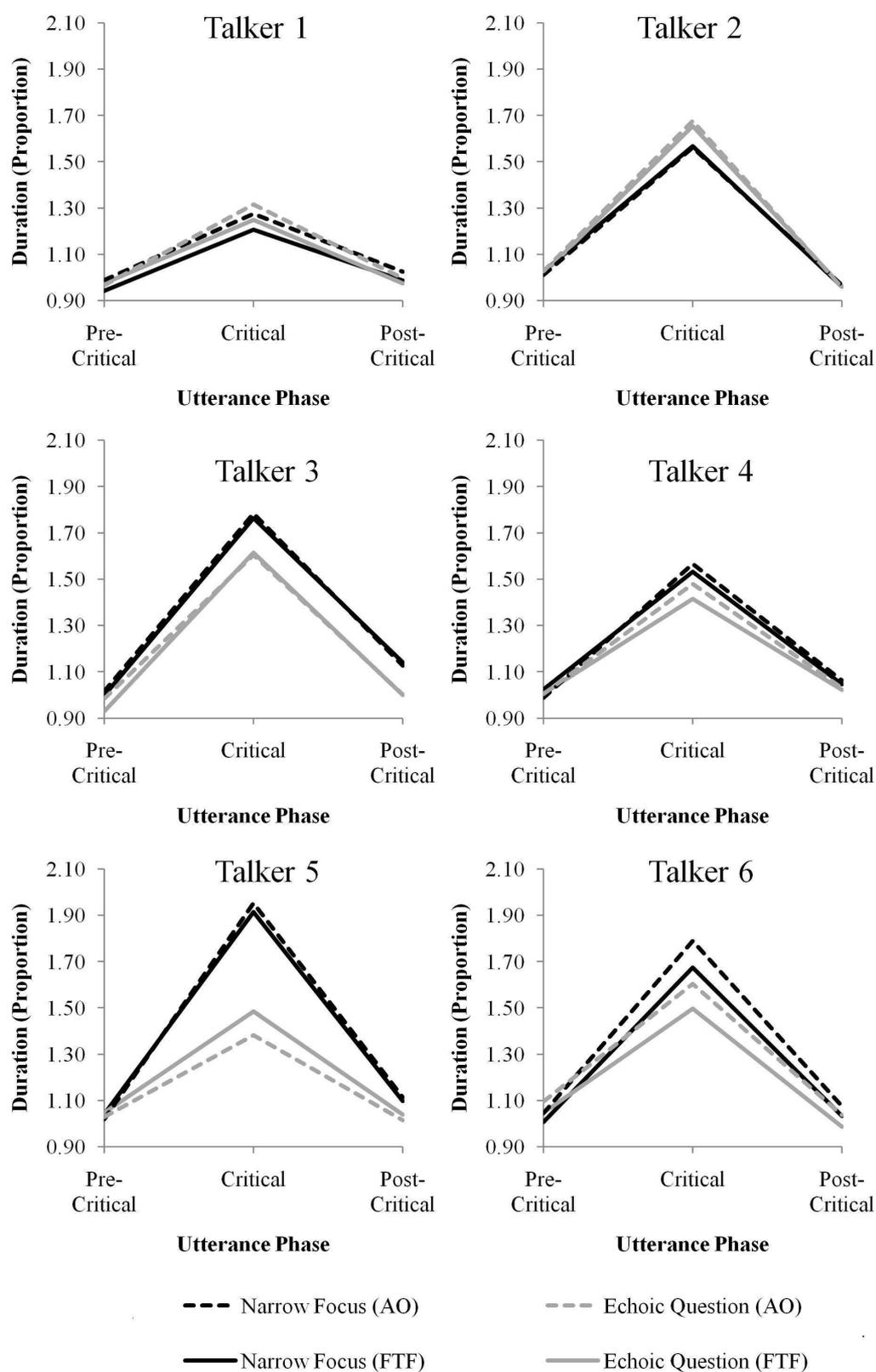


Figure 5.4. Mean syllable duration (represented as a proportion of the broad focused rendition) produced by each talker.

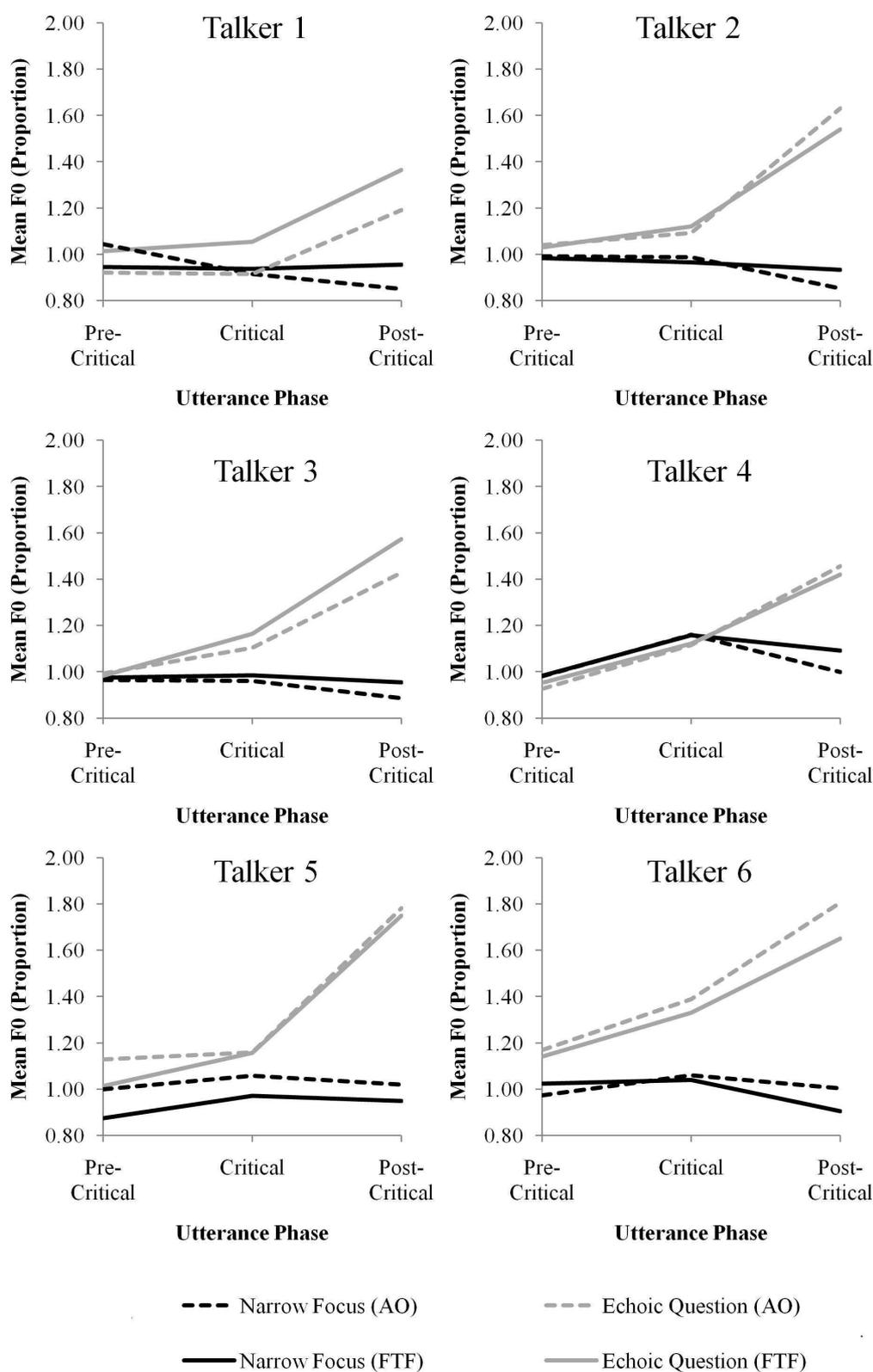


Figure 5.5. Mean fundamental frequency (represented as a proportion of the broad focused rendition) produced by each talker.

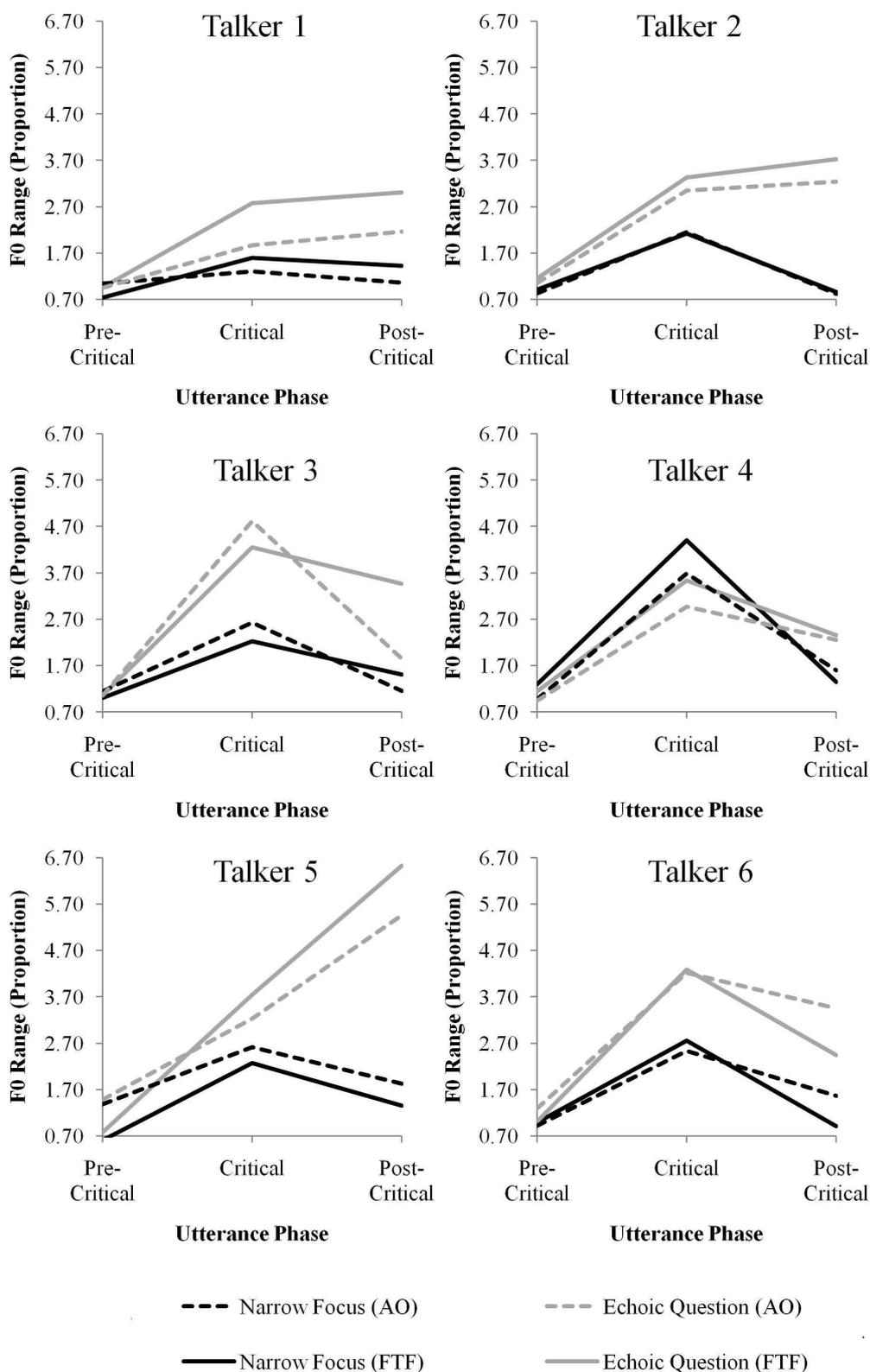


Figure 5.6. Fundamental frequency range (represented as a proportion of the broad focused rendition) produced by each talker.

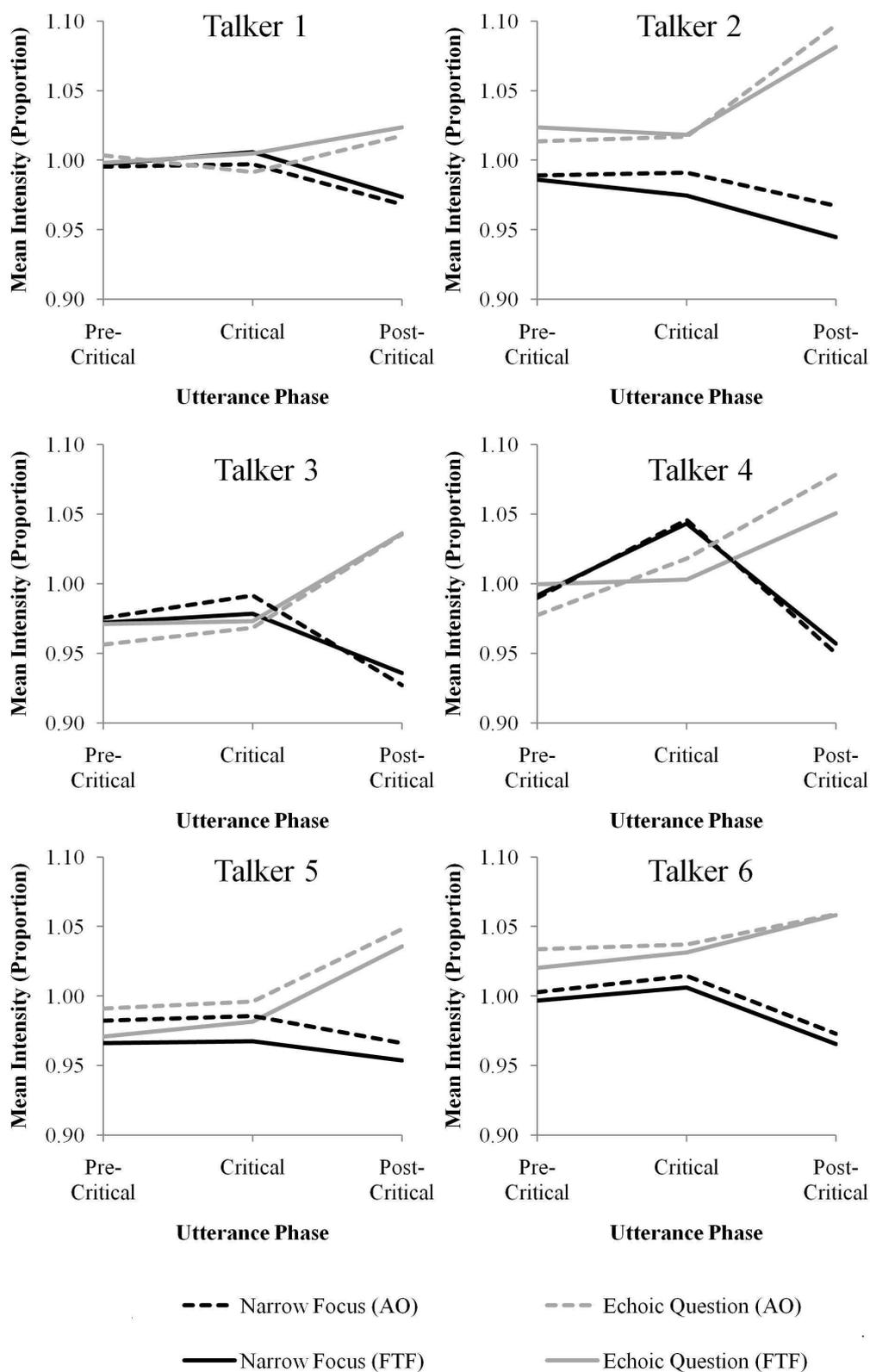


Figure 5.7. Mean intensity (represented as a proportion of the broad focused rendition) produced by each talker.

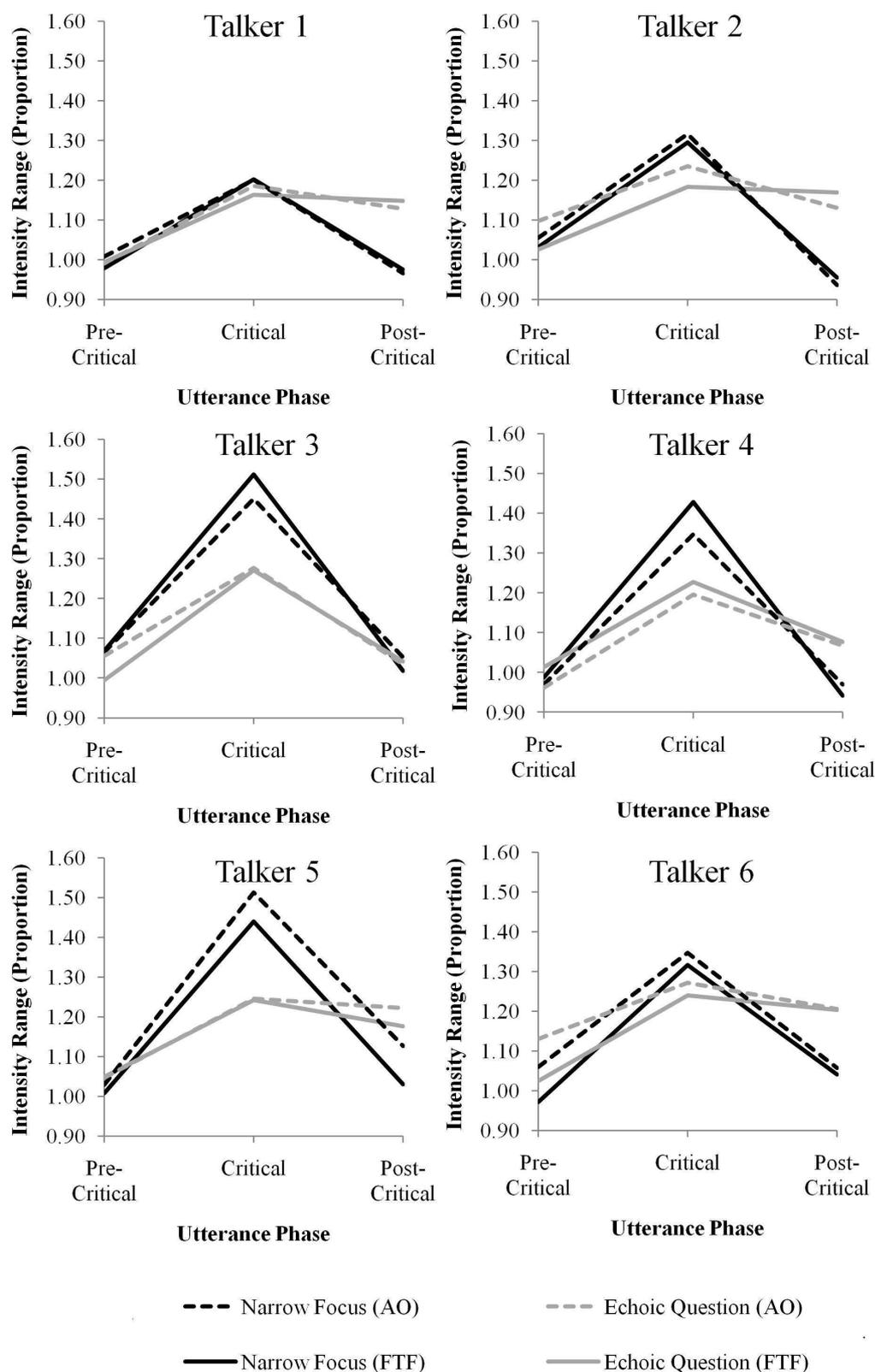


Figure 5.8. Intensity range (represented as a proportion of the broad focused rendition) produced by each talker.

5.3. Spectral Analysis⁷

5.3.1. Data Selection

In addition to measuring broad acoustic properties, the fine-grained acoustic-phonetic characteristics across prosodic conditions and interactive settings were also examined by determining vowel space properties. To measure vowel space, a sub-set of nine of the original 30 utterances were selected in which the first vowel of the critical constituent was one of the corner-most vowels of the corpus [i.e., /æ, ɪ, ɔ:/].

These sentences (and their associated vowels) are provided in Table 5.1. As each sentence was recorded twice within each prosodic and interactive condition, a total of 18 vowel tokens were used to calculate the vowel space for each talker within each condition.

⁷ A preliminary version of these analyses appeared in: Cvejic, E., Kim, J., & Davis, C. (2010). Modification of prosodic cues when an interlocutor cannot be seen: The effect of visual feedback on acoustic prosody production. *Proceedings of the 20th International Congress on Acoustics, Sydney, Australia*, Paper ID: 521, pp. 1-7.

Table 5.1. Sub-set of sentences within the recorded corpus containing the corner vowels within the critical item. Critical words appear in brackets. IPA glossing is of standard Australian English.

Vowel	Utterance	IPA Glossing of Critical Word
/æ/	The weight of the [package] was seen on the high scale	/pækɪdʒ/
	It was hidden from sight by a [mass] of leaves and shrubs	/mæs/
	Hold the [hammer] near the end to drive the nail	/hæmə:/
/ɪ:/	It is a band of [steel] three inches wide	/sti:l/
	The lobes of her ears were [pierced] to hold rings	/pɪəst/*
	This is a grand [season] for hikes on the road	/si:zən/
/ɔ:/	Clams are round, [small], soft and tasty	/smɔ:l/
	A [small] creek cut across the field	/smɔ:l/
	The set of china hit the [floor] with a crash	/flɔ:r/

* Only the /ɪ/ segment of the diphthong /ɪə/ was examined in the analysis.

5.3.2. Spectral Feature Extraction and Processing

The phonemically aligned acoustic signals were used to extract the critical phonemes listed in Table 5.1. The steady state values for the first and second formants ($F1$ and $F2$ respectively) were then determined in Praat (Boersma, 2001) using the procedure outlined in Munhall, MacDonald, Byrne and Johnsrude (2009). The acoustic signal was initially down sampled to 10 kHz, before calculating formant frequencies by applying a 25ms sliding window (with steps of 1ms) to the signal, with the steady state value being determined by averaging 40% of the formant estimates between 40

and 80 percent of the total vowel duration. The obtained values were then converted to the perceptually-motivated Mel scale (Fant, 1973) using (1):

$$M = \left(\frac{1000}{\log 2}\right) \times \left(\log\left(\frac{F}{1000}\right) + 1\right) \quad (1)$$

where M and F are frequency values expressed as Mels and Hertz, respectively (Bradlow, Toretta & Pisoni, 1996). Each talker's vowel space for broad focus, narrow focus and echoic questions in both AO and FTF interactive settings was then represented by the location of the 18 vowel measurements within the $F1$ by $F2$ space.

The vowel triangle area was then calculated by determining the Euclidian area covered by the triangle defined by the mean of each vowel category, using (2):

$$\text{VT Area (Mels}^2\text{)} = \sqrt{(S \times (S - A_L) \times (S - B_L) \times (S - C_L))} \quad (2)$$

where $S = \frac{A_L + B_L + C_L}{2}$ and A_L , B_L , and C_L are the Euclidean distances in Mels between vowel category centres /æ/ to /ɪ/, /ɪ/ to /ɔ:/ and /ɔ:/ to /æ/, respectively.

To calculate the between-category dispersion, the mean of the Euclidean distance between the $F1$ - $F2$ vowel space midpoint, and each of the 18 recorded vowel tokens was determined for each talker. This measure indicates the overall expansion (or compaction) of the vowel space (Bradlow et al., 1996), with a greater dispersion value suggesting that the vowel categories were produced to be perceptually more distinct from each other. Within-category dispersion was calculated in a similar way, by determining the Euclidean distance between the midpoint for each vowel category, and each measured token within that category (with the mean obtained for these values). This measure provides an indication of individual vowel category dispersion, indicating consistency (or variability) of

individual vowel productions across repetitions within each prosodic category and interactive condition. In addition to vowel triangle area, within-category dispersion and between-category dispersion, the $F1$ and $F2$ range was also measured by calculating the difference between the formant maxima and minima. All calculations were carried out in Matlab using custom written scripts.

5.3.3. Analysis Results

Table 5.2 displays the mean spectral properties collapsed across talkers for each of the prosodic and interactive conditions. A series of paired-samples t -tests were carried out in order to compare the spectral properties of broad to narrowly focused renditions, and broad focus to echoic questions within the AO condition, as well as comparing the spectral properties of each prosodic condition between the AO and FTF interactive settings.

The vowel triangle areas were larger for both narrowly focused [$t_{AO}(5) = 4.34, p = 0.007$] and echoic question productions [$t_{AO}(5) = 3.05, p = 0.028$] relative to broad focused renditions (see Figure 5.9). No differences were observed across interactive settings, with the size of vowel space expansion across prosodic conditions the same regardless of the visual availability of the interlocutor. This suggests that when critical vowels are either focused or questioned, vowel categories are made perceptually more distinct from each other. This is further supported by the measure of between-category displacement, showing that for both narrow focus [$t_{AO}(5) = 5.00, p = 0.004$] and echoic questions [$t_{AO}(5) = 3.75, p = 0.013$] vowel category centroids were located a greater distance away from the vowel space midpoint than when produced in a broad focused context (Figure 5.10). As with the

measure of vowel triangle area, no differences between the interactive settings were found for between-category displacements.

The measure of within-category displacement (i.e., displacement of each individual vowel token from the vowel category midpoint) was similar across all three prosodic contexts and interactive settings. As expected from an increase in vowel space area, the $F1$ and $F2$ range was greater for both narrow focus [$F1: t_{AO}(5) = 4.56, p = 0.006; F2: t_{AO}(5) = 3.53, p = 0.017$] and echoic question contexts [$F1: t_{AO}(5) = 3.86, p = 0.012; F2: t_{AO}(5) = 3.15, p = 0.026$] compared to broad focused renditions, with no effect of interactive setting observed.

Table 5.2. Spectral properties of critical vowels (with standard deviations) for the corner vowel subset as a function of interactive condition and prosodic context, collapsed across talkers, $n = 6$.

Prosodic Context	Vowel Triangle Area (Mels²)	Between-Category Dispersion (Mels)	Within-Category Dispersion (Mels)	F1 Range (Mels)	F2 Range (Mels)
<u>Auditory Only (AO) Interactive Setting</u>					
Broad Focus	100778.07 (44825.63)	323.01 (60.89)	61.68 (24.10)	368.85 (49.29)	876.36 (115.39)
Narrow Focus	151058.14 (65737.88)	389.71 (71.08)	65.58 (16.09)	460.72 (79.91)	1017.19 (143.34)
Echoic Question	138932.51 (71262.04)	373.56 (88.25)	65.47 (37.09)	436.31 (66.74)	958.70 (158.05)
<u>Face-to-Face (FTF) Interactive Setting</u>					
Broad Focus	99352.36 (32999.22)	321.16 (44.25)	57.48 (17.15)	366.76 (43.53)	871.45 (86.21)
Narrow Focus	139769.26 (55779.15)	380.81 (66.28)	77.26 (33.11)	455.79 (80.24)	1020.24 (148.99)
Echoic Question	134311.80 (54514.60)	373.21 (61.60)	66.77 (41.53)	430.35 (54.07)	979.60 (118.05)

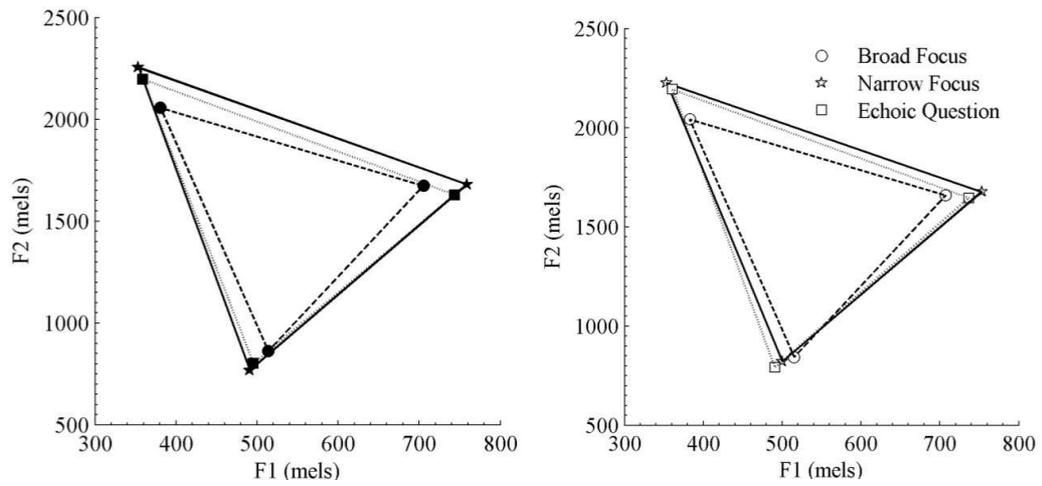


Figure 5.9. Vowel triangles for broad, narrow focus and echoic question renditions in the AO (left) and FTF interaction settings (right) collapsed across talkers.

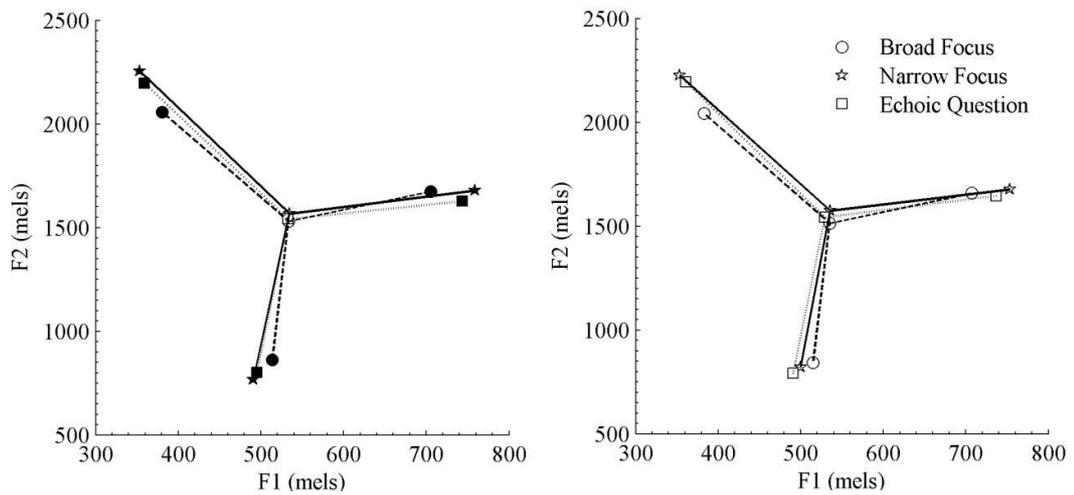


Figure 5.10. Between-category displacements for broad focus, narrow focus and echoic question utterance renditions in the AO (left) and FTF interaction settings (right), collapsed across talkers.

5.4. Summary

5.4.1. Realisation of Prosodic Contrasts

Overall, the properties of the narrow focussed utterances and the echoic questions corresponded to those typically described in the literature. Furthermore, most of

these characteristics were present in both the sentence and talker analyses, suggesting that there was a degree of consistency (i.e., typical cues used) across talkers in the way that prosodic contrasts were realised.

Relative to broad focused renditions, narrow focus utterances were produced with elongated syllables for the critical and post-critical phases. Although the mean $F0$ did not differ between focus conditions on critical words, a post-focal reduction in $F0$ was observed in narrow focused renditions. Similarly, there was an increased range of $F0$ covered during critical and post-critical utterance phases for narrow focused productions. Although the focused words were not produced at increased intensity levels, the content that followed was produced with decreased intensity, coupled with greater intensity range on the critical constituent; this would serve to make the focused word more salient. Spectrally, the initial vowels within focused constituents were produced over a larger vowel space (compared to the broad focused case), making them more distinct from each other.

For echoic questions, the critical words were generally produced with increased syllable durations, with a greater mean $F0$ and $F0$ range than declarative statements (i.e., broad focused renditions). This increase in mean $F0$ and $F0$ range was maintained for post-critical utterance phases. Intensity and intensity range was also greater in post-critical echoic questions than in statements. This combination of intensity and $F0$ manipulations likely serve as an indicator that some form of a response from the interlocutor is required (as opposed to statements that tend to signal finality with a declination on post-critical intensity and $F0$). As for the narrow focus case, echoically questioned vowels were produced to be more distinct from

each other (i.e., expanded vowel space with more dispersion between categories) than when the same vowels were produced in broad focused statement renditions.

Some differences in acoustic properties were apparent as a function of the utterance length and the location of the critical constituent within the utterance. Although utterance length and critical constituent location have been previously reported to play a role in modulating acoustic properties, the current results were not entirely compatible with those previously reported (e.g., Cooper et al., 1985; Eady & Cooper, 1986; Pell, 2001). For example, Pell (2001) reported that longer (but not shorter) narrowly focused sentences tended to be produced with a pre-focal reduction in syllable duration when the narrowly focused constituent occurred late within the utterance. In contrast, the results of the current analyses showed that the location of the critical constituent (regardless of utterance length) impacted on pre-critical duration, with both short and long utterances receiving pre-focal shortening when the critical constituent occurred late in the utterance, and pre-focal anticipatory lengthening when the focused word occurred in the first half of the utterance.

Eady and Cooper (1986) reported that when the critical constituent occurred in sentence initial or medial position (rather than sentence-final) the critical constituent had greater lengthening in both narrow focus and echoic question renditions (relative to broad focused ones). In comparison, the analysis of the current corpus showed no difference for the critical word regardless of utterance type, but did find differences in the content that followed the critical constituent (i.e., post-critical content). That is, compared to sentences where the critical constituent occurred early, post-critical content in sentences with a late occurring critical constituent showed greater durational lengthening in focused and questioned

renditions (relative to broad focused statements). This pattern could be accounted for by assuming that after prosodically marking a critical constituent using increased syllable duration, the talker's speech rate gradually returns to that exhibited pre-critically; in addition, for utterances with a late critical constituent, it could be assumed that there is insufficient time to complete this readjustment process.

One reason for the differences between the current and previous studies may be due to how utterance length has been classified. Pell (2001) defined short utterances as containing six syllables, while sentences containing ten syllables were deemed to be long. In contrast, the current study defined a short sentence as anything less than ten syllables, and as such, some of the short utterances (e.g., those with nine syllables) may have more in common with utterances classified as "long" in previous studies. Furthermore, the distinction between "early" and "late" critical constituents in the current analyses was based across sentences of differing segmental content rather than moving which word within the same sentence was prosodically marked. These differences, when considered along with the idiosyncrasies observed across talkers in their prosodic realisations, may explain why the same patterns across utterance types were not observed.

5.4.2. Differences as a function of Interactive Setting

Although the examined properties revealed (the expected) differences between the prosodic focus and phrasing contrasts, only a few differences were apparent as a function of the interactive setting (AO vs. FTF). For the realisation of narrow focus, mean syllable duration, mean relative intensity and intensity range of critical constituents, along with the syllable duration of post-critical content were greater when the interlocutor could not be seen (i.e., in the AO condition) when compared to

the FTF condition. For phrasing, the only difference observed was in the mean relative intensity of post-critical utterance content, with a higher intensity level for AO recordings.

It should be noted that these effects across interactive conditions pertained predominantly to the sentence analysis collapsed across individual talkers. That is, these effects were not significant in the talker analysis (when collapsed across utterances) and as such need to be interpreted with caution. Furthermore, it is yet to be determined whether these modifications have any perceptual consequences for how linguistic prosody is conveyed (this is explored in the next chapter), or whether such changes are accompanied by changes in the visual signal (see Chapter 7).

5.4.3. Talker Idiosyncrasies

Before moving on to the next chapter, it is worth pointing out that as well as the acoustic properties that were consistently used by talkers to contrast narrow from broad focused statements, and echoic questions from declarative statements, some idiosyncratic strategies in how these contrasts are signalled were also employed. This is also apparent when comparing how talkers change the produced speech signal depending on whether or not they can see their conversational partner. Whether these differences are associated with variation across talkers in the visual prosodic cues used, or impact on the perception of prosody, is further explored in the following chapters.

CHAPTER 6.
PERCEPTUAL RATING OF AUDITORY PROSODY

Chapter 6. Perceptual Rating of Auditory Prosody⁸

In Chapter 5, a range of acoustic differences were found between broad and narrow focused utterances, and between phrasing contrasts (i.e., statements and echoic questions), with some of these differences being greater in the AO interactive setting than the FTF one. This chapter investigated how these differences in acoustic measures between the two interactive settings might relate to the perception of prosody by using perceptual measures (i.e., subjective ratings of the degree of focus, or clarity of the statement-question contrast).

6.1. Experiment 7: Effects of Seeing the Interlocutor on the Production of Prosodic Contrasts

The ability of a perceiver to see the talker in a face-to-face situation facilitates the perception of both content (Summerfield, 1992) and prosody (Foxton et al., 2010) compared to when only the auditory signal is available. Furthermore, talkers appear to attune the production of their speech signals not only to the prevailing auditory conditions (Cooke & Lu, 2010) but also take into account whether the person they are conversing with can be seen or not. For example, when speaking in noise, talkers modify both their auditory speech and their visual speech in an attempt to make the produced speech more distinct (with visual speech changes particularly in FTF communication, see Fitzpatrick et al., 2011; Garnier et al., 2010).

Here, the effect of the talker being able to see (or not see) an interlocutor on the production of prosodic focus and phrasing is examined. One reason why the

⁸ Parts of this chapter (i.e., the key aspects of the results) appear in: Cvejic, E., Kim, J., & Davis, C. (revision under review). Effects of seeing the interlocutor on the production of prosodic contrasts. *Journal of the Acoustical Society of America*.

expression of prosody might differ across FTF and AO situations comes from a straightforward extension of Lindblom's Hyper-Hypospeech theory (Lindblom, 1990, 1996). In essence, Lindblom proposed that talkers dynamically tune speech output to be distinctive enough for the listener to achieve lexical access. This tuning was conceived in terms of a push-pull process in which production moves away from a default low-cost mode when forced to do so by the constraints imposed by the environmental or communicative setting. Constraints on speech output are often considered in terms of noise in the environment but can include any factor that affects the effectiveness to which the speech information can specify intended meaning. In this regard, it is proposed that not being able to see the interlocutor imposes a constraint on speech production (as this situation reduces the amount of speech information available) and as such, speech produced in AO settings would typically be more salient compared to that produced in FTF settings.

Lindblom's theory was proposed in terms of speech production being regulated by whether the listener can achieve lexical access (based upon phonetic discrimination) but production could equally be regulated with respect to the perception of prosody (as this too affects the meaning of an utterance). Given this, it was assumed that talkers in an AO setting would compensate for visual prosody information no longer being available to the conversational partner by making auditory cues to prosody more salient. This idea was initially tested in the previous chapter by comparing the acoustic characteristic of linguistic prosodic contrasts produced in AO and FTF settings. For narrowly focused utterances, the critical constituent was produced with greater syllable durations and mean intensity, while post-focal content was produced with elongated syllables and a greater intensity

range in conditions where the talker could no longer see the interlocutor (i.e., AO setting) compared to the FTF setting, whereas echoic questions recorded in the AO setting were produced with only an increase in post-critical intensity (relative to FTF renditions). In Experiment 7, the perceptual effect of these differences was explored by collecting subjective ratings of the perceived strength of these contrasts.

6.1.1. Method

6.1.1.1. Participants

Ten postgraduate psychology students at UWS ($M_{\text{Age}} = 29.2$ years, 5 females) participated in the prosodic rating tasks. All participants were native English listeners with self-reported normal hearing, no known communicative deficits, and no explicit phonetic training. These participants were naïve to the fact that tokens were recorded across differing interactive settings.

6.1.1.2. Materials

Nine of the sentences produced by all six talkers across prosodic conditions and interactive settings were selected from the recorded corpus for use as stimuli. These sentences were the ones examined in the spectral analysis detailed in Chapter 5 (listed below in Table 6.1), each of which contained one of the corner-most vowels of the corpus in the initial consonant-vowel (CV) syllable of the critical constituent. A sub-set of utterances were used (instead of the full corpus) to keep the total duration of the experimental tasks to a manageable minimum, with these particular utterances selected based on the expectation that some spectral differences may have been apparent across interactive setting (this however was not the case as elucidated in Chapter 5).

Table 6.1. Stimuli sentences used in the prosody rating tasks. The critical constituent is italicised.

Sentence	Segmental Content
1	The weight of the <i>package</i> was seen on the high scale
2	It was hidden from sight by a <i>mass</i> of leaves and shrubs
3	Hold the <i>hammer</i> near the end to drive the nail
4	It is a band of <i>steel</i> three inches wide
5	The lobes of her ears were <i>pierced</i> to hold rings
6	This is a grand <i>season</i> for hikes on the road
7	Clams are round, <i>small</i> , soft and tasty
8	A <i>small</i> creek cut across the field
9	The set of china hit the <i>floor</i> with a crash

Acoustic analyses were conducted on these utterances to ensure that they exhibited the general properties found in the previous chapter. The data for each acoustic parameter (for each utterance phase) for focus and phrasing contrasts were independently subjected to a series of repeated measures ANOVAs, with prosodic condition as a within-items factor, for both sentence data (item analysis, F_I ; collapsed across talkers) and talker data (subject analysis, F_S ; collapsed across sentences). The critical constituents of narrow focused tokens (relative to broad focus) were produced with greater mean syllable duration, $F_I(1,8) = 127.32$, $p < 0.001$, $\eta_p^2 = 0.941$; $F_S(1,5) = 49.69$, $p = 0.001$, $\eta_p^2 = 0.909$, and covered a higher $F0$ range, $F_I(1,8) = 37.65$, $p < 0.001$, $\eta_p^2 = 0.825$; $F_S(1,5) = 23.94$, $p = 0.005$, $\eta_p^2 = 0.827$, and intensity range, $F_I(1,8) = 13.83$, $p = 0.006$, $\eta_p^2 = 0.634$; $F_S(1,5) = 122.09$, $p < 0.001$, $\eta_p^2 = 0.961$. Post-critical content showed a lower mean intensity than the same content in broad focused renditions, $F_I(1,8) = 269.30$, $p < 0.001$, $\eta_p^2 = 0.971$; $F_S(1,5) = 10.62$, $p = 0.022$, $\eta_p^2 = 0.690$.

For phrasing contrasts, in comparison to statement renditions the critical constituents of the echoic question renditions were produced with longer syllable durations, $F_1(1,8) = 128.74, p < 0.001, \eta_p^2 = 0.941$; $F_S(1,5) = 77.38, p < 0.001, \eta_p^2 = 0.939$, higher mean F_0 , $F_1(1,8) = 39.34, p < 0.001, \eta_p^2 = 0.831$; $F_S(1,5) = 7.40, p = 0.042, \eta_p^2 = 0.597$, greater F_0 range, $F_1(1,8) = 46.66, p < 0.001, \eta_p^2 = 0.854$; $F_S(1,5) = 39.48, p = 0.002, \eta_p^2 = 0.888$, and intensity range, $F_1(1,8) = 6.38, p = 0.035, \eta_p^2 = 0.444$; $F_S(1,5) = 40.89, p = 0.001, \eta_p^2 = 0.891$. Post-critical content was produced with significantly greater mean F_0 , $F_1(1,8) = 210.22, p < 0.001, \eta_p^2 = 0.963$; $F_S(1,5) = 24.62, p = 0.004, \eta_p^2 = 0.831$, mean intensity, $F_1(1,8) = 72.45, p < 0.001, \eta_p^2 = 0.901$; $F_S(1,5) = 25.95, p = 0.004, \eta_p^2 = 0.835$, larger F_0 range, $F_1(1,8) = 39.48, p = 0.002, \eta_p^2 = 0.888$; $F_S(1,5) = 30.96, p = 0.001, \eta_p^2 = 0.795$, and intensity range, $F_1(1,8) = 35.64, p < 0.001, \eta_p^2 = 0.817$; $F_S(1,5) = 12.26, p = 0.017, \eta_p^2 = 0.710$, than equivalent content produced in a broad focused statement context.

Across interactive settings, data for each parameter (for each phase of the utterance) for narrow focus and echoic question tokens were independently subjected to a series of repeated measures ANOVAs, with interactive setting (AO; FTF) as a within-items factor. Narrowly focused tokens were produced with critical constituents of greater mean intensity, $F_1(1,8) = 7.65, p = 0.024, \eta_p^2 = 0.489$; $F_S(1,5) = 1.37, p = 0.295$, and post-critical content with a greater intensity range in the AO than FTF setting, $F_1(1,8) = 7.77, p = 0.024, \eta_p^2 = 0.493$; $F_S(1,5) = 4.11, p = 0.098$. For phrasing, two features were found to differ across the interactive settings. The mean syllable duration of the critical word was greater (relative to baseline) for AO than FTF recordings, $F_1(1,8) = 6.52, p = 0.034, \eta_p^2 = 0.449$; $F_S(1,5) = 23.92, p =$

0.005, $\eta_p^2 = 0.827$, as was the mean intensity of post-critical utterance content, $F_{I(1,8)} = 13.08$, $p = 0.007$, $\eta_p^2 = 0.621$; $F_{S(1,5)} = 26.60$, $p = 0.004$, $\eta_p^2 = 0.842$.

As the intensity level of the tokens differed across talkers and interactive settings, the mean intensity of each token was normalised to 65dB using Praat (Boersma, 2001).

6.1.1.3. Procedure

The utterances were presented to listeners in two perceptual tasks: subjective rating of the degree of focus, and the perceptual clarity of the statement-question contrast. For the focus rating task, participants were initially presented with the critical word printed in text on-screen, followed by an auditory token of an utterance, and then asked to rate the degree of focus received on the critical constituent within the token using a 7-point Likert scale (with a response of “1” indicating that the constituent received no focus, and “7” indicating that the word was clearly focused). In total, 162 stimulus items were presented, comprising of a single repetition of each of the nine sentences produced as a broad focused rendition (in the FTF setting), a narrow focus rendition recorded in the FTF setting, and a narrow focus rendition recorded in the AO setting, from each talker (i.e., 6 talkers \times 9 sentences \times 3 conditions). Presentation of items was blocked by talker with presentation order between- and within-blocks randomised by the presentation software (DMDX; Forster & Forster, 2003).

The phrasing rating task was similar to the focus rating task except that participants were asked to rate the utterance on a continuum of “statement” (by a response of “1”) to “clearly phrased question” (by a response of “7”) for broad focused statement renditions and echoic questions recorded in the FTF and AO

settings. The broad focus token used was always a different one from that which appeared in the focus rating task, so participants were never exposed to the same token more than once. For both tasks, participants were informed that there was no “correct” answer, and were encouraged to use the complete range of the rating scale responses. Furthermore, the participants were not instructed in any way as to what features to base their judgements on. The order of task completion was counter-balanced. DMDX was used for stimuli presentation, with participants hearing the speech items binaurally over Senheiser HD650 stereo headphones.

6.1.2. Results

The ratings were subjected to a series of repeated measures ANOVAs for each perceptual task; a subject analysis (F_S) with prosodic condition, talker and sentence as within-items factors; and an item analysis (F_I) with prosodic condition, talker and rater as within-items factors.

6.1.2.1. Perceptual Rating Scores for Focus Contrasts

The mean ratings (collapsed across sentences) of each talker’s production of focus for the AO and FTF conditions are presented in Figure 6.1. Significant main effects were found for prosodic condition, $F_S(2,18) = 761.81, p < 0.001, \eta_p^2 = 0.99$; $F_I(2,16) = 898.63, p < 0.001, \eta_p^2 = 0.99$, as well as talker, $F_S(5,45) = 14.20, p < 0.001, \eta_p^2 = 0.61$; $F_I(5,40) = 9.95, p < 0.001, \eta_p^2 = 0.55$. The talker by prosodic condition interaction was also significant, $F_S(10,90) = 10.79, p < 0.001, \eta_p^2 = 0.55$; $F_I(10,80) = 7.32, p < 0.001, \eta_p^2 = 0.48$. Sidak post hoc comparisons showed that broad focused renditions were rated significantly lower than narrow focus productions in both the FTF [$M_{Diff} = 3.71$, Sidak 95% CI: 3.33 – 4.09] and AO [$M_{Diff} = 4.25$, Sidak 95% CI: 3.84 – 4.65] conditions. A pairwise comparison of key contrast between the AO and

FTF productions confirmed that AO narrow focus renditions were rated significantly higher (i.e., as having stronger focus on the critical word) than those renditions produced in the FTF condition [$M_{\text{Diff}} = 0.54$, Sidak 95% CI: 0.32 – 0.76].

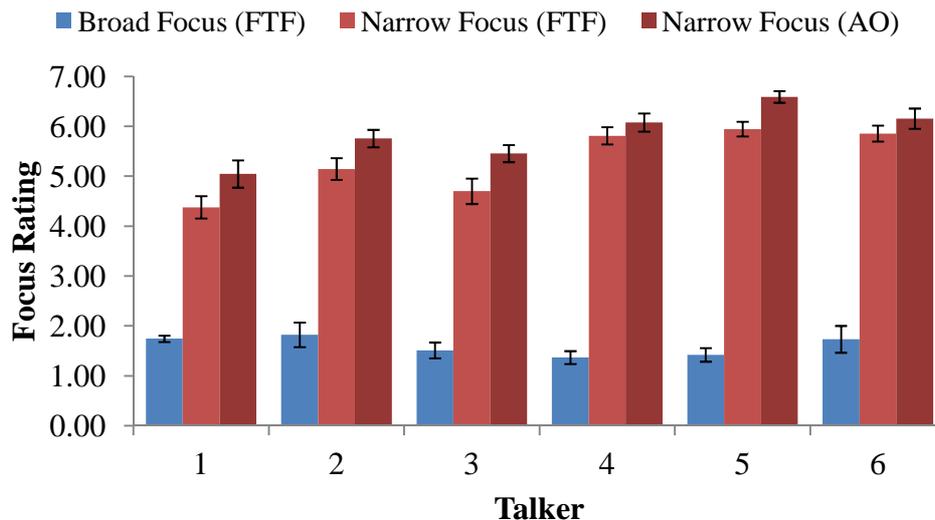


Figure 6.1. Mean ratings of focus (collapsed across sentences and raters) as a function of talker for the AO and FTF conditions. Error bars indicate the standard error of the mean, $n = 90$ observations per column.

To determine whether the differences between talkers occurred, a series of post-hoc repeated measures ANOVAs were conducted separately for each prosodic condition, with talker and sentence as the within-subjects factors in the subject analysis; and talker and rater as within-items factors in the item analysis (both interpreted with a Bonferroni adjusted α of 0.017 for multiple comparisons). No main effect of talkers was observed for ratings of broad focused renditions, $F_S(5,45) = 2.37$, $p = 0.054$, $\eta_p^2 = 0.21$; $F_I(5,40) = 2.11$, $p = 0.085$, $\eta_p^2 = 0.21$; however significant main effects occurred for ratings of FTF narrow focus, $F_S(5,45) = 16.15$, $p < 0.001$, $\eta_p^2 = 0.64$; $F_I(5,40) = 7.49$, $p < 0.001$, $\eta_p^2 = 0.48$; and AO narrow focus

items, $F_S(5,45) = 14.99$, $p < 0.001$, $\eta_p^2 = 0.63$; $F_I(5,40) = 20.31$, $p < 0.001$, $\eta_p^2 = 0.72$. Despite these differences, the pattern of rating data appears to be consistent across all six talkers (i.e., broad focus FTF < narrow focus FTF < narrow focus AO; see Figure 6.1). Furthermore, this pattern of data was also consistent across all nine sentences (see Figure 6.2) and all ten raters (Figure 6.3).

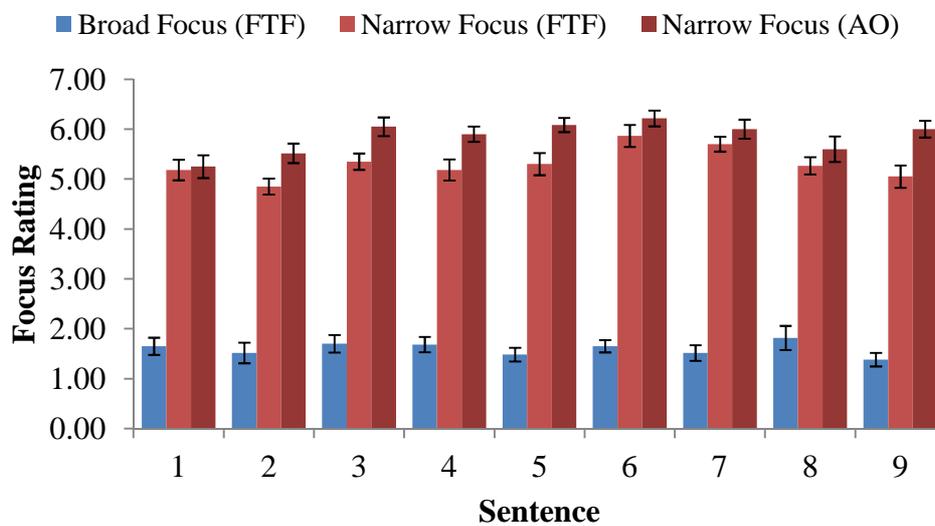


Figure 6.2. Mean ratings of focus (collapsed across talkers and raters) as a function of sentence for the AO and FTF conditions. Error bars indicate the standard error of the mean, $n = 60$ observations per column.

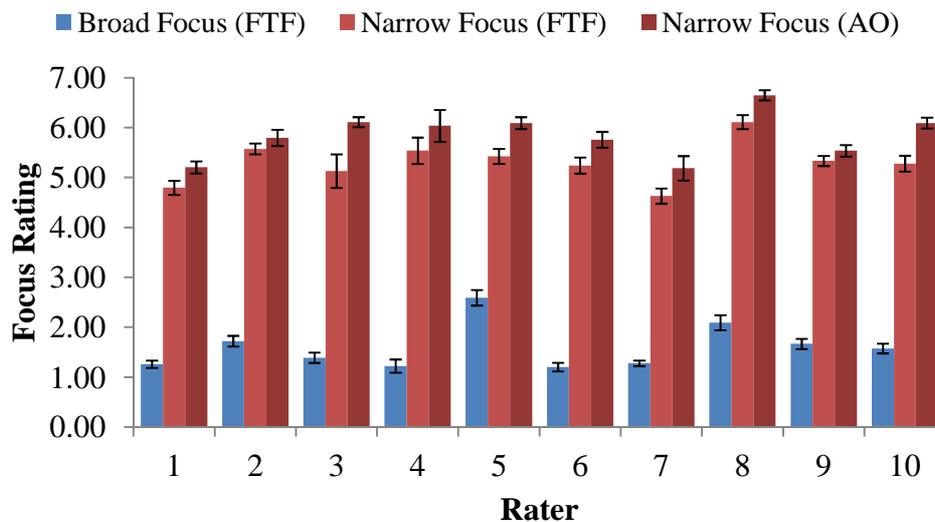


Figure 6.3. Mean ratings of focus (collapsed across talkers and sentences) as a function of rater for the AO and FTF conditions. Error bars indicate the standard error of the mean, $n = 54$ observations per column.

6.1.2.1. Perceptual Rating Scores for Phrasing Contrasts

The mean ratings (collapsed across sentences) of each talker's production of phrasing for the AO and FTF conditions are presented in Figure 6.4. The ANOVA of the phrasing rating task yielded similar outcomes to the focus task. The main effects of prosodic condition, $F_S(2,18) = 953.85$, $p < 0.001$, $\eta_p^2 = 0.99$; $F_I(2,16) = 2512.59$, $p < 0.001$, $\eta_p^2 = 0.99$; and talker $F_S(5.45) = 9.41$, $p < 0.001$, $\eta_p^2 = 0.51$; $F_I(5,40) = 7.40$, $p < 0.001$, $\eta_p^2 = 0.48$, were both statistically significant, as was the interaction, $F_S(10,90) = 19.14$, $p < 0.001$, $\eta_p^2 = 0.68$; $F_I(10,80) = 9.25$, $p < 0.001$, $\eta_p^2 = 0.54$. As expected, utterances phrased as statements were rated significantly lower (i.e., more statement-like) than echoic question productions in both FTF [$M_{\text{Diff}} = 4.54$, Sidak 95% CI: 4.09 – 4.98 – 4.09] and AO [$M_{\text{Diff}} = 4.91$, Sidak 95% CI: 4.47 – 5.34] conditions. The key contrast showed that echoic questions recorded in the FTF

interactive condition were rated significantly lower than those recorded during the AO interaction [$M_{\text{Diff}} = 0.37$, Sidak 95% CI: 0.25 – 0.49].

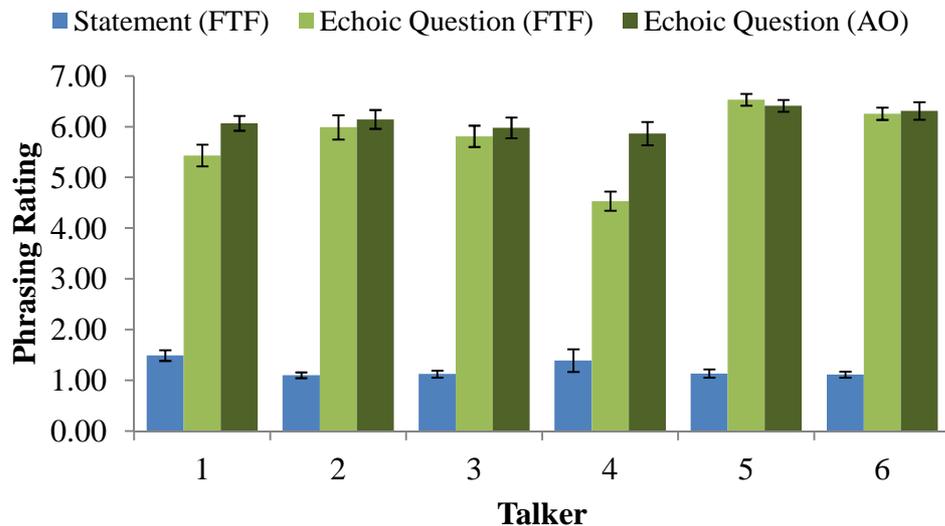


Figure 6.4. Mean ratings of phrasing (collapsed across sentences and raters) as a function of talker for the AO and FTF conditions. Error bars indicate the standard error of the mean, $n = 90$ observations per column.

To further examine whether talkers differed, a series of post-hoc repeated measures ANOVAs were conducted separately for each prosodic condition, with talker and sentence as the within-subjects factors in the subject analysis; and talker and rater as within-items factors in the item analysis (interpreted with a Bonferroni adjusted α of 0.017 for multiple comparisons). No differences were apparent across talkers for ratings of statement renditions in the subject analysis, $F_S(5,45) = 2.57$, $p = 0.040$, $\eta_p^2 = 0.22$; however a difference was found in the item analysis, $F_I(5,40) = 4.10$, $p = 0.004$, $\eta_p^2 = 0.339$. Significant main effects occurred for ratings of FTF echoic questions, $F_S(5,45) = 27.97$, $p < 0.001$, $\eta_p^2 = 0.76$; $F_I(5,40) = 10.25$, $p < 0.001$, $\eta_p^2 = 0.56$; and AO echoic question renditions, $F_S(5,45) = 3.54$, $p = 0.009$, $\eta_p^2 = 0.56$.

$= 0.28$; $F_1(5,40) = 3.65$, $p = 0.008$, $\eta_p^2 = 0.31$. With the exception of Talker 5, the remaining talkers echoic questions produced in AO settings were rated higher compared to FTF settings (see Figure 6.4). As with the ratings of focus, the pattern of data (i.e., declarative statement FTF < echoic question FTF < echoic question AO) was consistent across all nine sentences (see Figure 6.5) and ten raters (Figure 6.6).

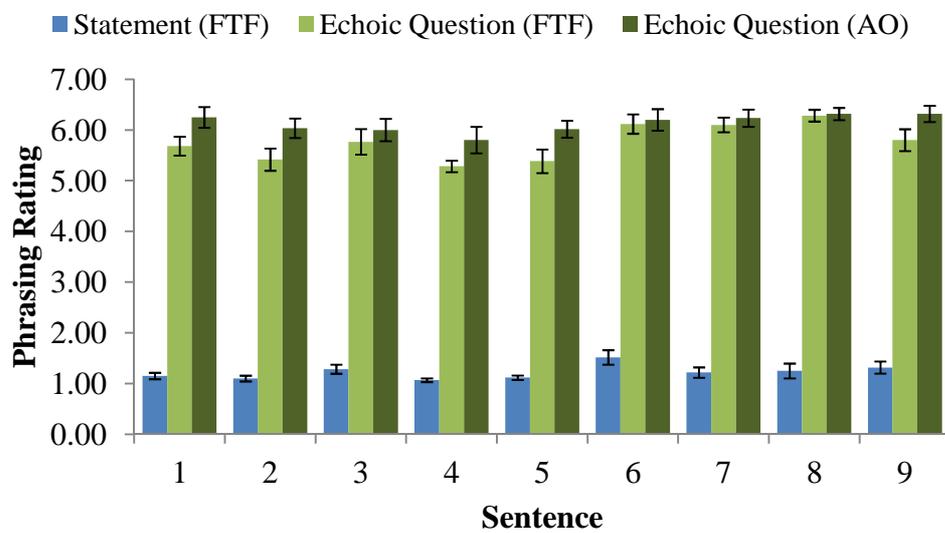


Figure 6.5. Mean ratings of focus (collapsed across talkers and raters) as a function of sentence for the AO and FTF conditions. Error bars indicate the standard error of the mean, $n = 60$ observations per column.

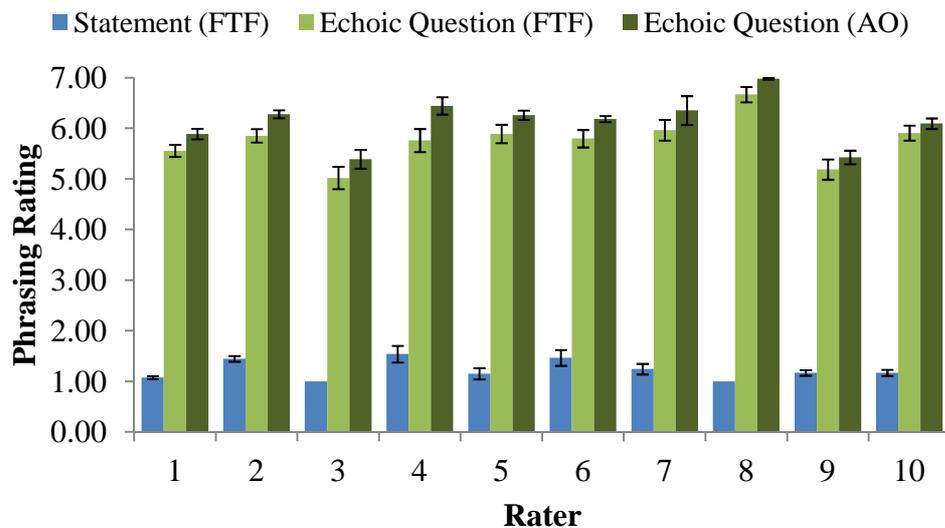


Figure 6.6. Mean ratings of phrasing (collapsed across talkers and sentences) as a function of rater for the AO and FTF conditions. Error bars indicate the standard error of the mean, $n = 54$ observations per column.

6.1.2.2. Regression Analyses between Acoustic and Perceptual Measures

To determine whether any particular acoustic features were able to explain the variance between the AO and FTF ratings, two separate standard multiple regression analyses were performed for the narrow focus and echoic question ratings. The mean subjective rating (across 10 listeners) for each auditory item (for the AO and FTF renditions) was the criterion and the extracted acoustic properties (syllable duration, mean $F0$, $F0$ range, mean intensity and intensity range during pre-critical, critical and post-critical utterance phases; from Chapter 5) were the predictor variables. The predictor variables showed no evidence of multicollinearity, and thus were all included in the regression model.

For the narrow focus ratings, the regression was significantly different from zero, $F(15,92) = 1.99$, $p = 0.024$, with $R = 0.50$, $R^2 = 0.25$, adjusted $R^2 = 0.12$, with

the mean syllable duration of the critical constituent contributing the greatest unique amount of variance explanation [$t(107) = 3.16, p = 0.002, sr^2 = 0.082$], followed by the syllable duration of pre-critical content [$t(107) = 2.57, p = 0.012, sr^2 = 0.054$] and mean intensity of the critical constituent [$t(107) = 2.04, p = 0.045, sr^2 = 0.034$]. Full statistical details are outlined in Table 6.2.

The regression for the echoic question ratings was also significantly different from zero, $F(15,92) = 3.44, p < 0.001$, with $R = 0.60, R^2 = 0.36$, adjusted $R^2 = 0.26$, with the greatest amount of variance being accounted for by the syllable duration of the critical constituent [$t(107) = 2.97, p = 0.004, sr^2 = 0.061$] and the mean $F0$ of post-critical phases [$t(107) = 2.75, p = 0.007, sr^2 = 0.052$]. Statistical values for all other predictors are shown in Table 6.3.

Table 6.2. Standard multiple regression of acoustic features on ratings of narrow focus.

Variable	Utterance Phase	Correlation (<i>r</i>) with Criterion (Mean Rating)	B	β	<i>sr</i>² (unique)
Syllable Length	Pre-Critical	0.153	1.484*	0.369	0.054
	Critical	0.280**	0.948**	0.360	0.082
	Post-Critical	0.085	-0.613	-0.095	0.004
Mean <i>F</i> 0	Pre-Critical	-0.045	-0.336	-0.093	0.001
	Critical	0.148	0.810	0.168	0.011
	Post-Critical	0.079	0.190	0.059	0.002
<i>F</i> 0 Range	Pre-Critical	-0.038	0.016	0.019	0.000
	Critical	0.172	0.029	0.068	0.003
	Post-Critical	-0.003	-0.049	-0.072	0.004
Mean Intensity	Pre-Critical	-0.033	-4.129*	-0.326	0.051
	Critical	0.115	4.043*	0.266	0.034
	Post-Critical	0.049	1.447	0.080	0.004
Intensity Range	Pre-Critical	0.032	-0.378	-0.176	0.015
	Critical	0.026	-0.166	-0.090	0.005
	Post-Critical	0.109	0.762	0.124	0.011
		Constant	1.037		

N = 108, ***p* < 0.01, **p* < 0.05.

Table 6.3. Standard multiple regression of acoustic features on echoic question ratings.

Variable	Utterance Phase	Correlation (<i>r</i>) with Criterion (Mean Rating)	B	β	<i>sr</i> ² (unique)
Syllable Length	Pre-Critical	0.082	0.970	0.170	0.019
	Critical	0.258**	0.855**	0.307	0.061
	Post-Critical	-0.053	-1.336	-0.158	0.020
Mean <i>F</i> ₀	Pre-Critical	0.158	0.640	.200	0.018
	Critical	0.135	-0.646	-0.188	0.013
	Post-Critical	0.375***	0.832**	0.351	0.052
<i>F</i> ₀ Range	Pre-Critical	0.102	-0.027	-0.029	0.001
	Critical	0.200*	0.018	0.054	0.002
	Post-Critical	0.240*	0.015	0.053	0.003
Mean Intensity	Pre-Critical	-0.210*	-2.529*	-0.226	0.030
	Critical	-0.173	-1.244	-0.078	0.003
	Post-Critical	-0.141	-1.196	-0.073	0.003
Intensity Range	Pre-Critical	-0.037	-.0362	-0.192	0.025
	Critical	-0.062	-0.440*	-0.221	0.031
	Post-Critical	0.093	0.362	0.079	0.005
		Constant	9.42***		

N = 108, *** *p* < 0.001, ***p* < 0.01, **p* < 0.05.

In sum, a significant effect of AO vs. FTF setting was found for the subjective ratings. Perceivers rated the renditions that were recorded when the talker could see their conversational partner as less focused (for narrow focus judgments) and less question like (for the echoic question judgments) compared to those recorded when the talker could not. The results of the regression analyses showed that variation in the narrow focus ratings across the AO and FTF conditions was accounted for by properties of the critical word (duration and intensity) and the

length of the pre-critical content, whereas the ratings of question phrasing were predicted by the length of the critical word and post-critical mean F_0 . It should be noted however that the above acoustic properties accounted for only a small amount of the variation in ratings across the AO and FTF settings.

6.1.3. Discussion

This experiment examined the production of narrow focus and echoic questions in an AO compared to a FTF setting. The rationale for comparing these communicative settings stemmed from the Hyper–Hypospeech theory of Lindblom (1990, 1996). The prediction was that the prosody produced in an AO setting would be more salient due to the increased auditory demands imposed on the talker/listener pair by the loss of visual prosody information.

The results of the prosody ratings supported this prediction (with higher ratings of AO utterances indicating that the produced prosody was clearer). Regression analyses examining the links between the AO and FTF ratings and acoustic properties showed that for the narrow focus ratings, the duration of the pre-critical and critical constituents made a contribution, as did the intensity of the critical constituent. For ratings of the echoic questions, the duration of the critical and the mean F_0 of post-critical constituent played a role. However, variation in these properties accounted for only a small amount of the difference between the AO and FTF ratings. One reason for why the acoustic properties do not fully account for the subjective ratings is that the ratings may reflect non-linear combinations of a variety of acoustic properties. Furthermore, the combination of properties that listeners used may have differed across tokens, and the listeners themselves may have varied in what properties they chose to exploit.

The results showed that when speech communication is limited to the auditory channel, talkers took care to make their auditory prosody clearer (compared to a FTF setting). This finding parallels those showing words spoken in FTF settings are less intelligible than those spoken from read lists (e.g., Anderson, Bard, Sotillo, Newlands & Doherty-Sneddon, 1997). However, the issue of precisely which visual cues modulate these changes in prosody may be difficult to pinpoint. Based upon the finding that interlocutors did not look at each other very often, Anderson et al. (1997) suggested that FTF conditions simply provide a global impression as to whether the communication is proceeding without difficulty. On the other hand, measures of direct FTF gaze may underestimate the availability of on-going speech-related information, as it has been shown that considerable visual speech information is available from the visual periphery (Kim & Davis, 2011); a situation more likely to be the case with some of the visual cues for prosody (e.g., large-scale rigid head motions, as in Chapter 2).

CHAPTER 7.
VISUAL ANALYSIS OF THE SPEECH PROSODY
CORPUS

Chapter 7. Visual Analysis of the Speech Prosody Corpus

The auditory analysis described in Chapter 5 indicated that the recorded tokens conformed to the acoustic characteristics of linguistic prosodic contrasts typically described in the literature, suggesting that the experimental task used to elicit the dialogue was effective. The manipulation of the interactive setting resulted in some changes to the produced acoustics, that is, when the talker was unable to see the interlocutor, selected acoustic features of narrow focused and echoic question tokens were produced with a larger degree of contrast from the broad focused baseline. Furthermore, the results of the perceptual measures of the prosodic characteristics (Chapter 6) showed that differences in acoustic measures between the AO and the FTF settings were associated with prosody perception to some extent, suggesting that when speech communication is limited to the auditory channel, talkers took care to make their auditory prosody clearer (compared to a FTF setting). Several questions arise from these results: What are the visual correlates of the prosodic characteristics? Will these visual properties also show differences across the two interactive settings? And what role does visual prosody play in prosodic perception? These questions were investigated in this and the following chapters.

First, in the current chapter, the visual properties associated with the prosodic contrasts were examined. To reiterate, the term “visual properties” is used as a proxy to refer to movements of the talker’s head and face that are likely to be visible to an interlocutor. Previous studies have quantified changes in the amplitude of articulatory gestures such as lip and jaw movements (Dohen et al., 2009), as well as increases in non-articulatory gestures such as eyebrow raises and rigid head

movements accompanying the production of linguistic prosodic contrasts (Cavé et al., 1996; Hadar, Steiner, Grant & Rose, 1983; Løevenbruck, Dohen & Vilain, 2009). For example, Scarborough et al. (2009) examined the spatial properties of visual cues to both lexical and phrasal stress produced by three native talkers of American English using optical tracking. In their analysis, individual markers of interest were examined along the vertical y-axis during the production of a designated syllable in order to identify the peak eyebrow displacement, head displacement, interlip distance (i.e., the distance between the upper and lower lip), as well as chin displacement for both opening and closing gestures (along with associated velocity measures). Differences in the maximal amplitude of each of these measures were then compared between stress conditions. To summarise their results, it was shown that lexically stressed syllables (i.e., dis-CHARGE vs. DIS-charge) when produced in isolation were accompanied by enhanced articulatory gestures and an overall increase in rigid head movements; but eyebrow movements did not differ between the two conditions. By contrast, when target words were placed within an utterance context, words that received phrasal stress (i.e., narrow focus relative to broad focus productions) were accompanied by larger raises of the eyebrows, along with greater rigid head motion, and larger and quicker articulatory gestures.

The analysis conducted by Scarborough and colleagues (2009) highlights the observation that gestures both intrinsically linked to articulation such as jaw and lip opening, as well as non-articulatory movements of the eyebrows can accompany the production of narrow focus. However, correlations between movement parameters were not examined, thus, the individual contribution of each of these movement features to conveying prosody was not determined (e.g., a difference in lip opening

may be driven primarily by changes in jaw opening). Also, such a rudimentary method of reducing the data (down to only single points on the vertical y-axis) may misestimate the magnitude of change in visible movements relative to the rest of the utterance. Furthermore, examination of visual cues only in relation to lexical/phrasal stress is somewhat limited since prosody can affect the interpretation of an utterance at the sentence level, and as such, changes may occur in the speech signal to content surrounding prosodically marked constituents (see Chapter 5).

A more comprehensive approach was employed by Dohen et al. (2006, 2009). Their analysis examined utterance phases that preceded (i.e., pre-critical) and followed (i.e., post-critical) the prosodically marked constituent (in addition to the critical constituent itself). As a measure of visual movements, the area under the amplitude curve over time was estimated for a range of visual speech features (i.e., inter-lip opening, inter-lip width, jaw height and upper lip protrusion, eyebrow raising and rigid head movements) as a function of focus contrasts (for five French talkers). It is worth noting that a normalisation technique was applied to allow for comparisons across segmentally varying sentences and utterances of different lengths. As reviewed in Chapter 3, this procedure allowed Dohen and colleagues to identify differing articulatory strategies corresponding to focus production: what they labelled an absolute contrast pattern, in which a focused word is produced over a longer duration and accompanied by hyperarticulated mouth and jaw movements; and a relative contrast pattern, where the focused constituent in the utterance is still enhanced (but to a smaller degree), with post-focal phases hypoarticulated thereby emphasizing the difference between the focused and unfocused constituents (with this latter articulatory strategy being more commonly used). Non-articulatory

gestures were also produced by some talkers, however their occurrence was not systematic, nor did they consistently accompany prosodically marked constituents. Similarly, correlated rigid head rotations were produced with prosodic focus by only one talker, but varied greatly in terms of amplitude and temporal alignment with the acoustic signal.

The essence of Dohen et al.'s (2006, 2009) findings is that visual changes beyond the segmental boundaries of prosodically marked constituents were associated with prosodic contrasts. On a technical matter, it appeared that comparisons of the normalised area under amplitude curve provide an efficient way to overcome durational and segmental variation across utterances, repetitions and talkers. In the current analysis, a similar approach to that used by Dohen and colleagues was adopted in order to examine both articulatory and non-articulatory gestures accompanying the production of focus and phrasing contrasts.

As with the auditory data of Chapter 5, visual speech properties were also examined as a function of the interactive setting. Given that talkers made adjustments to their acoustic production of prosody when they were unable to see the interlocutor, it was expected that modification to articulatory gestures should also be observed (i.e., enhanced articulatory movements correspond with greater acoustic output, see Edwards et al., 1991; Erickson, 2002; Huber & Chandrasekaran, 2006; Schulman, 1989). Additionally, there is evidence to suggest that talkers modify speech related non-articulatory gestures when they are aware that these will be seen (e.g., see Alibali, Heath & Myers, 2001; Cohen, 1977; Cohen & Harrison, 1973; Mol, Kraemer, Maes & Swerts, 2012). Since movements of the eyebrows and rigid head motion are not strictly tied to articulation, talkers may use such cues to a larger

extent when they know that the interlocutor will be able to see them (i.e., in FTF settings).

As some idiosyncratic talker specific properties were identified in the acoustic data, it was also expected that talkers would differ in the visual realisation of prosody, both in terms of articulatory movements and the use of non-articulatory gestures (as previously shown by Dohen et al., 2006, 2009).

7.1. Data Preparation

7.1.1. Face Shape Normalisation

To allow for comparisons to be made across talkers, each recorded token was first shape-normalized onto an “average head”. To achieve this, a unique motion database was calculated for each individual talker to ascertain the talker-dependant average marker configuration (Figure 7.1, left). The mean of these marker configurations across all talkers was then used to determine the normalised average head model (Figure 7.1, right). For each recorded token, the deviation away from the talker-dependant average marker configuration per frame was calculated and reprojected onto the normalised average head.

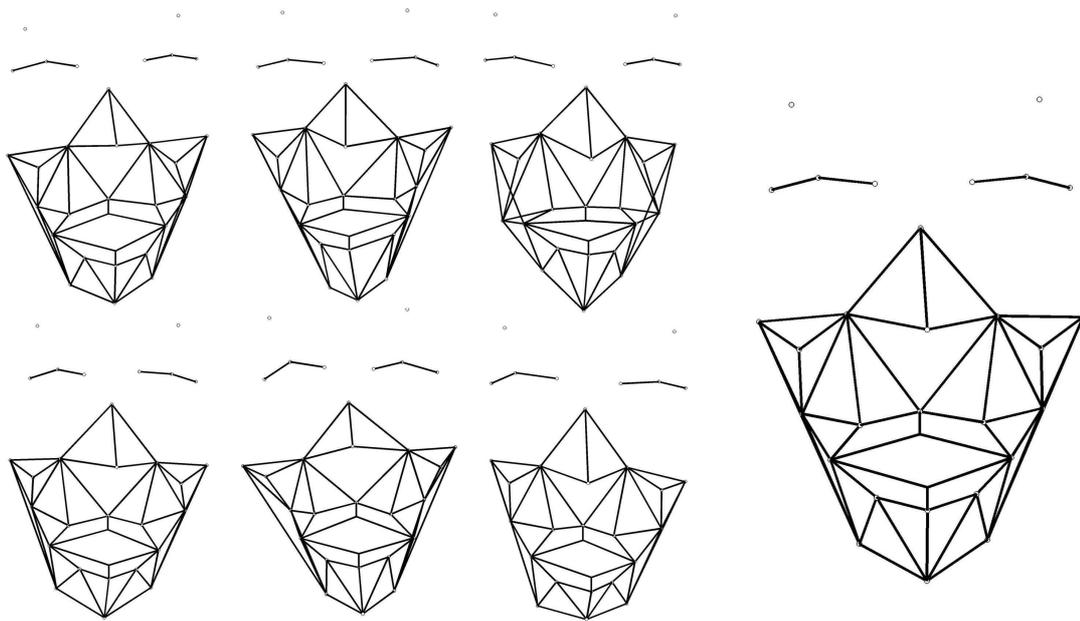


Figure 7.1. The mean talker face shapes (left) were used to generate the normalised average face model (right). Bones (lines between markers) have been added to assist in interpretation.

7.1.2. Dimensionality Reduction

Due to the high dimensionality of the visual motion dataset (i.e., $38 \text{ markers} \times 3$ movement axes per frame of recording at 60fps), dimensionality reduction was performed. The typical approach to achieve this is to apply a principal component analysis (PCA) to the data, deriving optimal orthogonal factors explaining the maximum amount of variance within the least number of components. However, when applied to visual speech data, the extracted components may be biomechanically complex or unfeasible (i.e., including multiple movement features on one component) making interpretation problematic (see Fagel & Madany, 2008; Kim et al., 2011, for a similar argument). An alternative utilised here is *guided principal component analysis* (gPCA; Badin, Bailly, Reveret, Baciú, Segebarth & Savariaux, 2002; Beautemps, Badin & Bailly, 2001; Maeda, 2005), which uses linear

decomposition to extract a set of a priori defined components representing biomechanically plausible articulatory control parameters (however, this may come at the possible cost of sub-optimal variance explanation compared to “standard” PCA). Six components are typically sufficient to explain the majority of articulatory data (see Bailly, Govokhina, Elisei & Breton, 2009), with several additional components specified for non-articulatory expressive gestures such as brow movements.

Thus, the shape-normalised motion data was processed using gPCA to reduce the dimensionality of the data set from 38 three-dimensional coordinates per frame to eight single-dimension non-rigid components, along with three rigid translations and three rigid rotations of the whole head. To minimise the overrepresentation of particular marker configurations (e.g., the neutral position at the start and end of each utterance), a database of unique movements was generated. The six rigid motion parameters around the estimated centre of rotation were determined (using the quaternion method, see Horn, 1987) and extracted from the database (Figure 7.2). The remaining non-rigid movements were then subjected to gPCA using the a priori parameters outlined in Table 7.1. Shape-normalized recordings were then reprojected into principal component (PC) space as deviations away from the average face.

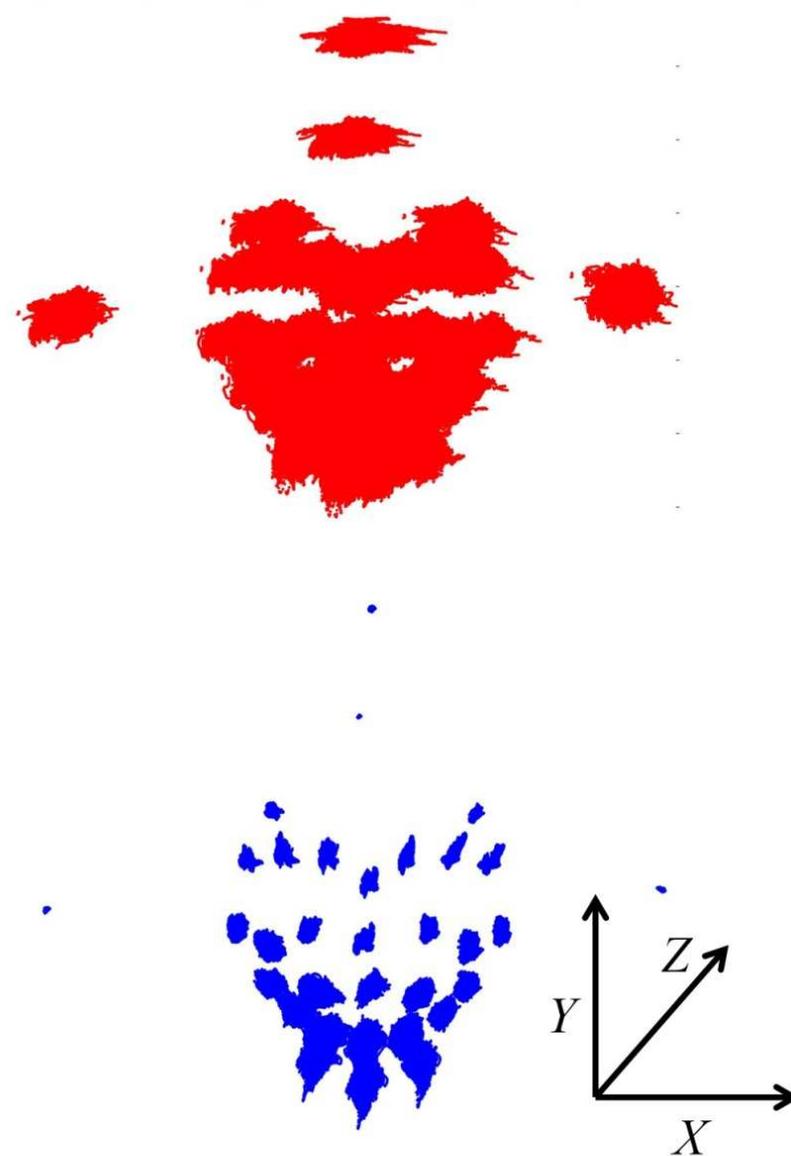


Figure 7.2. A database of unique motion was created (upper panel) to avoid over-representation of particular marker configurations. The rigid movements of the head were then calculated and removed from the database (lower panel). Also shown are the directions of the X , Y , and Z axes.

Table 7.1. A priori non-rigid components used to drive the gPCA, and the rigid movement parameters that were extracted based on rotation and translation around the centre of rotation.

Principal Component	Assigned Label	Axes of Movement	Corresponding Markers
<u>Non-Rigid Movement Parameters (from gPCA)</u>			
1	Jaw Opening	<i>Y</i>	14, 15, 23, 38
2	Lip Opening	<i>Y</i>	24, 25, 26, 27, 28, 29, 30, 31
3	Lower Lip Movement	<i>Y</i>	24, 28, 29, 30, 31
4	Upper Lip Movement	<i>Y</i>	24, 25, 26, 27, 28
5	Lip Rounding	<i>XYZ</i>	24, 25, 26, 27, 28, 29, 30, 31
6	Jaw Protrusion	<i>Z</i>	14, 15, 23, 38
7	Eyebrow Raising	<i>Y</i>	5, 6, 7, 16, 17, 18
8	Eyebrow Pinching	<i>XY</i>	5, 6, 7, 16, 17, 18
<u>Rigid Movement Parameters</u>			
R1	Pitch Rotation	<i>X</i>	1, 2, 3, 4
R2	Roll Rotation	<i>Z</i>	1, 2, 3, 4
R3	Yaw Rotation	<i>Y</i>	1, 2, 3, 4
T1	Fwd / Bwd Translation	<i>Z</i>	1, 2, 3, 4
T2	Left / Right Translation	<i>X</i>	1, 2, 3, 4
T3	Up / Down Translation	<i>Y</i>	1, 2, 3, 4

7.2. Visual Analysis

7.2.1. Guided Principal Component Analysis (gPCA)

Figure 7.3 shows the amount of variance explained by each of the non-rigid principal components from the unique movement database. With only eight non-rigid components, in excess of 96% of face motion variance was able to be recovered. As expected, a large proportion of variance (~65%) is explained by opening and closing

of the jaw and lips. The parameters are visualized in Table 7.2⁹. To objectively evaluate the accuracy of these extracted parameters, the mean euclidian error (i.e., the residual difference between the originally recorded and recovered marker locations) was calculated for the unique movement database, resulting in an average of 0.661 mm across 122215 frames of data. Indeed, some of this error was due to the small proportion of residual variance left unexplained by the extracted components¹⁰.

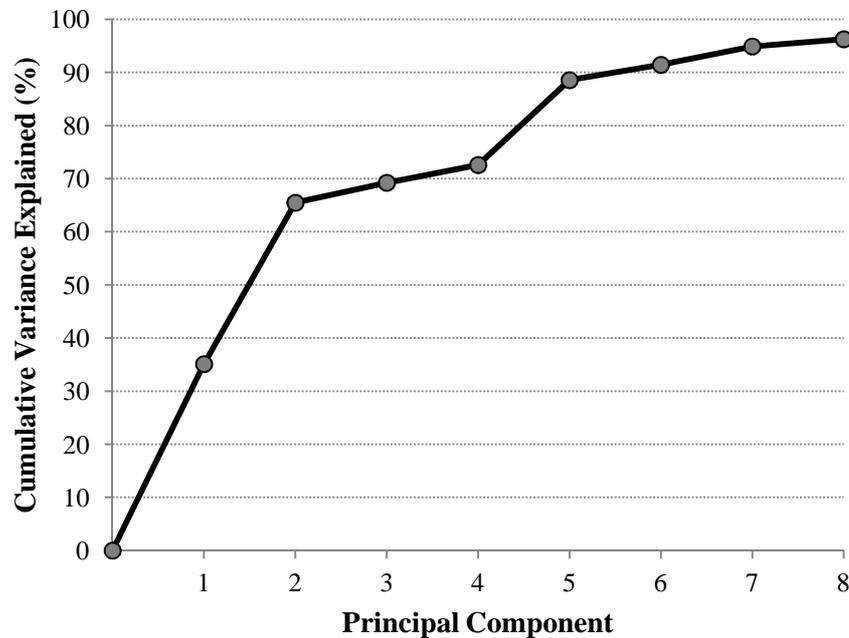
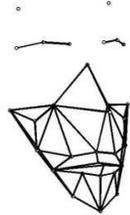
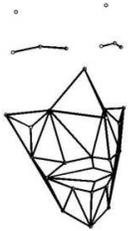
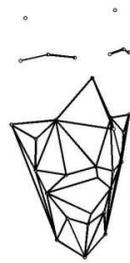
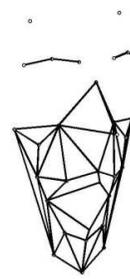
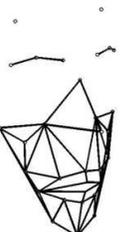
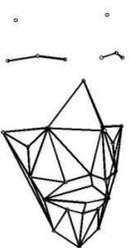
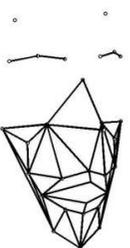


Figure 7.3. Cumulative percent of variance explained by each component from the guided principal components analysis.

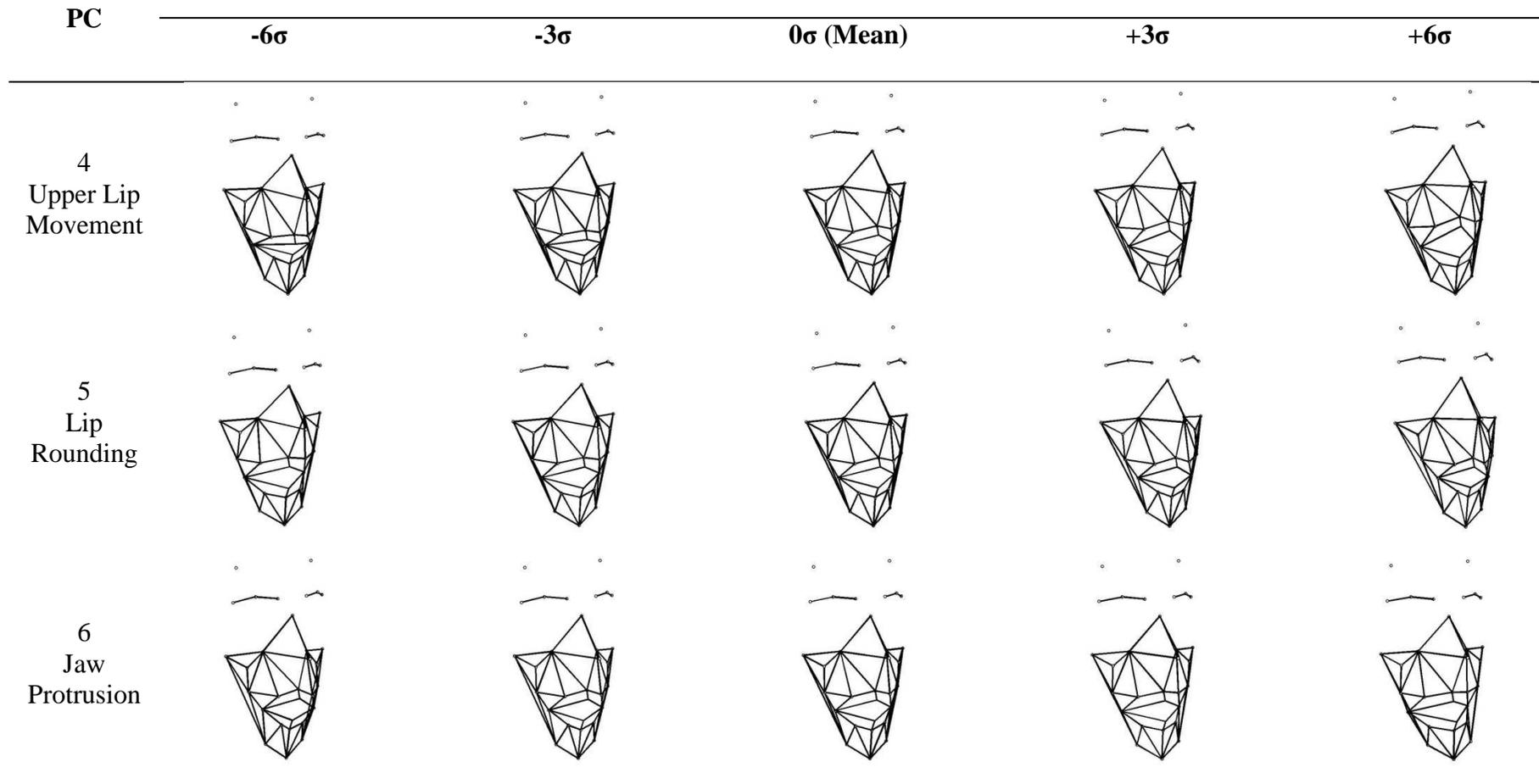
⁹ Animated renditions of the parameters can also be viewed in Appendix B.6.

¹⁰ Indeed, a “standard” PCA could be applied to the remaining variance; however the extracted components would no longer represent biomechanically interpretable gestures.

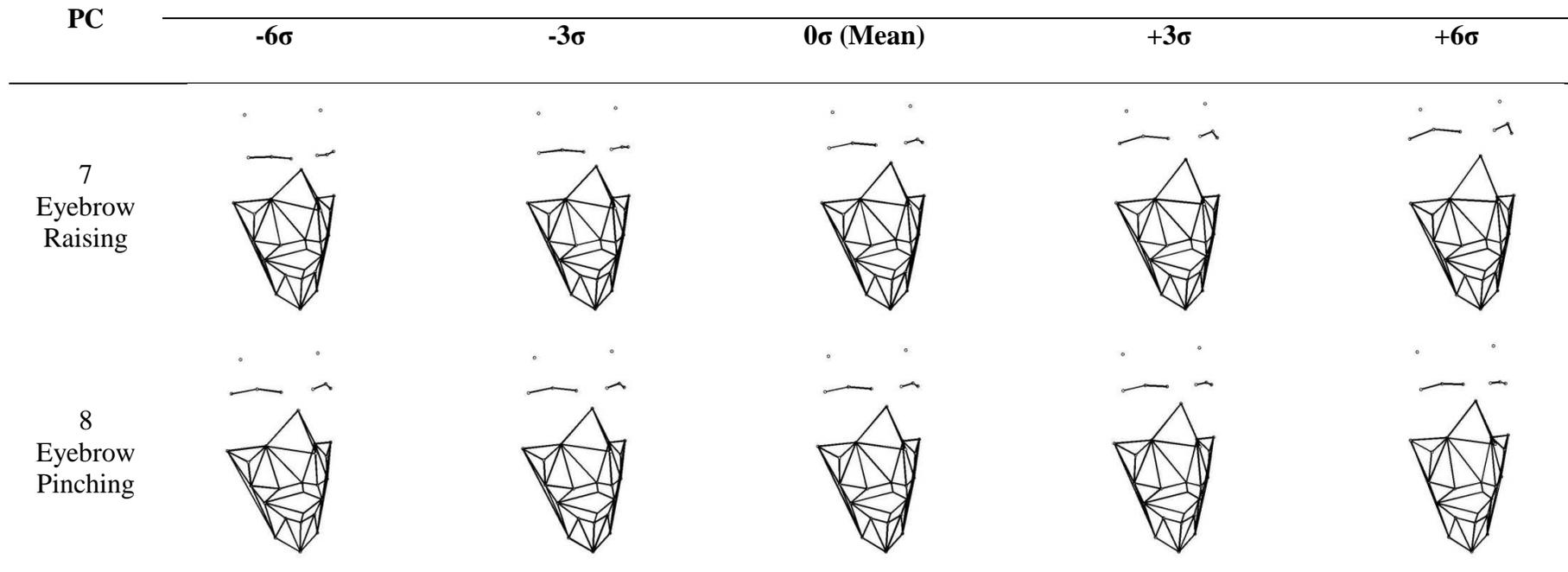
Table 7.2. Visualisation of non-rigid principle components, derived from guided principal components analysis.

PC	Standard Deviations (σ) away from Mean				
	-6σ	-3σ	0σ (Mean)	$+3\sigma$	$+6\sigma$
1 Jaw Opening					
2 Lip Opening					
3 Lower Lip Movement					

Standard Deviations (σ) away from Mean



Standard Deviations (σ) away from Mean



7.2.2. Correlation amongst Extracted Components

Pearson's product-moment correlations were conducted to determine if any relationships were present between extracted principal components and rigid parameters¹¹ for the database of 122215 unique recorded frames. As shown in Table 7.3, the non-rigid movement components extracted using gPCA were uncorrelated. There was some evidence of correlation between rigid pitch rotations (R1, i.e., rotations around the x -axis) and the lip-opening component (PC 2), which may relate to the idea that non-verbal gestures assist in the segmentation of speech signals (see Davis & Kim, 2006; Munhall et al., 2004), as this would result in such movements occurring with some synchronicity.

Table 7.3. Pearson's r correlation values within the principal components extracted using gPCA and rigid parameters for the unique movement database.

	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	R 1	R 2
PC 1	-.001	-.001	-.005	-.001	-.005	.004	.005	.384	.151
PC 2		.011	-.027	-.009	-.004	-.008	.003	.553	.165
PC 3			-.106	.070	-.074	.007	.038	-.012	.033
PC 4				-.055	.057	.005	-.037	.055	-.058
PC 5					-.044	.008	.019	.086	-.158
PC 6						-.023	-.028	.047	-.085
PC 7							.005	-.016	.059
PC 8								.154	.037
R 1									.015

¹¹ Yaw rotations (R 3) and translational movements (T 1 to 3) were not considered, as these more likely reflect postural changes by the talker while seated rather than being related to speech production.

7.2.3. Area under PC Amplitude Curves¹²

7.2.3.1. Time Normalisation

To allow for comparisons across repetitions, talkers, sentences and prosodic conditions for visual parameters, each recorded utterance was time-normalised to 1.2 times the longest utterance rendition (per sentence across talkers, prosodic conditions and repetitions) using linear spline interpolation in Matlab, and projected onto a new time series from 0 to 3 (see Figure 7.4). As can be seen from the figure, the normalisation procedure changes the overall length but not the characteristic shape of the components over time, so that comparisons made are based on the shape of the curve.

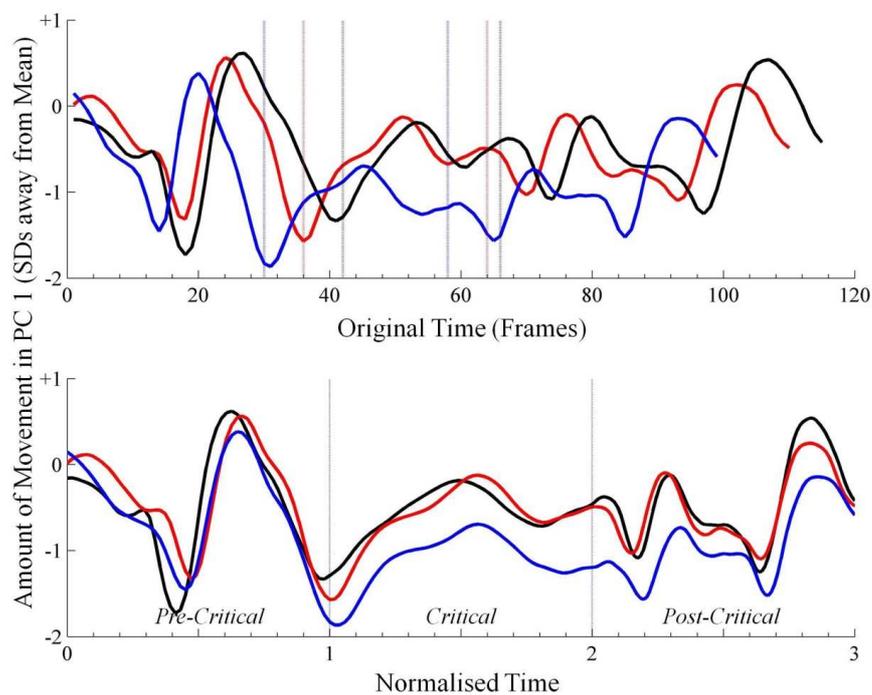


Figure 7.4. Linear spline interpolation was used to normalisation the time of each recorded token, allowing for comparisons across talkers and repetitions to be made.

¹² A preliminary version of this data analysis appeared in: Cvejic, E., Kim, J., Davis, C. & Gibert, G. (2010). Prosody for the eyes: Quantifying visual prosody using guided principal components analysis. *Interspeech 2010*, pp. 1433-1436.

7.2.3.2. Area Calculation and Normalisation

The area under the time-normalised non-rigid principal component (PCs 1 to 8) and rigid rotation parameter curves (R 1 and 2) over time were calculated for each utterance phase (i.e., pre-critical, critical and post-critical) of each recorded token. As the mathematical function of the principal component curves was unknown, the trapezoidal rule was used to estimate the definite integral. This estimation technique fits a series of linear functions between consecutive frames of the utterance, and calculates the area of the generated trapezoid; the sum of these areas for the total length of the utterance gives an accurate estimate of the area under the curve. This procedure generated three values for each token, representing the summed amplitude of the particular parameters for each utterance phase (i.e., pre-critical, critical and post-critical). The mean of the broad focused renditions for each sentence (per talker and interactive condition) were then calculated, and used to normalise the area values for the narrow focus and echoic question renditions (in their respective interactive conditions) relative to the broad focused rendition. As with the auditory analysis (Chapter 5), a value of 1 indicates no difference relative to the broad focused statement rendition, a value greater than 1 corresponds to an increased amount of movement, whereas a value below 1 indicates a reduction in the amplitude of movements. These values for each principal component (PC) for narrow focus and echoic question renditions are displayed in Figure 7.5 and 7.6.

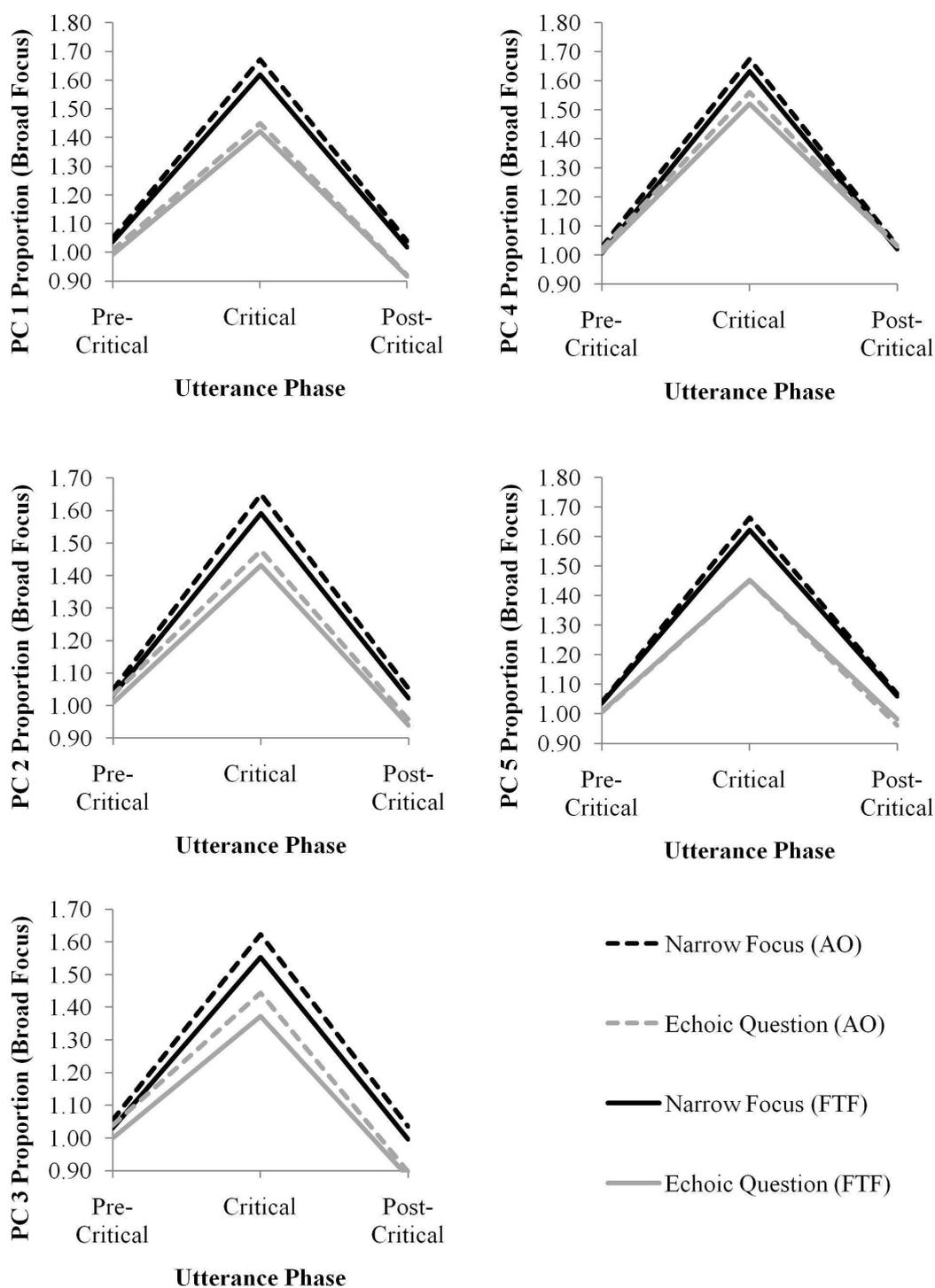


Figure 7.5. Mean area under curve (collapsed across talkers and sentences) of principal components 1 to 5, as a function of prosodic condition represented as a proportion of the broad focused rendition.

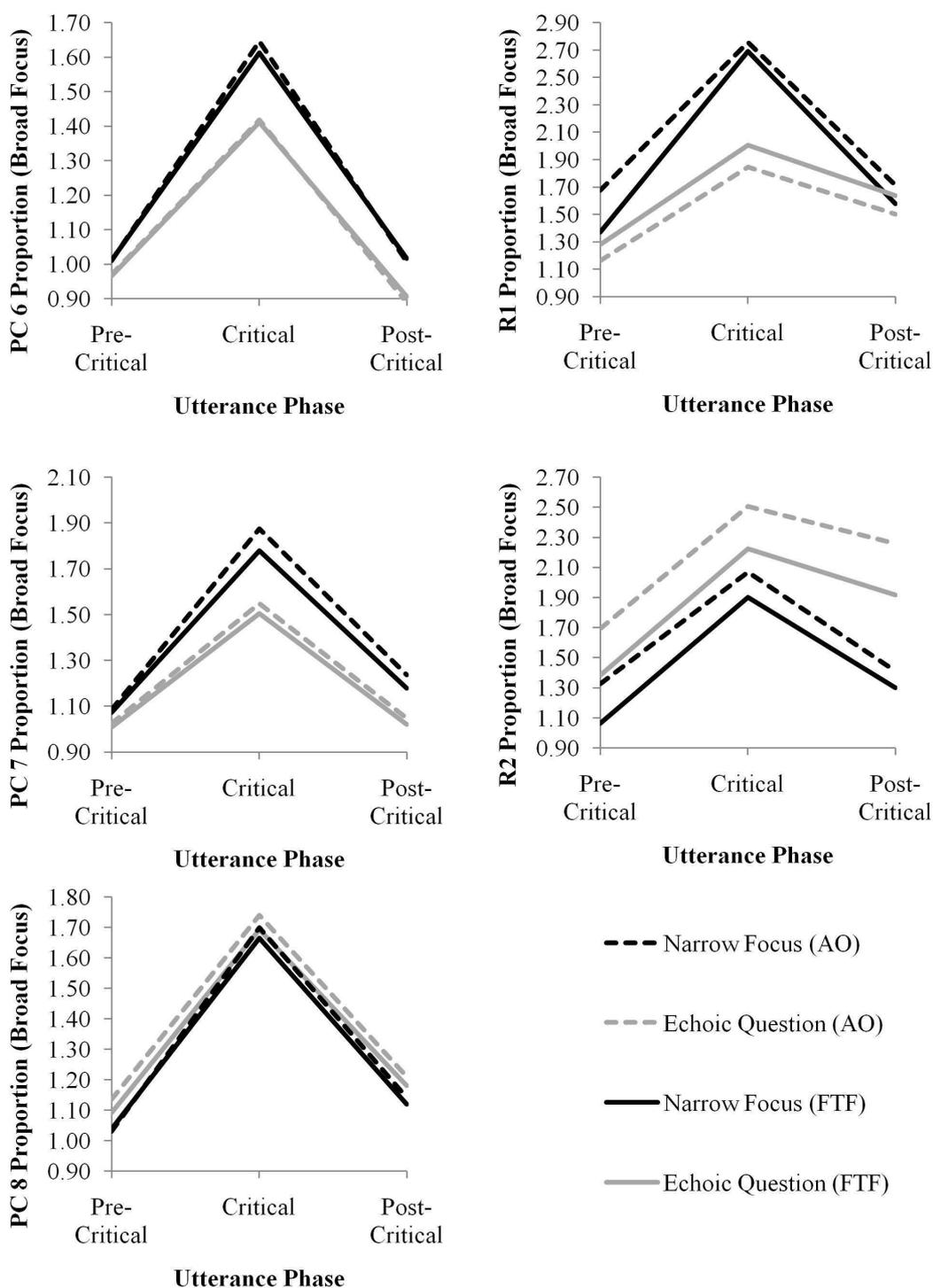


Figure 7.6. Mean area under curve (collapsed across talkers and sentences) of principal components 6 to 8, and R1 and R2, as a function of prosodic condition represented as a proportion of the broad focused rendition.

The resulting values were compared for each principal component at each utterance phase (pre-critical; critical; post-critical) in a series of mixed repeated measures ANOVAs comparing the broad focus and narrow focus renditions (in the AO interactive condition), and the broad focus and echoic question productions (in the AO interactive condition) to determine the visual correlates of realising the prosodic contrasts with prosody as the within-items factor, and utterance type (S/E; S/L; L/E; L/L) as a between subjects factor. Due to the number of analyses conducted, a conservative α level of 0.0125 was selected. For each comparison, an analysis was conducted for both the sentence data (item analysis, F_I ; collapsed across talkers) and talker data (subject analysis, F_S ; collapsed across sentences). The results of these analyses are summarised in the text, however full statistical details can be found in Appendix C.2.

7.2.3.3. Visual Realisation of Prosodic Contrasts

In considering the visual correlates of prosodic focus and phrasing contrasts, the main effect of prosody in the sentence and talker analyses were examined. To streamline reporting, the significant statistical values are presented in Table 7.4. For prosodic focus contrasts, the critical phases of narrow focused tokens (relative to broad focus renditions) were accompanied by an overall increase in jaw opening (PC 1), lip opening (PC 2), lower lip (PC 3) and upper lip (PC 4) movement. Narrow focused constituents (relative to the same content produced in a broad focused context) were also produced with greater lip rounding (PC 5), more jaw protrusion (PC 6), substantially more eyebrow raising (PC 7), and an increase in eyebrow pinching (PC 8). In sum, all eight non-rigid principal components showed an increase for the critical constituent in narrow focused relative to broad focus

renditions across both item and subject analyses. Pre-critical and post-critical utterance content showed no consistent differences between the focus contrasts.

Table 7.4. Significant main effects of prosody for the critical constituent during the production of prosodic focus contrasts.

Feature	Analysis Source*	<i>F</i> Value	<i>p</i> Value	η_p^2
PC 1	Sentence	428.74	< 0.001	0.943
	Talker	42.54	< 0.001	0.431
PC2	Sentence	390.87	< 0.001	0.938
	Talker	44.08	0.001	0.898
PC3	Sentence	264.80	< 0.001	0.911
	Talker	28.64	0.003	0.851
PC4	Sentence	231.12	< 0.001	0.891
	Talker	22.18	0.005	0.816
PC5	Sentence	356.39	< 0.001	0.932
	Talker	42.59	0.001	0.895
PC6	Sentence	261.61	< 0.001	0.910
	Talker	34.86	0.002	0.875
PC7	Sentence	412.09	< 0.001	0.941
	Talker	23.54	0.005	0.825
PC8	Sentence	326.46	< 0.001	0.926
	Talker	55.31	0.001	0.917

*Note: Sentence analyses (F_I) were interpreted with 1 between and 26 error degrees of freedom; the talker analyses (F_S) were interpreted with 1 within and 5 error degrees of freedom.

In the realisation of phrasing contrasts, similar effects were observed (see Table 7.5 for statistical values). Relative to broad focused renditions, the critical constituent of echoic questions were produced with greater jaw movement (PC 1)

and lip openings (PC 2), increased lip rounding (PC 5), and greater jaw protrusion (PC 6). Non-articulatory gestures also differed, with greater brow pinching (PC 8) and more rigid pitch rotations (R 1) of the head during production of the critical constituent of echoic questions compared to broad focused statement renditions.

Table 7.5. Significant effects of prosody for the phrasing contrasts for the critical constituents.

Feature	Analysis Source*	<i>F</i> Value	<i>p</i> Value	η_p^2
PC 1	Sentence	226.09	< 0.001	0.897
	Talker	18.44	0.008	0.787
PC2	Sentence	240.49	< 0.001	0.902
	Talker	31.52	0.002	0.863
PC5	Sentence	236.13	< 0.001	0.901
	Talker	94.42	< 0.001	0.950
PC6	Sentence	210.88	< 0.001	0.890
	Talker	17.07	0.009	0.773
PC8	Sentence	252.89	< 0.001	0.907
	Talker	30.44	< 0.001	0.845
R1	Sentence	43.47	< 0.001	0.626
	Talker	20.11	0.006	0.801

* Sentence analyses (F_T) were interpreted with 1 between and 26 error degrees of freedom; the talker analyses (F_S) were interpreted with 1 within and 5 error degrees of freedom.

These results suggest that non-articulatory gestures along with movements of the articulators both appear to be involved in differentiating prosodic contrasts to some degree; a greater amount of eyebrow raising was produced with narrow focus tokens (relative to broad focus), whereas an increase in brow pinching and rigid pitch rotations quantitatively differentiated echoic questions from statement renditions.

7.2.3.4. *Effects of Utterance Type*

The significant interactions between prosody and utterance type in the sentence analysis were examined further using a series of post-hoc, univariate between-subjects ANOVAs (with utterance type as the between-items factor) individually for each PC and utterance phase. Sidak pairwise comparisons (with 98.75% confidence intervals) were used to identify where significant differences occurred between utterance types.

7.2.3.4.1. *Effect of utterance type on focus contrasts*

For focus contrasts, there were no significant differences between utterance types for pre-critical phases or critical constituents. That is, regardless of the location of the critical constituent, or its position within the utterance, the realisation of pre-focal and focused content was consistent. However, effects of utterance type were found for post-critical phases across all eight non-rigid principal components, and the rigid pitch rotation (R 1) parameter; the statistical values are represented in Table 7.6.

For jaw opening (PC 1), S/E utterances had significantly less jaw movement than S/L utterances [$M_{\text{Diff}} = 0.23$, Sidak 98.75% CI: 0.10 - 0.35], as did L/E utterances compared to L/L utterances [$M_{\text{Diff}} = 0.13$, Sidak 98.75% CI: 0.01 - 0.25]. Thus, when the critical constituent occurred in the first half of the utterance (as in S/E and L/E utterances), there appears to be a post-focal *reduction* in jaw movement compared to broad focused renditions, whereas the opposite is true when the critical constituent occurs in the latter half of the utterance (see Figure 7.7).

Similar effects were observed for post-focal mouth opening (PC 2): S/E utterances had less mouth movement than S/L utterances [$M_{\text{Diff}} = 0.20$, Sidak 98.75% CI: 0.10 - 0.30]; L/E utterances showed less movement than L/L utterances

[$M_{\text{Diff}} = 0.12$, Sidak 98.75% CI: 0.03 - 0.21]; while S/L had greater movement than L/L utterances [$M_{\text{Diff}} = 0.12$, Sidak 98.75% CI: 0.01 - 0.22]. As with jaw movement, the earlier the critical constituent occurred, the less mouth movement was apparent in post-focal utterance phases. A similar pattern was observed during post-focal utterance phases for lower lip movement (PC 3), upper lip movement (PC 4), and jaw protrusion (PC 6). When the narrowly focused critical constituent occurred early in the utterance, there was a reduction in movement of each of these parameters relative to broad focus, whereas the converse was observed when the critical constituent occurred late in the utterance (i.e., there was an increase in overall movement).

Post-focal content in S/E utterances were also produced with less lip rounding (PC 5) than in S/L utterances [$M_{\text{Diff}} = 0.17$, Sidak 98.75% CI: 0.07 - 0.26], and less in L/E than L/L utterances [$M_{\text{Diff}} = 0.13$, Sidak 98.75% CI: 0.04 - 0.22]. When the critical constituent occurred late in the utterance, there was an increase in post-focal lip rounding (relative to broad focus renditions); this however was not the case when the critical constituent occurred in the first half of the utterance (see Figure 7.7).

In terms of non-articulatory gestures, the eyebrow raising component (PC 7) and eyebrow pinching component (PC 7) also showed differences during post-critical phases between utterance types. The post-critical phase of S/E utterances were produced with less eyebrow raising than S/L utterances [$M_{\text{Diff}} = 0.22$, Sidak 98.75% CI: 0.04 - 0.40]. This effect was also mirrored for eyebrow pinching, with post-critical phase of S/E utterances produced with less than S/L utterances [$M_{\text{Diff}} = 0.23$, Sidak 98.75% CI: 0.02 - 0.45]. Finally, although the rigid pitch rotation

showed a significant effect of utterance type for narrow focused renditions, no pairwise comparisons were significant.

Table 7.6. Significant interactions between prosody and utterance type for the post-critical utterance phase of narrow focus tokens. Analyses were interpreted with 3 between, and 26 error degrees of freedom.

Feature	<i>F</i> Value	<i>p</i> Value	η_p^2
PC 1	17.34	< 0.001	0.667
PC 2	23.91	< 0.001	0.734
PC 3	10.76	< 0.001	0.554
PC 4	8.47	< 0.001	0.494
PC 5	19.22	< 0.001	0.689
PC 6	5.31	0.005	0.380
PC 7	9.23	< 0.001	0.516
PC 8	6.37	0.002	0.424
R 1	4.50	0.011	0.342

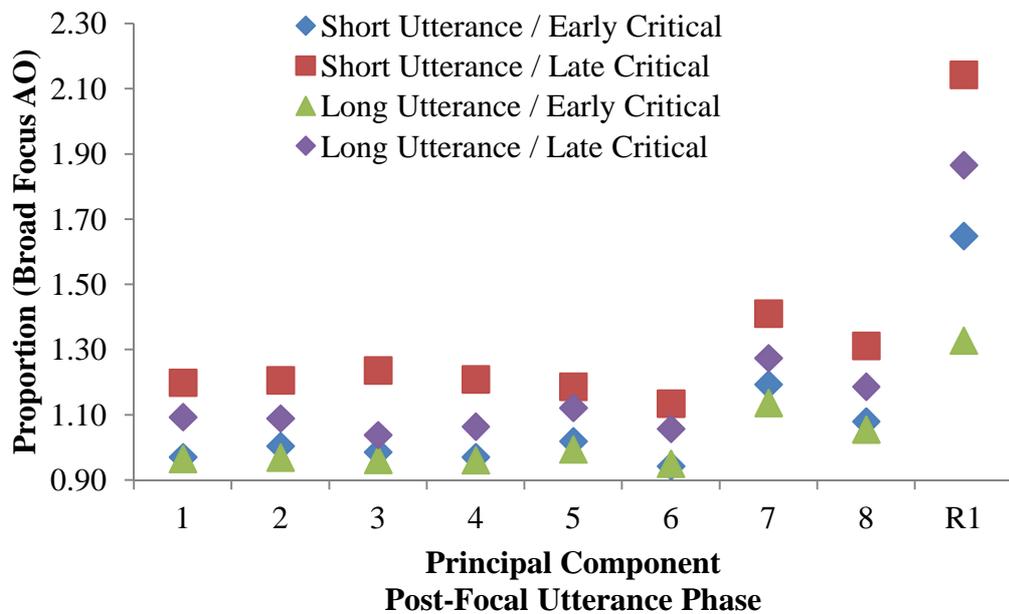


Figure 7.7. Proportion values (relative to broad focus renditions) for post-focal utterance phases in narrow focus renditions, as a function of utterance type.

7.2.3.4.2. *Effect of utterance type on phrasing contrasts*

For phrasing contrasts, jaw opening (PC 1) and jaw protrusion (PC 6) differed as a function of utterance type during pre-critical phases (see Figure 7.8). For jaw opening (PC 1), $F_1(3,26) = 4.55$, $p = 0.011$, $\eta_p^2 = 0.344$, there was a tendency for echoic questions with late-occurring critical constituents to be produced with a reduction in jaw movement (relative to broad focused renditions), whereas an increase in jaw movement was observed when the critical constituent occurred in the first half of the utterance. In particular, the pairwise comparisons revealed that L/E utterances showed significantly greater jaw movement in pre-critical phases than L/L utterances [$M_{\text{Diff}} = 0.12$, Sidak 98.75% CI: 0.03 - 0.22]. This pattern was replicated for jaw protrusion (PC 6), $F_1(3,26) = 4.88$, $p = 0.008$, $\eta_p^2 = 0.360$, with L/E utterances displaying significantly more jaw protrusion than L/L utterances during pre-critical utterance phases [$M_{\text{Diff}} = 0.13$, Sidak 98.75% CI: 0.02 - 0.24].

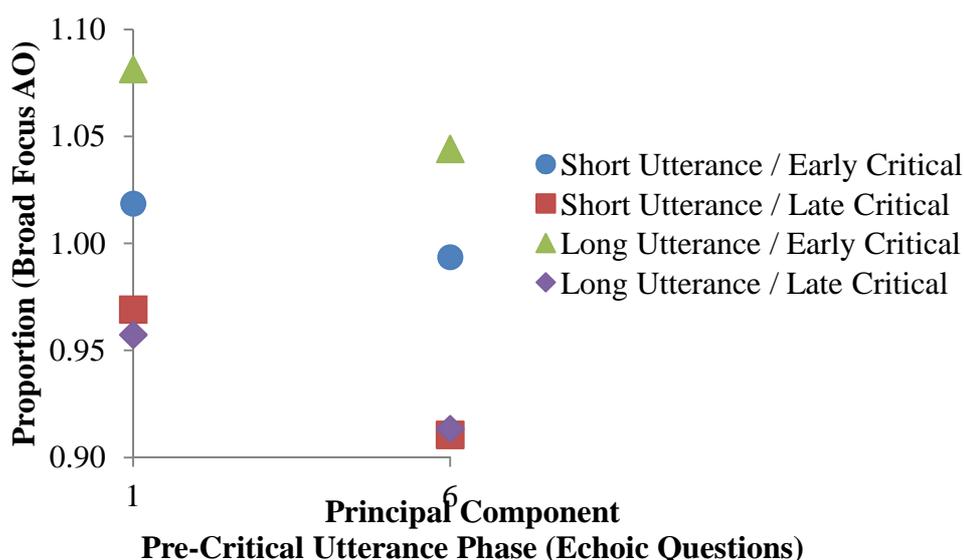


Figure 7.8. Proportion values (relative to broad focus renditions) for pre-critical utterance phases for echoic questions, as a function of utterance type.

Utterance type also had a significant effect for echoic questions during post-critical phases for PCs 1 through to 6 (see Table 7.7 for statistical values). A reduction in jaw movement (PC 1) was observed for echoic questions when the critical constituent occurred earlier on in the utterance; this was reflected by significant pairwise differences between S/E and S/L utterances [$M_{\text{Diff}} = 0.18$, Sidak 98.75% CI: 0.08 - 0.27]; and between L/E and L/L [$M_{\text{Diff}} = 0.09$, Sidak 98.75% CI: 0.00 - 0.18]. This pattern was mirrored for lip opening (PC 2), with greater lip opening in post-critical question content when the critical constituent occurred early on in the utterance (compared to in the later half) for both short [$M_{\text{Diff}} = 0.16$, Sidak 98.75% CI: 0.08 - 0.24] and long utterances [$M_{\text{Diff}} = 0.08$, Sidak 98.75% CI: 0.00 - 0.16].

For lower lip movement (PC 3), both S/E [$M_{\text{Diff}} = 0.02$, Sidak 98.75% CI: 0.08 - 0.26] and L/L [$M_{\text{Diff}} = 0.11$, Sidak 98.75% CI: 0.01 - 0.20] utterances were produced with less movement than S/L utterances. A similar pattern was found for upper lip movement (PC 4); when the critical constituent of echoic questions occurred in the first half the utterance, upper lip movement was the same as for statement renditions, however, when the critical constituent occurred later (particularly when the utterance was short), there was more post-focal lip movement. Both S/E [$M_{\text{Diff}} = 0.21$, Sidak 98.75% CI: 0.08 - 0.34] and L/L utterances [$M_{\text{Diff}} = 0.17$, Sidak 98.75% CI: 0.04 - 0.311] showed significantly less movement than S/L utterances (see Figure 7.9).

For lip rounding (PC 5) and jaw protrusion (PC 6), S/E utterances were produced with a reduction in movement than S/L utterances [PC 5: $M_{\text{Diff}} = 0.14$, Sidak 98.75% CI: 0.02 - 0.25; PC 6: $M_{\text{Diff}} = 0.14$, Sidak 98.75% CI: 0.03 - 0.25].

Table 7.7. Significant interactions between prosody and utterance type for the post-critical utterance phase of echoic question tokens. Analyses were interpreted with 3 between, and 26 error degrees of freedom.

Feature	<i>F</i> Value	<i>p</i> Value	η_p^2
PC 1	10.93	< 0.001	0.558
PC 2	12.28	< 0.001	0.586
PC 3	10.32	< 0.001	0.544
PC 4	7.12	0.001	0.451
PC 5	4.84	0.008	0.358
PC 6	4.74	0.009	0.353

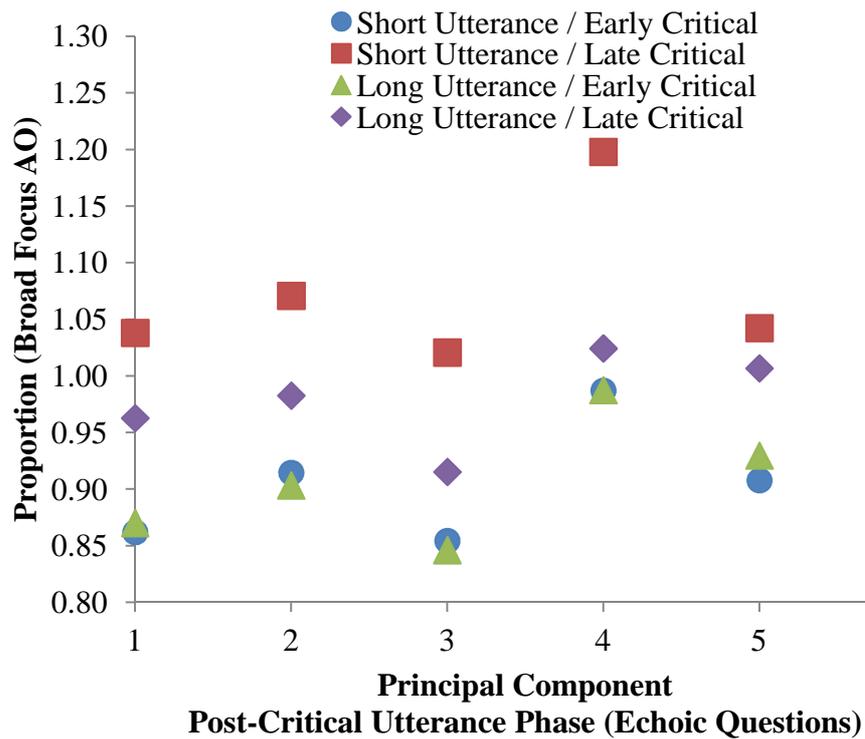


Figure 7.9. Proportion values (relative to broad focus renditions) for post-focal utterance phases in echoic question renditions, as a function of utterance type.

7.2.3.5. Differences across Interactive Settings

A series of repeated measures ANOVAs comparing the production of the prosodically marked tokens (i.e., narrow focus and echoic questions) between the AO and FTF interactive settings were also conducted (with interactive setting treated as a within-items factor) in order to determine whether the visual realisation of prosody varied as a function of whether or not the talker could see the interlocutor. For each comparison, two analyses were conducted, one for the sentence data (F_I) and for the talker data (F_S). The statistical summaries of these comparisons are shown in Table 7.8 (for focus contrasts) and Table 7.9 (for phrasing contrasts).

For narrow focused tokens, a number of components showed an increased amount of movement in the AO setting (i.e., when the talker could not be seen) compared to the FTF one. However, these effects were only significant in the sentence analysis (i.e., collapsed across talkers), not in the talker analysis. For pre-critical utterance phases, there was more movement in the AO narrow focus renditions than in the FTF ones for both rigid pitch rotations (R 1) and rigid roll rotations (R 2). For the critical constituent of narrowly focused utterances, there was more movement in the AO than FTF setting for lip opening (PC 2) and eyebrow raising (PC 7). These effects were maintained in post-critical phases, with lip opening (PC 2), eyebrow raising (PC 7), *and* lower lip movement (PC 3) greater in the AO than FTF renditions.

Fewer movement differences were found between the interactive conditions for the production of echoic questions, and as for the narrow focus comparisons, these differences were only secure in the sentence analysis. Pre-critically, echoic questions produced in the AO setting had a greater amount of lower lip movement

(PC 3) as well as more eyebrow pinching (PC 8) than recordings made in the FTF rendition. There was also an increase for the critical constituent within echoic question renditions in the AO recordings for lower lip movement (PC 3). Post-critically, the AO recordings of echoic questions were produced with greater movement relative to FTF settings in terms of lip opening (PC 2) and eyebrow pinching (PC 8).

Table 7.8. Main effects of interactive setting on the production of narrow focus.

Utterance Phase	Feature	Analysis Source*	<i>F</i> Value	<i>p</i> Value	η_p^2
Pre-Critical	R 1	Sentence	9.71	0.004	0.251
		Talker	1.79	0.238	0.264
Critical	R 2	Sentence	9.82	0.004	0.253
		Talker	0.34	0.586	0.064
	PC 2	Sentence	9.20	0.005	0.241
		Talker	5.29	0.070	0.514
Post-Critical	PC 7	Sentence	14.18	0.001	0.328
		Talker	3.02	0.143	0.377
	PC2	Sentence	14.41	0.001	0.332
		Talker	9.21	0.029	0.648
	PC3	Sentence	10.37	0.003	0.263
		Talker	1.77	0.241	0.261
PC7	Sentence	21.76	< 0.001	0.429	
	Talker	3.26	0.131	0.395	

* Sentence analyses (F_I) were interpreted with 1 between and 29 error degrees of freedom; the talker analyses (F_S) were interpreted with 1 within and 5 error degrees of freedom.

Table 7.9. Main effects of interactive setting on the production of echoic questions.

Utterance Phase	Feature	Analysis Source*	<i>F</i> Value	<i>p</i> Value	η_p^2
Pre- Critical	PC 3	Sentence	13.31	0.001	0.315
		Talker	1.18	0.327	0.191
Critical	PC 8	Sentence	11.67	0.002	0.287
		Talker	1.33	0.301	0.210
Critical	PC 3	Sentence	9.53	0.004	0.247
		Talker	5.70	0.063	0.532
Post- Critical	PC2	Sentence	11.62	0.002	0.286
		Talker	1.52	0.273	0.233
Critical	PC8	Sentence	8.04	0.008	0.217
		Talker	1.83	0.234	0.268

* Sentence analyses (F_I) were interpreted with 1 between and 29 error degrees of freedom; the talker analyses (F_S) were interpreted with 1 within and 5 error degrees of freedom.

Although some modifications to the visual signal were observed across interactive settings, these did not appear to be produced consistently across talkers. Indeed, some of these modifications likely relate to the changes necessary to shape the acoustic signal, for example increases in lip opening (PC 2) during the production of the narrowly focused critical constituent may correspond to the increased mean intensity difference reported in Chapter 5. Similarly, the increased mean intensity observed for post-critical phases likely corresponds to the increase in lip opening observed in the visual analysis.

Although differences in visual articulatory parameters may be accounted for by considering acoustic changes as a function of the interactive setting, it is interesting that some non-articulatory gestures that contribute little to the production of the acoustic signal (i.e., eyebrow and rigid head movements) were also found to

be more prominent in the AO setting, despite that these were not able to be seen by the interlocutor. It does however appear to be common for people to still produce gestures despite not seeing with who they are conversing. For example, according to Bavelas, Gerwing, Sutton and Prevost (2008) when talking on the telephone talkers produce a range of hand gestures (although to a lesser extent than when FTF with a conversational partner). Krauss and colleagues (Krauss, Chen & Chawla, 1996; Krauss, Dushay, Chen & Rauscher, 1995; Krauss, Morrel-Samuels & Colasante, 1991) propose that these gestures still occur because they assist the talker with speech production processes (i.e., assist in achieving lexical access). A similar interpretation could be made with regards to the production of non-articulatory visual gestures during prosodic focus and phrasing contrasts (this notion will be expanded upon in the discussion).

7.2.3.6. Talker Idiosyncrasies

As with the auditory analysis presented in the previous chapter, it is evident from both the analysis of prosodic realisations and the examination of the effect of differing interactive settings that not all talkers appear to be using the same pattern of visual features to contrast focus and phrasing, nor across interactive settings. This is reflected by the absence of an effect in the talker analysis (when data is collapsed across sentences) despite the sentence analysis (collapsed across talkers) showing significant differences. To investigate this further, a series of post-hoc analyses were conducted individually for each talker, comparing the realisation of prosodic focus and phrasing, and differences in the realisation of these contrasts across interactive conditions. These within-items ANOVAs were interpreted with regard to a Bonferroni adjusted α of 0.00625 due to the large number of comparisons being

made. The visual properties of each talker's realisations of prosodic focus and phrasing are presented in Figures 7.10 to 7.15.

7.2.3.6.1. Variable visual realisation of prosodic focus

In the visual realisation of prosodic focus, a total of 14 features differed significantly between broad and narrow focused renditions in the sentence analyses, but failed to show consistent differences across talkers, with the majority of these features relating to movements that occurred post-critically (i.e., after the critical constituent had already been produced). These features were analysed individually for each talker in a repeated measures ANOVA, with prosodic condition (broad focus AO; narrow focus AO) treated as the within-items factor. The significant results of these analyses are presented in Table 7.10.

Table 7.10. Idiosyncratic talker realisations of prosodic focus. Analyses were interpreted with 1 between and 29 error degrees of freedom.

Feature	Utterance Phase	Talker	<i>F</i> Value	<i>p</i> Value	η_p^2	
PC 2	Post-Critical	3	16.19	< 0.001	0.358	
		4	22.12	< 0.001	0.433	
		6	9.83	0.004	0.253	
PC 3	Post-Critical	3	23.58	< 0.001	0.448	
PC 4	Post-Critical	5	13.28	0.001	0.314	
		6	6.75	0.015	0.189	
PC 5	Post-Critical	2	11.53	0.002	0.285	
		3	8.61	0.006	0.229	
		4	12.28	0.002	0.297	
		5	32.53	< 0.001	0.529	
		6	9.49	0.005	0.246	
PC 7	Pre-Critical	1	15.23	0.001	0.344	
		2	15.55	< 0.001	0.349	
		5	18.63	< 0.001	0.391	
		Post-Critical	1	68.01	< 0.001	0.701
			2	51.82	< 0.001	0.641
	3		122.76	< 0.001	0.809	
	4		75.96	< 0.001	0.724	
	5	66.74	< 0.001	0.697		
	PC 8	Post-Critical	1	75.42	< 0.001	0.722
3			70.91	< 0.001	0.710	
4			23.42	< 0.001	0.447	
5			19.37	< 0.001	0.400	
R 1	Pre-Critical	2	43.91	< 0.001	0.602	
		3	9.28	0.005	0.242	
		5	14.04	0.001	0.326	
		6	13.35	0.001	0.315	
	Critical	1	15.00	0.001	0.341	

Feature	Utterance Phase	Talker	<i>F</i> Value	<i>p</i> Value	η_p^2
R 1	Critical	2	85.78	< 0.001	0.747
		3	30.19	< 0.001	0.510
		5	31.87	< 0.001	0.524
		6	24.92	< 0.001	0.462
	Post-Critical	2	16.72	< 0.001	0.366
		3	15.72	< 0.001	0.351
		5	16.41	< 0.001	0.361
		6	69.49	< 0.001	0.706
R 2	Pre-Critical	5	14.35	0.001	0.331
		6	28.46	< 0.001	0.495
	Critical	1	11.43	0.002	0.283
		2	16.77	< 0.001	0.366
		3	17.07	< 0.001	0.371
		4	11.82	0.002	0.289
		6	30.76	< 0.001	0.515
		6	19.14	< 0.001	0.394
Post-Critical	5	11.05	0.002	0.276	
	6	19.14	< 0.001	0.394	

The talkers varied in the amount of post-critical jaw movement (PC 1) relative to the baseline broads focus condition; some displayed greater amounts of movements, some showed similar amounts of jaw movement between the two conditions, and some showed a reduction during the production of post-focal phases. However, none of these differences managed to achieve significance at the adjusted α level.

For lip opening movements (PC 2), Talkers 3, 4 and 6 produced consistently greater movement in post-critical phases of narrow focus renditions compared to broad focused productions. Similarly, post-focal phases were produced with greater

lower lip movement (PC 3) by Talker 3 and more upper lip movement by Talker 5 (with Talker 6 approaching significance also). Four of the six talkers (i.e., Talker 3, 4, 5 and 6) displayed an increase in lip rounding (PC 5) during the production of post-critical phases of narrow focused renditions relative to broad focus. However, Talker 2 displayed the opposite pattern, with a reduction in lip rounding during the production of post-focal content.

There was also evidence of idiosyncratic visual realisations of narrow focus for non-articulatory movements of the eyebrows and rigid head movements across utterance phases. For eyebrow raises (PC 7), Talker 1 and Talker 5 both produced more pre-focal eyebrow raises during the narrow than broad focused renditions. However, Talker 2 produced pre-critical phases with less eyebrow movement in the narrow focus condition. Post-critically, four of the talkers (i.e., Talkers 1, 3, 4 and 5) produced more eyebrow raises (PC 7) and pinching movements (PC 7) during the narrow focused than in the broad focused condition, whereas Talker 2 produced post-focal content with a reduction in eyebrow raising movements.

Rigid pitch rotations (R 1) showed a general trend of being increased across the entire utterance for narrow focused renditions (compared to broad focused tokens), however varied in the degree of use by each talker. Four talkers (i.e., Talker 2, 3, 5 and 6) demonstrated an increase in movement across all three utterance phases, whereas Talker 1 increased such movement only for the critical constituent. For rigid roll rotations (R 2), Talker 5 reduced the amount of roll rotation in narrow focus (compared to broad focus) productions in both pre-critical and post-critical utterance phases. In contrast, Talker 6 showed a significant increase across all three

utterance phases. The remaining four talkers all showed an increase in roll rotations during the production of the critical constituent in narrow focused renditions.

7.2.3.6.2. Variable visual realisation of prosodic phrasing

As with the production of focus contrasts, the realisation of phrasing varied across talkers, with several features showing significant differences in the sentence analysis but not in the talker analysis. Thus, these features were analysed individually for each talker in a repeated measures ANOVA, with prosodic condition (broad focus AO; echoic question AO) treated as the within-items factor. Statistical values for these comparisons are shown below in Table 7.11.

Table 7.11. Idiosyncratic talker realisations of prosodic phrasing. Analyses were interpreted with 1 between and 29 error degrees of freedom.

Feature	Utterance Phase	Talker	<i>F</i> Value	<i>p</i> Value	η_p^2
PC 1	Post-Critical	1	10.70	0.003	0.270
		3	31.79	< 0.001	0.523
		4	8.91	0.006	0.235
		5	149.77	< 0.001	0.838
PC 2	Post-Critical	5	104.33	< 0.001	0.782
PC 3	Post-Critical	1	19.76	< 0.001	0.405
		4	16.37	< 0.001	0.361
		5	229.80	< 0.001	0.888
PC 4	Post-Critical	3	39.94	< 0.001	0.579
		4	13.95	0.001	0.325
		5	60.61	< 0.001	0.676
PC 6	Post-Critical	1	11.57	0.002	0.285
		3	117.01	< 0.001	0.801
		4	22.64	< 0.001	0.438
		5	83.17	< 0.001	0.741
PC 7	Critical	1	100.85	< 0.001	0.777
		2	113.55	< 0.001	0.797
		3	260.61	< 0.001	0.900
		4	158.08	< 0.001	0.845
		5	67.18	< 0.001	0.698
		6	100.85	< 0.001	0.777
	Post-Critical	1	63.51	< 0.001	0.687
		2	96.20	< 0.001	0.768
		3	243.34	< 0.001	0.894
		4	34.00	< 0.001	0.540
		5	1190.24	< 0.001	0.976
		6	100.85	< 0.001	0.777
PC 8	Pre-Critical	3	13.42	0.001	0.316

Feature	Utterance Phase	Talker	<i>F</i> Value	<i>p</i> Value	η_p^2
PC 8	Pre-Critical	5	51.30	< 0.001	0.639
		6	9.33	0.005	0.243
	Post-Critical	3	140.12	< 0.001	0.829
		4	17.53	< 0.001	0.377
		5	219.58	< 0.001	0.883
R 1	Pre-Critical	3	8.88	0.006	0.234
	Post-Critical	2	10.39	0.003	0.264
		3	9.29	0.005	0.243
		4	9.13	0.005	0.239
		5	18.01	< 0.001	0.383
R 2	Pre-Critical	4	11.66	0.002	0.287
		5	27.58	< 0.001	0.487
		6	31.82	< 0.001	0.523
	Critical	2	13.16	0.001	0.312
		3	21.83	< 0.001	0.430
		5	12.83	0.001	0.307
		6	30.48	< 0.001	0.512
	Post-Critical	3	44.80	< 0.001	0.607
		4	11.68	0.002	0.287
		5	207.612	< 0.001	0.877
		6	35.39	< 0.001	0.550

Talkers 1, 3, 4 and 5 showed a significant decrease in the amount of jaw movement (PC 1) during post-critical utterance phases of echoic questions compared to the baseline broad focus condition. Talker 5 also showed a large reduction in post-critical lip opening (PC 2) during echoic question production (however, this talker was the only one to use this feature). A reduction in lower lip movement (PC 3) was produced in the post-critical phase of echoic questions (relative to broad focus

renditions) by Talker 1, 4 and 5. Talker 3 decreased the amount of upper lip movement during the post-critical phase of echoic questions (relative to broad focus. By contrast, Talker 4 and 5 produced more upper lip movement (PC 4) for echoic questions. Talkers 1, 3, 4 and 5 also reduced jaw protrusion (PC 6) during the post-critical phase of echoic questions.

As with the focus contrasts, talkers also varied in their use of non-articulatory gestures to contrast phrasing types. This was evident during the critical constituent for the production of eyebrow raises (PC 7): Talker 5 reduced the amount of movement (i.e., lowered their brows), while the remaining five of the talkers all raised their eyebrow more during the production of echoic questions. The use of eyebrow raising also differed between talkers for post-critical content in echoic questions, with Talkers 2, 5 and 6 producing significantly less eyebrow raises compared to broad focused renditions, whereas the other three talkers produced significantly more eyebrow raises during the post-critical phase. Similarly, some talkers used more brow pinching (PC 8) in their realisation of echoic questions, with Talkers 3, 5 and 6 producing more of this movement in the pre-critical utterance phase, and Talkers 3, 4 and 5 increasing these movements during post-critical utterance phases of echoic questions, relative to broad focused renditions.

Rigid head movements (R 1 and R 2) were somewhat more variable across talkers for phrasing realisation. For example, Talker 3 decreased rigid pitch rotations during the production of pre-critical echoic questions, whereas all other talkers produced similar amounts as in broad focused renditions. For post-critical utterance content, there was a general trend for echoic questions to be produced with greater rigid pitch rotations, with four talkers (i.e., Talker 2, 3, 4 and 5) displaying this

pattern of data. For rigid roll rotations (R 2), Talker 5 produced all three utterance phases in echoic question renditions with less movement than in broad focused renditions, whereas Talker 6 produced all three phases with an increase in roll rotation. Talker 2 only increased roll rotations during the production of the critical constituent in echoic questions, whilst Talker 3 increased these movements during critical and post-critical phases. Finally, Talker 4 reduced such movements in pre-critical phases, however produced post-critical phases of echoic questions with more rigid roll rotations than in broad focused statement renditions.

From these analyses, it is evident that talkers utilise a wide range of idiosyncratic strategies to visually mark prosodic contrasts. Furthermore, these individual idiosyncrasies occur both in articulatory movement and non-articulatory gestures (as previously observed by Dohen et al., 2006, 2009).

7.2.3.6.3. Idiosyncratic prosody production across interactive conditions

To determine whether talkers differed in their use of particular movement features dependant on the interactive setting (i.e., whether or not they could see the talker), the features identified as being significant in the sentence analysis (but not in the talker analysis) were further examined for each talker in a repeated measures ANOVA, with prosodic condition (AO rendition; FTF rendition) treated as the within-items factor. The statistical comparisons are reported in Table 7.12.

Table 7.12. Idiosyncratic talker realisations of narrow focus and echoic questions across interactive settings. Analyses were interpreted with 1 between and 29 error degrees of freedom.

Utterance Phase	Feature	Talker	F Value	p Value	η_p^2
<u>Narrow Focus</u>					
Pre-Critical	R 1	2	46.17	< 0.001	0.614
		4	17.44	< 0.001	0.376
	R 2	1	10.93	0.003	0.274
		5	9.80	0.004	0.253
Post-Critical	PC 2	6	40.08	< 0.001	0.580
		2	12.66	0.001	0.304
		2	27.71	< 0.001	0.489
	PC 3	3	9.30	0.005	0.243
		1	31.47	< 0.001	0.520
	2	36.29	< 0.001	0.556	
<u>Echoic Questions</u>					
Pre-Critical	PC 3	2	7.79	0.009	0.212
		3	9.89	0.004	0.254
		3	12.93	0.001	0.308
Critical	PC 3	6	8.79	0.006	0.233
Post-Critical	PC 2	2	11.95	0.002	0.292
		4	14.76	0.001	0.337
		5	10.20	0.003	0.260
	PC 8	2	11.33	0.002	0.281
		3	10.04	0.004	0.257

For the production of narrow focus renditions, both Talker 2 and Talker 4 produced pre-critical content with more rigid pitch rotations (R 1) in the AO than FTF condition. Similarly, Talker 6 produced an increased amount of rigid roll movements (R 2) during the pre-critical phase in AO settings. However, other talkers

produced *more* roll rotation movement when they were able to be seen by the interlocutor (i.e., in the FTF setting).

For the narrowly focused critical constituent, an effect was found across sentences but not across talkers for lip opening (PC 2). All but one of the talkers produced these movements at similar levels across both AO and FTF settings, while the remaining talker produced a greater amount of lip opening movement in the FTF setting (however, this difference was not statistically significant at the adjusted α level). Post-critically, narrow focus renditions in the AO setting were produced with a greater amount of lip opening (PC2) by Talker 2; and with a greater amount of lower lip movement by both Talker 2 and 3. Eyebrow raising (PC 7) during post-critical utterances also showed variability across talkers, with Talker 1 producing more movements in the AO relative to FTF condition, whereas Talker 2 produced more of these movements when they could be seen by an interlocutor in the FTF condition.

For the realisation of echoic questions, pre-critical content produced in the AO setting was produced with a greater amount of lower lip movement (PC 3) by Talker 2 and Talker 3. Similarly, Talker 3 produced pre-critical content in AO settings with more eyebrow pinching (PC 8). For the critical constituent, only Talker 6 consistently produced a greater amount of lower lip (PC 3) movement in the AO compared to FTF renditions. For utterance content in echoic questions following the critical constituent, Talker 2 and Talker 4 produced more lip opening (PC 2) in the AO renditions, whereas Talker 5 produced greater movements in the FTF renditions. Eyebrow pinching (PC 8) during post-critical phases was also greater in AO than

FTF settings for Talker 2, and Talker 3 despite the fact it could not be seen in these settings.

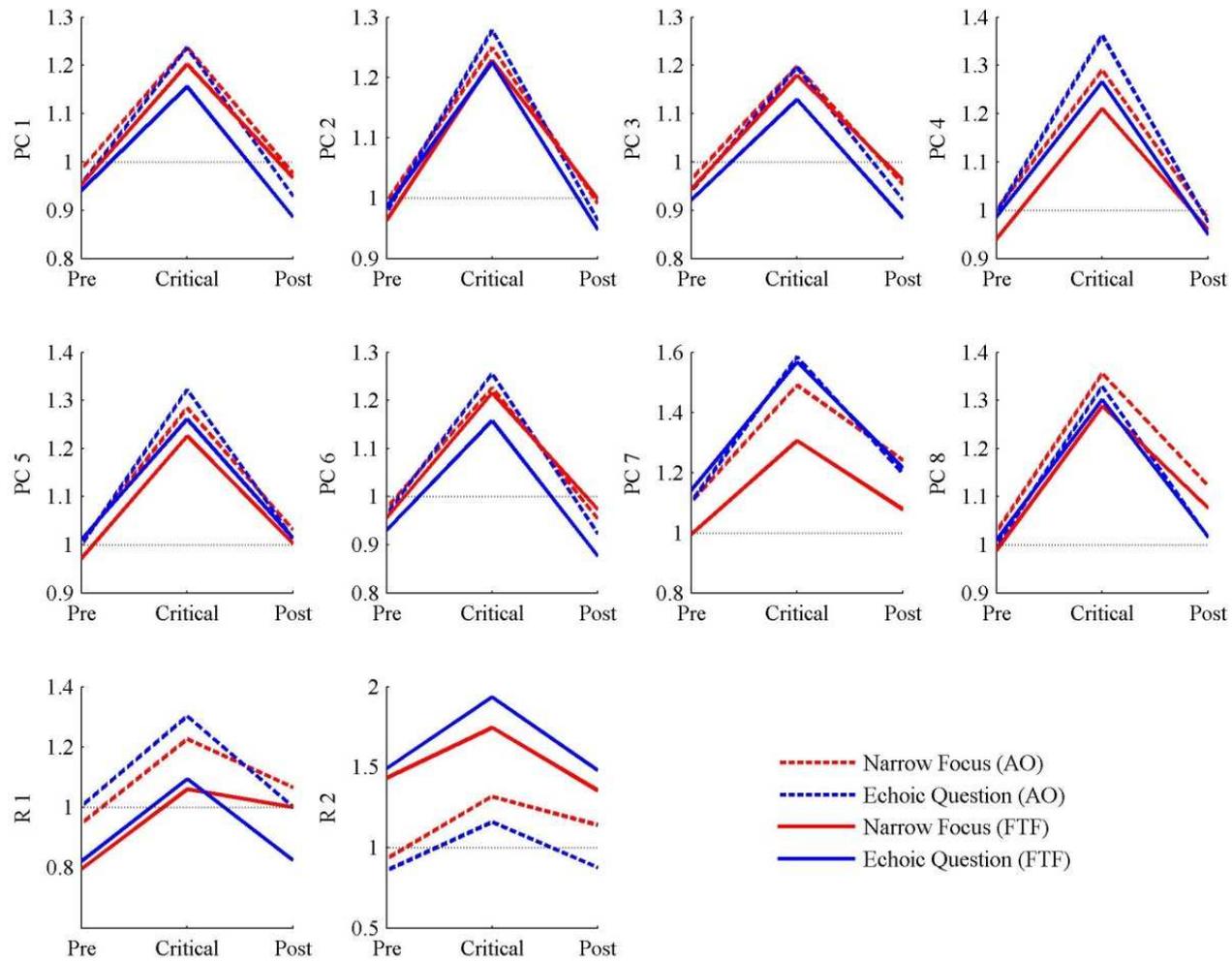


Figure 7.10. Visual realisation of prosodic contrasts (represented as a proportion of the broad focused rendition) of utterances produced by Talker 1.

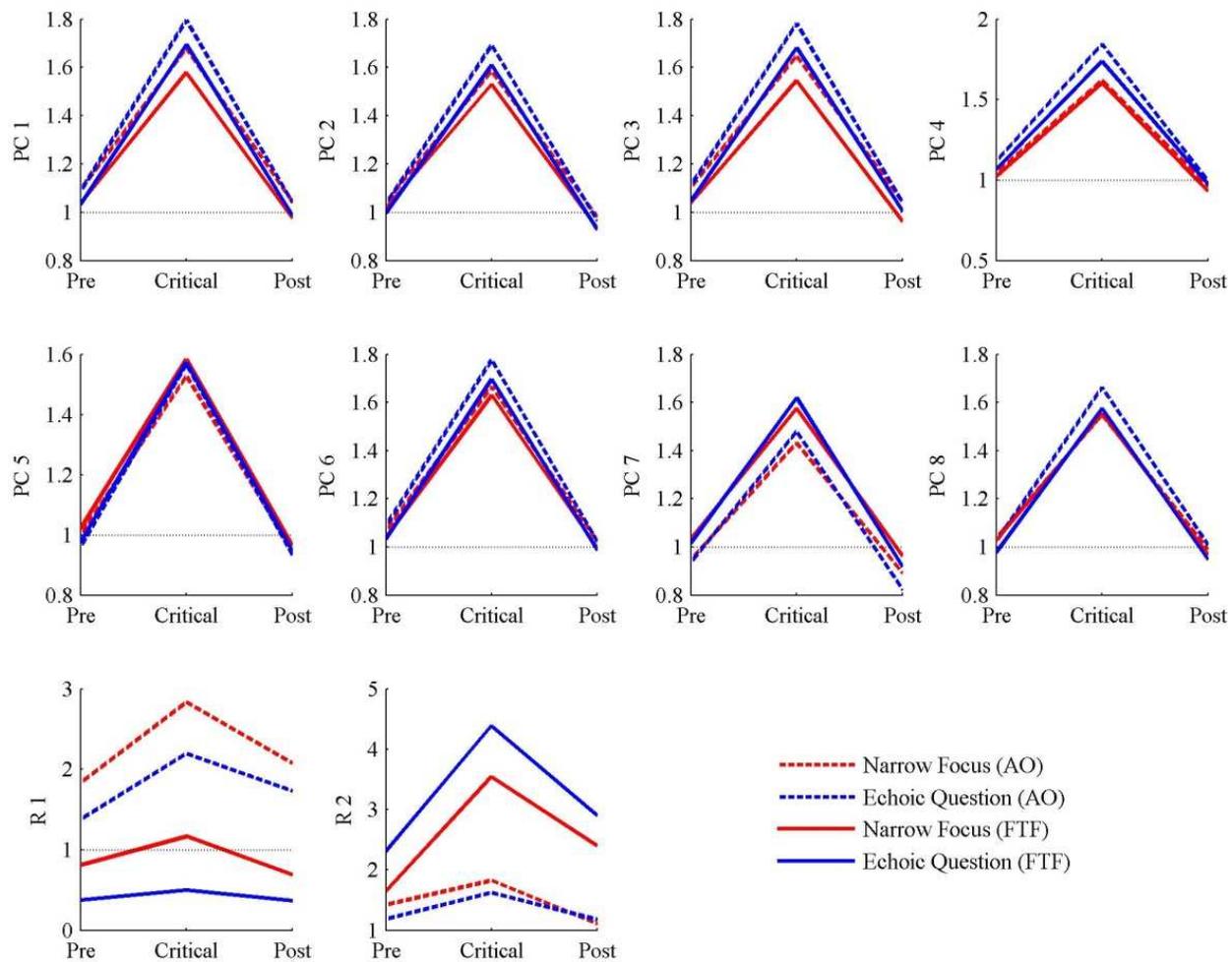


Figure 7.11. Visual realisation of prosodic contrasts (represented as a proportion of the broad focused rendition) of utterances produced by Talker 2.

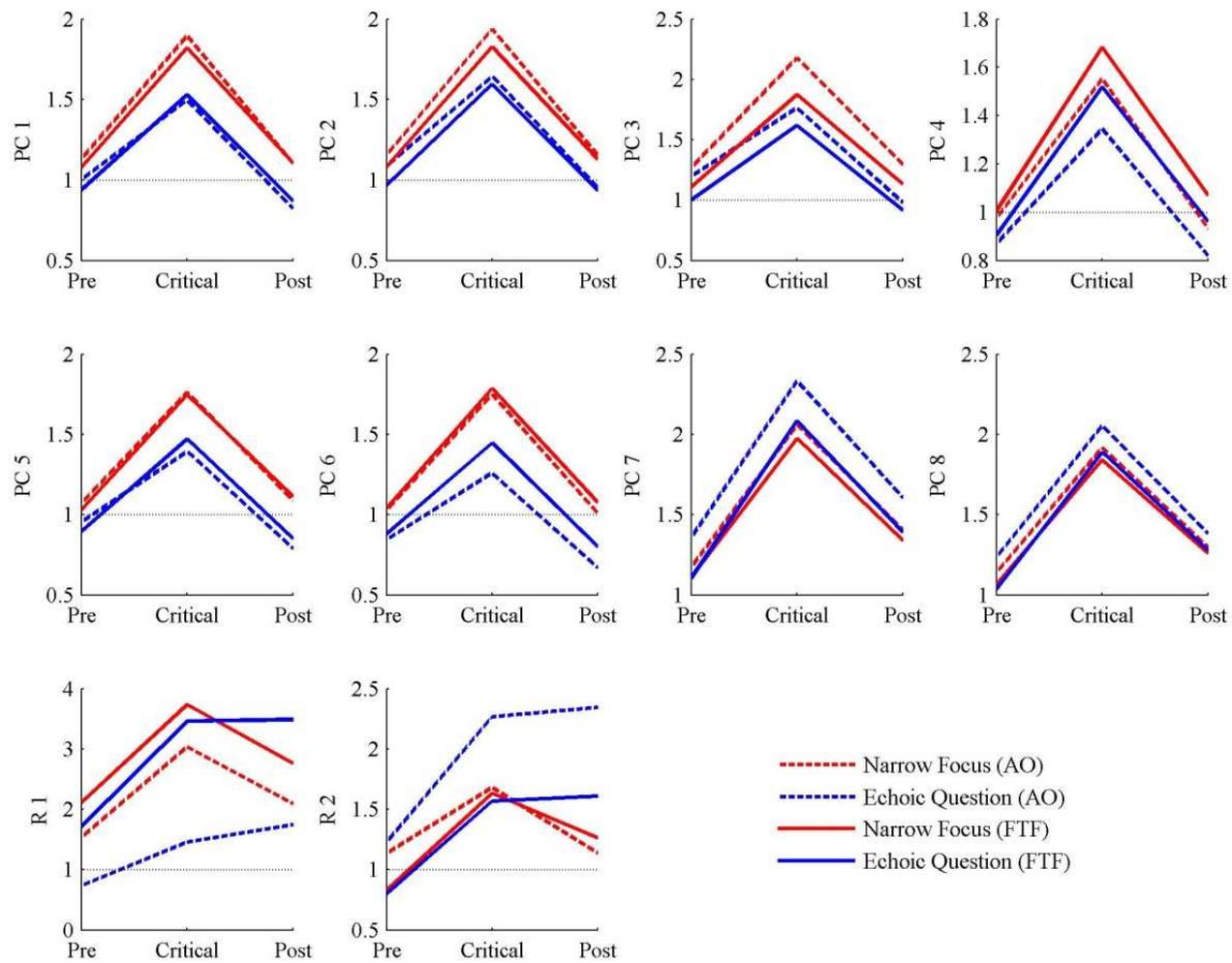


Figure 7.12. Visual realisation of prosodic contrasts (represented as a proportion of the broad focused rendition) of utterances produced by Talker 3.

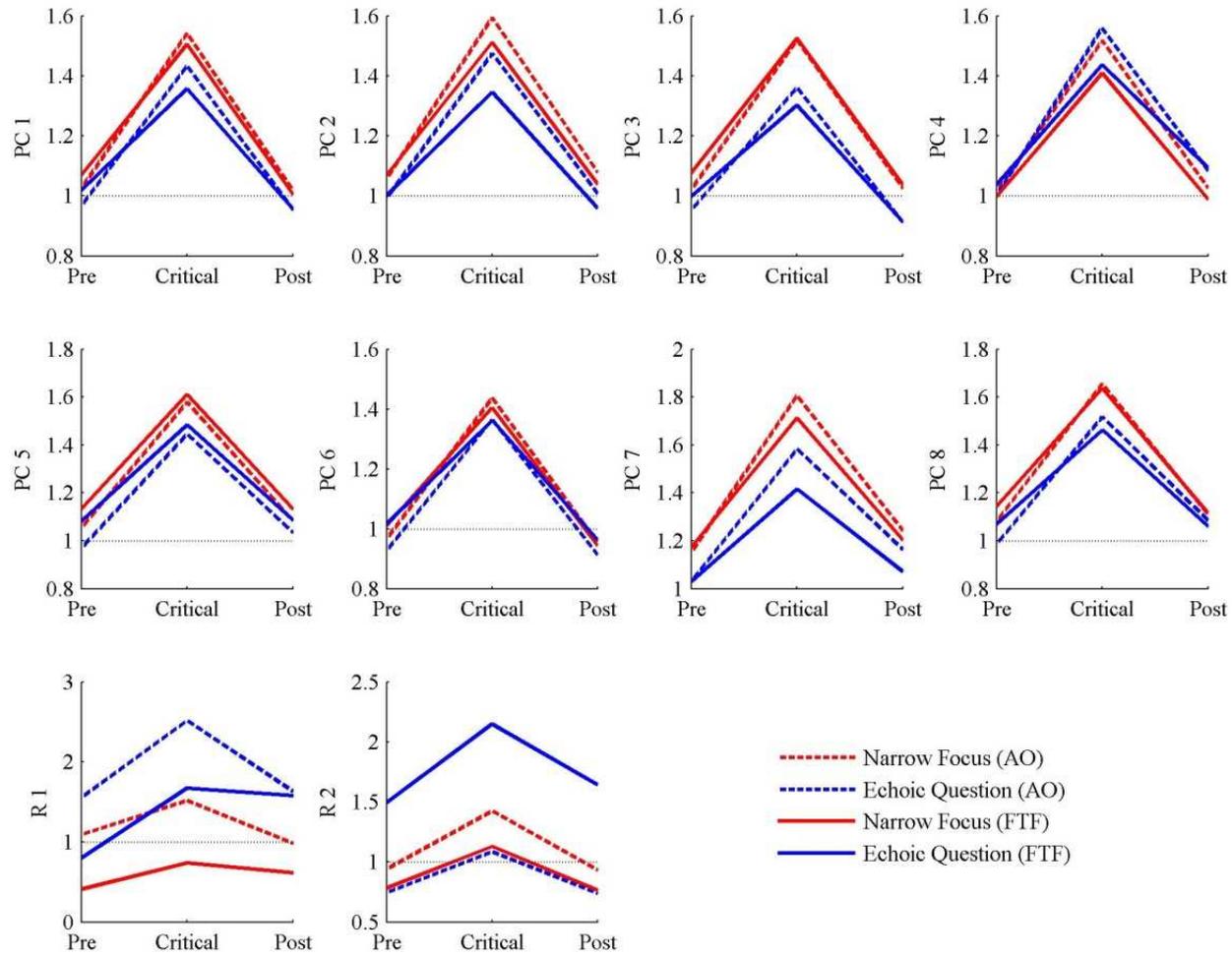


Figure 7.13. Visual realisation of prosodic contrasts (represented as a proportion of the broad focused rendition) of utterances produced by Talker 4.

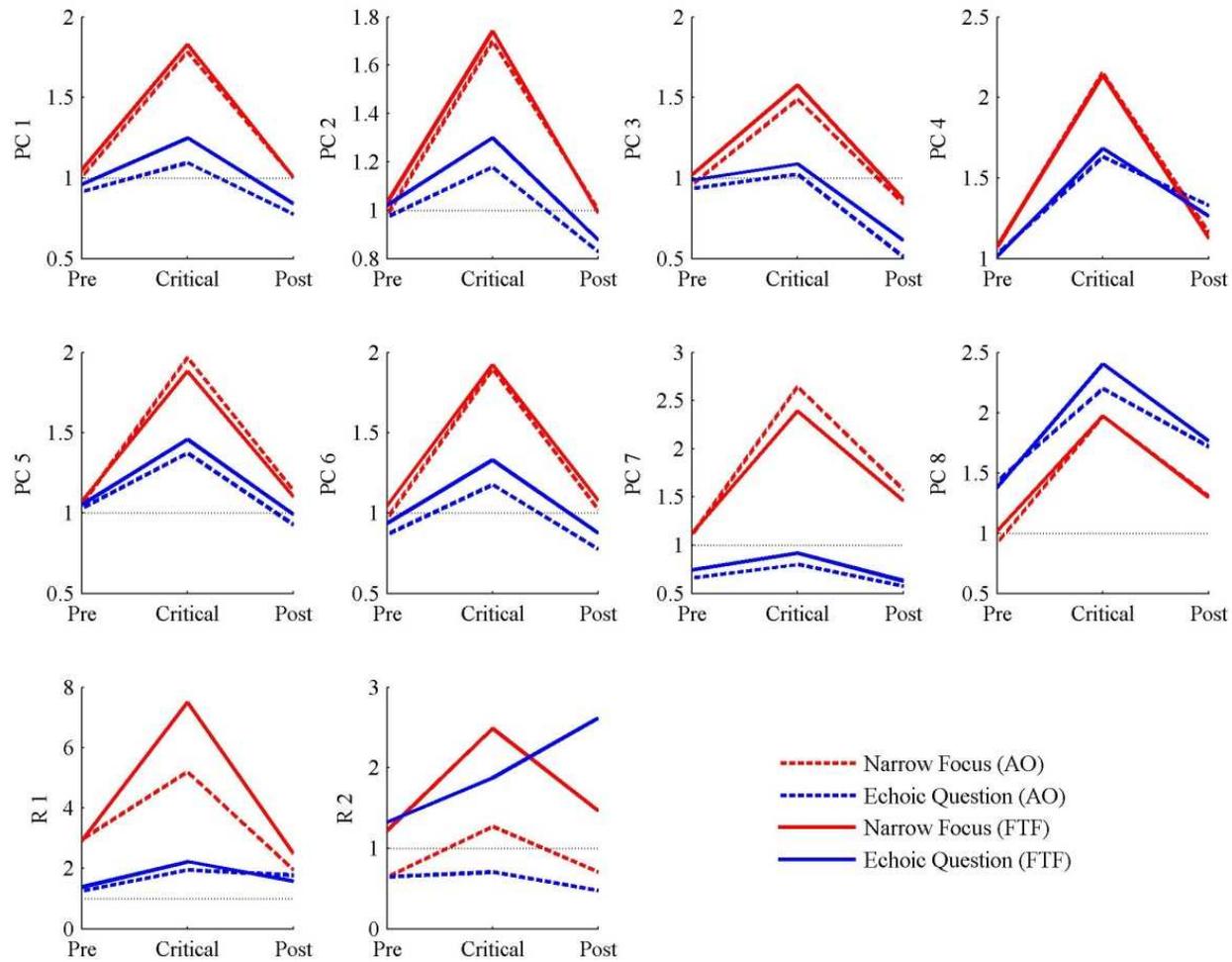


Figure 7.14. Visual realisation of prosodic contrasts (represented as a proportion of the broad focused rendition) of utterances produced by Talker 5.

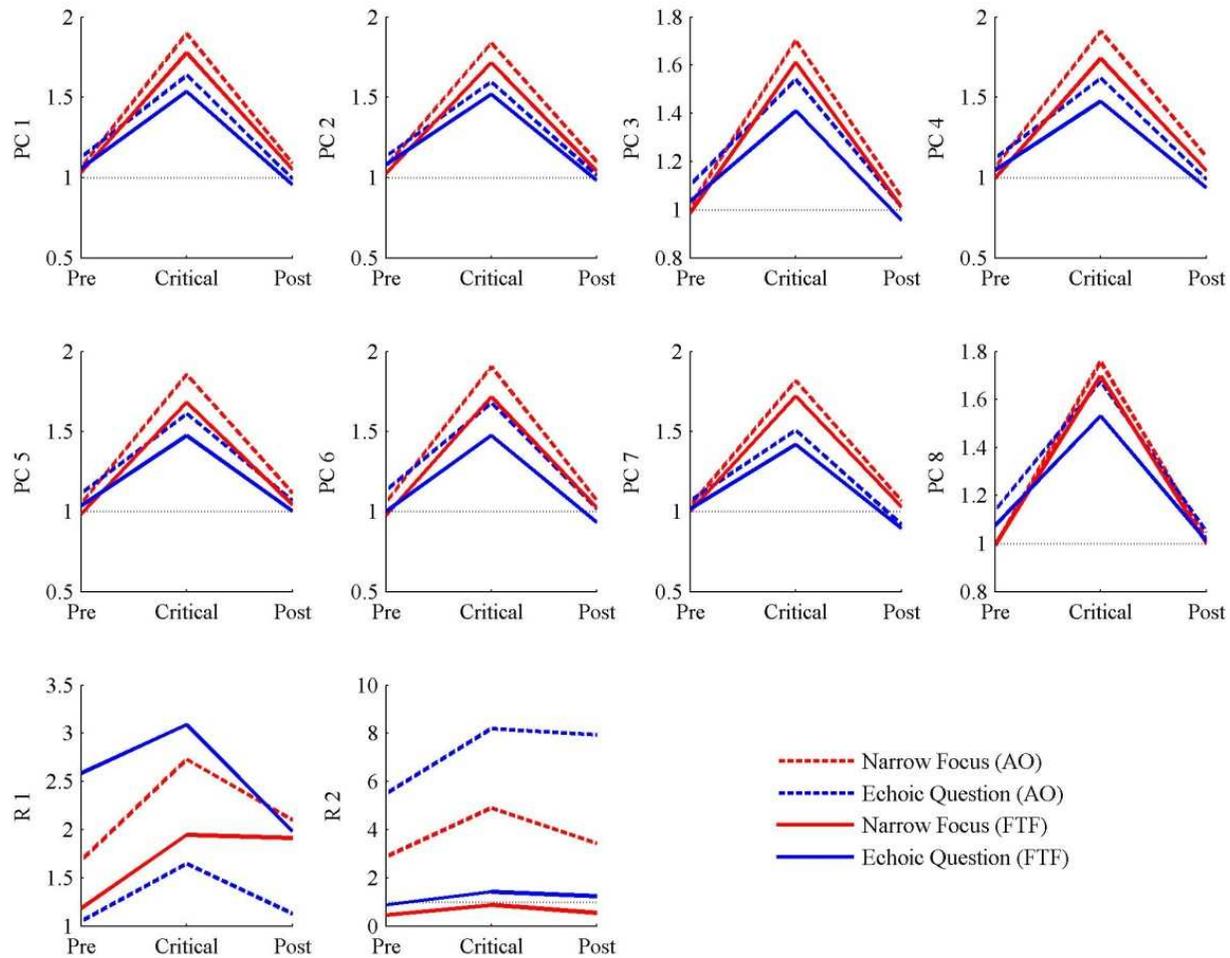


Figure 7.15. Visual realisation of prosodic contrasts (represented as a proportion of the broad focused rendition) of utterances produced by Talker 6.

7.2.3.6.4. Use of rigid head motion across interactive settings

Whereas the analysis of many of the visual features revealed only small differences between the interactive conditions (many of which likely relate to the articulatory processes necessary to produce the acoustic effects highlighted in Chapter 5), rigid movements of the head appeared to clearly differentiate the interactive settings (although this occurred in differing ways across talkers). Given that these movements are not directly or exclusively tied to articulation, they are more able to exhibit greater variability across talkers, sentences and repetitions (and as such may not be consistently produced enough to show significant effects when the data is collapsed across sentences and talkers). From Figures 7.10 to 7.15 that show individual talkers' visual strategies for marking prosody, it appears that some talkers are indeed sensitive to whether or not their gestures can be seen by an interlocutor. The statistical values of these comparisons are shown in Table 7.13.

For the critical constituent in narrow focus renditions, Talker 1 and Talker 2 both produced more rigid roll rotations (R 2) in the FTF condition (this difference however was not significant). Similarly, Talker 5 produced significantly more rigid roll rotations in the FTF condition. Additionally, Talker 3 and Talker 5 both produced more (but not significantly more) rigid pitch rotations (R 1) in the FTF condition. In FTF settings, the post-focal utterance phase was produced with more roll rotation by both Talker 2 and Talker 5.

For echoic questions, Talker 3 produced all three utterance phases when the interlocutor could be seen with an increase in rigid pitch rotations (R 1). With the exception of Talker 6, the remaining four talkers produced significantly more rigid

roll rotations (R 2) in the AO setting during pre-critical, critical and post-critical utterance phases.

Table 7.13. Idiosyncratic talker uses of rigid head motion during the production of narrow focus and echoic questions across interactive settings. Analyses were interpreted with 1 between and 29 error degrees of freedom.

Utterance Phase	Feature	Talker	F Value	p Value	η_p^2	
<u>Narrow Focus</u>						
Critical	R 2	5	9.77	0.004	0.252	
Post-Critical	R 2	2	7.49	0.010	0.205	
		5	15.53	< 0.001	0.349	
<u>Echoic Questions</u>						
Pre-Critical	R 1	3	15.22	0.001	0.344	
		1	13.12	0.001	0.312	
		2	8.72	0.006	0.231	
		4	24.14	< 0.001	0.394	
		5	17.84	< 0.001	0.381	
Critical	R 1	3	11.01	0.002	0.275	
		1	18.84	< 0.001	0.394	
		2	18.73	< 0.001	0.392	
		4	24.95	< 0.001	0.462	
		5	19.24	< 0.001	0.399	
Post-Critical	R 1	3	13.21	0.001	0.313	
		R 2	1	20.85	< 0.001	0.418
			2	9.62	0.004	0.249
			4	40.18	< 0.001	0.581
			5	14.53	0.001	0.334

7.2.4. Discussion

The outcome of the current analyses suggests that both articulatory and non-articulatory gestures are involved in contrasting broad from narrowly focused utterances, and statements from echoic questions. The greater amount of movement on PCs 1 through to 6 during the production of the critical constituents for narrow focus and echoic question renditions (relative to the broad focus baseline) likely stem from the movements required to shape the acoustic signal, however for the gestures less directly tied to articulation, such as eyebrow and rigid head movement, it may be that these are used by talkers to convey additional suprasegmental content to perceivers. Furthermore, the type of movement utilised also differed between contrasts types: narrow focus tokens were produced with more eyebrow raising (PC 7) and pinching (PC 8) consistently across sentences and talkers; for phrasing, eyebrow pinching and rigid pitch rotations (R 1) were systematically used to contrast statements from questions. The perceptual relevance of these visual differences is further examined in Chapters 9 and 10.

As with the acoustic properties, some visual parameters differed as a function of utterance length and the location of the critical constituent within the utterance. Most of these differences occurred for post-critical content in both the focus and phrasing contrasts. Utterance length only appeared to matter when the critical constituent occurred late in the utterance; this was true for the post-focal phase of both narrow focus and echoic questions, with greater amounts of movement occurring for short utterances with late critical constituents. This pattern could be interpreted in a similar fashion as the acoustic data. That is, after prosodically marking a critical constituent using increased movement, the amount of movement

gradually returns to that which was exhibited pre-critically; when the critical constituent occurs late in the utterance (particularly when the utterance is short), there is insufficient time to readjust.

Some interesting effects were observed as a function of interactive settings (i.e., whether or not the talker was able to see the interlocutor). It was expected that talkers would only enhance non-articulatory visual cues in situations where they would be visible to an interlocutor (in the FTF condition), however the opposite was observed: greater movements, both articulatory and non-articulatory, occurred in the AO setting when the talker could only hear their conversational partner. For example, despite the fact that the eyebrow movements were not able to be seen and have little to do with the shaping of the acoustic signal, they were exaggerated (relative to the FTF settings) across critical and post-critical phases for narrow focus renditions.

In interpreting this result, it is important to consider the various proposals for the functions of visual prosody. The “muscular synergy” account (Guaïtella et al., 2009) suggests that non-articulatory gestures occur simply as a by-product of the speech production process, with any communicative benefit being epiphenomenal. One explanation for why such gestures are produced despite the fact they will not be seen would be if non-articulatory cues cannot be decoupled from speech production. However, if this were the case, some degree of correlation would be expected between articulatory and non-articulatory parameters (this was not observed in the correlation analysis between principal components; Section 7.2.2), or between the produced auditory signal and non-articulatory gestures (this is examined in the following chapter).

An alternate view is that gestures not involved in articulation are produced intentionally to serve some communicative function (Flecha-Garcia, 2010; Swerts & Krahmer, 2010), either by enhancing the perceptual salience of a prosodic contrast (with non-articulatory gestures being temporally aligned with the prosodically marked constituent, i.e., “alignment” hypothesis), or to signal to a perceiver that a prosodically marked constituent is about to occur in the auditory stream (with such gestures occurring before the critically marked constituent, i.e., “signalling” hypothesis). In the case that visual cues do serve a signalling function, an increase in the amount movement may have occurred for utterances recorded in the FTF setting, but due to them being temporally shifted and thus occurring during the pre-critical utterance phase, any differences as a function of the interactive setting may have been washed out (i.e., due to these movements being collapsed across the rest of the pre-critical utterance content, an increase in movement at the end of the utterance phase may not be large enough to generate a statistically significant difference in the mean area under amplitude curve measure). Indeed, there was an increase in at least one rigid head motion parameter for critical constituents produced in the FTF setting when talkers were examined independently, suggesting that different movement parameters may have differential functions. To disentangle these hypotheses, it is necessary to examine the temporal alignment between prosodically marked constituents and the onset of non-articulatory visual cues (such an examination is conducted in Chapter 8), and to evaluate whether the presence of visual markers increases the overall perceptual salience of the prosodic contrasts (supporting the “alignment” hypothesis, see Chapter 9).

An alternative (but not necessarily exclusive) account is that non-articulatory movements of the face may serve a purpose for the talkers themselves, assisting in the conceptualisation of the spoken message. That is, these movements may form part of the talker's mental representation of prosody, and thus be produced regardless of whether or not they will be seen. Indeed, this notion has previously been proposed to account for talkers' production of manual gestures (e.g., arm and hand movements) despite the fact they are not visible to conversational partners (Alibali et al., 2001; Alibali, Kita & Young, 2000; Hostetter, Alibali & Kita, 2007; Kita, 2000; Krauss, 1998).

It should also be noted before moving on that the increased articulatory movements as well as the increased intensity level found for utterances recorded in the AO setting appear to be at odds with the results reported by Fitzpatrick et al. (2011). In their study, talkers produced speech in noise with significantly lower intensity levels, but with greater articulatory movements in a FTF setting compared to speech produced in AO settings. There are several key differences between Fitzpatrick et al.'s (2011) study and the current one; for example, Fitzpatrick et al.'s study involved the production of speech in noise and a more interactive dialogue exchange task. Given this, further investigation is required to determine whether the degree of interactivity, or the presence of noise (or both) was responsible for the difference.

CHAPTER 8.
THE RELATIONSHIP BETWEEN AUDITORY AND
VISUAL PROSODY

Chapter 8. The Relationship between Auditory and Visual Prosody

In Chapters 5 and 7, the auditory and visual properties of the recorded speech prosody corpus were independently examined. As it turned out, differences across prosodic contrasts were found in both modalities. This suggests that in addition to the auditory properties typically assigned a prosodic function there are visual properties that can also be regarded as prosodic correlates or cues. However, some questions remain to be answered: First, what is the precise nature of the relationship between auditory and visual signals? One assumption might be that the relationship between these signals would be stronger in situations where a constituent is prosodically marked (i.e., in narrow focus and echoic question renditions, compared to broad focused ones). Also, what role do the visual correlates play with respect to conveying prosodic information to perceivers? If there is some form of auditory-visual benefit for prosody (e.g., an increase in the perceptual salience of prosodic contrasts when accompanied by visual information), might this effect be greatest when the relationship between signal modalities is strongest?

Here in the current chapter, the relationship between the signal modalities was explored (while the functional roles that these visual prosodic correlates have for perception will be examined in the chapters that follow). The relationship between auditory and visual cues was determined by first examining the correlation between the acoustic features and the visual parameters. Secondly, the temporal relationship between the onset of the critical word and the occurrence of non-articulatory visual cues (i.e., eyebrow raises and rigid pitch rotations) was investigated. Finally, the

relationship between rises in F_0 in the auditory signal and the occurrence of eyebrow and rigid head movements was explored.

8.1. Correlation between Auditory and Visual Properties

Given that many properties of auditory and visual speech originate from the same spatio-temporal event (i.e., speech production), it is expected that articulatory (i.e., lip and jaw opening) and closely related movements (e.g., cheek motion) will bear a reasonably close relationship with those aspects of the produced acoustics that are used to signal prosody (Yehia et al., 1998). Indeed, in order to produce a speech sound over an extended duration (a property found for narrowly focused and echoically questioned syllables), the talker must maintain the configuration of the articulators for this amount time (de Jong, 1995). Similarly, increases in amplitude are likely to be accompanied by more dynamic jaw movements that end in a lower jaw position (Edwards et al., 1991; Summers, 1987).

Other visual cues to prosody, although not strictly coupled with the articulatory process, have also been shown to share a dynamic relationship with auditory signal properties (e.g., between intensity modulation and rigid head movement, Hadar et al., 1983; F_0 modulation and eyebrow movements, Cavé et al., 1996; Guaïtella et al., 2009). For example, strong associations have been found between F_0 modulations and rigid head motion, with Yehia, Kuratate and Vatikiotis-Bateson (2002) showing that a large amount of variance in F_0 (88% for an American English talker, and 73% for a talker of Japanese) could be estimated from rigid head motion. Although these correlations are high, there was a negligible relationship between F_0 and rigid head movements for some tokens in the recorded corpus, suggesting that the coupling of the two parameters may be functional, rather than a

necessary one. Additionally, the amount of head motion that could be inferred by F_0 was substantially lower (i.e., 50% for the talker of American English, and only 25% for the Japanese talker) when the estimation was calculated in the opposite direction (Yehia et al., 2002).

The lack of evidence of a direct coupling between F_0 and head movements leaves open the possibility that the association reflects particular communicative strategies of the talker; strategies that are likely to vary according to the prosodic nature of the utterance. If this is the case, then the relationship between auditory and visual signals is likely to be intermittent and possibly non-linear. As such, the current analysis examined whether there was a significant relationship between the auditory and visual signals in the recorded corpus by determining the strength of the correlations between the auditory (i.e., intensity and F_0 contours) and visual parameters (principal component amplitude curves) obtained by the analyses detailed in the previous chapters. In addition to determining these values across more than 2000 utterances, the design of the speech prosody corpus allows for the examination of whether such relationships were greater for utterances that contained a prosodically marked constituent (i.e., narrow focus and echoic questions) compared to the baseline broad focused ones (this was done particularly for those auditory and visual properties that showed differences across the prosodic contrasts), with it expected that the relationship between signal modalities would be strongest in situations where a constituent was prosodically marked.

8.1.1. Method

Although the acoustic properties of each utterance had previously been determined (see Chapter 5), the sampling rate was substantially higher (i.e., 44.1kHz) than that

of the visual motion data. As such, the acoustic parameters for each utterance were re-calculated using Praat (Boersma, 2001). The $F0$ contour was extracted for each utterance at time steps of $1/60^{\text{th}}$ of a second (~ 16.667 ms, to match the sampling rate of the visual data of 60Hz), with a pitch floor of 65 Hz and ceiling of 300Hz. Octave jumps were removed from the resulting contour, the curve was interpolated over voiceless content, and smoothed with a 10Hz filter. The intensity profile of each utterance was also extracted at time steps of $1/60^{\text{th}}$ of a second.

Pearsons correlations were then calculated independently for each utterance between the extracted acoustic features, and the visual parameters (represented as principal component values over time, see Chapter 6) and the rigid pitch (R 1) and roll (R 2) parameters, using custom-written scripts in Matlab. The resulting correlations for every utterance in the corpus can be found as Appendix D.

It should be noted that, by using cross-correlations, Grant and Seitz (2000) found that the strongest relationship between intensity and area of mouth opening occurred when the auditory signal lagged behind the visual one by approximately 33ms (i.e., the video signal preceded the auditory signal by one frame). However, this lag value was obtained by examining only two sentences, and also varied in a subsequent analysis examining different segmental content. As such, the current analyses were conducted using zero-lag correlations that test the strength of the correlation between auditory and visual properties without any time delay. It is acknowledged that this way of testing the relationship between the signals is likely to underestimate the strength of their correlation¹³; however it has the virtue that it was

¹³ It is possible that some of the variables are associated by a non-linear relationship. To evaluate this, Spearman rank (r_s) correlations were conducted between the acoustic features and visual parameters. The outcome of this analysis was only negligibly different to the Pearsons correlations. The results of these analyses are included in Appendix E.

applied equally for all the different types of measured properties and so was unlikely to favour one against the other.

8.1.2. Results

8.1.2.1. Global Correlations

The global correlation properties between acoustic and visual properties, collapsed across utterances and prosodic conditions are displayed in Table 8.1. As expected, jaw opening (PC 1) and lip opening (PC 2) were both moderately correlated with the intensity of the produced acoustic signal (i.e., there is a causal relationship between the movement of the articulators and the produced speech output). However, even for these movements, the relationship with auditory properties varied greatly across utterances. For example, the correlation between lip opening (PC 2) and intensity ranged between r values of -0.81 and 0.85. In contrast, there was no consistent strong relationship with $F0$ for any of the visual parameters. As with intensity, the relationship was highly variable between the visual parameters and $F0$, for example, rigid pitch rotations ranged from -0.98 to 0.97 across utterances.

Table 8.1. Global correlation properties collapsed across utterances, prosodic contrasts and talkers, $n = 2160$.

Visual Feature	Mean r	Standard Deviation	Median r	Minimum r Value	Maximum r Value
Intensity					
PC 1	0.395	0.201	0.402	-0.390	0.810
PC 2	0.390	0.296	0.458	-0.808	0.847
PC 3	0.278	0.310	0.344	-0.785	0.847
PC 4	0.026	0.297	0.018	-0.812	0.809
PC 5	0.155	0.272	0.152	-0.706	0.836
PC 6	-0.122	0.283	-0.145	-0.784	0.771
PC 7	0.115	0.307	0.153	-0.771	0.791
PC 8	0.004	0.292	0.030	-0.758	0.649
R 1	0.187	0.322	0.224	-0.680	0.851
R 2	-0.076	0.353	-0.061	-0.839	0.782
Fundamental Frequency					
PC 1	-0.015	0.343	-0.027	-0.904	0.894
PC 2	0.004	0.346	-0.008	-0.883	0.875
PC 3	0.019	0.392	0.021	-0.867	0.908
PC 4	0.000	0.353	-0.003	-0.911	0.932
PC 5	0.016	0.403	0.007	-0.926	0.967
PC 6	-0.009	0.417	0.008	-0.948	0.952
PC 7	-0.019	0.418	-0.035	-0.954	0.962
PC 8	-0.028	0.387	-0.031	-0.933	0.900
R 1	0.128	0.503	0.200	-0.978	0.974
R 2	0.009	0.521	-0.009	-0.974	0.966

8.1.2.2. Correlations as a function of Prosodic and Interactive Settings

The correlation values showed large variation across utterances. One method of using such variable data is to determine whether the central tendencies of these

correlations differed over the prosodic conditions and interactive settings, as this would provide evidence that AV relationships were affected by these variables.

For intensity, the relationship with jaw movement (PC 1), lip opening (PC 2), lower lip movement (PC 3) and rigid pitch rotations (R 1) were examined. The articulatory parameters were selected on the basis that they showed the greatest median correlation value across all conditions (and as addressed previously, there is an a priori expectation that articulatory movements are linked with the properties of speech output). The rigid parameter was chosen as it has been previously suggested that such movements may assist listeners with the segmentation of the continuous speech stream into individual word units (cf. Davis & Kim, 2004, 2006; Hadar et al., 1983; Munhall et al., 2004). For *F0*, the relationship with eyebrow raising (PC 7), eyebrow pinching (PC 8), rigid pitch rotations (R 1), and rigid roll rotations (R 2) were examined. These features were selected on the basis of previous research that proposed a strong link between *F0* and eyebrow and rigid head movements (Cavé et al., 1996; Granström & House, 2005; Guaitella et al., 2009; Yehia et al., 1998, 2002).

The median correlation values between these acoustic and visual features for the entire corpus (as a function of prosodic condition and interactive setting) are provided in Table 8.2. Histograms of the distributions of these correlation values are shown in Figure 8.1 to Figure 8.6.

Chapter 8: The Relationship between Auditory and Visual Prosody

Table 8.2. Median correlation (r) values between acoustic properties and visual components, as a function of the prosodic condition and interactive setting. $n = 360$ per cell. Median correlation values above 0.30 are presented in bold.

Component	<u>AO Interactive Setting</u>			<u>FTF Interactive Setting</u>		
	Broad Focus	Narrow Focus	Echoic Question	Broad Focus	Narrow Focus	Echoic Question
Intensity						
PC 1	0.437	0.399	0.357	0.419	0.421	0.396
PC 2	0.431	0.465	0.451	0.468	0.466	0.451
PC 3	0.352	0.355	0.373	0.307	0.300	0.356
R 1	0.274	0.182	0.112	0.291	0.274	0.199
Fundamental Frequency						
PC 7	-0.083	0.073	-0.060	-0.053	-0.006	-0.038
PC 8	-0.058	0.051	-0.052	-0.064	0.016	-0.063
R 1	-0.341	-0.226	-0.142	-0.318	-0.196	0.085
R 2	-0.062	0.118	-0.108	0.011	0.091	-0.114

Chapter 8: The Relationship between Auditory and Visual Prosody

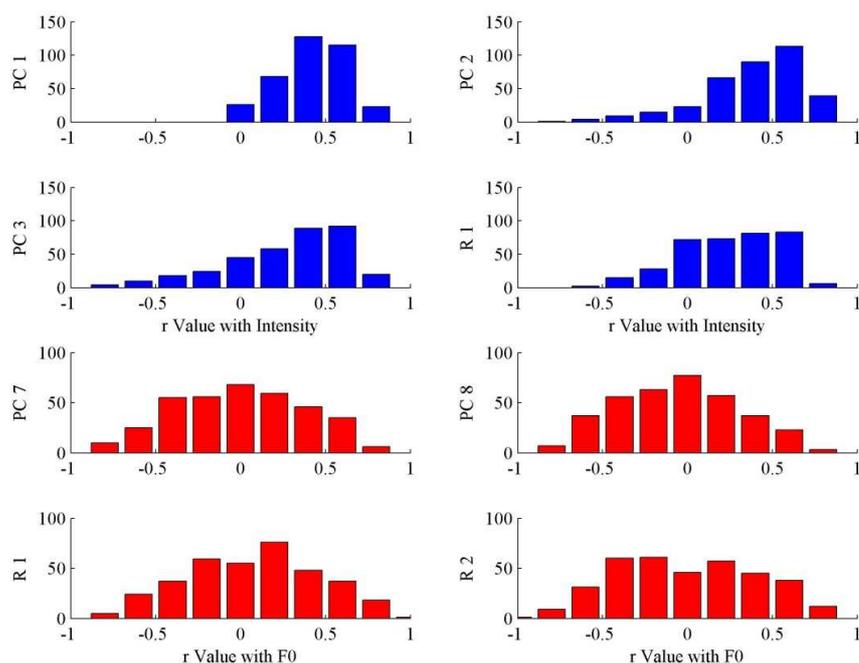


Figure 8.1. Histograms of the distribution of correlation coefficients (calculated between auditory and visual parameters) for broad focus tokens recorded in the AO interactive setting (blue shows the distribution of the movement and intensity correlations; red shows the F0 correlations).

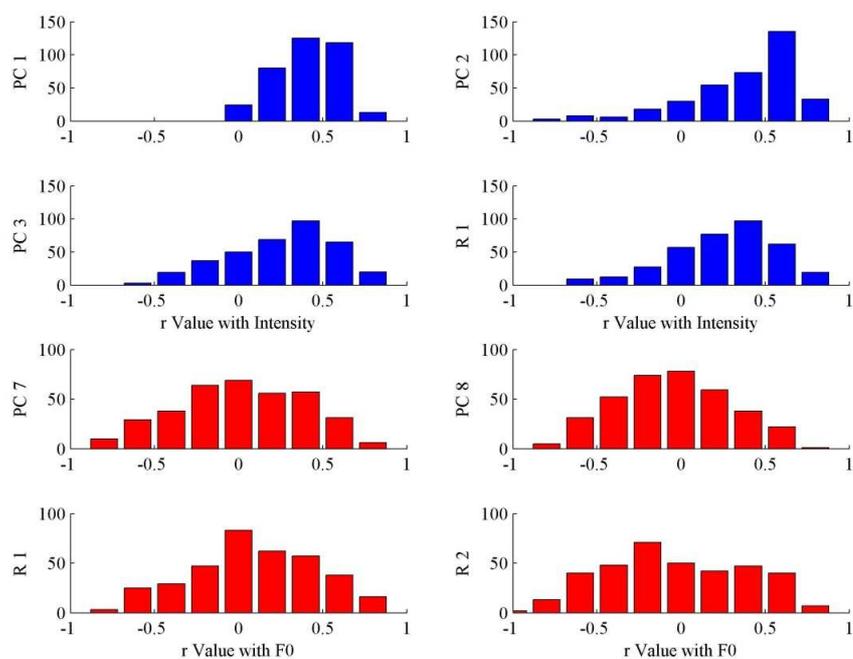


Figure 8.2. Histograms of the distribution of correlation coefficients (calculated between auditory and visual parameters) for broad focus tokens recorded in the FTF interactive setting.

Chapter 8: The Relationship between Auditory and Visual Prosody

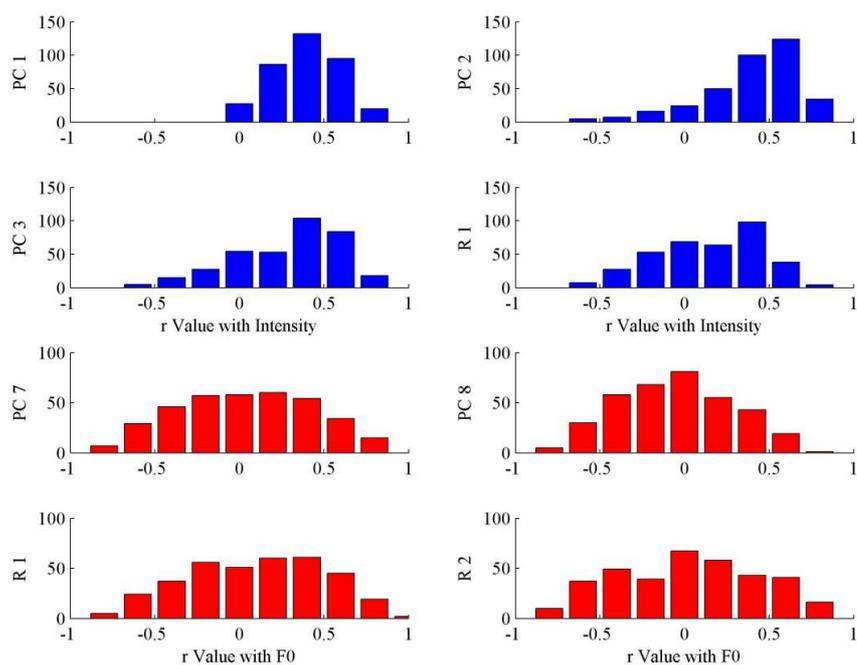


Figure 8.3. Histograms of the distribution of correlation coefficients (calculated between auditory and visual parameters) for narrow focus tokens recorded in the AO interactive setting.

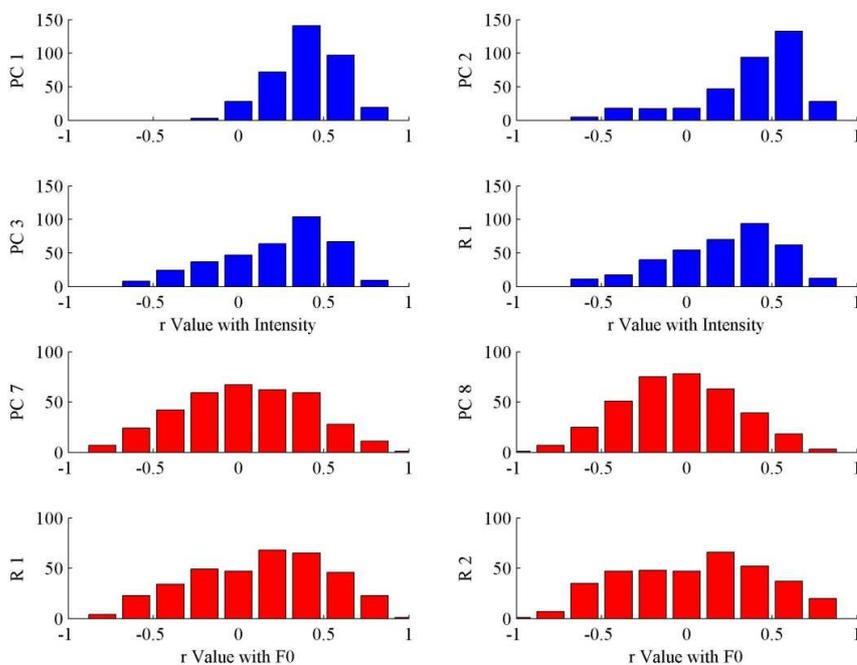


Figure 8.4. Histograms of the distribution of correlation coefficients (calculated between auditory and visual parameters) for narrow focus tokens recorded in the FTF interactive setting.

Chapter 8: The Relationship between Auditory and Visual Prosody

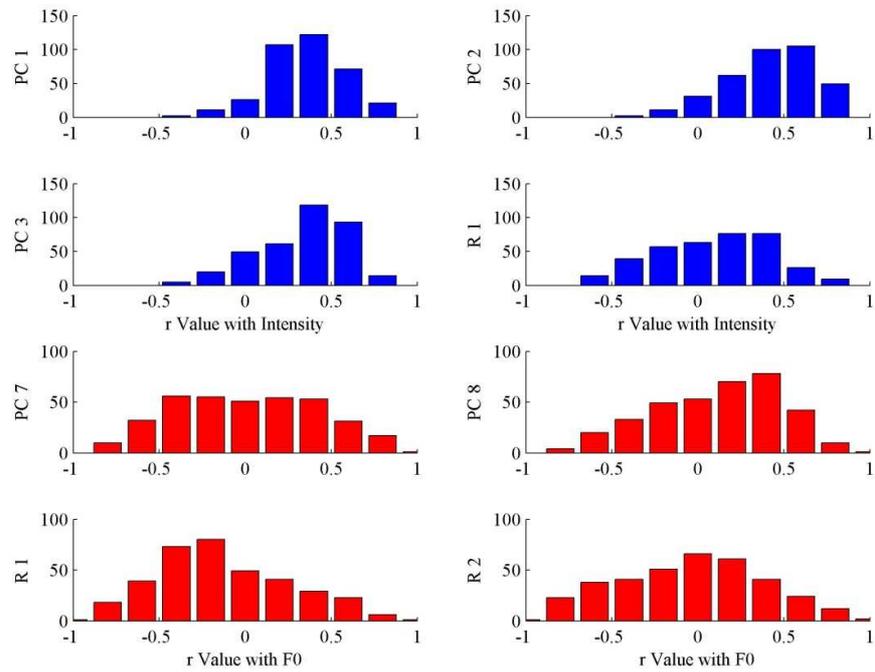


Figure 8.5. Histograms of the distribution of correlation coefficients (calculated between auditory and visual parameters) for echoic question tokens recorded in the AO interactive setting.

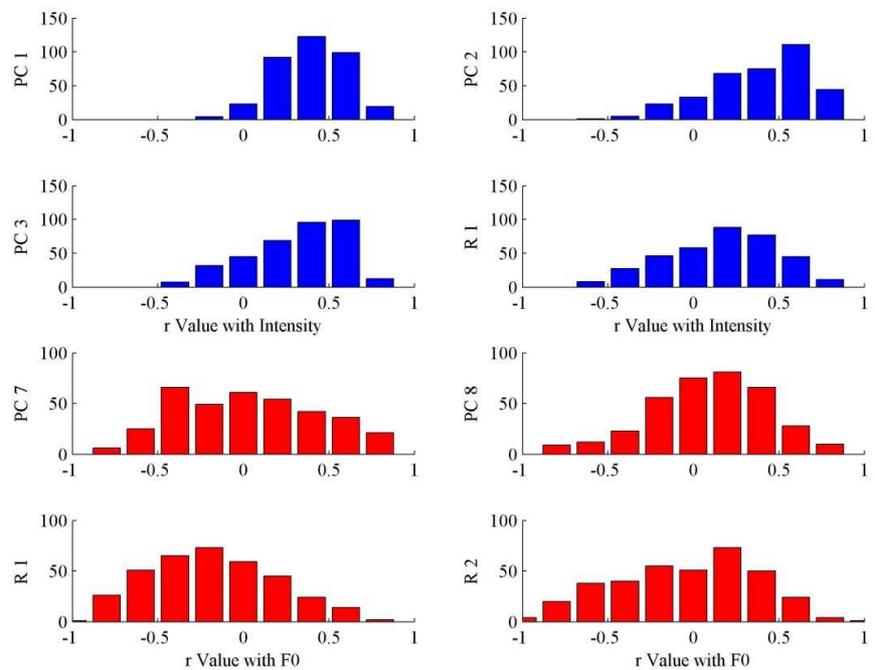


Figure 8.6. Histograms of the distribution of correlation coefficients (calculated between auditory and visual parameters) for echoic question tokens recorded in the FTF interactive setting.

The correlation values were examined in a series of between-subjects ANOVAs. The first set of analyses compared these values as a function of the prosodic contrast (i.e., focus: broad vs. narrow; phrasing: statement vs. echoic question), while the other compared the production of each prosodic condition between interactive settings (i.e., AO vs. FTF). Due to the number of comparisons, these analyses were interpreted with an α level of 0.0125.

8.1.2.2.1. Correlations across focus contrasts

For intensity, the only difference as a function of focus condition was for rigid pitch rotations (R 1), $F(1,718) = 16.96, p < 0.001, \eta_p^2 = 0.023$, with the strength of the relationship between head movement and intensity weaker in the narrow focus compared to broad focus condition. In terms of $F0$, a difference was found between broad and narrow focus for eyebrow raising (PC 7), $F(1,718) = 15.92, p < 0.001, \eta_p^2 = 0.022$; eyebrow pinching (PC 8), $F(1,718) = 8.69, p = 0.003, \eta_p^2 = 0.012$; and rigid roll rotations (R 1), $F(1,718) = 8.29, p = 0.004, \eta_p^2 = 0.011$. However, for all three contrasts, the mean r value was $< |0.10|$, so although there was a significant difference between conditions, it reflects a change from weak correlation to no correlation at all.

8.1.2.2.2. Correlations across phrasing contrasts

Phrasing contrasts showed significant differences in the relationship between intensity and jaw opening (PC 1), $F(1,718) = 19.25, p < 0.001, \eta_p^2 = 0.026$; lower lip movement (PC 3), $F(1,718) = 9.20, p = 0.003, \eta_p^2 = 0.013$; and rigid pitch rotations (R 1), $F(1,718) = 42.75, p < 0.001, \eta_p^2 = 0.056$. The strength of the relationship between both articulatory parameters and intensity was weaker for echoic questions compared to statement renditions; this was also the case for pitch rotations (R 1). For

F_0 , phrasing contrasts showed a difference in the relationship only for rigid pitch rotations, $F(1,718) = 11.13$, $p = 0.001$, $\eta_p^2 = 0.015$, with a weaker relationship for the echoic question renditions.

8.1.2.2.3. *Correlations across interactive settings*

For the comparison of the strength of the relationship between auditory and visual properties for prosodic conditions across interactive settings, no difference was observed for intensity or F_0 for broad focus renditions (i.e., the mean correlation across all parameters was equivalent regardless of the visual availability of the interlocutor).

For narrow focus conditions, a difference in correlation means was observed between intensity and rigid pitch rotations (R 1), $F(1,718) = 6.99$, $p = 0.008$, $\eta_p^2 = 0.010$, with a stronger correlation between the auditory and visual properties for the FTF condition (i.e., when the movements could be seen by an interlocutor). No differences as a function of the interactive setting were found for F_0 relationships.

The effect across interactive settings for intensity observed for narrow focus was replicated for echoic questions, $F(1,718) = 9.18$, $p = 0.003$, $\eta_p^2 = 0.013$, with the strength of the relationship between intensity and rigid pitch rotations (R 1) increasing in the FTF setting (compared to the AO one). The relationship between F_0 and rigid pitch rotations also differed across interactive settings for echoic questions, $F(1,718) = 11.29$, $p = 0.001$, $\eta_p^2 = 0.015$, however the relationship weakened in the FTF setting.

8.1.2.2.4. *Correlations during critical utterance phases*

An interesting outcome of the analyses so far was the *weakening* of the overall relationship between auditory and visual properties in sentences when the critical

constituent was prosodically marked (i.e., narrowly focused or echoically questioned). A possible explanation for this reduced correlation is that for these sentences, the relationship between auditory and visual signals became more non-linear, with the timing of pre-critical and post-critical utterance phases affected by talkers attempting to bring the auditory and visual signals more into alignment for the critical utterance phase. To explore this possibility, an additional set of correlation analyses were conducted between auditory and visual signal properties, but only for the critical constituent (instead of the complete utterance), as differences in these values would be more straightforward to interpret. The median values of these correlations as a function of the prosodic and interactive condition are displayed in Table 8.3.

Table 8.3. Median correlation (r) values between acoustic properties and visual components for the critical constituent of each utterance, as a function of the prosodic condition and interactive setting. $n = 360$ per cell.

Component	<u>AO Interactive Setting</u>			<u>FTF Interactive Setting</u>		
	Broad Focus	Narrow Focus	Echoic Question	Broad Focus	Narrow Focus	Echoic Question
Intensity						
PC 1	0.275	0.338	0.337	0.286	0.342	0.269
PC 2	0.239	0.337	0.342	0.221	0.299	0.254
PC 3	0.126	0.200	0.182	0.116	0.195	0.149
R 1	0.176	0.049	-0.108	0.199	0.096	0.051
Fundamental Frequency						
PC 7	0.062	0.252	0.088	0.045	0.137	-0.091
PC 8	0.043	0.067	-0.071	-0.027	-0.015	-0.179
R 1	0.263	0.146	-0.005	0.262	0.216	0.174
R 2	-0.087	0.038	0.314	-0.071	0.037	0.115

For the critical constituent during focus contrasts, the strength of the relationship between intensity and jaw movement (PC 1), $F(1,718) = 51.35$, $p < 0.001$, $\eta_p^2 = 0.067$; intensity and lip opening (PC 2), $F(1,718) = 16.30$, $p < 0.001$, $\eta_p^2 = 0.022$; $F0$ and eyebrow raising (PC 7), $F(1,718) = 18.32$, $p < 0.001$, $\eta_p^2 = 0.025$; and $F0$ and eyebrow pinching (PC 8), $F(1,718) = 10.24$, $p = 0.001$, $\eta_p^2 = 0.014$, all increased in the narrow focus condition (relative to the broad focused utterances).

For phrasing contrasts, a significant difference in the strength of correlations was found between intensity and jaw opening (PC 1), $F(1,718) = 19.70$, $p < 0.001$, $\eta_p^2 = 0.027$, lip opening (PC 2), $F(1,718) = 25.50$, $p < 0.001$, $\eta_p^2 = 0.034$, and rigid pitch movements (R 1), $F(1,718) = 57.28$, $p < 0.001$, $\eta_p^2 = 0.074$, for the critical constituent in echoic question renditions. The relationship between articulatory parameters and intensity increased in strength for the echoic question renditions (compared to statement renditions), whereas the rigid movement parameter shifted from being slightly correlated to practically showing no correlation. No variations in the strength of correlation were observed for the relationship between $F0$ and movement parameters for the critical constituent in phrasing contrasts, nor was there any effect of interactive setting across any of the prosodic conditions. Thus, it appears when the critical word is important, the relationship between modalities (for some parameters) is enhanced, but this only occurs for the critical word, not the entire utterance.

8.1.3. Discussion

The current analyses examined whether there was a significant correlation between the auditory and visual properties of the recorded speech corpus, and whether the strength of these relationships differed as a function of the prosodic condition. For

both articulatory and non-articulatory movements, the relationship with acoustic properties was highly variable. In terms of articulatory movements, the variability in the strength of the relationship may be due in part to the parcelling of the data into principal components. That is, to achieve a greater mouth opening (which correlates with increases in signal intensity), talkers can use a greater amount of jaw motion (PC 1), or they could achieve this by increasing the opening of their mouth (PC2, independent of jaw motion), or by moving the upper (PC 4) or lower lip independently (PC 3), with each of these movements represented by different components. Thus, if talkers are inconsistent in the type of movement they use to achieve articulatory movements, so too will be the correlation between such movements and the generated acoustic signal.

To explore this hypothesis further, a post-hoc analysis was conducted by calculating the Euclidean distance between the upper middle lip and lower middle lip markers from the visual data, and examining the correlation of this value with the intensity profile for each utterance. The median r value collapsed across utterances of 0.41 ($M = 0.41$, $SD = 0.17$, Range: 0.00 – 0.80) was comparable to the figure observed for the correlation conducted with the principal components (i.e., between intensity and PC 1, $r = 0.40$, and between intensity and PC 2, $r = 0.46$). Thus, it is more likely that the variability in the correlation values across utterances reflects some non-linear and intermittent relationship between the signals.

For non-articulatory movements, the relationship between auditory and visual properties was much more variable than that previously reported by Yehia et al. (2002). Similar degrees of variability (to that found here) have been reported for the relationship between brow movements and F_0 , with Cavé et al. (1996), and more

recently Guïtella et al. (2009) acknowledging that although a large proportion of brow raises were accompanied by raises in F_0 , this is not the case for *every* occurrence of an F_0 rise. Whereas talkers are believed to be fairly consistent in their use of F_0 to signal prosodic contrasts (this however is evaluated in Section 8.3), there seems a large degree of behavioural variability both within and between talkers in the use of non-articulatory movements, supporting a functional role for these gestures rather than that they occur due to some automatic coupling with the process of speech production.

When the critical constituent was examined independent of its surrounding context, the strength of the relationship between intensity, and PC 1 and 2 (jaw and lip opening) increased in conditions where the critical word was prosodically marked (i.e., in narrow focus and echoic question renditions) relative to the baseline broad focused tokens. Similarly, the relationship between F_0 and eyebrow raises was stronger for the critical constituent in narrowly focused renditions. The perceptual effect of this increased strength of relationship will be explored further in Chapter 9.

In sum, the relationship between the auditory and visual prosodic signals appears to be highly variable, with little evidence to suggest a simple one-to-one coupling between the modalities. However, a limitation of the current analysis is the assumption that the signals are linearly related and occur at the same time (not at some delay). Given that non-articulatory gestures are not involved in the production of the acoustic signal, they are free to vary in their form but also in their timing, and as such may exhibit a variable and non-linear relationship with auditory signal properties. Given this, an alternate approach to exploring the relationship between

the modalities is to examine the timing of how non-articulatory gestures relate to segmental content. Such an analysis is detailed in the following section.

8.2. Temporal Alignment between Auditory and Visual Prosody¹⁴

As the relationship between non-articulatory visual features and the auditory signal may be in part non-linear, the current analyses explored aspects of the auditory-visual relationship that might not be adequately captured by examining the correlation values. To do this, the temporal aspects of the visual prosodic cues were examined in relation to the timing of the critical utterance phase.

The examination of the timing of auditory and visual cues to prosody is also important because it has been specifically proposed that the occurrence of eyebrow raises or a rigid tilts of the head align with cues in the auditory stream in order to enhance the “perceptual strength” of a prosodically marked constituent. That is, it has been argued that visual cues will occur in synchrony with, or in close temporal proximity to, auditory markers of prosody, and by doing so generate a more salient prosodic percept (i.e., the “alignment hypothesis”). Indeed, the results of several recent studies appear to provide some support for this hypothesis. For example, in a study by Flecha-Garcia (2010), pairs of participants were audio-visually recorded engaging in a fairly unconstrained face-to-face dialogue task. The auditory and visual recordings were then independently annotated offline for the occurrence of pitch accents (i.e., the syllable was prominent in the utterance) in the auditory stream, and of eyebrow raises in the visual signal (defined as any upward movement from a neutral baseline position of at least one eyebrow). More than 80% of eyebrow

¹⁴ A preliminary version of this analysis has previously appeared in: Cvejic, E., Kim, J., & Davis, C. (2011). Temporal relationship between auditory and visual prosodic cues. *Interspeech 2011*, pp. 981-984.

movements started within 330ms of the nearest pitch accent, with the average brow raise occurring 60ms before the onset of a pitch accent.

Further evidence supporting the alignment hypothesis is provided by Swerts and Kraemer (2010) in their analysis of non-articulatory gestures (i.e., eyebrow raises and rigid head movements) that co-occurred with pitch accents in the speech of four Dutch newsreaders. Perceivers were presented with 60 auditory-only tokens of sentences produced by the newsreaders and asked to identify the word (or words) that were clearly emphasised (with items identified as being emphasised by more than half of the perceivers being classified as having a “strong” accent, and words with less than 50% agreement across raters being labelled as possessing a “weak” accent). Two observers independently annotated the occurrence of rigid head movements (in any direction) and rapid eyebrow movements (i.e., movement of at least one eyebrow in an upward or downward direction) in visual only displays, with these annotations compared to the perceptually rated pitch accents. Their analysis revealed that 70% of strong pitch accents were accompanied by an eyebrow raise, while 89% were accompanied by a rigid head movement. In contrast, weakly accented words were accompanied by head movements on 40% of occasions, and by brow movements for only 37% of occurrences. Some brow movements were still present on non-accented words (i.e., 23.2%), but very little rigid head movement was apparent. These results suggest that talkers align the occurrence of non-articulatory visual prosodic cues with auditory correlates of prosody, in an attempt to maximize the strength of the prosodic contrast that is conveyed to the perceiver.

An alternate explanation to the alignment hypothesis for the function of visual prosody is the “signalling hypothesis” (cf. Schwartz, Berthommier &

Savariaux, 2004) which proposes that visual prosodic cues act to alert perceivers to upcoming content in the auditory speech stream that is of greatest informational relevance. If this were the case, the temporal onset and/or peak of visible gestures would occur sometime *before* the start of the prosodically marked constituent in the auditory stream. In this regard, visual cues serve to direct attention to the most informative part of the spoken message, rather than having a direct influence on the salience of the prosodically marked constituent.

The proposal that auditory-visual information is combined to mark prosody is important for models of speech production (indicating a need to take the visual modality into account) as well as for the construction of synthetic conversational agents (in order to make them more natural, see Al Moubayed, Beskow & Granström, 2010; Al Moubayed, Beskow, Granström & House, 2011). However, the studies to date that have examined the temporal relationship between the visual and auditory prosodic cues have typically used raters to designate auditory prominence and eyebrow/head motion from offline recording, with the speech data coming from relatively unconstrained procedures. The structure of the speech prosody corpus recorded in Chapter 4 however allows for the spatiotemporal properties of eyebrow and rigid head motion to be measured more objectively where the timing and production of a prosodic event can be more clearly defined (i.e., the temporal location of the critical constituent onset). Thus, the occurrence of rigid head movements and brow raises were determined from the corpus obtaining three measures: the degree of alignment between the visual and auditory prosodic cues; the temporal distribution of such cues; and how these movements varied across prosodic conditions.

8.2.1. Method

The phoneme boundaries from the auditory transcriptions (Appendix B.3) were used to locate the critical constituent (i.e., the word within the utterance that was either corrected or questioned during the dialogue exchange task in Chapter 4) within each visual token. From the auditory analysis reported in Chapter 5, the production of these constituents was characterised by an increase in syllable duration, intensity range and *F0* range. To determine how brow raises relate to an auditory prosodic event, PC 7 (corresponding to brow raising in the *Y*-axis) was examined around the auditory onset of the critical constituent by measuring the temporal displacement between the onset of the eyebrow movement, and the onset of the critical constituent. Brow raises were chosen (as opposed to lowering) as they appear to represent the majority of brow movements (see Swerts & Kraemer, 2010; Chapter 7). The direction of brow movement at the temporal onset of the critical constituent was first determined; if the eyebrows were moving in an upward direction, the next “peak” was located, before finding the temporal onset of this upward movement (see Figure 8.7). Conversely, if the eyebrows were returning to a neutral position at the time of the critical constituent, the previously occurring “peak” was temporally located, as well as the onset of that movement. A movement was considered to have occurred if the distance covered between the onset and peak of was equivalent to or greater than 3mm (Cavé et al., 1996).

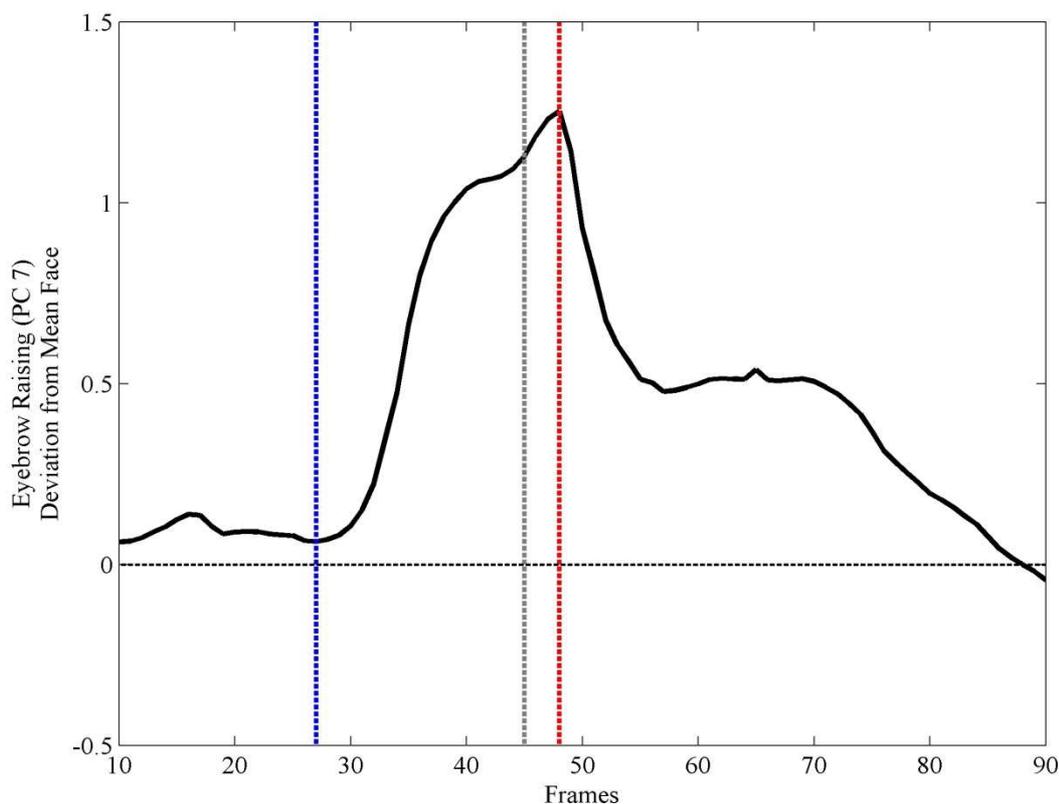


Figure 8.7. An example of temporal location of the onset (blue line) and peak (red line) of eyebrow movements in the vicinity of the critical constituent (grey line) of an utterance. A value of “0” indicates the average face position.

To determine how rigid head movement may relate to an auditory prosodic event, R1 was examined (i.e., pitch rotations around the X -axis, or “head nodding”, that accounted for the majority of rigid motion in Swerts & Kraemer, 2010). Here, the displacement between the start of the critical constituent and the *peak* of the pitch rotation was measured. A movement was only considered to have occurred if the rotation between the peak and the end of the preceding downward rotation covered a minimum of 4° (Srinivasan & Massaro, 2003).

8.2.2. Results

8.2.2.1. *Eyebrow Raising (PC 7)*

Of the 360 utterances recorded per prosodic condition and interactive setting, many failed to show any movements reaching the 3mm criteria. In the AO setting, the greatest number of criteria achieving movements occurred in the narrow focus condition ($N = 239$, 66.39%), followed by the echoic question condition ($N = 208$, 57.78%), while the broad focus condition had the fewest ($N = 175$, 48.61%). This pattern was also mirrored in the FTF setting, with the greatest amount of eyebrow raises occurring for narrow focused renditions ($N = 213$, 59.17%), followed by the echoic questions ($N = 205$, 56.94%), while the broad focus condition had the fewest ($N = 171$, 47.50%).

These distributions (for each interactive setting) were analysed with a series of one-way chi-squares (with α set to 0.05), showing no statistically significant difference in the number of utterances accompanied by brow raises across the three prosodic conditions in the FTF setting, $\chi^2(2, N = 589) = 5.07, p = 0.079$. A difference was however apparent for tokens recorded in the AO setting, $\chi^2(2, N = 622) = 9.88, p = 0.007$. It should be noted that although a large proportion of echoic question renditions showed no raising movements across both interactive settings, this type of utterance phrasing is often considered to demonstrate uncertainty, characterised by an overall smaller degree of eyebrow movements than conditions where a talker is certain or issuing confirmation (see Beskow et al., 2006; Flecha-Garcia, 2010).

Figure 8.8 displays the distribution of movements that occurred for each of the prosodic conditions and interactive settings, as a function of the time between

brow movement onset and the start of the critical constituent. A series of one-way chi-squares (with an adjusted α of 0.025 for multiple comparisons) showed that the interactive setting played no role in the number of brow raises that occurred within prosodic conditions. That is, the number of brow raises was comparable between AO and FTF settings for broad focus [χ^2 (1, $N = 346$) = 0.05, $p = 0.830$], narrow focus [χ^2 (1, $N = 452$) = 1.50, $p = 0.221$], and echoic question renditions [χ^2 (1, $N = 413$) = 0.02, $p = 0.883$].

A series of two-way chi-square analyses (with α set to .0125 for multiple comparisons) were then used to determine if there was any relationship between prosodic conditions and the temporal distribution of brow movements. For the AO setting, this relationship was significant, χ^2 (8, $N = 621$) = 25.98, $p = 0.001$; as was it for tokens recorded in the FTF setting, χ^2 (8, $N = 588$) = 28.98, $p < 0.001$.

A series of separate one-way chi-square were then used to further examine these distributions (with α set to 0.0061) for each prosodic condition and interactive setting. In the AO setting, the majority of movement onsets occurred more than 150ms before the start of the critical constituent for broad focus [χ^2 (4, $N = 175$) = 29.83, $p < 0.001$], narrow focus [χ^2 (4, $N = 239$) = 58.13, $p < 0.001$] and echoic question renditions [χ^2 (4, $N = 207$) = 73.85, $p < 0.001$]. The average brow movement onset time across the three conditions ranged between 80 and 95ms before the onset of the critical constituent. Compared to the distribution of movements in the broad focused prosodic condition (which contained no explicit point of informational focus), both the narrow focus [χ^2 (4, $N = 414$) = 17.00, $p = 0.002$] and echoic question renditions [χ^2 (4, $N = 382$) = 16.99, $p = 0.002$], contained

a greater number of criteria achieving movements earlier than 90ms before the onset of the critical constituent.

This pattern of data was also observed in FTF setting. The majority of movement onsets occurred more than 150ms before the start of the critical constituent for broad focus [$\chi^2(4, N = 171) = 25.64, p < 0.001$], narrow focus [$\chi^2(4, N = 212) = 49.08, p < 0.001$] and echoic question renditions [$\chi^2(4, N = 205) = 81.17, p < 0.001$]. Across all three prosodic conditions, the average onset time of brow raises was between 85 and 95ms before the start of the critical word. Relative to the distribution of movements in the broad focused prosodic condition, the narrow focus [$\chi^2(4, N = 212) = 21.28, p < 0.001$] and echoic question [$\chi^2(4, N = 205) = 27.54, p < 0.001$] conditions contained more eyebrow raises with temporal onsets earlier than 90ms before the start of the critical word.

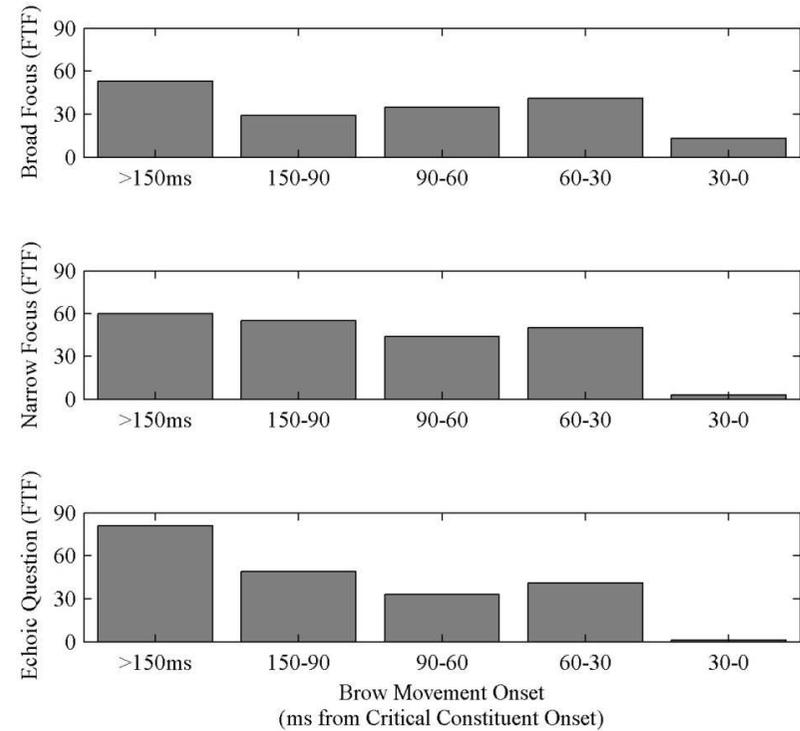
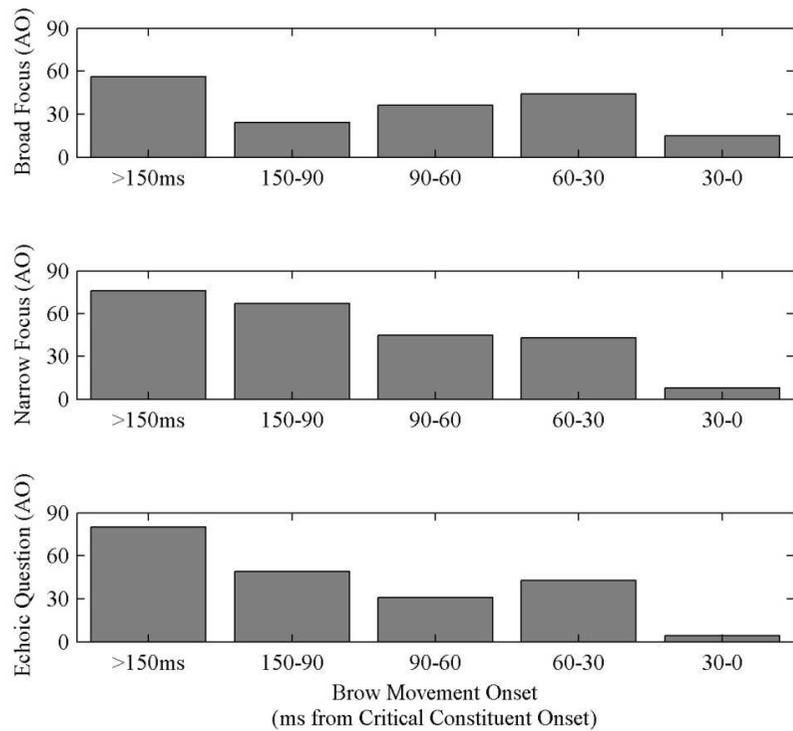


Figure 8.8. Distribution of brow raises as a function of temporal onset of movement preceding the start of the critical constituent for each prosodic condition and interactive setting ($n = 360$ in each condition).

8.2.2.2. Rigid Pitch Rotations (R 1)

As with brow movements, only a small proportion of utterances were accompanied by pitch rotations. In the AO interactive setting, the greatest amount occurred in the narrow focus condition ($N = 179, 49.72\%$), followed by the echoic questions ($N = 131, 36.39\%$), with the least occurring for broad focused utterances ($N = 47, 13.06\%$). A one-way chi-square revealed that the difference in number of utterances displaying pitch rotations between prosodic conditions in the AO condition was significant, $\chi^2(2, N = 357) = 75.03, p < 0.001$. In the FTF setting, the pattern of data was the same, with the most amount of rigid pitch movements occurring for narrow focused tokens ($N = 160, 44.44\%$), followed by echoic questions ($N = 139, 38.61\%$), with the least occurring in the broad focus condition ($N = 40, 11.11\%$). Between prosodic conditions, the number of criteria-achieving movements was significantly different, $\chi^2(2, N = 339) = 72.69, p < 0.001$.

The distribution of pitch rotations for each prosodic condition and interactive setting, as a function of the time between the peak in rotation and the start of the critical constituent are shown in Figure 8.9. Overall, there was no effect of the interactive setting in the number of criteria-achieving movements for broad focus [$\chi^2(1, N = 87) = 0.56, p = 0.453$], narrow focus [$\chi^2(1, N = 339) = 1.07, p = 0.302$], or echoic question renditions [$\chi^2(1, N = 270) = 0.24, p = 0.626$].

A series of two-way chi-squares indicated that the distribution of rigid movements peaks significantly differed as a function of the prosodic condition in both AO, $\chi^2(14, N = 333) = 71.21, p < 0.001$; and FTF settings, $\chi^2(14, N = 329) = 68.39, p < 0.001$. A series of one-way chi-squares revealed that the distributions were significantly different across the time for both interactive settings for narrow

focus [AO: $\chi^2(7, N = 179) = 49.26, p < 0.001$; FTF: $\chi^2(7, N = 160) = 38.50, p < 0.001$] and echoic question renditions [AO: $\chi^2(6, N = 131) = 49.13, p < 0.001$; FTF: $\chi^2(5, N = 139) = 41.22, p < 0.001$], but not for broad focus renditions [AO: $\chi^2(6, N = 47) = 11.09, p = 0.050$; FTF: $\chi^2(5, N = 40) = 8.90, p = 0.113$]. For narrow focus renditions, pitch rotation peaks occurred most frequently at the start of the critical constituent or only slightly before, making it possible that these movements functions to alert the perceiver. However, the peak movements were also distributed between 150ms before and 150ms after the onset (at which time the “important” part of the message has already begun); in this case the downward movement that follows the rotation peak may be contributing to transmitting suprasegmental content (i.e., reinforcing the auditory markers of focus). In echoic questions, pitch rotation peaks tended to occur before the start of the critical word, but were distributed more evenly across these time ranges.

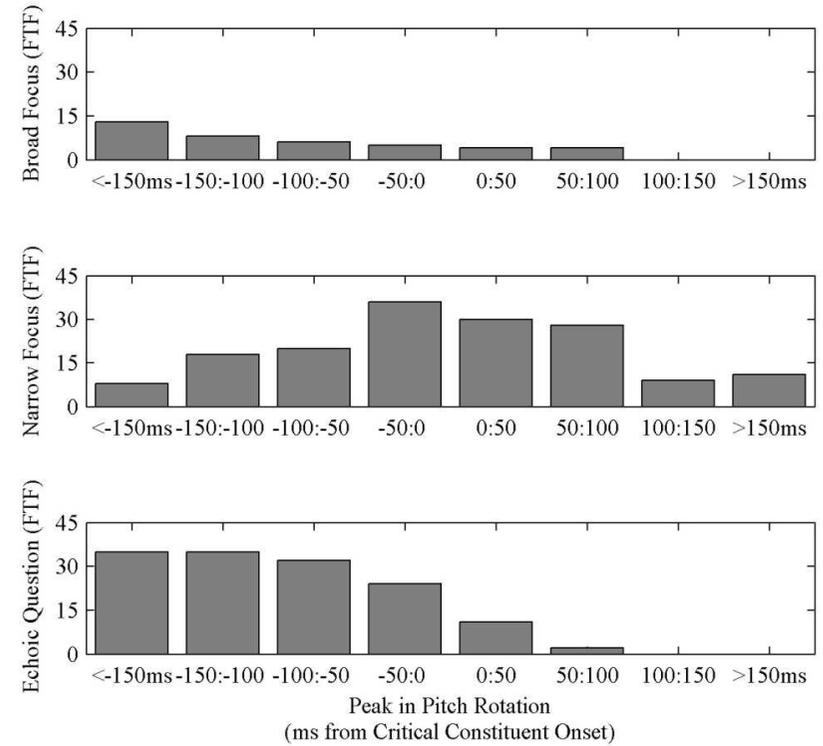
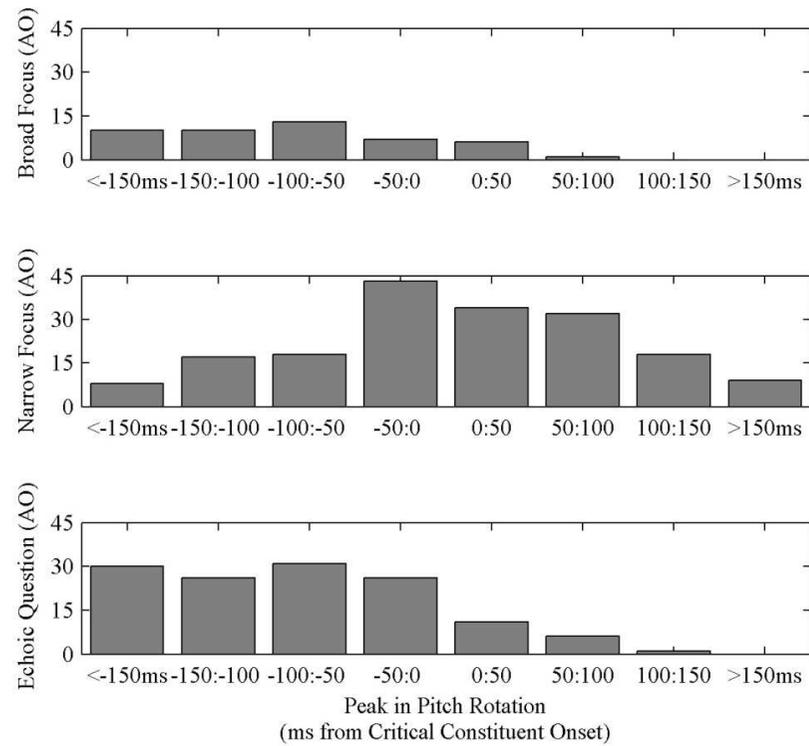


Figure 8.9. Distribution of pitch rotation peaks around the start of the critical constituent for each prosodic condition and interactive setting ($n = 360$ in each condition).

8.2.2.3. *Co-occurrence of eyebrow and rigid head movements*

Given that only a small proportion of utterances were accompanied by brow movements and rigid head movements, it is possible that such features are used interchangeably by talkers across utterances. That is, talkers may choose to visually mark a constituent in an utterance with either an eyebrow raise, or by increasing the rigid head movement. Alternatively, they may use a combination of these gestures. Indeed, in the previously mentioned study conducted by Krahmer and Swerts (2010), 67.2% of accents produced by newsreaders that were perceptually rated as being “strong” were accompanied by both eyebrow raises and rigid head movements. As such, the current analysis compared the number of non-articulatory movements that accompanied each utterance, as a function of the prosodic condition and interactive setting (i.e., whether there were no criteria-achieving movements, a single feature, or both eyebrow and rigid head movements accompanying the critical constituent). The distribution across interactive settings and prosodic conditions is displayed in Figure 8.10.

No difference was observed across interactive settings for broad focus [$\chi^2(2, N = 720) = 2.84, p = 0.241$], narrow focus [$\chi^2(2, N = 720) = 4.52, p = 0.104$] or echoic question conditions [$\chi^2(2, N = 720) = 1.43, p = 0.489$]. A series of two-way chi-squares showed that there was differences in the distribution of how many non-articulatory features accompanied the critical word as a function of the prosodic condition for both interactive settings, with more utterances being accompanied by both eyebrow raises and rigid pitch rotations in the narrow focus than broad focus condition [AO: $\chi^2(2, N = 720) = 117.30, p < 0.001$; FTF: $\chi^2(2, N = 720) = 100.30, p$

< 0.001], and for echoic questions compared to broad focused statements [AO: $\chi^2(2, N = 720) = 50.29, p < 0.001$; FTF: $\chi^2(2, N = 720) = 77.92, p < 0.001$].

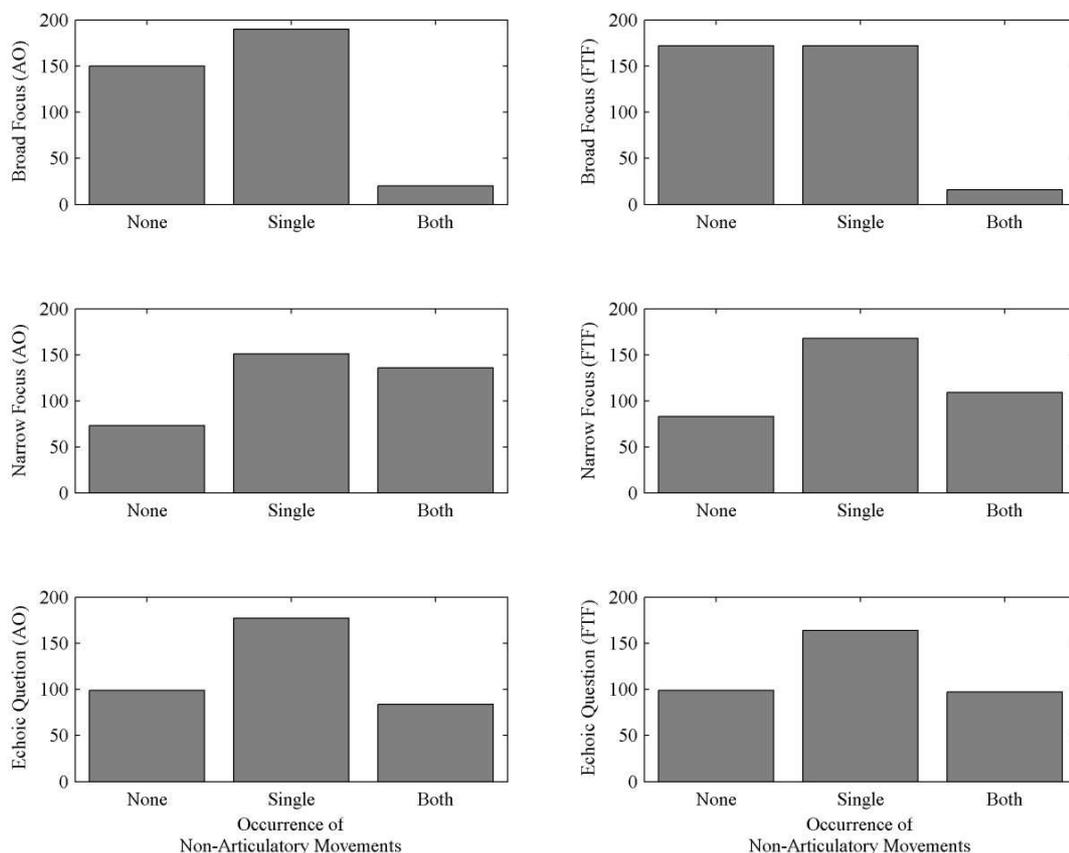


Figure 8.10. Occurrence of non-articulatory features accompanying the production of the critical constituent within utterances across prosodic conditions and interactive settings ($n = 360$ in each condition).

Although some of these movements co-occurred (primarily for the narrow focus and echoic question renditions), they may not have taken place at the same time. To examine whether this was the case, the temporal distribution of brow movement onsets and the peak in rigid pitch rotations around the onset of the critical constituent (when both non-articulatory features accompanied the utterance

production) were examined using a series of two-way chi squares. These distributions are shown in Figure 8.11.

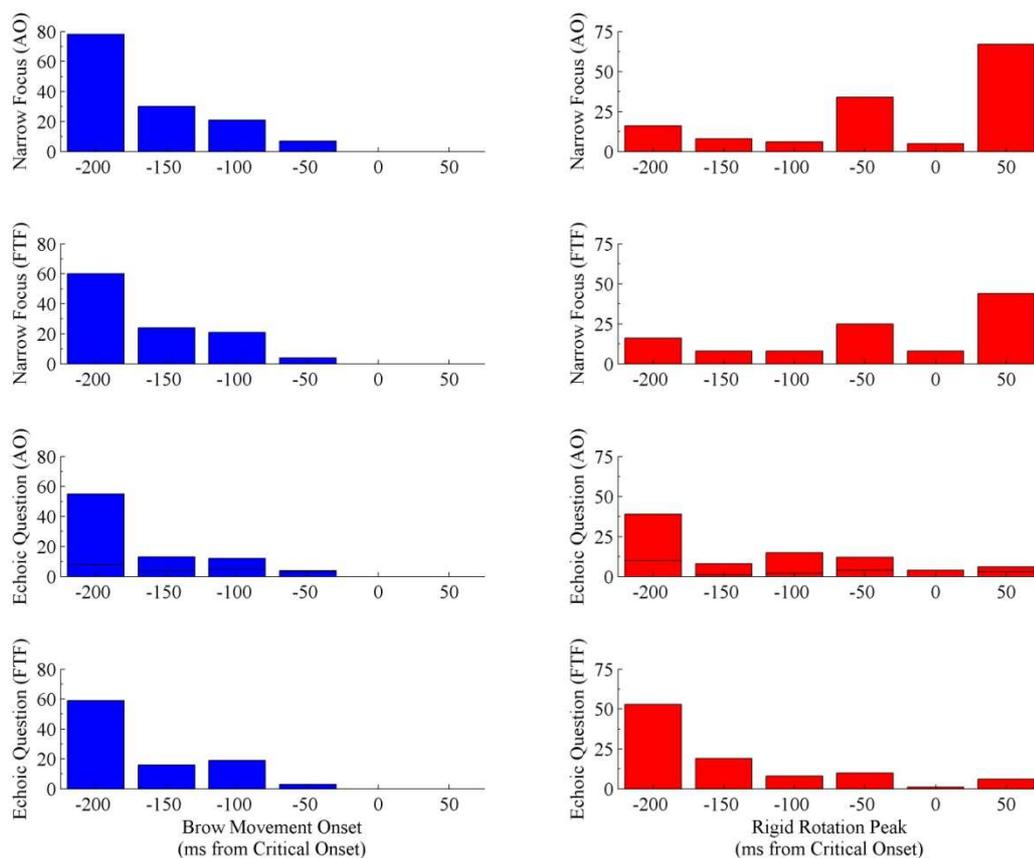


Figure 8.11. Temporal distribution of co-occurring brow movements (blue) and pitch rotation peaks (red) around the start of the critical constituent, as a function of prosodic condition and interactive setting.

When eyebrow raises and rigid pitch rotations co-occurred for narrow focused utterances, the timing of the non-articulatory movements was significantly different [AO: $\chi^2(5, N = 272) = 151.85, p < 0.001$; FTF: $\chi^2(5, N = 218) = 106.51, p < 0.001$], with the majority of brow raises occurring at least 200ms before the onset of the critical constituent, whereas the majority of rigid pitch rotations peaked

between 50ms after the word had begun to be uttered. Furthermore, these results were consistent across both AO and FTF interactive settings.

The temporal distribution of the non-articulatory movement features when they co-occurred also showed significant differences for echoic questions across both interactive settings [AO: $\chi^2(5, N = 168) = 18.25, p = 0.003$; FTF: $\chi^2(5, N = 194) = 15.83, p = 0.007$]. Whereas the eyebrow movements showed a similar pattern to the narrow focus renditions (i.e., the majority of brow movement onsets preceded the start of the critical constituent by at least 200ms), the rigid pitch rotations were distributed more evenly in time in echoic question contexts.

8.2.3. Discussion

In these analyses, the temporal relationship between visual prosodic cues (brow raising and rigid head tilts) and auditory speech were examined. With regards to brow raises, the results showed that many utterances failed to display motion that reached the minimum movement criteria, even when the contrast contained an important word (i.e., narrow focus and echoic question renditions). In cases where eyebrow raises did occur, the majority of movements began 150 ms or more before the onset of the critical constituent; this finding was not consistent with that reported by Flecha-Garcia (2010) or with the description that such events occur “in tandem” (Swerts & Kraemer, 2010), but rather suggest that such movements may function to alert perceivers to upcoming information. (Of course, it may be that the newsreaders in the Swerts and Kraemer study adopted an exaggerated style of facial expression in order to maintain the viewers’ attention).

For rigid head (pitch) rotations, the temporal location of the movement peak was also variable, extending 150ms either side of the critical constituent onset. In the

narrow focus and echoic question conditions, the majority of movement peaks occurred slightly before the onset of the critical word, making it possible that such a cue has an alerting function. However, the occurrence of a peak indicates that the head rotation changed in direction (i.e., during the critical constituent) and as such, the downward movements of the talker's head may serve to reinforce the auditory signal (however, further examination of this proposal is required). From this analysis, it may be that eyebrow and head motion cues to prosody act in different ways: with eyebrows acting to alert perceivers to, and head motion acting as confirmation of, a noteworthy event. Furthermore, it is possible that these cues are used interchangeably across utterances, with their functional roles differing dependant on when they occur within the speech signal. If so, it would seem a practical next step to examine how the occurrence of an eyebrow raise or head motion (particularly around the time of the critical utterance phase) relates to prosody perception. This is explored in Chapter 9.

8.3. *F*₀ Rises and Visual Prosodic Markers

So far, the occurrence of non-articulatory visual markers of prosody has been explored in terms of their temporal relationship to the onset of the segmental content designated as the critical constituent in the auditory stream. However, the onset of the segmental content may not reflect the timing of an auditory marker of prosody (such as a rise in *F*₀, as in Swerts & Kraemer, 2010). That is, the start of a rise in *F*₀ may not be aligned with the start of the word itself; a rise may start before the critical constituent has even begun to be uttered, or occur after the initial phoneme has already been produced (see Dilly, Ladd & Schepman, 2005; Ladd, Faulkner, Faulkner & Schepman, 1999; Ladd & Schepman, 2003). In this analysis, the

occurrence of *F0* rises around the critical constituent was examined for each utterance, and it was determined how the timing of non-articulatory gestures (if any) related to these rises.

8.3.1. Method

The typical way of examining the modulation of the *F0* contour for the purpose of prosody research is to use a standardised annotation scheme such as ToBI (Tone and Break Indices; Silverman, Beckman, Pitrelli, Ostendorf, Wightman, Price et al., 1992). However, employing such annotation systems are labour intensive and rely on trained human listeners to make subjective judgements based on what they hear (see Hirst, 2005; Wightman, 2002). Given that the dialogue task used to collect the data for the current corpus was well structured with the constituent expected to receive a prosodic marker known a priori (i.e., the critical word), an automatic pitch rise detection technique (similar to that used in Section 8.2 to detect eyebrow raises and rigid pitch rotations) was utilised to locate the temporal onset of rises and peaks in the extracted *F0* contour for each utterance.

The *F0* contour at either side of the temporal onset of the critical constituent was first examined to determine whether the *F0* was rising or falling. In the case that the contour was falling, the next occurring minimum was identified before identifying the next nearest *F0* peak. By contrast, if the *F0* was already rising at the temporal onset of the critical constituent, the start of this movement was traced back to the nearest minimum, before identifying the temporal location of the *F0* peak (Figure 8.12). As human detection of *F0* change is quite accurate (Flanagan & Saslow, 1958), an *F0* rise was considered to occur if the difference between the *F0* minimum and peak values was at least 15 Hz (corresponding to around 10-15 % of

the F_0 range covered by the average male talker, e.g., see Nootboom, 1997). The occurrence and temporal relation of brow raises (PC 7) and rigid pitch rotations (R 1) to F_0 rises were then examined.

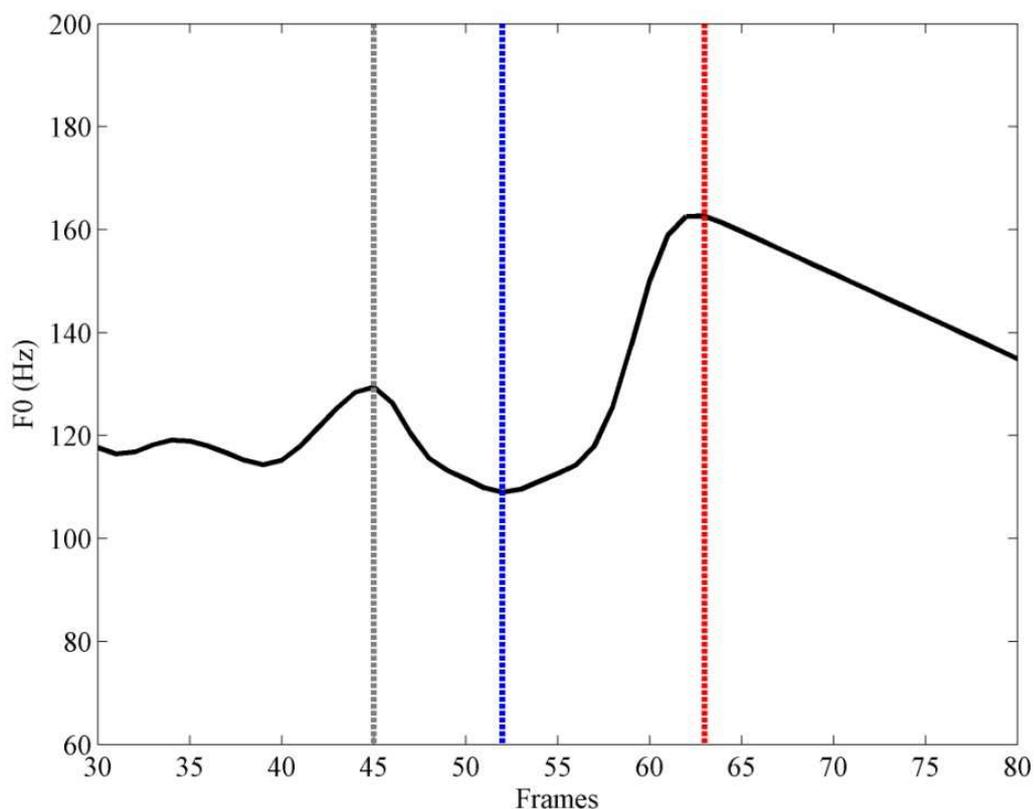


Figure 8.12. F_0 rises were detected automatically by examining the F_0 contour around the temporal onset of the critical constituent (grey line). A pitch rise was considered to have occurred if the difference between the F_0 minimum (blue line) and peak (red line) was at least 15 Hz.

8.3.2. Results and Discussion

Table 8.4 shows the distribution of non-articulatory gestures accompanying the production of the critical constituent, dependant on the occurrence of an F_0 rise. As expected, a substantial proportion of broad focused tokens were not accompanied by a rise in F_0 during the production of the critical utterance phase (i.e., 69% in the AO

Chapter 8: The Relationship between Auditory and Visual Prosody

setting, and 77% in the FTF setting), however more than 50% of these utterances were still produced with an eyebrow raise or rigid pitch rotation even in the absence of an *F0* rise. Similarly, although more than 60% of narrow focused and echoic question utterances contained an *F0* rise, around 70% of those that did not were still accompanied by some form of visual marker of prosody (with between 20 and 30% being accompanied by both an eyebrow raise and a rigid pitch rotation).

Table 8.4. Distribution of non-articulatory gestures accompanying the production of the critical constituent, dependant on the occurrence of an *F0* rise, as a function of the interactive setting and prosodic condition.

Prosodic Condition	<i>F0</i> Modulation	<i>F0</i> Properties		No Movement		Brow Raising Movement		Rigid Pitch Rotation		Both Movement Features	
		Count	%	Count	%	Count	%	Count	%	Count	%
AO Interactive Setting											
Broad	<i>F0</i> Rise	110	30.56	53	48.18	48	43.64	6	5.45	3	2.73
	No <i>F0</i> Rise	250	69.44	97	38.80	115	46.00	21	8.40	17	6.80
Narrow	<i>F0</i> Rise	221	61.39	52	23.53	70	31.67	24	10.86	71	32.13
	No <i>F0</i> Rise	139	38.61	21	15.11	38	27.34	19	13.67	61	43.88
Echoic	<i>F0</i> Rise	235	65.28	61	25.96	89	37.87	32	13.62	53	22.55
	No <i>F0</i> Rise	125	34.72	38	30.40	41	32.80	15	12.00	31	24.80
FTF Interactive Setting											
Broad	<i>F0</i> Rise	84	23.33	49	58.33	30	35.71	3	3.57	2	2.38
	No <i>F0</i> Rise	276	76.67	123	44.57	118	42.75	21	7.61	14	5.07
Narrow	<i>F0</i> Rise	217	60.28	59	27.19	71	32.72	32	14.75	55	25.35
	No <i>F0</i> Rise	143	39.72	24	16.78	46	32.17	19	13.29	54	37.76
Echoic	<i>F0</i> Rise	220	61.11	52	23.64	75	34.09	25	11.36	68	30.91
	No <i>F0</i> Rise	140	38.89	47	33.57	47	33.57	17	12.14	29	20.71

The temporal distribution of visual prosodic markers in relation to the onset of *F0* rises is shown in Table 8.5. In situations where an *F0* rise was accompanied by non-articulatory movement, the majority of movements (~30% across all prosodic conditions and interactive settings) were eyebrow raises, the onset of which preceded the start of the *F0* rise (see Figure 8.13). These temporal distributions were examined in a series of one-way chi-squares for each prosodic condition and interactive setting. For broad focused utterances, the majority of eyebrow raises preceded the *F0* rise by 200ms in both AO and FTF settings [AO: $\chi^2(5, N = 48) = 19.25, p = 0.002$; FTF: $\chi^2(5, N = 30) = 15.60, p = 0.008$]. The majority of brow raises for narrow focused utterances preceded *F0* rises by at least 100ms [AO: $\chi^2(5, N = 70) = 42.11, p < 0.001$; FTF: $\chi^2(5, N = 71) = 27.62, p < 0.001$]. A similar observation was made for echoic questions [AO: $\chi^2(5, N = 89) = 139.07, p < 0.001$; FTF: $\chi^2(5, N = 75) = 70.52, p < 0.001$], with the onset of eyebrow raises occurring in excess of 200ms before the start of a rise in *F0*. Across all the prosodic conditions, very few eyebrow raises occurred after the onset of an *F0* rise; this further supports the proposal that eyebrow movements act as a signalling device to alert perceivers that the upcoming information in the auditory signal is likely to be important.

When an *F0* rise was accompanied only by a rigid pitch rotation (i.e., no eyebrow raising), the temporal distribution differed depending on the prosodic context (see Figure 8.14). For broad focused utterances, there were very few criteria-achieving movements (i.e., less than 6%). For narrow focus, the peak of these movements tended to occur more than 50ms after the *F0* rise had already begun [AO: $\chi^2(4, N = 24) = 28.50, p < 0.001$; FTF: $\chi^2(3, N = 32) = 48.25, p < 0.001$], suggesting that these movements may occur somewhat in tandem with the rising of

F_0 , acting to reinforce the auditory content. Conversely, the peak in rigid pitch rotations for echoic questions tended to precede the onset of an F_0 rise, with the majority of these movements occurring at least 200ms beforehand [AO: $\chi^2(4, N = 32) = 31.75, p < 0.001$; FTF: $\chi^2(5, N = 25) = 25.16, p < 0.001$]. In this case, the rigid head movement may serve a similar functional role to eyebrow movements.

Finally, if an F_0 rise was accompanied by an eyebrow raise and a rigid pitch rotation (as was the case for ~25% of narrow focus and echoic question renditions), the timing of these movements tended to differ as a function of the prosodic condition (Figure 8.15). That is, for narrow focused items, the brow movements typically occurred 200ms before the rise in F_0 , whereas the rigid head movement peaked more than 50ms after the F_0 had already risen [AO: $\chi^2(5, N = 150) = 78.03, p < 0.001$; FTF: $\chi^2(5, N = 110) = 37.16, p < 0.001$]. In contrast, the timing of brow and rigid head movements for echoic questions was similar [AO: $\chi^2(5, N = 106) = 6.88, p = 0.230$; FTF: $\chi^2(5, N = 136) = 10.62, p = 0.059$], with the majority of both movements occurring more than 200ms before the start of a pitch rise. The function of these relationships will be explored further in the perception study reported in Chapter 9.

Table 8.5. Temporal distribution of non-articulatory gestures in relation to the onset of an *F0* rise, as a function of the interactive setting and prosodic condition. Values within brackets indicate the number of movements expressed as a percentage value.

Prosodic Condition	Number of <i>F0</i> Rises	No Movement	Brow Raising Movement		Rigid Pitch Rotation Only		Both Movement Features		
			Only		Preceding <i>F0</i> Rise	Following <i>F0</i> Rise	Both Preceding	One Preceding	Both Following
			Preceding <i>F0</i> Rise	Following <i>F0</i> Rise					
AO Interactive Setting									
Broad	110	53 (48.2)	42 (38.2)	6 (5.5)	5 (4.5)	1 (0.9)	2 (1.8)	1 (0.9)	0 (0.0)
Narrow	221	52 (23.5)	61 (27.6)	9 (4.1)	8 (3.6)	16 (7.2)	52 (23.5)	18 (8.1)	1 (0.5)
Echoic	235	61 (26.0)	85 (36.2)	4 (1.7)	29 (12.3)	3 (1.3)	50 (21.3)	3 (1.3)	0 (0.0)
FTF Interactive Setting									
Broad	84	49 (58.3)	26 (31.0)	4 (4.8)	3 (3.6)	0 (0.0)	2 (2.4)	0 (0.0)	0 (0.0)
Narrow	217	59 (27.2)	63 (29.0)	8 (3.7)	7 (3.2)	25 (11.5)	41 (18.9)	13 (6.0)	1 (0.5)
Echoic	220	52 (23.6)	72 (32.7)	3 (1.4)	20 (9.1)	5 (2.3)	68 (30.9)	0 (0.0)	0 (0.0)

Chapter 8: The Relationship between Auditory and Visual Prosody

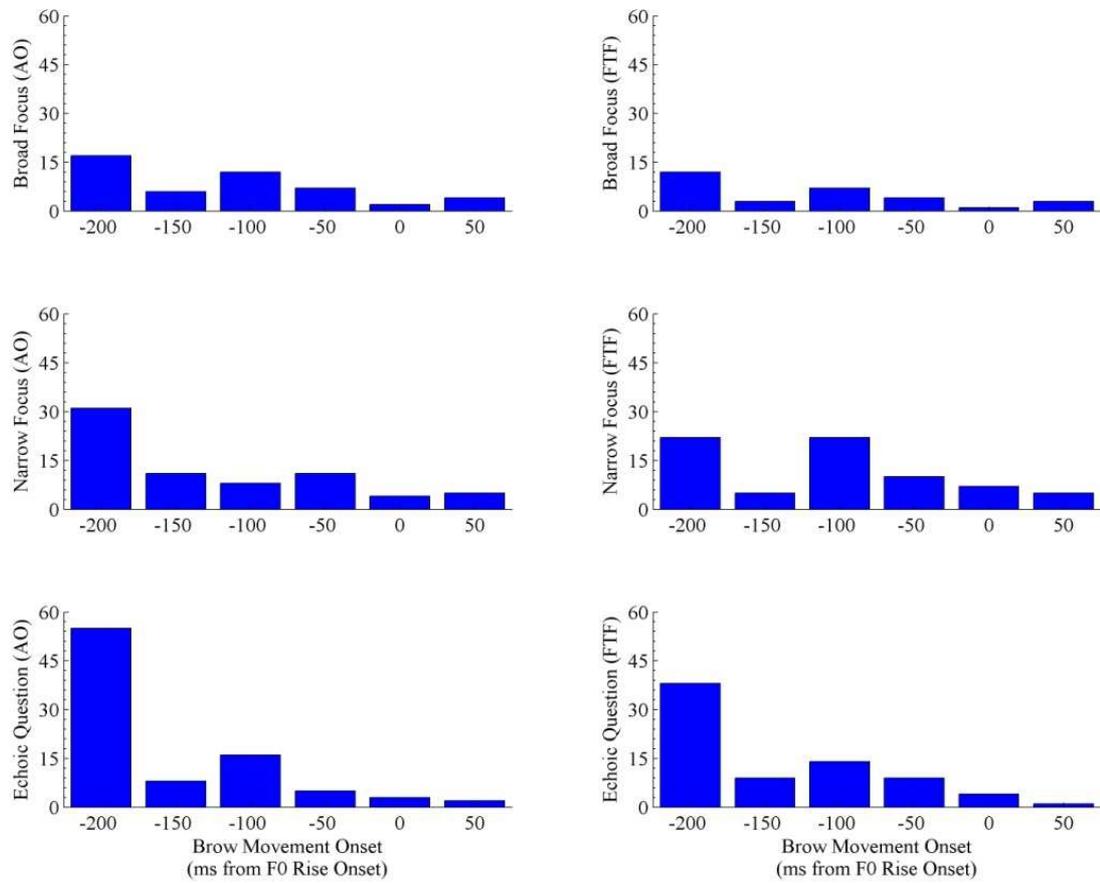


Figure 8.13. Temporal distribution of brow movement onsets co-occurring with an *F0* rise, presented as a function of the prosodic condition and interactive setting.

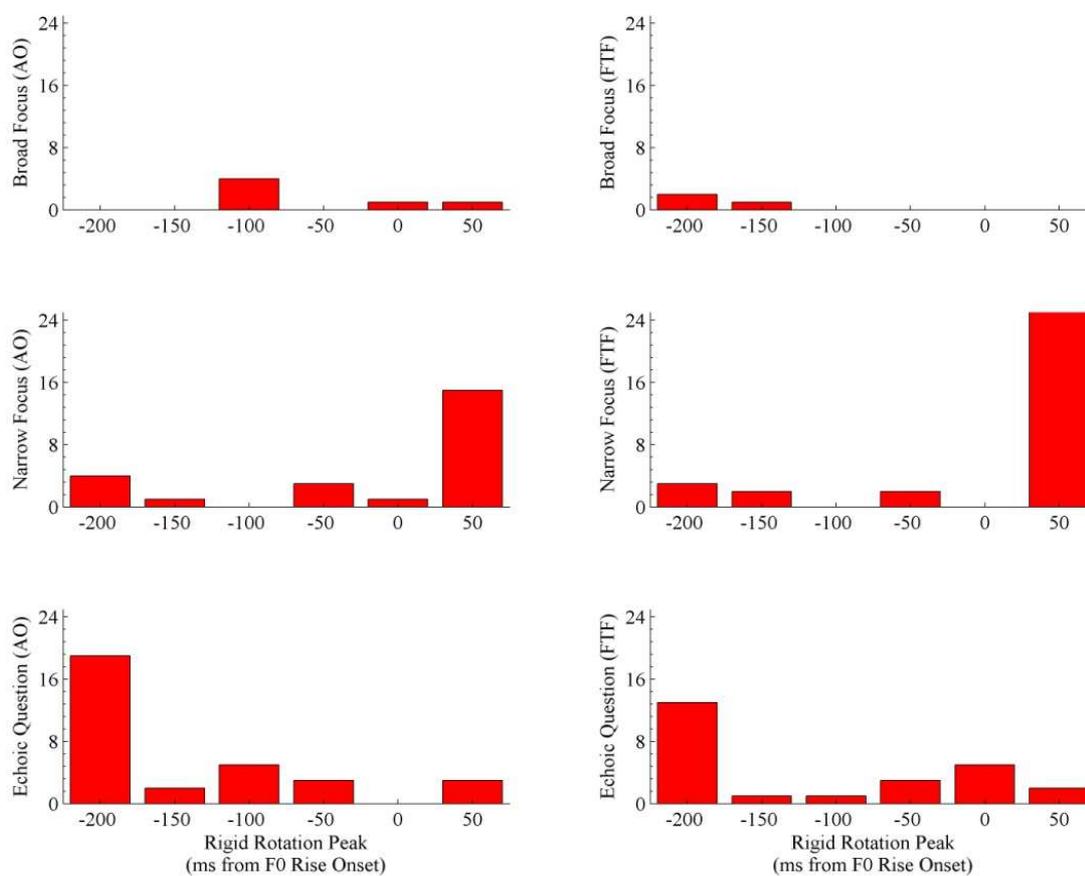


Figure 8.14. Temporal distribution of pitch rotation peaks co-occurring with an F_0 rise, presented as a function of the prosodic condition and interactive setting.

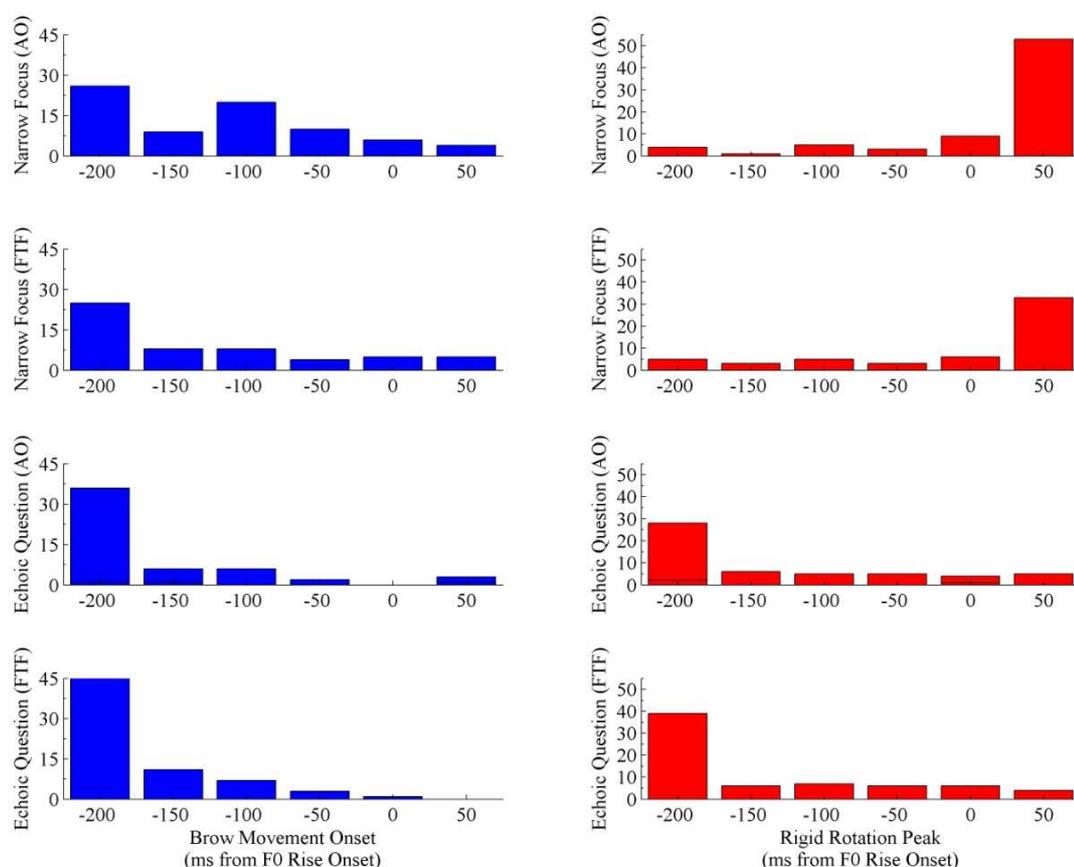


Figure 8.15. Temporal distribution of brow movement onsets (blue) and rigid rotation peaks (red) when both non-articulatory features accompany an F_0 rise, presented as a function of the prosodic condition and interactive setting.

The above temporal distributions provide further evidence that the production of non-articulatory visual prosodic markers is disengaged from auditory prosodic features. Furthermore, the occurrence of visible gestures in the absence of an auditory prosodic marker suggests that auditory and visual features may be used interchangeably to signal equivalent content; in the absence of one feature (i.e., in this case an acoustic rise in F_0), a visual marker can be used instead. This idea can be partially tested by examining whether visual prosody itself can be perceived as well as auditory prosody in a similar rating task to the one used in Chapter 6, and

whether any of the variance in rating data can be accounted for by the occurrence (or co-occurrence) of auditory and visual prosodic markers.

8.4. Summary

In this chapter, the relationship between the auditory and visual properties of linguistic prosodic contrasts (recorded in Chapter 4) was examined. The outcome of these analyses suggests that there are relationships between auditory and visual prosodic signals but these are highly variable. This was the case for the relationship between acoustic properties (i.e., intensity and *F0* contours) and articulatory movements (jaw opening, lip opening and lower lip movement), even though these movements are directly involved in the production of the segmental content.

There was a relationship between non-articulatory gestures (eyebrow and rigid head movement) and acoustic features and this too was highly variable across utterances. Given this variability, it is unlikely that the production of non-articulatory visual prosodic correlates occurs as a general by-product of speech production.

Examination of the temporality of these movements around the onset of the critical word, and the occurrence of an acoustic prosodic marker (*F0* rise) showed that although the temporal nature of non-articulatory gestures was also variable, the majority of eyebrow movements began before the start of the critical word (and before the onset of *F0* rises). This result is consistent with the idea that these movements function to prepare perceivers for upcoming information that may be important in the auditory stream. Rigid pitch rotations of the head seem to play a similar role for echoic questions, with the peak in movement occurring before a critical constituent had been uttered. By contrast, the peak rotation in movement for narrow focused renditions occurred sometime after the *F0* had risen, suggesting that

Chapter 8: The Relationship between Auditory and Visual Prosody

such movements may function to reinforce the prosody conveyed in the auditory modality.

CHAPTER 9.
PERCEPTUAL RATING OF AUDITORY-VISUAL
PROSODY

Chapter 9. Perceptual Rating of Auditory-Visual Prosody

In the previous chapters, the auditory and visual correlates of prosody, and their relationship with each other, were examined. Signal differences were found in both modalities across prosodic contrasts and the interactive settings; however the spatial and temporal relationship between the two modalities was highly variable across utterances. Although differences were found across prosodic conditions and interactive settings at the signal level, this does not mean that these should be perceptually salient or contribute in any way to the perception of prosody.

To examine this, the experiments reported in the current chapter investigated the way that the visual prosodic features (visual cues, as measured in Chapter 7) related to the perception of prosody. This was done by determining how well prosody was perceived when presented in auditory alone (AA), visual only (VO), and auditory-visual (AV) conditions by means of a subjective rating task similar to the one used in Chapter 6, then finding out how this performance measure was associated with the visual cues. A straightforward hypothesis to begin with is that the perceptual salience of prosodic contrasts would be enhanced when both auditory and visual information was available. This would be reflected by higher rating scores when utterances were presented in AV conditions (compared to AA presentations). To further explore this hypothesis, a series of multiple regression analyses were conducted on the rating scores obtained for VO and AV presentations.

Rather than real video displays, Experiment 8 made use of point-light representations of the talker's speech-related head and face movements for the visual stimuli. The reason for using point-light displays is that these offer a high degree of

experimental control, while providing an accurate representation of the underlying motion (hence their tradition of use in studies of biological motion, gesture and speech; see Hill & Pollick, 2000; Johansson, 1973; Rosenblum, Johnson & Saldaña, 1996). In addition, point-light displays lack textural information (e.g., skin wrinkling, eye widening or closing, nostril flaring) and although these features might play some role in signalling prosody, such features would not be represented in the movement data (as measured by OPTOTRAK in Chapter 4). Thus, given that it is indeed the motion information that is responsible for the conveying prosody in the visual modality (and not textural details), then the point light representations should be able to provide sufficient information for perceivers to visually discriminate between the prosodic conditions. That is, perceivers are sensitive to the visual expression of prosody, as shown in Experiments 1 to 6 (Cvejic, Kim & Davis, 2010, in press; Dohen & Lævenbruck, 2009; Foxtan et al., 2010; Lansing & McConkie, 1999; Srinivasan & Massaro, 2003), so if motion per se is sufficient to represent prosodic information, then it was expected that the ratings would differ significantly across focus contrasts (broad vs. narrow focus) and phrasing types (statements vs. echoic questions), even when perceivers were presented with only point-light visual stimuli.

In addition, as the results of Experiment 7 showed differences in subjective ratings across prosodic contrasts on the basis of auditory properties, it was expected that this result would be replicated (i.e., significantly different ratings would be obtained between broad and narrow focus, between statements and echoic questions, and between AO versus FTF when presented in the AA condition).

9.1. Experiment 8: Perceptual Rating of Prosody across Modalities

9.1.1. Method

9.1.1.1. Participants

Twenty undergraduate psychology students from UWS took part in the experiment in exchange for course credit. All participants were fluent talkers of English, and self-reported normal or corrected-to-normal vision, normal hearing, and no known communicative deficits. None of these participants had taken part previously in any of the reported experiments.

9.1.1.2. Materials

Ten sentences were selected from the audio-visual speech prosody corpus for use as stimuli (Chapter 4), corresponding to the items used in the original experimental series (i.e., Experiments 1 to 6). Broad focus, narrow focus, and echoic question renditions produced by all six talkers in both AO and FTF interactive settings were used (see Table 9.1).

Table 9.1. Stimuli sentences used in each task version of Experiment 8. The critical constituent of each utterance is italicised.

Segmental Content	Set 1 (Talkers)	Set 2 (Talkers)
<u>Focus Rating Task</u>		
It is a band of <i>steel</i> three inches wide	1, 4, 6	2, 3, 5
The pipe ran almost the <i>length</i> of the ditch.	1, 4, 6	2, 3, 5
It was hidden from sight by a <i>mass</i> of leaves and shrubs.	1, 4, 6	2, 3, 5
The weight of the <i>package</i> was seen on the high scale.	1, 4, 6	2, 3, 5
Wake and rise, and <i>step</i> into the green outdoors.	1, 4, 6	2, 3, 5
<u>Phrasing Rating Task</u>		
The green light in the <i>brown</i> box flickered.	2, 3, 5	1, 4, 6
The brass <i>tube</i> circled the high wall.	2, 3, 5	1, 4, 6
The lobes of her ears were <i>pierced</i> to hold rings.	2, 3, 5	1, 4, 6
Hold the <i>hammer</i> near the end to drive the nail.	2, 3, 5	1, 4, 6
Next <i>Sunday</i> is the twelfth of the month.	2, 3, 5	1, 4, 6

Stimuli were then created in three presentation conditions: auditory-alone (AA), visual only (VO) and auditory-visual (AV). For the AA stimuli, the mean intensity of each utterance (produced by each talker in both AO and FTF settings) was normalised to approximately 65dB using Praat (Boersma, 2001).

To create the VO stimuli, the shape normalised optical marker locations (Chapter 7) were converted to point-light representations using custom-written scripts in Matlab. Marker positions that were missing due to occlusion or drop-out were first recovered using native *b*-spline interpolation functions in Matlab. The positions of the optical markers were represented in three-dimensional space by solidly filled white dots on a black background. To aid participants in the interpretation of visual stimuli, the point lights were augmented with animated lips.

This was achieved by first creating “phantom” marker positions below and slightly in front of the mid-upper lip marker, and above and slightly in front of the mid-lower lip marker, before superimposing a series of colour-filled triangles to join the lip markers. Eyebrows were also added by connecting the outer to mid brow, and mid brow to inner brow markers with solid lines. A “nose” was also added by joining the nose bridge marker to the nose tip marker. To elicit a three-dimensional percept, and to make movements in the z -axis (e.g., head and lip protrusions) more apparent, the point light talkers were presented “looking” approximately 30° to the left. An example frame of the point-light talkers with visual features is shown in Figure 9.1 (with an animated rendition included as Appendix F.1). Videos were created at 60Hz to match the original recording resolution.

The AV stimuli were created by dubbing the VO items with the corresponding auditory token in VirtualDub (Lee, 2008). This was always a different repetition to the one used for the AA and VO item, so that participants were never presented with the same auditory or video item more than once.

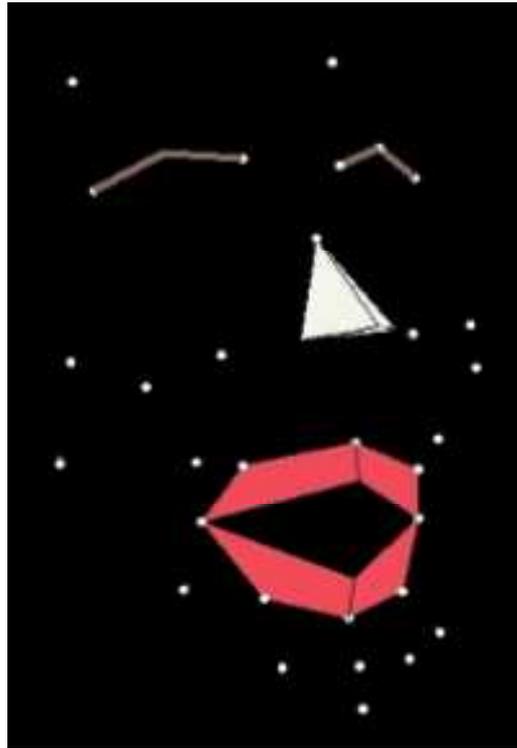


Figure 9.1. Example frame of a point-light talker used as stimuli in the VO and AV presentation modalities. An animated rendition is included as Appendix F.1.

9.1.1.3. Procedure

The stimuli were presented to participants in two rating tasks (of similar design to those used in Experiment 7), one rating the degree of focus, and the other the perceptual clarity of the statement-question contrast, with each item presented in all three presentation modalities (i.e., AA, VO and AV). Considering the number of conditions (3 [presentations] x 2 [interactive settings] x 6 [talkers]), five sentences were used for the focus rating task, while the remaining five were used for the phrasing rating task. To minimise any talker effect, two sets of the task were created dividing the talkers (see Table 9.1). Each participant completed both the focus and phrasing rating tasks in counter-balanced order with one of the stimuli sets (i.e., if a participant completed the focus task with stimuli produced by Talkers 1, 4 and 6,

they then completed the phrasing task with stimuli produced by Talker 2, 3 and 5, or vice versa).

DMDX was used for stimulus presentation and collection of responses (Forster & Forster, 2003), with auditory stimuli presented over Senheiser HD650 stereo headphones. For the focus rating task, participants were initially presented with the segmental content of the sentence printed in text along the top of the screen (with the critical constituent clearly indicated in bold typeface and underlined) for 1500ms, followed by the stimulus item in one of the presentation modalities (AA, VO, or AV). During presentation of the stimulus item, the segmental content remained on screen. Participants were then asked to rate the degree of focus received on the critical constituent within the token using a 7-point Likert scale (with a response of “1” indicating no focus and “7” indicating that the critical constituent was clearly focused). In each stimuli set, 180 items were presented (comprising of five sentences in three presentation modalities across two interactive settings [AO; FTF], in two focus conditions [broad; narrow] produced by three talkers). Presentation of items was blocked by talker with presentation order between- and within-blocks randomised by the presentation software.

The phrasing task was similar to the focus rating task, except that participants were requested to rate the utterance on a continuum of “statement” (with a response of “1”) to “clearly phrased question” (by responding with “7”). No segmental content was provided on screen. A total of 180 items were presented (5 sentences \times 3 presentation modalities \times 2 interactive settings \times 2 phrasings [statement; echoic question] \times 3 talkers). For both tasks, participants were informed that there was no “correct” answer, and were encouraged to use the complete range of the rating scale

responses. No explicit instruction was given to perceivers as to what features to base the rating judgement on.

9.1.2. Results

The rating data was subjected to a series of repeated measures ANOVAs for a subject analysis (F_S , collapsed across talkers and sentences) with prosodic condition, interactive settings and presentation modality as within-items factors; and an item analysis (F_I , collapsed across raters and talkers).

9.1.2.1. Ratings of Focus across Modalities

The mean ratings of focus contrasts, collapsed across sentences and talkers, are shown in Figure 9.2. Given that the results of Experiment 7 showed a difference between interactive settings (i.e., AO recordings were rated higher than those recorded in FTF setting) for narrow focused tokens presented in an AA condition, a series of paired-samples t -tests were conducted between interactive settings for broad and narrow focused tokens for each presentation modality. The outcome of this analysis revealed that regardless of the presentation modality, broad focus renditions were rated similarly across AO and FTF settings [$t_{AA}(9) = 0.85, p = 0.417$; $t_{VO}(9) = 0.36, p = 0.725$; $t_{AV}(9) = 0.11, p = 0.918$]. In contrast, a significant difference was found between AO and FTF ratings of narrow focus productions in the AA presentation modality consistent with the results of Experiment 7 [$t_{AA}(9) = 4.87, p = 0.001$], with AO renditions being rated significantly higher. However, this effect was not maintained in the VO [$t_{VO}(9) = 0.02, p = 0.985$] or AV [$t_{AV}(9) = 2.06, p = 0.070$] presentation modalities.

The ANOVA yielded a significant main effect of prosody, $F_S(1,9) = 242.32, p < 0.001, \eta_p^2 = 1.00$; $F_I(1,4) = 729.89, p < 0.001, \eta_p^2 = 0.995$, with broad focused

renditions (collapsed across presentation modalities and interactive settings) being rated significantly lower (i.e., less focus on the critical constituent) than narrow focused tokens, as expected. There was no main effect of interactive setting, $F_S(1,9) = 2.89, p = 0.123, \eta_p^2 = 0.331$; $F_I(1,4) = 2.02, p = 0.228, \eta_p^2 = 0.336$, but a significant main effect of presentation modality was found, $F_S(2,18) = 12.41, p < 0.001, \eta_p^2 = 0.988$; $F_I(2,8) = 24.51, p < 0.001, \eta_p^2 = 0.860$. The interaction was significant between prosody and presentation modality, $F_S(2,18) = 39.29, p < 0.001, \eta_p^2 = 1.00$; $F_I(2,8) = 62.91, p < 0.001, \eta_p^2 = 0.940$ but not for the prosody by interactive setting, $F_S(1,9) = 1.90, p = 0.202, \eta_p^2 = 0.234$; $F_I(1,4) = 0.36, p = 0.582, \eta_p^2 = 0.082$; and the interactive setting by presentation modality, $F_S(2,18) = 1.46, p = 0.259, \eta_p^2 = 0.271$; $F_I(2,8) = 0.93, p = 0.435, \eta_p^2 = 0.188$. The three-way interaction was also not significant, $F_S(2,18) = 0.99, p = 0.391, \eta_p^2 = 0.195$; $F_I(2,8) = 0.78, p = 0.489, \eta_p^2 = 0.164$.

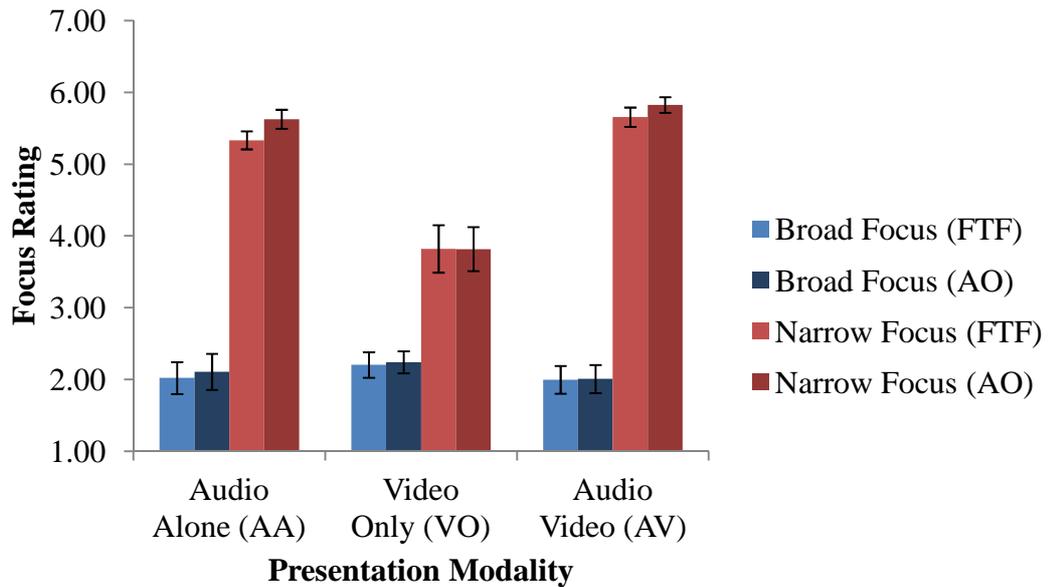


Figure 9.2. Mean rating of focus (collapsed across sentences and talkers) as a function of prosodic condition, interactive setting and presentation modality. Error bars indicate the standard error of the mean.

To interpret the interaction between prosody and presentation modality, a series of post-hoc repeated measures ANOVAs were conducted for each prosodic condition (separately for each interactive setting). For the broad focused renditions recorded in the AO setting, there was no significant differences across presentation modalities, $F(2,18) = 1.58$, $p = 0.234$, $\eta_p^2 = 0.149$. Similarly, rating scores were not different across presentation modalities for broad focus tokens recorded in the FTF setting, $F(2,18) = 1.04$, $p = 0.372$, $\eta_p^2 = 0.104$.

For narrow focused AO renditions, there was a significant effect of presentation modality, $F(2,18) = 29.18$, $p < 0.001$, $\eta_p^2 = 0.764$. Sidak pairwise comparisons (with 97.5% confidence intervals) revealed that presentations in the AA modality were rated significantly higher than VO presentations [$M_{\text{Diff}} = 1.81$, Sidak 97.5% CI: 0.74 – 2.89]; while AV presentations also resulted in significantly higher

ratings than VO presentations [$M_{\text{Diff}} = 2.01$, Sidak 97.5% CI: 0.79 – 3.23]. However, there was no difference in ratings between AA and AV presentations [$M_{\text{Diff}} = 0.20$, Sidak 97.5% CI: -0.66 – 0.26].

For narrow focused tokens recorded in the FTF setting, there was a significant effect of presentation modality, $F(2,18) = 19.08$, $p < 0.001$, $\eta_p^2 = 0.679$. Sidak pairwise comparisons (with 97.5% confidence intervals) revealed that, as for the narrow focus AO tokens, AA presentations resulted in significantly greater ratings than VO presentations, [$M_{\text{Diff}} = 1.51$, Sidak 97.5% CI: 0.35 – 2.67]. This was also the case for AV compared to VO presentations [$M_{\text{Diff}} = 1.84$, Sidak 97.5% CI: 0.46 – 3.22]; but no difference was observed between AA and AV presentations [$M_{\text{Diff}} = 0.32$, Sidak 97.5% CI: -0.73 – 0.08].

9.1.2.2. Ratings of Phrasing across Modalities

The mean ratings of the phrasing contrasts, collapsed across sentences and talkers, are shown in Figure 9.3. As with the focus ratings, the results of Experiment 7 showed a difference between interactive settings (i.e., AO recordings were rated higher than those recorded in FTF setting) for echoic question tokens presented in an AA condition. A series of paired-samples t -tests were conducted between interactive settings for statements and echoic questions for each presentation modality. The outcome of this analysis revealed that, regardless of the presentation modality, statements were rated similar between AO and FTF settings [$t_{\text{AA}}(9) = 0.16$, $p = 0.874$; $t_{\text{VO}}(9) = 1.08$, $p = 0.307$; $t_{\text{AV}}(9) = 0.74$, $p = 0.480$]. Consistent with the results of Experiment 7, a significant difference was once again found between AO and FTF ratings of echoic questions when presented in the AA modality [$t_{\text{AA}}(9) = 3.99$, $p = 0.003$], with AO renditions being rated as better renditions of questions. However,

this effect was not observed for the VO [$t_{VO}(9) = 0.66, p = 0.295$] or AV [$t_{AV}(9) = 0.06, p = 0.951$] modalities.

The results of the ANOVA showed a significant main effect of prosody, $F_S(1,9) = 445.05, p < 0.001, \eta_p^2 = 0.980$; $F_I(1,4) = 13440.15, p < 0.001, \eta_p^2 = 1.00$, with statement renditions (collapsed across presentation modalities and interactive settings) being rated significantly more “statement-like” than echoic questions. No main effect of interactive setting was found, $F_S(1,9) = 0.75, p = 0.409, \eta_p^2 = 0.077$; $F_I(1,4) = 0.90, p = 0.395, \eta_p^2 = 0.184$. The main effect of presentation modality was significant in the subject analysis, $F_S(2,18) = 4.86, p = 0.020, \eta_p^2 = 0.351$; but not in the item analysis, $F_I(2,8) = 1.80, p = 0.226, \eta_p^2 = 0.310$. The prosody by presentation modality interaction was significant, $F_S(2,18) = 311.79, p < 0.001, \eta_p^2 = 0.972$; $F_I(2,8) = 697.24, p < 0.001, \eta_p^2 = 0.994$. The interaction between interactive setting and presentation modality was significant in the subject analysis, $F_S(2,18) = 5.09, p = 0.018, \eta_p^2 = 0.361$, but not for the item analysis, $F_I(2,8) = 2.37, p = 0.156, \eta_p^2 = 0.372$. Finally, no significant effects were observed for the prosody by interactive setting interaction, $F_S(1,9) = 3.57, p = 0.091, \eta_p^2 = 0.284$; $F_I(1,4) = 1.27, p = 0.323, \eta_p^2 = 0.240$, or the three-way interaction, $F_S(2,18) = 2.05, p = 0.581, \eta_p^2 = 0.186$; $F_I(2,8) = 1.56, p = 0.267, \eta_p^2 = 0.281$.

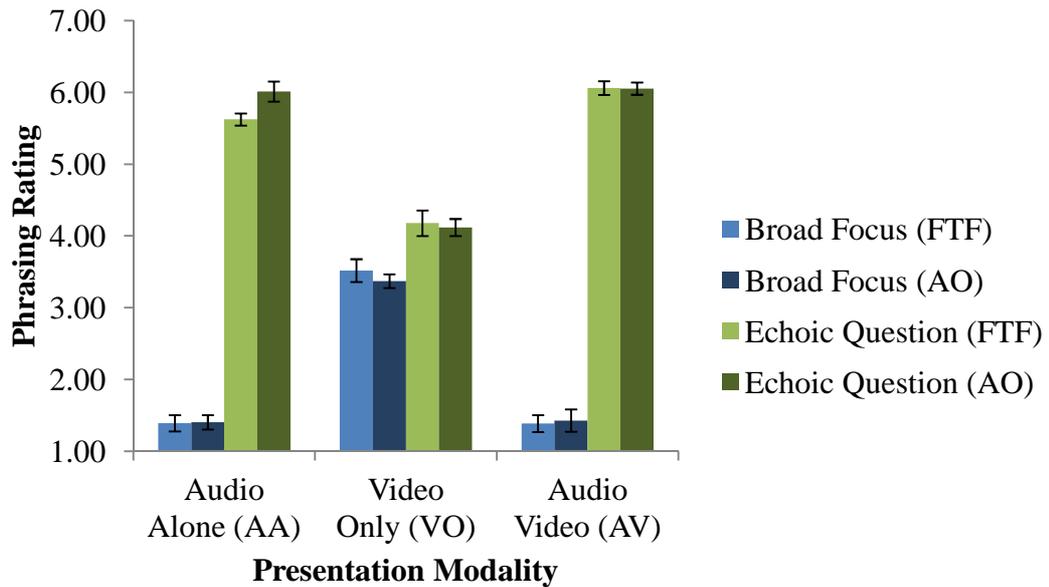


Figure 9.3. Mean rating of phrasing (collapsed across sentences and talkers) as a function of prosodic condition, interactive setting and presentation modality. Error bars indicate the standard error of the mean.

To determine the source of the interaction between prosody and presentation modality, a series of post-hoc repeated measures ANOVAs were conducted for each prosodic condition (separately for each interactive setting). Statements recorded in the AO setting showed a difference across presentation modalities, $F(2,18) = 114.55$, $p < 0.001$, $\eta_p^2 = 0.927$. AA presentations resulted in significantly lower ratings than VO presentations [$M_{\text{Diff}} = 1.97$, Sidak 97.5% CI: 1.36 – 2.57]; this was also the case for AV presentations [$M_{\text{Diff}} = 1.94$, Sidak 97.5% CI: 1.34 – 2.55]. No difference was observed between AA and AV presentations [$M_{\text{Diff}} = 0.02$, Sidak 97.5% CI: -0.17 – 0.12].

These effects were mirrored for statements recorded in the FTF settings, $F(2,18) = 205.17$, $p < 0.001$, $\eta_p^2 = 0.958$. Ratings were significantly lower in AA than VO presentations [$M_{\text{Diff}} = 2.13$, Sidak 97.5% CI: 1.68 – 2.57]; and similarly

lower in AV than VO presentations [$M_{\text{Diff}} = 2.13$, Sidak 97.5% CI: 1.60 – 2.67]; whereas no difference was observed between AA and AV presentations [$M_{\text{Diff}} = 0.01$, Sidak 97.5% CI: -0.11 – 0.12].

For the AO echoic question renditions, there was a significant effect of presentation modality, $F(2,18) = 174.51$, $p < 0.001$, $\eta_p^2 = 0.951$. Sidak pairwise comparisons showed that AA ratings were significantly greater than when items were presented in the VO condition [$M_{\text{Diff}} = 1.89$, Sidak 97.5% CI: 1.39 – 2.40]; AV ratings were also significantly greater compared to VO presentations [$M_{\text{Diff}} = 1.94$, Sidak 97.5% CI: 1.61 – 2.26]. No difference was observed between AA and AV presentations [$M_{\text{Diff}} = 0.04$, Sidak 97.5% CI: -0.30 – 0.38].

For the FTF echoic question renditions, there was also a significant effect of presentation modality, $F(2,18) = 118.28$, $p < 0.001$, $\eta_p^2 = 0.929$. VO presentation ratings were significantly lower than both AA [$M_{\text{Diff}} = 1.45$, Sidak 97.5% CI: 1.03 – 1.87]; and AV presentations [$M_{\text{Diff}} = 1.88$, Sidak 97.5% CI: 1.37 – 2.40], while AV presentations were significantly greater than AA presentations [$M_{\text{Diff}} = 0.44$, Sidak 97.5% CI: 0.11 – 0.77].

9.1.2.3. Regression of VO and AV Rating Scores

On average, narrow focused tokens were rated as having a stronger degree of focus, and echoic questions were rated as more question-like than broad focused statement renditions in the both the VO and AV modalities. However, there was some variability across items. For narrow focused items collapsed across interactive settings, the ratings ranged between 1.80 and 5.50 ($M = 3.82$, $SD = 0.92$) in the VO modality, and between 3.90 and 6.80 ($M = 5.74$, $SD = 0.75$) in the AV modality.

Similarly, the echoic questions in the VO presentation modality ranged between 2.55

and 5.78 ($M = 4.18$, $SD = 0.64$), and between 3.09 and 6.90 ($M = 6.07$, $SD = 0.68$) in the AV modality.

To determine whether the amount of visible movement accompanying the production of a prosodically marked constituent (as measured in Chapter 7) was able to explain the variance in the VO rating data, two separate standard multiple regression analyses were performed for the narrow focus and echoic question ratings (collapsed across interactive settings, as no differences were found in the ANOVAs). The mean subjective rating (across ten perceivers) for each of the VO items was the criterion for both analyses. For the regression of VO narrow focus ratings, the area under the principal component amplitude curve (represented as a proportion of the mean broad focused rendition) for jaw opening (PC 1), lip opening (PC 2), lower lip movement (PC 3), upper lip movement (PC 4), lip rounding (PC 5), jaw protrusion (PC 6), eyebrow raising (PC 7) and eyebrow pinching (PC 8) during the production of the critical utterance phase were treated as predictor variables. For the regression of VO echoic question ratings, the predictor variables were the area under PC amplitude curves for jaw opening (PC 1), lip opening (PC 2), lip rounding (PC 5), jaw protrusion (PC 6), eyebrow pinching (PC 8) and rigid pitch rotations (R 1). These components were selected as they were the ones that showed significant differences across the prosodic contrasts (i.e., in comparison to broad focused tokens) in the visual analysis reported in Chapter 7. Table 9.2 displays the properties of both regression analyses for the VO ratings.

For narrow focus ratings, the regression was significantly different from zero, $F(8,51) = 2.35$, $p = 0.031$, with $R = 0.52$, $R^2 = 0.27$, adjusted $R^2 = 0.16$, however no individual predictor made a significant unique contribution. The regression for

ratings of echoic questions presented in the VO modality was also significantly different from zero, $F(6,53) = 3.90$, $p = 0.003$, with $R = 0.55$, $R^2 = 0.31$, adjusted $R^2 = 0.23$. The only feature that made a significant unique contribution to explaining the variance in VO echoic question rating data was lip opening (PC 2).

Table 9.2. Standard multiple regression of the magnitude of visible movements on ratings of focus and questions in the VO presentation modality.

Predictor	Correlation (r) with Criterion (Mean Rating)	B	β	sr^2 (unique)
<i>Focus Ratings</i>				
Jaw Opening (PC 1)	0.454***	0.775	0.339	0.021
Mouth Opening (PC 2)	0.217*	-0.373	-0.110	0.006
Lower Lip Movement (PC 3)	0.284*	-0.079	-0.036	0.000
Upper Lip Movement (PC 4)	0.327**	-0.687	-0.256	0.019
Lip Rounding (PC 5)	0.378**	-0.159	-0.078	0.001
Jaw Protrusion (PC 6)	0.431***	1.041	0.483	0.029
Eyebrow Raising (PC 7)	0.373**	0.309	0.150	0.010
Eyebrow Pinching (PC 8)	0.356**	-0.013	-0.005	0.000
	Constant	2.34**		
<i>Question Ratings</i>				
Jaw Opening (PC 1)	0.096	-0.190	-0.105	0.003
Mouth Opening (PC 2)	-0.151	-1.349**	-0.628	0.159
Lip Rounding (PC 5)	0.303**	-0.573	0.302	0.030
Jaw Protrusion (PC 6)	-0.019	0.272	0.126	0.004
Eyebrow Pinching (PC 8)	0.293*	0.634	0.451	0.048
Rigid Pitch Rotation (R 1)	0.060	0.001	0.009	0.000
	Constant	4.372***		

$N = 60$, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

To determine whether the occurrence of an auditory prosodic cue (i.e., amount of syllable duration elongation, or the occurrence of an intensity increase or *F0* rise during the critical constituent) or non-articulatory visual prosodic marker was able to explain the variance in the AV rating data, two separate standard multiple regression analyses were performed for the narrow focus and echoic question ratings collapsed across interactive settings. The mean subjective rating for each of the AV items was the criterion; an increase in the mean intensity of the critical constituent, the occurrence of an *F0* rise before or after the start of the critical constituent, the occurrence of an eyebrow raise before or after the onset of an *F0* rise, and the occurrence of a rigid pitch rotation peak before or after the onset of an *F0* rise were treated as dichotomous predictor variables, and mean syllable duration (as a proportion of the mean broad focused rendition) as a continuous predictor variable. Table 9.3 displays the properties of both regression analyses for the AV ratings.

For the narrow focus ratings in the AV condition, the regression was significantly different from zero, $F(8,51) = 4.28, p = 0.001$, with $R = 0.63, R^2 = 0.40$, adjusted $R^2 = 0.31$, with the degree of syllable elongation during the critical constituent contributing the largest amount of unique variance explanation, $t(59) = 3.72, p = 0.001$. The occurrence of a pitch rotation peak preceding the onset of an *F0* rise, $t(59) = -3.25, p = 0.002$, contributed a small amount of unique variance explanation, as did the occurrence of an eyebrow movement preceding an *F0* rise, $t(59) = 2.90, p = 0.006$. As with the VO ratings, the regression for ratings of echoic questions presented in the AV modality was not significantly different from zero, $F(7,52) = 1.20, p = 0.322$, with $R = 0.37, R^2 = 0.14$, adjusted $R^2 = 0.02$.

Table 9.3. Standard multiple regression of the relationship between auditory prosodic markers and non-articulatory movements on ratings of focus and questions in the AV presentation modality.

Predictor	Correlation (<i>r</i>) with Criterion (Mean Rating)	B	β	<i>sr</i> ² (unique)
<i>Focus Ratings</i>				
Duration	0.318**	1.380**	0.483	0.162
Intensity Increase	0.070	0.089	0.057	0.003
F0 Rise Before Onset	0.062	0.376	0.254	0.038
F0 Rise After Onset	-.313**	-0.487	-0.144	0.015
Brow Raise Before F0 Rise	0.318**	0.549**	0.368	0.099
Brow Raise After F0 Rise	-0.006	0.390	0.132	0.014
Rigid Rotation Before F0 Rise	0.220*	-1.310**	-0.386	0.124
Rigid Rotation After F0 Rise	-0.081	-0.331	-0.218	0.032
Constant		3.136***		
<i>Question Ratings</i>				
Duration	0.021	-0.010	-0.004	0.000
Intensity Increase	-0.315**	-0.434*	-0.292	0.075
F0 Rise Before Onset	0.124	-0.025	-0.018	0.000
F0 Rise After Onset	-0.91	-0.230	-0.155	0.014
Brow Raise Before F0 Rise	-0.020	-0.060	-0.044	0.002
Rigid Rotation Before F0 Rise	-0.198	0.304	0.169	0.024
Rigid Rotation After F0 Rise	0.014	-0.098	-0.056	0.002
Constant		6.289***		

$N = 60$, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

9.1.3. Discussion

In this experiment, the perception of prosodic focus and phrasing contrasts in various presentation modalities was explored. Consistent with the rating results reported in Chapter 6, the narrow focused renditions were rated as having a greater degree of

focus on the critical word than broad focused ones, and echoic questions were consistently rated as being more question-like than broad focused statements, regardless of the modality of presentation (i.e., auditory alone, visual only, or auditory-visual).

The prosodic ratings of tokens presented in the AA modality replicated the results found in Experiment 7 using different stimuli and raters, with auditory presentations of tokens recorded in AO interactive settings (where the talker was not able to see the interlocutor) resulting in higher ratings of both focus and phrasing, compared to when the talker could see the interlocutor (i.e., in the FTF setting). This result further supports the idea that when speech communication is limited to the auditory channel, talkers take additional care to ensure that the prosody is clearly conveyed auditorily due to the loss of visual prosody information.

Although the differences in ratings between prosodic contrasts were still observed in visual only (VO) presentations, the degree of difference was substantially smaller than when compared to AA or AV presentations, particularly for the phrasing contrasts. Narrow focused tokens presented in the VO condition were rated as conveying less focus than both AA and AV presentation. This was also the case for echoic question renditions, with VO presentations being rated as less question-like than in AA and AV presentations. Furthermore, the VO presentations of statements in the phrasing task were rated as being less statement-like than in AA and AV presentations.

Overall, the visual correlates measured in Chapter 7 appear to be linked to perception. Despite this, AV ratings were not particularly better than presentations in the AA modality. That is, all of the prosodic conditions and interactive settings (with

the exception of echoic questions recorded in the FTF setting) failed to show any AV benefit with similar ratings attained across both presentation modalities. The lack of an additive effect when visual information was available does not necessarily indicate that visual prosodic correlates play no role in the conveying meaning, or that there is no relationship between the signal modalities. Given that AA ratings for narrow focus and echoic questions were already high in their respective tasks, there was little room for improvement in ratings. Indeed, a similar effect was reported by Dohen and Lœvenbruck (2009) for focus detection; an AV advantage was not observed for the detection of focus in conversational speech due to ceiling effects for AA presentations. In contrast, the accuracy and speed of focus detection in AV conditions was superior (i.e., more accurate and faster) compared to AA and VO presentations when whispered speech was used (which degrades some of the acoustic cues). Similarly, Srinivasan and Massaro (2003) found minimal AV effects for the discrimination of utterance phrasing due to the robust effect of AA information. Thus, the results obtained in the current experiment may have underestimated the contribution of visual prosody due to a certain lack of sensitivity in the rating task. To investigate whether this is the case, a different type of perceptual task involving the cross-modal matching of auditory to visual prosodic displays was used in Chapter 10.

In sum, the current experiment showed that the recorded tokens in the auditory visual speech prosody database are generally perceived both auditorily and visually as conveying the intended prosodic contrasts as expected from the signal-level differences uncovered in the auditory and visual analyses. However, in contrast to the current relatively small effects of visual prosody, previous studies (e.g.,

Krahmer & Swerts, 2004; Lansing & McConkie, 1999) have demonstrated that visual-only perception of prosody is in general pretty good for both focus and phrasing contrasts. The difference might be due to the fact that the current experiment used augmented point-light representations rather than real videos of talkers. For instance, it has been proposed that expressive “eye flashes” lasting approximately 750ms that occur independently of eyebrow raising can assist in highlighting important information in the auditory signal (see Massaro & Beskow, 2002; Walker & Trimboli, 1983). Similar textural features may also assist in conveying phrasing contrasts. Thus, in visual-only point-light displays when such information is no longer present, the contrasts may be more difficult to distinguish from each other.

Furthermore, the outcome of the regression analysis that explored whether particular visual gestures could account for the variability in the VO and AV rating scores yielded no definitive results. As such, the problem remains in deciphering exactly which cues are responsible for providing suprasegmental content to perceivers. In Chapter 10, the role that different types of visible movements may play in the conveyance of prosody is explored by using a method that allows different cue types (e.g., rigid head motion only, articulatory movements only) to be presented in combination or separately.

CHAPTER 10.
PERCEIVING PROSODY FROM AUGMENTED POINT
LIGHT DISPLAYS

Chapter 10. Perceiving Prosody from Augmented Point Light Displays

In Chapter 9, the prosody ratings showed that the focus and phrasing contrasts were able to be perceived from either the auditory or the point-light visual tokens. In the current chapter, the role that different types of visible movements may play in conveying prosody is explored by using a method that allows different cue types (e.g., rigid head motion only, articulatory movements only) to be presented in combination or separately, with the effects of this display restriction on the perception of prosody evaluated. That is, here the cross-modal prosody matching task (as used in Experiments 2 to 4) was used as this task has been shown to be a sensitive measure of the extent to which participants can perceive and relate auditory and visual cues to prosody.

In brief, the current chapter followed up some issues raised in the previous chapter by investigating which visual motion cues may be responsible for conveying prosodic content to perceivers. To achieve this, a series of experiments were conducted in which different movement information was presented alone or in combination. First, in Experiment 9, all of the motion features (i.e., whole head rigid movements and non-rigid face movements) were included in the visual stimuli to confirm that the cross-modal matching task could be completed with point-light stimuli. Then in Experiment 10, stimuli were presented with only the rigid movement of the head, only the non-rigid movement of the face, or only the movements of the articulators, to determine whether these movement types presented in isolation were sufficient to convey prosodic information.

10.1. Experiment 9: Cross-Modal Prosody Matching using Point-Light Displays¹⁵

Although previous studies have found that prosodic information appears to be transmitted differentially across face regions, evidence as to whether a face region may be better for conveying a particular prosodic type remains mixed. For instance, Swerts and Kraemer (2008) paired visual cues to prosody with a monotonic acoustic rendition of a spoken Dutch sentence and required people to identify the word within the utterance that received focus. Even though auditory prosodic information was absent, perceivers were highly accurate in detecting the narrowly focused word. When the video displays were restricted to show only the lower face of the talker, performance for identifying the focused constituent was substantially poorer than either the full face or upper face conditions (suggesting that the lower face alone is not sufficient to convey cues for focus).

In contrast, the study conducted by Lansing and McConkie (1999) used a video editing technique to still the movement of the talker's upper head. Compared to presentation with full motion information available, this manipulation hardly affected performance in identifying segmental aspects or identifying the narrowly focused constituent within an utterance, however the manipulation had a marked detrimental effect on performance in identify the phrasal nature (statement or echoic question) of the utterances.

The experiments presented in Chapter 2 and 3 (Cvejic et al., 2010, in press) also employed restricted visual displays (showing either the upper or lower face

¹⁵ A preliminary version of this experiment appeared as: Cvejic, E., Kim, J., & Davis, C. (2011). Perceiving prosody from point light displays. *Proceedings of the 10th International Conference on Auditory-Visual Speech Processing (AVSP2011)*, pp. 15-20.

only) but in a cross-modal matching task, where matching accuracy exceeded 80% across all prosodic types (i.e., broad focus, narrow focus, and echoic questions). Regardless of the face area provided, sufficient contrastive detail was available allowing for accurate prosodic discrimination. Furthermore, equivalent levels of matching accuracy were observed in Experiment 2 when upper face stimulus videos were filtered to remove eyebrow and skin deformations from the video displays (suggesting that rigid head motions that remained intact also provided sufficient cues to reliably match prosodic contrasts).

By restricting the amount of visual information provided to perceivers, the abovementioned studies have shown that visual prosody is conveyed by different face regions; however, what these studies have not revealed is the contribution of *particular* motion cues to the perception of prosody. For example, presenting perceivers with only the upper face still provides information about rigid head motion, non-rigid movements of the eyebrows and cheeks, as well as eye widening and textural deformations. Moreover, these studies used small numbers of tokens and talkers, so it is unclear the extent to which their results can be generalised.

To address this issue of whether particular prosody cues convey relatively specific information, the current experimental series used augmented point-light displays so that certain visual speech cues for prosody could be presented in isolation or in combination. The motion of these displays was derived from the motion capture data for both articulatory (e.g., lip opening, lip protrusion, jaw opening height) and non-articulatory movements (e.g., eyebrow raises and rigid head movements) that had been processed using guided principal components analysis (gPCA, see Chapter 7). This analysis procedure generated a set of independent and

uncorrelated parameters representing biomechanically feasible movements, and thus permits the creation of stimuli with full control over individual rigid and non-rigid movement features. To address the issue of the talker and item generalisation, the point-light displays were generated from six different talkers (a factor that will allow the degree of variability in how cues to visual prosody are realised to be confirmed) and for ten segmentally different sentences.

The current study adopted the matching paradigm used in Experiments 2 to 4 as it has been shown to provide a sensitive index of the extent to which participants can perceive the prosodic cues. In Experiment 9, it was first determined whether the visual speech movements represented in the point-light displays provide sufficient information to allow perceivers to cross-modally match auditory to point-light video tokens on the basis of prosody alone. Given the results of Experiment 8 in which perceivers differentiated between prosodic contrasts when presented with visual only presentations of point-light talkers, it was expected that perceivers would attain performance levels greater than chance. These results provided a baseline performance measure for Experiment 10, where individual movement features were systematically removed from the point-light displays in order to determine the importance of these movements for perception of prosodic contrasts.

10.1.1. Method

10.1.1.1. Participants

Thirty-three undergraduate psychology students ($M_{\text{Age}} = 24.19$ years) from UWS participated for course credit. All participants self-reported normal or corrected-to-normal vision, normal hearing, and were fluent talkers of English. None had taken part in any of the previously reported experiments.

10.1.1.2. Materials

Ten sentences were selected from the audio-visual speech prosody corpus (Chapter 4) for use as stimuli (see Table 10.1). These items were the same as the ones used in Experiments 2 to 4; given that the task was the same, the results can be compared to give rough estimate of how well point-light displays convey prosodic information. Two repetitions of the broad focus, narrow focus, and echoic question renditions produced by all six talkers in the FTF interactive setting¹⁶ were used.

Table 10.1. Stimuli sentences used in Experiment 9. The critical constituent of each utterance is italicised.

Sentence	Segmental Content
1	It is a band of <i>steel</i> three inches wide
2	The pipe ran almost the <i>length</i> of the ditch.
3	It was hidden from sight by a <i>mass</i> of leaves and shrubs.
4	The weight of the <i>package</i> was seen on the high scale.
5	Wake and rise, and <i>step</i> into the green outdoors.
6	The green light in the <i>brown</i> box flickered.
7	The brass <i>tube</i> circled the high wall.
8	The lobes of her ears were <i>pierced</i> to hold rings.
9	Hold the <i>hammer</i> near the end to drive the nail.
10	Next <i>Sunday</i> is the twelfth of the month.

The auditory tokens were created by normalizing the mean intensity of each utterance to 65dB using Praat (Boersma, 2001). To create the visual stimuli, the shape-normalised motion capture data for each utterance that had been processed using gPCA (Chapter 7) was reprojected into three-dimensional space, represented

¹⁶ Tokens from the FTF setting were chosen as they were produced in a context where the movements of the face were able to be seen. Furthermore, the results from the rating tasks in the previous chapter found no effect of the interactive setting.

by solidly-filled white dots on a black background. As for the visual data in Experiment 8, the point-lights were augmented with animated lips, eyebrows and a nose, and presented “looking” approximately 30° to the left to assist in eliciting a three-dimensional percept and so that protruding movements along the z -axis of the jaw, lips and head were clear. For Experiment 9, movement on all eight non-rigid (i.e., jaw opening, lip opening, upper lip movement, lower lip movement, lip rounding, jaw protrusion, eyebrow raising and eyebrow pinching) and six rigid components (three axial rotations and three axial translations) were presented in the augmented point-light displays (i.e., “All Motion”, see Appendix F.1).

10.1.1.3. Procedure

Each participant completed two sets of tasks: a cross-modal prosody matching task and a set of auditory prosody rating tasks. For both sets of tasks, participants were tested individually in a sound-attenuated booth, with video stimuli presented on a 17” LCD computer monitor at 60fps and auditory stimuli presented binaurally over Senheiser HD650 stereo headphones. DMDX (Forster & Forster, 2003) was used to control stimuli presentation and response collection.

10.1.1.3.1. Cross-modal matching task

For the cross-modal matching task, stimuli were presented in a two-interval, alternate forced choice (2AFC) task as used in Experiments 2 to 4. Participants were informed that they would be presented two pairs of stimuli, each consisting of an audio-only and a video-only item, and that their task was to select the pair in which the visual display of prosody matched the auditory token (Figure 10.1). To avoid instance-specific strategies, the matching items in the correct pair were always taken from a different recorded token. The non-matching pair of stimuli consisted of

utterances that were segmentally identical but produced as one of the alternate prosodic types (i.e., the non-matching items for half of the echoic question trials were broad focused renditions and narrow focus renditions for the remaining half). The same auditory item was used as the first item of each pair, and was the standard against which the visual stimuli were to be matched. Participants indicated their response as to which pair had the same prosody via a selective button press. The order of correct response pair was counter-balanced, occurring equally in the first and second pair.

Two versions of the cross-modal matching task were created, each requiring a total of 90 matching judgments (5 sentences \times 6 talkers \times 3 prosodic conditions). The 5 sentences that were presented differed between the task versions, with participants completing only one version of the task. Presentation was blocked by talker, with between- and within-block randomization controlled by the presentation software. In total, the cross-modal matching task took approximately 30 minutes to complete, including six practice trials and several short breaks between talker blocks.

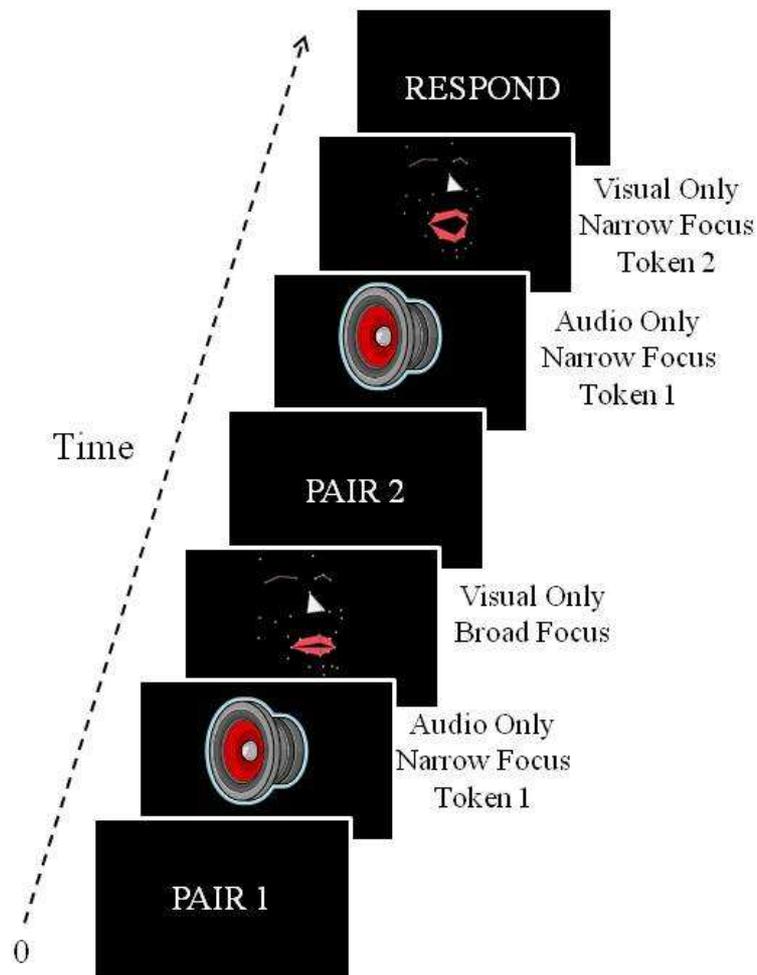


Figure 10.1. Schematic representation of the 2AFC cross-modal matching task used in Experiment 9. The same auditory token appeared first for both pairs, and was the standard that the matching judgment was to be made on. The matching item within pairs was always taken from a different recorded token, and non-matching items were the same sentence produced as a different prosodic type.

10.1.1.3.2. Auditory prosody rating task

In addition to the cross-modal matching task, participants also completed two perceptual rating tasks of the auditory stimulus; one rating the degree of focus and the other the perceptual clarity of the statement-question contrast. Procedurally, these

tasks were identical to those outlined in Experiment 7. Participants completed these tasks in a counter-balanced order.

For both rating tasks, a total of 60 stimulus items were presented. In the focus rating task, the items comprised of a single repetition of five sentences produced as a broad focused rendition, and five narrow focus renditions, from each talker. In the phrasing rating task, the items consisted of a single repetition of five sentences produced as a broad focused statement, and five as an echoic question rendition from each talker. For both tasks, item presentation was blocked by talker, with presentation order between- and within-blocks randomized by the presentation software.

Two versions of the task were made, so that participants never heard the same auditory token more than once across the three tasks (i.e., the auditory items presented to participants were the tokens that they had not yet been exposed to in the cross-modal matching task). Similarly, the broad focus token used was always a different repetition across both rating tasks. For both tasks, participants were informed that there was no “correct” answer, and were encouraged to use the complete range of the rating scale responses. All other procedural details are identical to those outlined for Experiment 7 (see Chapter 6).

10.1.2. Results and Discussion

10.1.2.1. Matching Point-Light Displays of Prosody across Modalities

A series of 6 (talker) \times 2 (task version) mixed repeated measures ANOVAs were initially conducted to test whether the two task versions differed for any of the prosodic conditions. No difference in performance accuracy was found across task versions for any of the prosodic conditions [broad focus: $F(1,31) = 2.52, p = 0.122,$

$\eta_p^2 = 0.075$; narrow focus: $F(1,31) = 0.23, p = 0.637, \eta_p^2 = 0.007$; echoic questions: $F(1,31) = 0.26, p = 0.617, \eta_p^2 = 0.008$], and thus the following analyses were conducted on data collapsed across the two task versions.

The percent of correct responses for each prosodic condition and talker (collapsed across sentences) are displayed in Figure 10.2. Collapsed across talkers, cross-modal matching performance was greater than chance (i.e., 50%) for all three prosodic conditions, as confirmed by a series of significant one-sample t -tests [broad focus: $t(32) = 3.34, p < 0.01$; narrow focus: $t(32) = 12.38, p < 0.001$; echoic questions: $t(32) = 8.05, p < 0.001$].

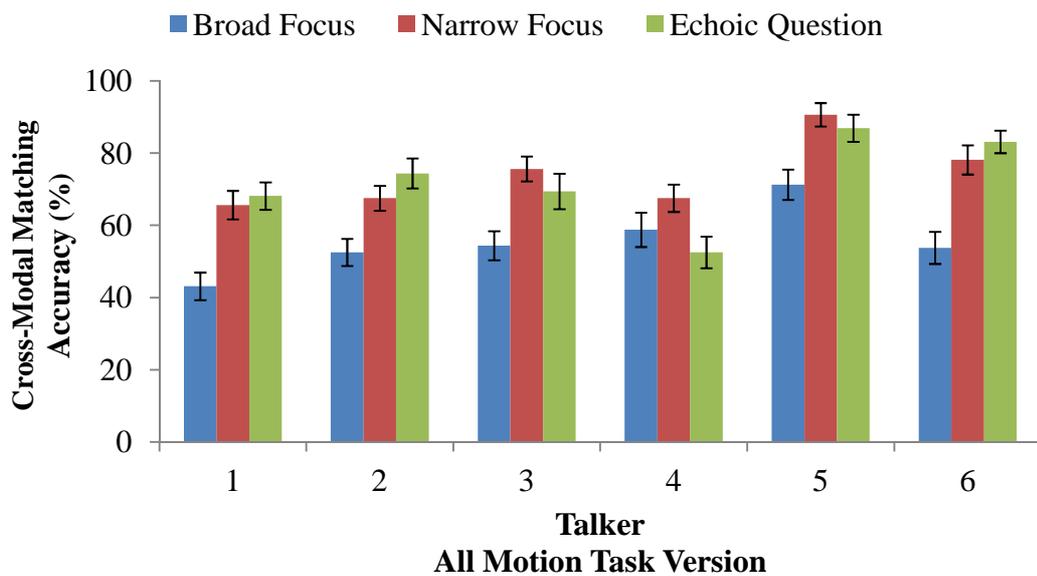


Figure 10.2. Mean percent of correct responses (with standard error) for the cross-modal prosody matching tasks as a function of prosodic contrast and talker (collapsed sentences) for the “all movement” task version.

These data were analysed in a 3 (prosodic condition: broad focus; narrow focus; echoic question) \times 6 (talker) repeated measures ANOVA. The main effect of

prosodic condition, $F(2,64) = 28.16$, $p < 0.001$, $\eta_p^2 = 0.468$, and talker, $F(5,160) = 16.67$, $p < 0.001$, $\eta_p^2 = 0.343$, were both statistically significant, and so was the interaction, $F(10,320) = 2.59$, $p = 0.005$, $\eta_p^2 = 0.075$. Sidak pairwise comparisons (interpreted with a 95% confidence interval) revealed that, collapsed across talkers, perceivers were significantly better at matching both narrow focus [$M_{\text{Diff}} = 17.88$, Sidak 95% CI: 10.96 – 24.80] and echoic question items [$M_{\text{Diff}} = 15.96$, Sidak 95% CI: 8.64 – 23.28] across modalities than broad focused ones.

Overall, the main effect of talker seems to be driven by the superior matching performance for items produced by Talker 5 compared to all other talkers [Talker 1: $M_{\text{Diff}} = 23.43$, Sidak 98.3% CI: 14.30 – 32.57; Talker 2: $M_{\text{Diff}} = 18.18$, Sidak 98.3% CI: 8.86 – 27.51; Talker 3: $M_{\text{Diff}} = 15.76$, Sidak 98.3% CI: 6.14 – 25.37; Talker 4: $M_{\text{Diff}} = 22.83$, Sidak 98.3% CI: 11.51 – 34.14; Talker 6: $M_{\text{Diff}} = 10.91$, Sidak 98.3% CI: 1.61 – 20.21]. Items produced by Talker 6 were also matched with significantly better accuracy than for Talker 1 [$M_{\text{Diff}} = 12.53$, Sidak 98.3% CI: 4.70 – 20.35] and Talker 4 [$M_{\text{Diff}} = 11.92$, Sidak 98.3% CI: 2.05 – 21.79].

To interpret the interaction, a post-hoc within-subjects ANOVA was conducted independently for each prosodic condition with talker as the repeated factor, interpreted with an adjusted α of 0.017 for multiple comparisons. For broad focused items, the effect of talker was significant, $F(5,160) = 4.66$, $p = 0.001$, $\eta_p^2 = 0.127$. Sidak pair-wise comparisons (interpreted with a 98.3% confidence interval) suggested that although items produced by Talker 5 were perceived with greater accuracy, the difference was only significant in comparison to Talker 1 [$M_{\text{Diff}} = 27.27$, Sidak 98.3% CI: 7.40 – 47.14].

Chapter 10: Perceiving Prosody from Augmented Point Light Displays

Narrow focus items also differed across talkers, $F(5,160) = 7.41$, $p < 0.001$, $\eta_p^2 = 0.188$, with pairwise comparisons indicating that this effect was driven primarily by perceivers superior performance for cross-modally matching items produced by Talker 5 compared to Talker 1 [$M_{\text{Diff}} = 24.85$, Sidak 98.3% CI: 6.98 – 42.72], Talker 2 [$M_{\text{Diff}} = 23.03$, Sidak 98.3% CI: 5.13 – 40.93], Talker 3 [$M_{\text{Diff}} = 15.15$, Sidak 98.3% CI: 0.89 – 29.41] and Talker 4 [$M_{\text{Diff}} = 23.03$, Sidak 98.3% CI: 5.68 – 40.38].

For echoic question items, the main effect of talker was once again significant, $F(5,160) = 10.84$, $p < 0.001$, $\eta_p^2 = 0.253$, with perceivers significantly better at matching auditory tokens to point-light videos when items were produced by Talker 5 and Talker 6 compared to Talker 1 [vs. Talker 5: $M_{\text{Diff}} = 18.18$, Sidak 98.3% CI: 2.52 – 33.84; vs. Talker 6: $M_{\text{Diff}} = 15.15$, Sidak 98.3% CI: 1.23 – 29.07], and Talker 4 [vs. Talker 5: $M_{\text{Diff}} = 32.12$, Sidak 98.3% CI: 11.98 – 52.26; vs. Talker 6: $M_{\text{Diff}} = 29.09$, Sidak 98.3% CI: 10.15 – 48.03].

That matching performance was better than chance indicates that the point-light displays have captured some prosodic information. However, performance levels overall were lower than those obtained for the same task using restricted video displays of the upper (Chapter 2) and lower face (Chapter 3) in which matching accuracy exceeded 80% across all three prosodic conditions. The reason for poorer performance may simply be that the prosody information transmitted by the point-light stimuli is degraded (e.g., it lacks textural information and changes in eye shape). However, it should also be kept in mind that the comparison between real videos and point-light displays also involves comparing different talkers and the current data clearly show that performance for the different talkers varied

considerably (with mean matching accuracy for individual talkers ranging between 58% and 82%).

An additional thing to note was that performance for the broad focused contrasts was particularly poor (in fact, the accuracy for items produced by Talker 5 is the only reason that such items were better than chance). As both narrow focused and echoic question tokens were used as the non-matching distracter item, a 2×2 mixed repeated measures ANOVA was conducted for the broad focused items, with foil type (narrow focus; echoic question) as the within-subjects factor, and task version as the between-subjects factor, to examine whether the prosodic type of the non-matching item influenced matching accuracy. However, there was no main effect of distracter type, $F(1,31) = 1.53$, $p = 0.226$, $\eta_p^2 = 0.047$, no main effect of task version, $F(1,31) = 1.07$, $p = 0.308$, $\eta_p^2 = 0.033$, and no significant interaction, $F(1,31) = 0.35$, $p = 0.560$, $\eta_p^2 = 0.011$. Thus, regardless of the prosodic type used as a distracter item, accuracy for cross-modal matching of broad focus tokens was generally poor. This suggests that participants may be using the initial auditory stimulus as a guide to what to look for in the subsequent visual displays. This strategy would work well for the narrow focus and echoic questions since these auditory stimuli provide positive cues (e.g., increased intensity and $F0$ range), however the broad focused statements do not; so a decision here would need to be based on negative evidence, a decision that is always less secure (see Repp & Crowder, 1990, for a similar argument).

10.1.2.2. Auditory Ratings of Prosodic Contrasts

Overall, the results of the auditory rating task replicated those found in Experiment 7 and 8. For each task, the rating scores were subjected to a series of repeated

measures ANOVAs for each perceptual task; a subject analysis (F_S , collapsed across sentences), and an item analysis (F_I , collapsed across raters), both with prosodic condition and talker as within-items factors. Due to technical error, the auditory rating data for one participant was not included in the analysis (i.e., $n = 32$).

For the focus rating task, the main effect of prosody was significant, $F_S(1,31) = 527.77, p < 0.001, \eta_p^2 = 0.944$; $F_I(1,9) = 3124.17, p < 0.001, \eta_p^2 = 0.997$.

Collapsed across talkers, the narrow focused utterances were rated as having a significantly greater degree of focus on the critical constituent than the broad focused renditions [$M_{\text{Diff}} = 3.71$, Sidak 95% CI: 3.55 – 3.85]. The main effect of talker, $F_S(5,155) = 23.42, p < 0.001, \eta_p^2 = 0.430$; $F_I(5,45) = 13.99, p < 0.001, \eta_p^2 = 0.609$, and the prosodic condition by talker interaction, $F_S(5,155) = 45.18, p < 0.001, \eta_p^2 = 0.593$; $F_I(5,45) = 13.53, p < 0.001, \eta_p^2 = 0.600$, were also significant.

To interpret the interaction, a series of paired samples t -tests (interpreted with a Bonferroni adjusted α of 0.025) were conducted between the broad and narrow focused items for each talker. For all six talkers, the prosodic effect was maintained, with broad focused utterances being rated as having less focus on the critical constituent than narrow focused renditions [Talker 1: $t(31) = 11.87, p < 0.001$; Talker 2: $t(31) = 16.00, p < 0.001$; Talker 3: $t(31) = 14.56, p < 0.001$; Talker 4: $t(31) = 26.80, p < 0.001$; Talker 5: $t(31) = 21.78, p < 0.001$; Talker 6: $t(31) = 21.37, p < 0.001$]. These ratings across talkers are shown in Figure 10.3.

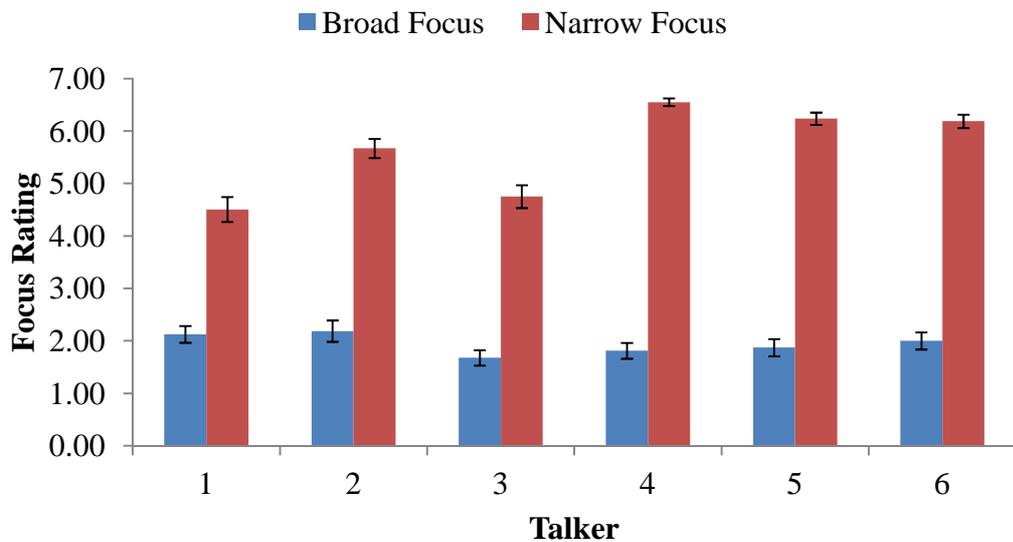


Figure 10.3. Mean ratings of focus (collapsed across sentences and raters) as a function of talker for broad and narrow focused utterances. Error bars indicate the standard error of the mean.

The ANOVA of the phrasing rating task yielded similar results. The main effect of prosody was significant, $F_S(1,31) = 288.03, p < 0.001, \eta_p^2 = 0.903$; $F_I(1,9) = 1345.25, p < 0.001, \eta_p^2 = 0.993$, with echoic questions being rated as being more “question-like” than broad focused statements [$M_{Diff} = 3.92$, Sidak 95% CI: 3.45 – 4.39]. The main effect of talker, $F_S(5,155) = 16.71, p < 0.001, \eta_p^2 = 0.350$; $F_I(5,45) = 8.17, p < 0.001, \eta_p^2 = 0.476$, and the prosodic condition by talker interaction, $F_S(5,155) = 22.26, p < 0.001, \eta_p^2 = 0.4183$; $F_I(5,45) = 10.94, p < 0.001, \eta_p^2 = 0.549$, were both significant.

Paired samples *t*-tests to interpret the significant interaction showed that, for all talkers, the statement renditions were perceived as being significantly more “statement-like” than the echoic question ones [Talker 1: $t(31) = 12.63, p < 0.001$; Talker 2: $t(31) = 18.58, p < 0.001$; Talker 3: $t(31) = 10.33, p < 0.001$; Talker 4: $t(31)$

= 11.45, $p < 0.001$; Talker 5: $t(31) = 24.39$, $p < 0.001$; Talker 6: $t(31) = 13.35$, $p < 0.001$]. These ratings for each talker are shown in Figure 10.4.

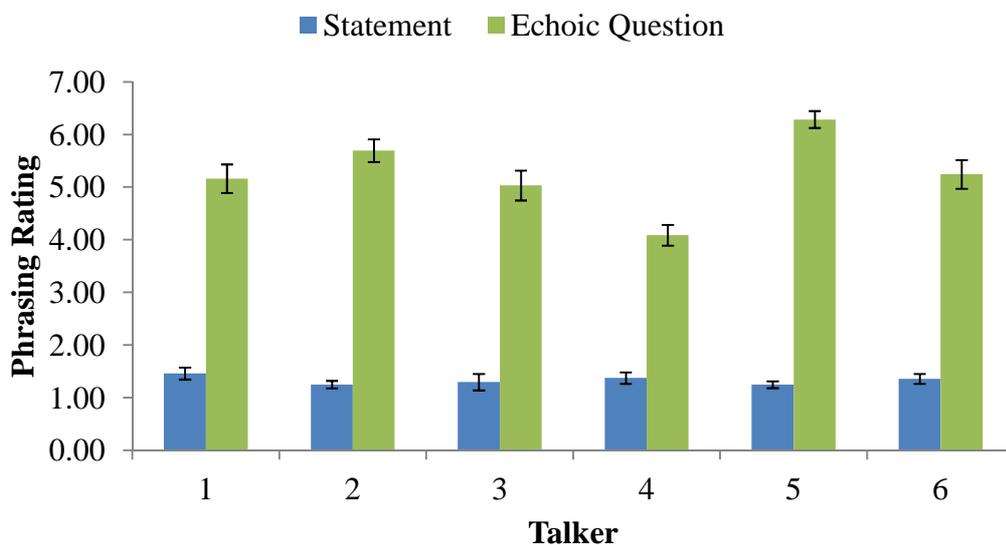


Figure 10.4. Mean ratings of phrasing (collapsed across sentences and raters) as a function of talker for the broad focused statement and echoic question utterances. Error bars indicate the standard error of the mean.

10.1.2.3. Relationship between Auditory Ratings and Item Accuracy

To determine if there was any relationship between the ratings for the narrow focus and echoic question items in the subjective auditory rating tasks, and item accuracy when these tokens were the auditory targets in the cross-modal matching task, a series of Pearson product-moment correlations were conducted. Using an α of 0.05, a small yet statistically significant positive correlations was found between item accuracy and ratings of focus strength, $r(58) = 0.24$, $p = 0.033$, with auditory items rated as having a *stronger* degree of focus production resulting in more accurate matching performance (Figure 10.5, upper panel). Similarly, ratings of phrasing for

the echoic question renditions of utterances were significantly correlated with item accuracy in the cross-modal matching task, $r(58) = 0.36, p = 0.002$, with utterances subjectively rated as being more “question-like” resulting in better matching accuracy (Figure 10.5, lower panel).

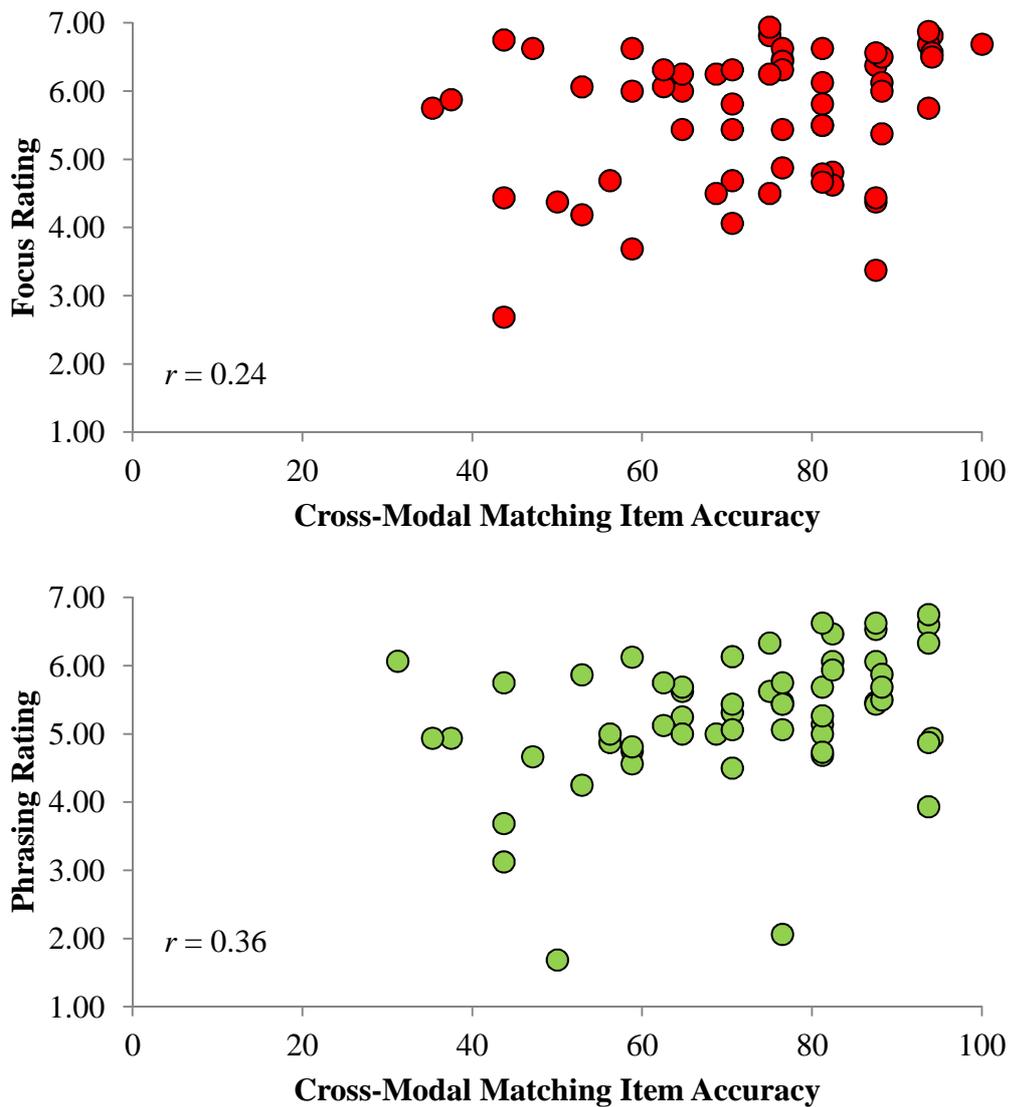


Figure 10.5. Scatter plot indicating the relationship between subjective rating of auditory tokens and item accuracy in the cross-modal matching task, for narrow focus (upper panel) and echoic question (lower panel) items.

Chapter 10: Perceiving Prosody from Augmented Point Light Displays

Although there was a significant relationship between the subjective auditory ratings and item accuracy in the cross-modal matching task, these correlations need to be interpreted with caution. For example, Talker 4's auditory productions of narrow focus items (rated across 10 sentences) were rated the highest of all six talkers (see Figure 10.3), however accuracy of perceivers in matching these contrasts to the corresponding visual point-light displays was lower than for most of the other talkers (Figure 10.2). Indeed, this corroborates the finding from Chapter 8 that the relationship between auditory and visual prosody is variable and potentially non-linear. Given that auditory tokens across all six talkers were perceived as conveying the intended prosodic contrasts, variability in the salience of visual cues produced by talkers is more likely the cause of differences in cross-modal matching accuracy. From Chapter 7, talkers displayed differences not only in the types of cues that they used to contrast between the prosodic types, but also in the overall amount of movement. Thus, those talkers that provided larger movements (or at least more salient visual cues) in conjunction with strong contrastive acoustic information are more likely to be matched with higher accuracy (i.e., when the correspondence between the auditory and visual signals is obvious).

In sum, Experiment 9 showed that a combination of non-rigid and rigid motion features presented in slightly augmented point-light displays provided perceivers with sufficient contrastive (prosodic) information to support cross-modal matching. These results can be used as a baseline measure for determining how well motion features can be perceived when presented in isolation in Experiment 10.

10.2. Experiment 10: Perceiving Prosody from Manipulated Point-Light Displays

To determine what type of movements are responsible for conveying prosodic information to perceivers, Experiment 10 used stimuli in which individual motion features were systematically removed from the point-light displays. Specifically, three stimuli conditions were presented to perceivers: (1) only the non-rigid movements of the face were presented (i.e., the rigid head movements were removed), (2) only the movements of the articulators (i.e., the talkers' lip and jaw movements) was presented (i.e., no eyebrow or rigid head movements were included), and (3) only the rigid movements of the whole head were made available (i.e., no face movements were included).

Based on the results found in Chapter 3 (Cvejic et al., in press), and in Lansing and McConkie (1999), performance for matching narrow focused items should be maintained when only articulatory information is available perceivers, whereas performance for identifying echoic questions is likely to decline when the eyebrow and rigid head movements are no longer available. In contrast, a different pattern of results is expected on the basis of Swerts and Krahmer's (2008) findings; given that their results indicated that the upper face held a greater cue value for identifying narrow focus, cross-modal prosody matching should *decline* for narrow focus items when upper face movement is no longer provided. Furthermore, on the basis of the results reported in Chapter 2 (Cvejic et al., 2010), then being provided with only the rigid movements of the talkers' head should allow perceivers to match both focus and phrasing contrasts across modalities.

10.2.1. Method

10.2.1.1. Participants

Forty-two undergraduate students ($M_{\text{Age}} = 22.5$ years) from UWS participated for course credit. All participants self-reported normal or corrected-to-normal vision and hearing, with no known communicative deficits. None had taken part in Experiment 9, or any of the other previously reported experiments.

10.2.1.2. Materials and Procedure

Three stimuli conditions were designed by removing specific movement features from the “All Motion” point-light displays in Experiment 9. The “Non-Rigid Movement Only” condition consisted of stimuli that presented the non-rigid lip, jaw and brow motion with head rotations and translations removed from the visual signal (i.e., movement of PCs 1 to 8, see Appendix F.2). The “Articulator Movement Only” condition consisted of stimuli that were similar to the “Non-Rigid Movement Only” ones except that the eyebrow movements (PCs 7 and 8) were also removed (but the static eyebrows were still shown, Appendix F.3). Finally, the “Rigid Movement Only” condition consisted of stimuli where rotations and translations of the whole head were shown (R 1 to R 6), while all other markers remained in the average face configuration moving in accordance with the rigid head motion (Appendix F.4). Participants were not presented with an “Eyebrow Only” display condition (Appendix F.5) since the results of a pilot study (with participants who were familiar with the task and the point-light displays) showed that only chance level performance with these stimuli could be achieved. This is not to say that there are no eyebrow movements produced at all by the talkers, but rather that people are less sensitive to these movements as prosodic cues when presented in isolation.

The procedure of Experiment 10 was identical to the cross-modal matching task in Experiment 9 (with the exception of the number of items), with the task once again requiring matching from auditory to video tokens. Each participant was randomly assigned to and completed the task in one of the stimuli conditions (i.e., all stimulus items were presented either as the non-rigid movement only, articulator movement only, or rigid movement only). Each version required 180 matching judgments to be made, consisting of a single repetition of each of the 10 sentences in the three prosodic conditions produced by all six talkers, with the matching audio and video pair always being based on different recorded tokens. The task took approximately 55 minutes to complete, including several short breaks and six practice trials.

10.2.2. Results and Discussion

The percent of correct responses for each stimuli version as a function of prosodic condition and talker (collapsed across sentences) are displayed in Figures 10.6 to 10.8. A series of one sample *t*-tests (the values of which are shown in Table 10.2) indicated that, collapsed across talkers, above chance performance was maintained for narrow focus and echoic question items across all task versions; however performance for broad focus items dropped to chance level. Furthermore, performance for individual talkers across the stimuli versions varied. That is, when rigid movements were removed from the visual stimuli (i.e., in the “Non-Rigid Movement Only” condition), better than chance performance was maintained for both narrow focus and echoic question items for all talkers. When only articulatory movements were made available, narrow focus items were matched better than chance for all talkers except for Talker 4, but echoic question items were matched

better than chance only when produced by Talker 2 and Talker 5. In contrast, when only the rigid movements of the head were provided (with no articulatory or eyebrow movements), matching was performed better than chance for echoic question items for all talkers except Talker 4, however narrow focus matching above chance was only for items produced by Talker 3, Talker 5 or Talker 6. Indeed, these differences clearly reflect differential strategies (i.e., different cue usage) across talkers in the visual marking of prosody, as shown in Chapter 7.

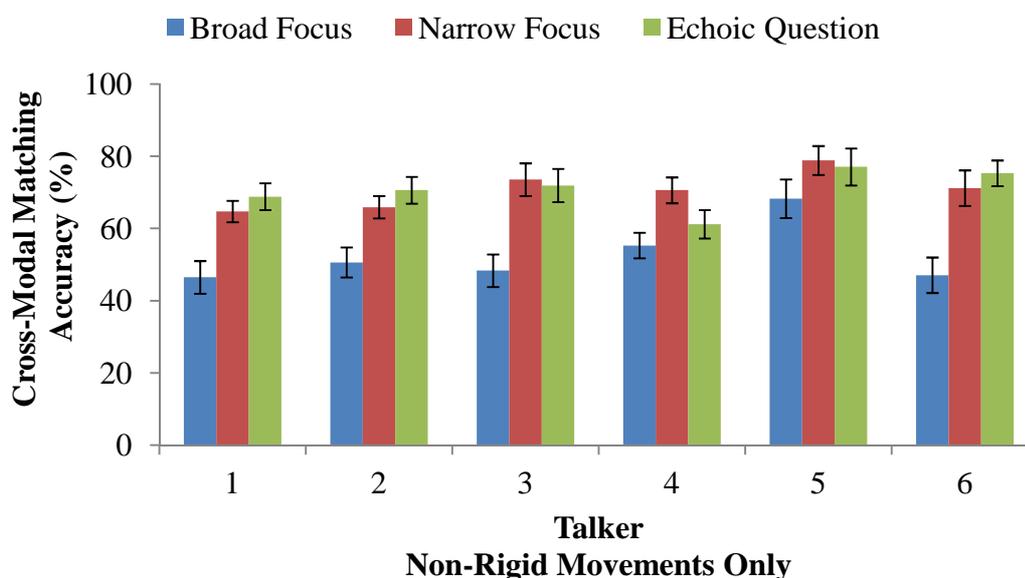


Figure 10.6. Mean percent of correct responses (with standard error) for the cross-modal prosody matching tasks as a function of prosodic contrast and talker (collapsed sentences) for the “non-rigid movement only” stimuli.

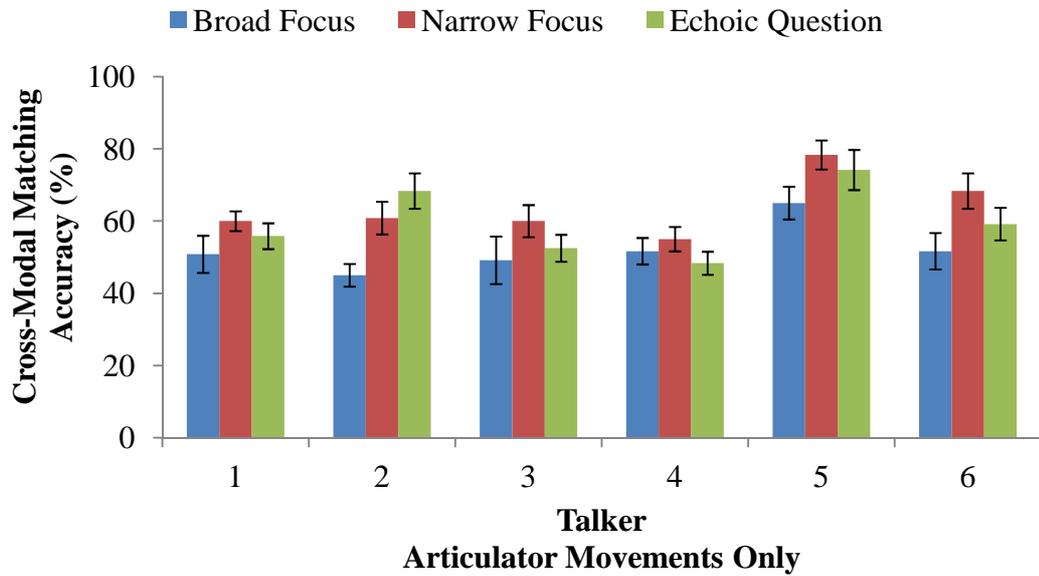


Figure 10.7. Mean percent of correct responses (with standard error) for the cross-modal prosody matching tasks as a function of prosodic contrast and talker (collapsed sentences) for the “articulator movement only” stimuli.

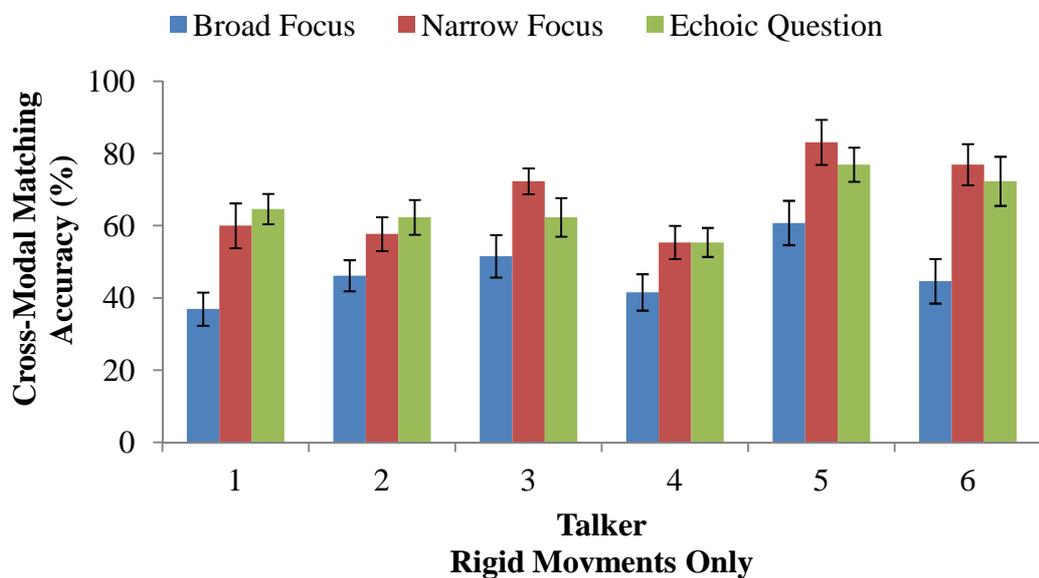


Figure 10.8. Mean percent of correct responses (with standard error) for the cross-modal prosody matching tasks as a function of prosodic contrast and talker (collapsed sentences) for the “rigid movement only” stimuli.

Table 10.2. Matching performance against chance for the cross-modal matching task of Experiment 10, as a function of talker and prosodic condition for each of the stimulus conditions.

Talker	Prosodic Condition	<i>t</i> -test Value vs. Chance (50%)		
		Non-Rigid Movement Only (<i>df</i> = 16)	Articulatory Movement Only (<i>df</i> = 11)	Rigid Movement Only (<i>df</i> = 12)
<i>Group</i>	Broad Focus	1.08	0.81	-0.81
	Narrow Focus	7.92***	8.24***	5.66***
	Echoic Question	8.45***	3.82**	5.67***
Talker 1	Broad Focus	-0.78	0.16	-2.85
	Narrow Focus	4.93***	3.63**	1.61
	Echoic Question	5.05***	1.63	3.50**
Talker 2	Broad Focus	0.14	-1.59	-0.89
	Narrow Focus	5.13***	2.40*	1.64
	Echoic Question	5.59***	3.74**	2.55*
Talker 3	Broad Focus	-0.37	-0.13	0.26
	Narrow Focus	5.19***	2.25*	6.18***
	Echoic Question	4.75***	0.67	2.31*
Talker 4	Broad Focus	1.49	0.46	-1.68
	Narrow Focus	5.74***	1.48	1.17
	Echoic Question	2.85*	-0.52	1.34
Talker 5	Broad Focus	3.40**	3.32**	1.75
	Narrow Focus	7.19***	6.99***	5.30***
	Echoic Question	5.28***	4.34**	5.70***
Talker 6	Broad Focus	-0.60	0.33	-0.87
	Narrow Focus	4.31**	3.74**	4.72***
	Echoic Question	7.13***	2.03	3.28**

Note: *** indicates $p < 0.001$, ** indicates $p < 0.01$, * indicates $p < 0.05$

These data (along with the results from Experiment 9) were compared in a $3 \times 6 \times 4$ mixed repeated measures ANOVA, with prosodic condition (broad focus; narrow focus; echoic question) and talker as repeated within-subjects factors, and stimuli condition (all motion; non-rigid movement only, articulator movement only; rigid movement only) as the between-subjects factor. The main effects of prosodic condition, $F(2,142) = 64.04, p < 0.001, \eta_p^2 = 0.474$, and talker, $F(5,355) = 28.67, p < 0.001, \eta_p^2 = 0.288$, were statistically significant, so was the between-subjects main effect of stimuli condition, $F(3,71) = 4.09, p = 0.010, \eta_p^2 = 0.147$. The interaction between talker and prosodic condition was significant, $F(10,710) = 3.28, p < 0.001, \eta_p^2 = 0.044$, but no significant interaction for the talker by stimuli condition was found, $F(15,355) = 1.00, p = 0.459, \eta_p^2 = 0.040$; nor was the prosody by stimuli condition significant, $F(6,142) = 1.09, p = 0.370, \eta_p^2 = 0.044$. There was no significant three-way interactions, $F(30,710) = 0.64, p = 0.936, \eta_p^2 = 0.026$.

Overall, matching performance was the greatest in the “All Motion” stimuli condition, when a combination of rigid and non-rigid motion was available in the visual signal. However, the only significant difference was between the “All Motion” and “Articulator Movement Only” conditions [$M_{\text{Diff}} = 8.07$, Sidak 95% CI: 0.73 – 15.40].

Across all four stimuli conditions, the main effect of prosody was maintained (as found in Experiment 9), with pairwise comparisons suggesting that matching performance for broad focused items was significantly lower than both narrow focus [$M_{\text{Diff}} = 17.04$, Sidak 95% CI: 12.70 – 21.39] and echoic question items [$M_{\text{Diff}} = 15.08$, Sidak 95% CI: 10.57 – 19.59]. Commensurate with Experiment 9, the main effect of talker across all four task versions seems to be driven by perceivers finding

it much easier to match items produced by Talker 5 than all other talkers [Talker 1: $M_{\text{Diff}} = 18.71$, Sidak 95% CI: 12.81 – 24.60; Talker 2: $M_{\text{Diff}} = 15.80$, Sidak 95% CI: 9.89 – 21.81; Talker 3: $M_{\text{Diff}} = 14.01$, Sidak 95% CI: 8.40 – 19.61; Talker 4: $M_{\text{Diff}} = 19.71$, Sidak 95% CI: 12.75 – 26.67; Talker 6: $M_{\text{Diff}} = 10.71$, Sidak 95% CI: 4.88 – 16.55], while items produced by Talker 6 were perceived much better than for Talker 1 [$M_{\text{Diff}} = 7.99$, Sidak 95% CI: 2.83 – 13.16] and Talker 4 [$M_{\text{Diff}} = 8.99$, Sidak 95% CI: 3.20 – 14.79].

To determine the source of the prosody by talker interaction, a series of 6 (talker) \times 4 (stimuli condition) mixed repeated measures ANOVAs were conducted independently for each of the prosodic conditions (and interpreted with an adjusted α of 0.017 for multiple comparisons). For broad focused items, the main effect of talker was maintained, $F(5,355) = 9.08$, $p < 0.001$, $\eta_p^2 = 0.113$, however the main effect of stimuli condition, $F(3,71) = 2.09$, $p = 0.110$, $\eta_p^2 = 0.081$, and the stimuli condition by talker interaction, $F(15,355) = 0.50$, $p = 0.939$, $\eta_p^2 = 0.021$, failed to reach significance.

For narrow focus items, the talker main effect was once again observed, $F(5,355) = 14.54$, $p < 0.001$, $\eta_p^2 = 0.170$. An effect of stimuli condition was found, $F(3,71) = 2.8$, $p = 0.043$, $\eta_p^2 = 0.107$, but this was not significant at the adjusted alpha level¹⁷. The interaction also failed to reach significance, $F(15,355) = 0.82$, $p = 0.655$, $\eta_p^2 = 0.034$. Similarly, the talker main effect was significant for echoic questions, $F(5,355) = 13.79$, $p < 0.001$, $\eta_p^2 = 0.163$, but the between-subjects effect

¹⁷ Examination of the pairwise comparisons suggests that this effect is driven by the difference between the all movement and articulator movement only stimulus conditions [$M_{\text{Diff}} = 9.48$, Sidak 98.3% C.I.: -1.20 – 20.17].

of stimuli condition was not significant at the adjusted alpha level¹⁸, $F(3,71) = 2.95$, $p = 0.039$, $\eta_p^2 = 0.111$, while the interaction also failed to achieve significance, $F(15,355) = 1.04$, $p = 0.417$, $\eta_p^2 = 0.042$.

Overall, these results suggest that matching prosody from the rigid (head) or non-rigid (articulatory and eyebrow movements) gestures in isolation is just as accurate as when these motion types are presented in combination (i.e., in the all motion condition). Indeed, no differences were found among the all motion condition, the non-rigid movement only, and the rigid movement only conditions, with the availability of more cues not necessarily leading to better prosodic perception (i.e., performance in the “All Motion” condition showed no evidence of a ceiling effect). In line with the proposal made in Chapter 3 to explain perceivers’ ability to accurately match prosody despite being provided with very different cues (i.e., different face areas or different talkers), this ability to efficiently use any of the cues likely stems from the visual signal conveying multiple cues to prosody, and that perceivers can make use of any available cue to determine the prosodic category (with matching performance performed at this abstract categorical level).

The results of the prosody matching from the non-rigid movement cues only indicated that there was a benefit from adding eyebrow movements to the articulatory ones. This is interesting since it appeared that eyebrow movements by themselves provided insufficient information to drive reliable matching performance. Upon examining performance at the individual talker level, this difference was mainly apparent for echoic question items. That is, narrow focus items were still matched at greater than chance levels regardless of the availability of eyebrow

¹⁸ As with narrow focus, the pairwise comparisons indicate that this effect is driven by the difference between the all movement and articulator movement only stimulus conditions [$M_{\text{Diff}} = 11.59$, Sidak 98.3% C.I.: -1.43 – 24.61].

movements (for all talkers except for Talker 4), whereas echoic questions were matched at greater than chance levels only for Talker 2 and 5. This is consistent with the results reported in Chapter 3 (Cvejic et al., in press) and by Lansing and McConkie (1999) that the movement information contained in the upper face is important for the perception of phrasing, whereas the most beneficial cues for focus are available from lower face motion (i.e., articulatory movements). Indeed, these findings go against those of Swerts and Kraemer (2008) who suggested that upper face movements are more important for determining focus.

Furthermore, although performance collapsed across talkers suggests that rigid motion is just as efficient as full head and face motion for conveying prosody (as found in Chapter 2), this was primarily the case for echoic question items. All talkers except for Talker 4 were matched at greater than chance levels in this condition, however this was not the case for narrow focus: only items produced by Talkers 3, 5 and 6 were matched above chance. Indeed, these talkers were the ones shown to utilise an increase in rigid pitch rotations to contrast narrow from broad focused tokens in the visual analysis reported in Chapter 7.

10.3. Summary

A talker's articulatory and non-articulatory (i.e., rigid head and eyebrow movements) gestures vary as a function of prosodic change (Chapter 7; Dohen et al., 2006, 2009; Scarborough et al., 2009), and in general people are sensitive to such visual cues and able to extract prosodic information (Chapters 2, 3 and 9; Cvejic et al., 2010, in press; Foxton et al., 2010). The current chapter examined how well these movement features were perceived as prosodic cues in isolation or in combination. This question has been of interest in previous studies (Cvejic et al., 2010, in press;

Lansing & McConkie, 1999; Srinivasan & Massaro, 2003; Swerts & Kraemer, 2008) but the research method used (i.e., by processing video recordings) has its limits in selecting and presenting particular movement features in isolation. Thus, the current study followed up these studies with a method that overcomes such limits, that is, by using an auditory-visual speech prosody corpus that includes three-dimensional motion tracking information (Chapter 4) to create stimuli of a talker's face animated by selected movements.

The current results showed that articulatory gestures alone (i.e., movements of the lips and jaw) convey prosodic information that can be perceived better than chance, however the perceptual salience of narrow focus and echoic question items is further enhanced when non-articulatory movements, such as eyebrow raises and whole head rigid movements, accompany movements of the articulators. Furthermore, rigid head movements in isolation were effective as prosodic cues (particularly for echoic questions, commensurate with the results reported in Chapter 2). In sum, it seems that the more visual cues available do not necessarily lead to more accurate perception of prosody; the benefit of additional cues may depend on whether they can be attended to without distracting from each other, with multiple cues giving the perceiver more choice as to what can be attended to rather than being additive.

Finally, the outcome from both Experiment 9 and 10 reaffirm the findings of the previous chapters that talkers vary in both the auditory and visual cues used to signal prosodic contrasts (in terms of which cues are used, and the perceptual salience of these cues).

CHAPTER 11.
SUMMARY AND CONCLUSIONS

Chapter 11. Summary and Conclusions

This thesis explored the auditory and visual properties of two types of spoken prosodic contrast: a contrast of prosodic focus, where narrow focus statements (containing an explicit point of informational focus) were contrasted with broad focus ones (where no individual constituent was given greater informational importance), and an utterance phrasing contrast, in which declarative statements were contrasted with echoic questions (that had the same segmental content but where a degree of uncertainty was implied).

A series of production and perception studies were carried out to address questions about the form, perceptibility and potential functions of visual prosody, as well as about the nature of the relationship between the auditory and visual prosodic correlates. In this chapter, the key outcomes of these studies are highlighted then considered. This is followed by a discussion of the limitations of the research program, and the presentation of some proposals for future work in the area of auditory-visual prosody research.

11.1. Perceiving Prosody

The initial series of six experiments reported in Chapters 2 and 3 explored the perception of visual prosodic cues that are available from the head and face of talkers using a two-interval, alternate forced choice (2AFC) matching task. Specifically, the aim of these studies was: to determine whether perceivers were sensitive to visual prosodic cues, to determine whether perceivers could relate the visual correlates to the auditory realisation of prosodic contrasts, to identify the types of visual movements that were of most benefit for contrasting prosodic types, and to explore

perceivers tolerance for variability across the auditory and visual prosodic signals. In what follows, each of these issues will be addressed in turn.

11.1.1. Perceptual Sensitivity to Visual Prosody

Experiments 1 to 3 examined whether perceivers were sensitive to the way that prosody was realised by talkers in the visual modality. In Experiments 1 and 2, visual displays showed only the talkers' upper head and face in two conditions; textured displays that showed a combination of rigid and non-rigid movement, and outline only displays where textural details such as eyebrows and skin wrinkling were removed (leaving only an outline of the talkers head and irises). In Experiment 3, only the lower half of the face was shown. This division of upper and lower face cues provided a neat way of separating out those cues that are directly tied to articulatory processes (visible from the lower half of the face) from those that are less causally related to speech production (such as eyebrow and rigid head movements). Two variants of the 2AFC matching task were employed: a within-modal matching task (requiring matching of video tokens that were produced with the same prosody), and a cross-modal matching task (involving the matching of auditory to video tokens).

The purpose of the within-modal task was to test whether the differences in visible movements from selective face areas were able to be picked up by perceivers. That is, performance on this task provided an indication of whether perceivers were able to use movement information as expressed across different tokens (as different recorded tokens were used for the matching pair), possibly by using overall or distinctive motion as a cue, (i.e., choosing the pair where there was some sort of conspicuous motion occurring in both stimuli). High levels of matching performance

were obtained, indicating that perceivers were sensitive to (and able to use) differences in visual cues from restricted displays. Of course, this on its own does not mean that performance can be attributed to sensitivity to prosodic type per se, but it does show that the visible differences across the prosody contrasts were large enough to be perceptually salient.

In contrast, the cross-modal matching task required perceivers to interpret the prosodic information from the auditory modality, and to find suitable correlates in the visual one, even though the different tokens may not be perfectly matched. The result that perceivers attained high performance levels in this task reflects not only that there are perceptually salient differences between the auditory and visual cues used to contrast prosodic types, but also that these differences are identified as being representative of particular prosodic categories.

Furthermore, perceivers were capable of performing these matching tasks regardless of the type of visual information that was presented. That is, above chance levels of performance were attained for textured upper face displays, outline upper face displays, and displays showing only the lower half of the talkers face. This outcome provided preliminary evidence that there are multiple (and potentially redundant) visual cues to prosody distributed across the face. In part the motivation for the recording of an auditory-visual speech prosody production corpus (Chapter 4) was to quantitatively explore this proposition. Additionally, these results not only suggested that perceivers were sensitive to the array of prosodic cues in the visual signal but were able to employ this information flexibly; i.e., that when one of these cues was no longer available (due to occlusion, image manipulation or simply not being produced), those that remained appeared sufficient to permit the underlying

prosodic category to be determined. This hypothesis was further explored in Experiments 4 to 6 (see Section 11.1.3 below).

11.1.2. Beneficial Face Areas for Specific Prosodic Contrasts

To determine whether a particular face area held a greater cue value for conveying specific prosodic contrasts to perceivers, the results of Experiments 1 to 3 were compared. Although performance was above chance across all presentation conditions and prosodic contrasts, displays of the lower face (showing articulatory movements) resulted in better discrimination for the contrasting of narrow focus from broad focus (differing from the reports of Swerts & Kraemer, 2008), whereas the phrasing contrasts (echoic questions vs. declarative statements) were better discriminated from upper face displays (commensurate with the results Lansing & McConkie, 1999).

Note though that this result was only found for the within-modal matching task, with no specific face area being of greater benefit for cross-modal matching. This was interpreted as being due to differences in how well prosody was specified by the initial item within a stimulus pair. Given that auditory perception of prosody is quite good (as shown in Chapters 6 and 9; Dohen & Lœvenbruck, 2009), the initial item in the cross-modal matching task (i.e., an auditory token) specifies the prosodic type equally well regardless of whether it is followed by an upper or lower face display. This specification can then be used to guide the perceiver to appropriate correlates in the subsequently presented video display.

11.1.3. Tolerating Variability in Prosodic Realisation

Experiments 4 to 6 examined the degree to which perceivers were able to tolerate signal-level differences (i.e., across face areas and talkers) in the realisation of

prosodic contrasts. This was indexed by the perceiver's ability to perform within-modal and cross-modal prosody matching. These experiments allowed for further investigation of the hypothesis that perceivers determine the underlying prosodic category from the presented tokens (regardless of modality, face area or the talker that produced it), and make their matching decision at this abstract (categorical) level.

Although performance was better when items within pairs were produced by the same talker, the results showed that perceivers were able match prosody from visual cues provided by the upper face to the lower face (and vice versa) across different talkers. Similarly, good matching performance was obtained for the cross-modal task when the initially presented auditory token was produced by one talker and the video token was produced by the other talker (regardless of the face area shown). This ability to match very different cues supported the proposal that matching was performed at an abstract level.

In order to begin to understand how variable cues might be mapped onto prosodic categories, models that have been proposed to deal with variability in speech recognition (i.e., models of how perceivers distinguishing phonemes) were considered. In this regard, cue-integration approaches, in particular the C-CuRE model (McMurray & Jongman, 2011; McMurray et al, 2011), appeared to be the most attractive framework to explain the results, since such models assume that there may be many cues that flexibly signal a linguistic property (in this case prosody) rather than a few invariant ones. Furthermore, such an approach seems more suited to coping with novel types of input (e.g., matching from the upper face of one talker to the lower face of a different one). Indeed, the C-CuRE style of approach has the

benefit of combining aspects of both invariance and exemplar approaches, by assuming that numerous cues are encoded by perceivers and that in combination, variability in any one of these can overcome by the other cues in the signal.

11.2. Producing Prosody

The results of the first experimental series (Experiments 1 to 6) suggested that there are multiple visual cues to prosody (distributed across upper and lower face areas), and that perceivers are sensitive to their occurrence. To determine a precise definition and a more comprehensive understanding of the function of visual prosody, a production study was conducted (Chapter 4) to quantify the properties of auditory prosody (Chapter 5) and their visual correlates (Chapter 7), the consistency of the manifestation of these cues across talkers, and the relationship between them (Chapter 8). The collection of three-dimensional data also allowed for sophisticated manipulation of the data for use in further perceptual studies (Chapter 9 and 10). Detailed below are the key findings from these analyses.

11.2.1. Auditory Correlates of Prosodic Focus and Phrasing Contrasts

Given that the auditory correlates of prosodic focus and phrasing contrasts have previously been well described in the literature (i.e., in terms of F_0 , intensity, duration and vowel space properties, Cooper et al., 1985; Eady & Cooper, 1986; Hay et al., 2006; Kochanski et al., 2005; Krahmer & Swerts, 2001; Nooteboom, 1997), the exploration of acoustic properties from the recorded corpus was conducted to confirm that the prosodic contrasts demonstrated the typical characteristics rather than to identify any new properties.

In general, both contrast types were realised with the expected differences in acoustic properties. For focus contrasts, the pre-critical content of narrow focused

renditions was produced with a lower mean intensity than equivalent content in a broad focused context. The prosodically marked constituent was produced with longer syllable durations, a greater $F0$ range, and larger intensity range, while utterance content following the focused constituent was also produced with longer syllable durations, a lower mean $F0$, a greater range of $F0$, and a lower mean intensity (compared to broad focused renditions).

In comparison to declarative statements, echoic questions were produced with longer mean syllable durations, increased mean $F0$, greater range of $F0$ and a larger intensity range during the critical constituent (i.e., the questioned item within the sentence), whereas the post-critical content was produced with a higher mean $F0$, a larger range of $F0$ and intensity, and a greater mean intensity.

11.2.2. Visual Correlates of Prosodic Focus and Phrasing Contrasts

To determine the visual correlates of the prosodic contrasts, the dimensionality of the motion capture data (representing the talkers' head and face movements) was reduced using a guided principal components analysis (Maeda, 2005) from 38 three-dimensional marker positions per frame, to three rigid rotation and three rigid translation parameters, and eight non-rigid movement parameters representing biomechanically plausible articulatory control parameters. For each utterance, the area under the principal component amplitude curves was then compared across prosodic contrasts.

In general, both articulatory (e.g., jaw and lip movements) and non-articulatory gestures (e.g., eyebrow and rigid head motion) were involved in contrasting focus and phrasing types. For focus contrasts, an increased amount of movement occurred on all eight non-rigid movement parameters (i.e., jaw opening,

lip opening, lower lip movement, upper lip movement, lip rounding, jaw protrusion, eyebrow raising and eyebrow pinching) during the production of the critical constituent in narrow focused utterances compared to broad focused ones.

Phrasing contrasts showed similar differences, with the echoically questioned critical constituent produced with more jaw movement, lip opening, increased lip opening and jaw protrusion, greater eyebrow pinching and more rigid pitch rotations (i.e., rotations around the x -axis) than the equivalent segmental content embedded within a statement context.

11.2.3. Variability in the Production of Prosodic Contrasts

The realisation of prosodic contrasts varied, both auditory and visually, as a function of three different factors: the utterance properties (i.e., the number of syllables in the utterance, and the location within the utterance of the critical constituent), the interactive setting in which the recording took place (i.e., whether or not the talker could see the interlocutor), and as a function of the talker.

In terms of utterance properties, the majority of differences occurred post-critically dependant on the location of the prosodically marked constituent in the utterance (i.e., whether the critical constituent occurred in the first half or second half of the utterance). A relatively straightforward account was proposed to explain such differences which made two basic assumptions: The first was to consider the prosodic marking of a constituent as a form of localised hyperarticulation (see de Jong, 1995, 2004; de Jong, Beckman & Edwards, 1993; Silbert & de Jong, 2008). That is, a constituent can be prosodically marked by enhancing any number of a range of auditory and visual signal properties (e.g., larger jaw movements and increased intensity). The second was to assume that change does not occur

immediately, i.e., the lead up to enhancement and the subsequent return are gradual. Due to this hysteresis, when the critical constituent occurs in the latter half of the utterance (particularly when the utterance is short), there may be insufficient time for the talker to readjust their articulation (and hence the signal) back to pre-critical levels.

Talkers also varied in their realisation of auditory and visual prosody as a function of the interactive setting (whether they could see who they were talking to). This was in part expected based on an extension of Lindblom's (1990, 1996) Hyper-Hypospeech theory, which proposes that when conversing talkers tend to expend only as much effort that allows the listener to maintain lexical access, and will shift away from such a low-cost mode based on factors that surround the interaction (such as noise of the communicative environment). Given perceivers' sensitivity to (and use of) visual prosody, it was predicted that situations where visual prosody was no longer made available to perceivers (i.e., in auditory-only interactions) would result in talkers compensating for the loss of this signal by enhancing auditory-based cues. Indeed, this prediction was supported in the acoustic analysis, with a small number of features being enhanced to a greater degree for narrow focus and echoic question renditions (relative to broad focused ones) in the AO setting than when compared to the FTF setting (the perceptual effects of which are discussed later on in Section 11.3.1).

To produce some of these acoustic enhancements requires an increased amount of articulatory movement, so it was not surprising to find larger lip openings in AO compared to FTF settings. However, for visual cues which are not directly involved in the shaping of the acoustic signal (e.g., eyebrow movements), one might

expect there to be an increase in movements only when they can be seen by an interlocutor (in FTF settings) and a reduction (or at most maintenance of these cues) in situations where they will not be visible. Interestingly, the opposite pattern was observed in visual analysis (Chapter 7): even though eyebrow movements were not able to be seen and have little to do with the shaping of the acoustic signal, they were exaggerated (relative to the FTF settings) across critical and post-critical phases for narrow focus renditions. The implication of these finding in relation to the potential functions of visual prosody is discussed in further detail in Section 11.4.

Finally, talkers varied in the acoustic and visual features used to contrast prosodic focus and phrasing types. There were some features that were more commonly used (and consistently produced by all talkers), but the degree to which these properties were enhanced, and the additional auditory and visual features that accompanied them, varied substantially across talkers. This difference in the way prosody is realised across talkers appears to also have perceptual consequences (as explored in Chapters 10 and 11).

11.2.4. Relationship between Auditory and Visual Prosodic Signals

The relationship between auditory and visual signals was explored in Chapter 8 by conducting a series of correlation analyses between the extracted auditory ($F0$ and intensity contours) and visual parameters (principal component curves), and by examining the temporal displacement between the onsets of auditory and non-articulatory visual prosodic markers.

Overall, the correlation between auditory and visual parameters, even for those involved in speech production (i.e., jaw and lip movement) was highly variable across utterances. Given the structured nature of the recorded corpus, it was possible

to compare the strength of the correlations between auditory and visual features in utterances containing a prosodically marked constituent (i.e., narrow focus and echoic question tokens) with broad focused ones. When the utterances were examined in their entirety, the relationship between modalities weakened in situations where a constituent was prosodically marked, suggesting that the nature of the relationship may be non-linear, with the action of prosodically marking a segment having consequences on auditory-visual alignment of pre-critical and post-critical utterance phases. Indeed, when the relationship between signal modalities was considered only for the critical constituent, the strength of the relationship was found to be greater compared to the broad focussed renditions.

Given the possibility of a non-linear relationship between auditory and non-articulatory visual cues, some aspects of the AV relationship may not have been adequately captured in the examination of correlation values. As such, the timing of eyebrow raises and rigid pitch rotations of the talkers' head in relation to the onset of the critical constituent was examined. For the utterances that were accompanied by an eyebrow raise, these occurred before the critical constituent had been uttered, regardless of the prosodic context. In contrast, the timing of rigid head movements varied across the prosodic conditions: for broad focus utterances there was no systematic temporal relationship, for narrow focus the rigid pitch rotation peaked after the critical constituent had been uttered, whereas echoically questioned critical constituents were preceded by the rigid head movement. Furthermore, this pattern of data was observed when both movements were present within an utterance, and when the timing of a rise in the F_0 was used as the starting point of the prosodically marked constituent. This outcome suggests that talkers have at least some degree of

control over the production of non-articulatory visual features, and that they can be decoupled from the production of acoustic features in order to serve different functions dependant on the prosodic context (these potential functions are discussed in more detail in Section 11.4).

11.3. Linking Production and Perception

The data recorded in Chapter 4 was used to generate the stimuli for a series of perceptual experiments designed to explore the link between prosody production and perception. Two sets of tasks were employed: Chapters 6 and 9 used a subjective rating task requiring perceivers to rate either the degree of focus or clarity of the statement-question contrast, whereas Chapter 10 employed the cross-modal prosody matching task (as used in Chapters 2 and 3) with visual stimuli showing augmented point-light representations of the talkers face movements.

11.3.1. Perceptual Effects of the Talker Seeing the Interlocutor

Chapter 6 explored the perceptual effect of talkers' acoustic modifications made to the speech signal when realising prosody in interactive settings where they could not see the interlocutor. By comparing the subjective rating scores across FTF and AO settings for narrow focus and echoic question tokens, both prosodic contrasts were rated higher (i.e., more emphasis for the narrow focused tokens, and more question-like for the echoic questions, relative to broad focused renditions) when they were recorded in AO settings. This effect was robust, as it was maintained across all talkers, sentences and raters, as well as being replicated in Chapter 9 where a different stimuli set and raters were used.

A series of multivariate linear regressions were used to explain the variance in the rating data, the outcome of which showed that some auditory features uniquely

explained a small amount of variance, but a substantial proportion remained unaccounted. This inability of a linear model to account for a large portion of the variance suggested several possibilities: that a non-linear combination of acoustic cues were used to enhance the expression of prosody, that additional signal-based modifications were made that were not adequately captured in the analysis, or that the performance of perceivers provides a more sensitive measure than signal based measurements for determining prosodic differences.

Note though that this effect of higher ratings for tokens produced in AO settings compared to FTF ones, although replicated in Chapter 9, was only observed for auditory-alone presentations. That is, when visual information recorded across the two interactive settings was presented to perceivers for rating, no differences as a function of the interactive setting were found (despite signal-level differences in the visual analysis detailed in Chapter 7). Similarly, auditory-visual stimulus ratings showed no difference as a function of the interactive setting for either narrow focus or echoic question items. This finding implies that perceivers may be more attuned to detecting auditory-based cues for prosody than visual ones (see Srinivasan & Massaro, 2003, for a similar argument).

11.3.2. Perceptual Effects of Seeing the Talker

As mentioned above, Chapter 9 also presented stimuli in visual only and auditory visual conditions for subjective rating. Given that the prior experiments had shown that perceivers were sensitive to visual prosodic correlates, it was expected that differences in ratings would be observed between broad and narrow focus, and between statements and echoic question renditions in the respective rating tasks. Furthermore, given that a proposed function of visual prosody is that it reinforces the

overall salience of prosodic content (Flecha-Garcia, 2010; Swerts & Krahmer, 2010), the ratings were compared between auditory alone (AA), video only (VO) and auditory visual (AV) presentation conditions to determine if any AV effect was apparent.

The rating data indeed showed that perceivers were sensitive to the visual correlates of prosody, with narrow focus renditions being rated as possessing a stronger degree of emphasis than broad focused utterances, and echoic questions being rated as more question-like on the statement-question continuum, in both the VO and AV modalities. However, there was no evidence of an AV effect, with the ratings in the AV task showing no difference in comparison to the AA task for any of the prosodic conditions (the implications of this outcome are considered in Section 11.4).

11.3.3. Movement Requirements for Perceiving Prosody

The final experimental series (Chapter 10) explored which motion cues (i.e., rigid movements, non-rigid movements or articulatory gestures) may be better for conveying prosodic information to perceivers. This was achieved by animating point-light displays with the movement types either in combination or in isolation, and presenting them to perceivers in a cross-modal prosody matching task.

When all of the motion types were presented in combination, perceivers were proficient (i.e., performed at levels greater than chance) in matching auditory tokens to point-light representations on the basis of prosody alone. However, the performance levels were substantially lower than performance in the experiments reported in Chapters 2 and 3 where restricted face displays that included textural details were presented. This finding suggests that, although motion cues do carry

prosodic content to some extent, textural details such as eye widening and skin wrinkling, also appear to be involved.

Finally, when the movement types were presented in isolation, it appears that articulatory movements alone provide sufficient contrastive detail (for five of the six talkers) to determine focus, whereas rigid motions of the whole head are beneficial (also for five from six talkers) for conveying phrasing contrasts. Collapsed across talkers, no differences were observed across the isolated presentation conditions, with the inclusion of a greater number of motion cue types resulting in similar levels of cross-modal matching performance.

11.4. Potential Communicative Functions of Visual Prosody

Throughout the thesis, several proposals were put forth as to the possible communicative functions that visual prosody may serve. These functions can be broadly grouped into two categories: talker-centric, in which the perceptual benefits are “epiphenomenal” (as they occur as consequence of some other process), and perceiver-centric, where the visual movements are intentionally produced to provide some benefit to those viewing them.

The first of the talker-centric functions proposed that visual prosody is merely an uncontrolled by-product of speech production, occurring as a consequence of articulatory processes (out of muscular synergy) rather than being intentionally produced by the talker. For this hypothesis to be supported, one would expect that the relationship between auditory and visual properties would be reasonably consistent across utterances, and that every occurrence of an auditory prosodic marker would be accompanied by a corresponding visual one. Similar to the results of Cavé et al. (1996), Guaitella et al. (2009) and Yehia et al. (1998), this was not

found to be the case. The correlations between the visual parameters and auditory properties were inconsistent, varying in strength across utterances and prosodic conditions. Furthermore, not every occurrence of an auditory marker of prosody was accompanied by a visual correlate, and when they did occur, the timing of such movements varied substantially. Overall, these results suggest that talkers do have some control over the production of visual prosodic correlates.

An alternate view to the muscular synergy proposal (although one still centred on the talker) is that visual prosody serves a purpose for the talker themselves, i.e., it assists in the conceptualisation of the spoken message by facilitating access to the mental representation of prosody. This proposal is based in the literature on the production of manual gestures during speech production that have suggested such movements are often produced despite the fact that they are not visible to an interlocutor (e.g., when talking on the telephone, Bavelas et al., 2008).

The main support for such a proposal was the finding in the visual analysis (Chapter 7) that non-articulatory visual prosodic cues, such as rigid head movements and eyebrow raises, although not involved in shaping the speech signal per se, were still produced despite the fact that they could not be seen by an interlocutor (i.e., in the AO interactive setting). Similarly, when examining the co-occurrence of auditory and visual prosodic markers and their temporal relationship (Chapter 8), no differences in the number of movements, or the timing of such gestures, were observed across the interactive settings. However, such a proposal should not be considered an exclusive account for the occurrence of all types of visual prosody, as some movements (e.g., rigid head movements) were enhanced to a greater extent in

the FTF setting when they were able to be seen. Further investigation is still required to explore this proposal.

Regardless of whether or not visual prosody is produced by the talker for themselves, perceivers are sensitive to its occurrence (as demonstrated in the perceptual rating, within-modal matching and cross-modal matching tasks). Two possibilities were considered for how visual prosody may benefit perceivers: the first was that visual cues may enhance the overall salience of a prosodic contrast due to both occurring at the same time (i.e., alignment hypothesis; Flecha-Garcia, 2010; Kraemer & Swerts, 2010). The second possibility was that visual cues act as a signalling mechanism, to indicate that important information (i.e., a prosodically marked constituent) is about to occur in the auditory stream (with visual cues preceding the auditory ones).

These two proposals were initially explored by examining the timing of non-articulatory visual cues in relation to auditory prosodic markers. While eyebrow movements consistently preceded the onset of the prosodically marked constituent in the auditory signal (lending support to the signalling hypothesis), the rigid pitch rotation movements varied across prosodic contexts, with such movements being aligned with the auditory prosodic markers for narrow focus tokens, but preceding the critical word in echoic question renditions. However, further evaluation using perceptual rating tasks showed no AV effect. That is, despite the occurrence of both auditory and visual prosodic cues in AV presentations, the overall ratings of narrow focus and echoic questions showed no increase when compared to presentations of only auditory information alone, lending further support to the signalling hypothesis.

It should be noted that none of these proposals need be exclusive accounts. Given the variable spatial and temporal relationship between auditory and visual signals, the function of visual cues may vary across utterances. That is, in some cases the auditory prosodic marker may be weak and be compensated for with a visual marker, whereas other tokens may be produced with such a strong auditory contrast that the talker does not need to produce a particularly salient visual marker. Furthermore, although the timing of these cues have been used as a way of disentangling their perceptual functions, human perceivers are able to tolerate asynchrony between auditory and visual speech signals (for perceptual integration) in the range of -30ms (i.e., auditory signal preceding the visual one) to +170ms (Conrey & Pisoni, 2006; Dixon & Spitz, 1980; van Wassenhove, Grant & Poeppel, 2007). Thus, perceivers' tolerance for asynchrony between auditory and visual prosodic signals still remains to be examined. It is also important to bear in mind that the acts of speech production and perception are not identical tasks, so the link between them may not be all that tight: speech production is a collective task involving the coordinated movement of many different anatomical structures (a large portion of which is planned, Dogil, Ackerman, Grodd, Haider, Kamp, Mayer, Riecker & Wildgruber, 2002; Tseng, Pin, Lee, Wang & Chen, 2005) in order to generate a communicative signal, whereas perception is interested first in the decomposition of the signal, followed by selection of those features deemed to be relevant by the perceiver (both processes of which are modulated by attention).

11.5. Limitations and Future Directions

Before concluding, it is important to consider the limitations of the current research program, and possibilities for future work in the area of auditory-visual speech

prosody. Firstly, the use of the 2AFC prosody matching tasks may provide a generous estimate of perceivers' ability to use visual prosodic information, with the presentation of minimal pairs making key differences more salient and potentially shaping correct responding. An alternative could be to use an identification task that requires perceivers to determine the word within the utterance that received a prosodic marker in unimodal and bimodal contexts, which may better reflect their ability to relate visual events to auditory prosodic cues (rather than a reliance on low-level differences between stimulus pairs such as absolute duration).

A similar identification paradigm could be used to further explore the tolerance for the timing between auditory and visual prosodic cues, by temporally displacing the visual cue onsets relative to the start of the critical word, and pairing such visual markers with broad focused auditory renditions (as used by Swerts & Kraemer, 2008). Given that the point-light representations of talkers' visual speech movements can convey suprasegmental information (with no evidence of floor or ceiling effects in the cross-modal matching task), such stimuli provide a suitable platform to conduct further experiments that investigate the perceptual consequences of manipulating the spatial and temporal properties of visual cues to prosody.

In the current study, the use of a highly constrained dialogue task allowed for ease of comparison between talkers, interactive settings and prosodic conditions, but may have come at the cost of more natural interactive behaviours (e.g., eye gaze, speech disfluencies and self-corrections). By contrast, free dialogue tasks such as those used by Fitzpatrick et al. (2011) generate natural interactive behaviours, but require substantially more processing post-recording to identify comparable tokens, with no guarantee that the targeted contrasts will be produced consistently across all

talkers or speech conditions. Thus, the use of a more natural (yet semi-structured) language task for eliciting the prosodic contrasts should also be considered, e.g., an error correction task with less predictability, or the “wizard of oz” style paradigm (Bulyko, Kirchhoff, Ostendorf & Goldberg, 2005; Burnham, Joeffry & Rice, 2010; Jefferson, 1974; Oviatt, Levow, Moreton & MacEachern, 1998).

The recording of the corpus involved the use of an active marker system (OPTOTRAK), with small optical markers placed directly on the head and face of the talker. While these systems provide highly accurate measurements, the presence of markers may interfere with natural speech production (Popat, Richmond, Benedikt, Marshall & Rosin, 2009; Stone, 1997). Alternate motion capture techniques could be considered for recording further data, particularly those that involve no markers (i.e., so-called “texture” based systems), such as the 4D Capture System (3dMD, as used by Popat, Henley, Richmond, Benedikt, Marshall & Rosin, 2010; Popat, Richmond, Marshall & Rosin, 2011) or MOVA’s Contour Reality Capture.

The analysis of the recorded motions could also be approached in an alternate way. While the area under curve approach (Dohen et al., 2009) indicates the spatial properties of the visual correlates to prosody, the temporal aspects cannot be accurately determined. One possibility is to use functional Analysis of Variance (fANOVA). In essence, the fANOVA procedure involves a series of one-way ANOVAs conducted at multiple time points, with *F*-values represented graphically as a function of time (Cuevas, Febrero & Fraiman, 2004). This approach has previously been used in speech contexts to compare lip movements and acceleration for multiple productions of four vowels in /bVb/ syllable contexts (Ramsay,

Munhall, Gracco & Ostry, 1996), and allows the identification not only of where differences in the visual signal occur, but also when they occur. However, fANOVA is based on token repetition, and thus when segmental variation is not of interest, a large number of analyses would be required (i.e., one set of analyses per sentence, where each set contains a separate analysis per principal component). Alternately, recording a smaller corpus of sentences but with a greater number of repetitions, or the use of reiterant speech (i.e., speech with the rhythm, intensity and *F0* properties of normal speech but with all segmental content replaced with simple CV syllables such as /ba/), could be used to overcome the large number of utterances required to examine the timing of visual prosody.

11.6. Summary

In sum, in addition to recording a multi-talker corpus of audiovisual speech prosody productions, this thesis has shown that:

1. Talkers produce linguistic prosodic contrasts (i.e., focus and phrasing) with both auditory and visual correlates, which are distributed across face areas.
2. Perceivers are sensitive to both the auditory and visual prosodic correlates, regardless of the face area which they occur in.
3. Despite variability in the realisation of prosody across talkers, perceivers tolerate signal-level difference by determining the underlying prosodic type.
4. The nature of the relationship between auditory and visual modalities is highly variable and likely non-linear.
5. Prosody can be conveyed by point-light representations of the talkers head and face movements despite lacking textural details.

6. Narrow focus is better conveyed by lower face articulatory movements, such as lip and jaw opening and protrusions.
7. Movement of the upper face and overall rigid head motion provides more beneficial information for determining an utterances phrasal nature.
8. The availability of multiple visual cues to prosody does not generate a stronger overall percept of prosody, but rather gives perceivers more choice as to what they can attend to.

References

- Al Moubayed, S., Beskow, J., & Granström, B. (2010). Auditory visual prominence: From intelligibility to behavior. *Journal on Multimodal User Interfaces*, 3, 299-309.
- Al Moubayed, S., Beskow, J., Granström, B., & House, D. (2011). Audio-visual prosody: Perception, detection, and synthesis of prominence. *Lecture Notes in Computer Science: Towards Autonomous, Adaptive and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues*, 6456, 55-71. doi:10.1007/978-3-642-18184-9_6 .
- Alibali, M. W., Heath, D. C., & Myers, H. J. (2001). Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen. *Journal of Memory and Language*, 44, 169-188. doi:10.1006/jmla.2000.2752.
- Alibali, M. W., Kita, S., & Young, A. J. (2000). Gesture and the process of speech production: We think, therefore we gesture. *Language and Cognitive Processes*, 15, 593-613. doi:10.1080/016909600750040571.
- Allen, J. (1976). Synthesis of speech from unrestricted text. *Proceedings of the IEEE*, 64, 433-443. doi:10.1109/PROC.1976.10152.
- Anderson, A. H., Bard, E. G., Sotillo, C., Newlands, A., & Doherty-Sneddon, G. (1997). Limited visual control of the intelligibility of speech in face-to-face dialogue. *Attention, Perception & Psychophysics*, 59, 580-592. doi:10.3758/BF03211866.

- Aubergé, V., & Cathiard, M. (2003). Can we hear the prosody of smile? *Speech Communication, 40*, 87-97. doi:10.1016/S0167-6393(02)00077-8.
- Badin, P., Bailly, G., Reveret, L., Baciú, M., Segebarth, C., & Savariaux, C. (2002). Three-dimensional linear articulatory modelling of tongue, lips and face, based on MRI and video images. *Journal of Phonetics, 30*, 533-553. doi:10.1006/jpho.2002.0166.
- Bailly, G., Govokhina, O., Elisei, F., & Breton, G. (2009). Lip-synching using speaker-specific articulation, shape and appearance models. *EURASIP Journal on Audio, Speech and Music Processing, 2009*, 1-11. doi:10.1155/2009/769494.
- Bavelas, J., Gerwing, J. J., Sutton, C. L., & Prevost, D. (2008). Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language, 58*, 495-520. doi:10.1016/j.jml.2007.02.004.
- Beach, C.M. (1991). The interpretation of prosodic patterns at points of syntactic structure ambiguity: Evidence for cue trading relations. *Journal of Memory and Language, 30*, 644-663. doi:10.1016/0749-596X(91)90030-N .
- Beautemps, D., Badin, P., & Bailly, G. (2001). Linear degrees of freedom in speech production: Analysis of cineradio- and labio-film data and articulatory-acoustic modeling. *Journal of the Acoustical Society of America, 109*, 2165-2180. doi:10.1121/1.1361090.
- Benoît, C., & Le Goff, B. (1998). Audio-visual speech synthesis from French text: Eight years of models, designs and evaluation at the ICP. *Speech Communication, 26*, 117-129. doi:10.1016/S0167-6393(98)00045-4 .

- Bernardis, P., & Gentilucci, M. (2006). Speech and gesture share the same communication system. *Neuropsychologia*, *44*, 178-190.
doi:10.1016/j.neuropsychologia.2005.05.007.
- Bernstein, L. E., Eberhardt, S. P., & Demorest, M. E. (1989). Single-channel vibrotactile supplements to visual perception of intonation and stress. *Journal of the Acoustical Society of America*, *85*, 397-405. doi:10.1121/1.397690.
- Beskow, J., Granstrom, B., & House, D. (2006). Visual correlates to prominence in several expressive modes. *Interspeech 2006*, pp.1272-1275.
- Boersma, P. P. G. (2001). Praat: A system for doing phonetics by computer. *Glottal International*, *5*, 341-345.
- Bolinger, D. (1972). Accent is predictable (if you're a mind reader). *Language*, *48*, 633-644.
- Bolinger, D. (1989). *Intonation and its uses*. Edward Arnold, London.
- Bradlow, A. R., Toretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, *20*, 255-272. doi:10.1016/S0167-6393(96)00063-5.
- Buchan, J. N., Paré, M., & Munhall, K. G. (2008). The effect of varying talker identity and listening conditions on gaze behaviour during audiovisual speech perception. *Brain Research*, *1242*, 162-171.
doi:10.1016/j.brainres.2008.06.083.
- Buchwald, A. B., Winters, S. J., & Pisoni, D. B. (2008). Visual speech primes open-set recognition of spoken words. *Language and Cognitive Processes*, *24*, 580-610. doi:10.1080/01690960802536357.

- Bulyko, I., Kirchhoff, K., Ostendorf, M., & Goldberg, J. (2005). Error-correction detection and response generation in a spoken dialogue system. *Speech Communication, 45*, 271-288. doi:10.1016/j.specom.2004.09.009.
- Burnham, D., Jeoffrey, S., & Rice, L. (2010). Computer- and human-directed speech before and after correction. *Proceedings of SST 2010*, 13-17.
- Burnham, D., Reynolds, J., Vignali, G., Bollwerk, S., & Jones, C. (2007). Rigid vs non-rigid face and head motion in phone and tone perception. *Interspeech 2007*, pp. 698-701.
- Cavé, C., Guaitella, I., Bertrand, R., Santi, S., Harlay, F., & Espesser, R. (1996). About the relationship between eyebrow movements and F0 variations. In *Proceedings of the Fourth International Conference on Spoken Language Processing*, pp. 2175-2178.
- Cohen, A. A. (1977). The communicative functions of hand illustrators. *Journal of Communication, 27*, 54-63. doi:10.1111/j.1460-2466.1977.tb01856.x.
- Cohen, A. A., & Harrison, R. P. (1973). Intentionality in the use of hand illustrators in face-to-face communication situations. *Journal of Personality and Social Psychology, 6*, 341-349. doi:10.1037/h0035792.
- Conrey, B., & Pisoni, D. B. (2006). Auditory-visual speech perception and synchrony detection for speech and nonspeech signals. *Journal of the Acoustical Society of America, 119*, 4065-4073. doi:10.1121/1.2195091.
- Cooke, M., & Lu, Y. (2010). Spectral and temporal changes to speech produced in the presence of energetic and information maskers. *Journal of the Acoustical Society of America, 128*, 2059-2069. doi:10.1121/1.3478775.

- Cooper, W. E., Eady, S. J., & Mueller, P. R. (1985). Acoustical aspects of contrastive stress in question-answer contexts. *Journal of the Acoustical Society of America*, *77*, 2142-2156. doi:10.1121/1.392372.
- Crystal, D. (1991). *A Dictionary of Linguistics and Phonetics* (3rd ed.). Oxford, UK: Blackwell Reference.
- Cuevas, A., Febrero, M., & Fraiman, R. (2004). An anova test for functional data. *Computational Statistics & Data Analysis*, *47*, 111-122. doi:10.1016/j.csda.2003.10.021.
- Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language & Speech*, *40*, 141-201. doi:10.1177/002383099704000203.
- Cvejic, E., Kim, J., & Davis, C. (2010). Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion. *Speech Communication*, *52*, 555-564. doi:10.1016/j.specom.2010.02.006.
- Cvejic, E., Kim, J., & Davis, C. (in press). Recognizing prosody across modalities, face areas and speakers: Examining perceivers' sensitivity to variable realizations of visual prosody. *Cognition*. doi:10.1016/j.cognition.2011.11.013.
- Dahan, D., & Bernard, J.-M. (1996). Interspeaker variability in emphatic accent production in French. *Language and Speech*, *39*, 341-374. doi:10.1177/002383099603900402.
- Davis, C., & Kim, J. (2004). Audiovisual interactions with intact clearly audible speech. *The Quarterly Journal of Experimental Psychology*, *57A*, 1103-1121. doi:10.1080/02724980343000701.

- Davis, C., & Kim, J. (2006). Audio-visual speech perception off the top of the head. *Cognition, 100*, B21-B31. doi:10.1016/j.cognition.2005.09.002.
- Dilly, L. C., Ladd, D. R., & Schepman, A. (2005). Alignment of L and H in bitonal pitch accents: testing two hypotheses. *Journal of Phonetics, 33*, 115-119. doi:10.1016/j.wocn.2004.02.003.
- Dixon, N. F., & Spitz, L. (1980). The detection of auditory visual desynchrony. *Perception, 9*, 719-721. doi:10.1068/p090719.
- Dodd, B. E. (1977). The role of vision in the perception of speech. *Perception, 6*, 31-40. doi:10.1068/p060031.
- Dogil, G., Ackerman, H., Grodd, W., Haider, H., Kamp, H., Mayer, J., Riecker, A., & Wildgruber, D. (2002). The speaking brain: A tutorial introduction to fMRI experiments in the production of speech, prosody and syntax. *Journal of Neurolinguistics, 15*, 59-90. doi:10.1016/S0911-6044(00)00021-X.
- Dohen, M., & Løevenbruck, H. (2005). Audiovisual production and perception of contrastive focus in French: A multispeaker study. *Interspeech 2005*, pp. 2413-2416.
- Dohen, M., & Løevenbruck, H. (2009). Interaction of audition and vision for the perception of prosodic contrastive focus. *Language and Speech, 52*, 177-206. doi:10.1177/0023830909103166.
- Dohen, M., Løevenbruck, H., Cathiard, M.-A., & Schwartz, J. (2004). Visual perception of contrastive focus in reiterant French speech. *Speech Communication, 44*, 155-172. doi:10.1016/j.specom.2004.10.009.

- Dohen, M., Løevenbruck, H., & Hill, H. (2006). Visual correlates of prosodic contrastive focus in French: description and inter-speaker variability. *Proceedings of Speech Prosody 2006*, pp. 221-224.
- Dohen, M., Løevenbruck, H., & Hill, H. (2009). Recognizing prosody from the lips: Is it possible to extract prosodic focus from lip features? In A.W.C Liew & S. Wang (eds.). *Visual Speech Recognition: Lip Segmentation and Mapping* (pp. 416-438). London, United Kingdom: IGI Global.
- Drahota, A., Costall, A., & Reddy, V. (2008). The vocal communication of different kinds of smiles. *Speech Communication*, 50, 278-287.
doi:10.1016/j.specom.2007.10.001.
- Eady, S. J., & Cooper, W. E. (1986). Speech intonation and focus location in matched statements and questions. *Journal of the Acoustical Society of America*, 80, 402-415. doi:10.1121/1.394091.
- Eady, S. J., Cooper, W. E., Klouda, G. V., Mueller, P. R., & Lotts, D. W. (1986). Acoustical characteristics of sentential focus: Narrow vs. broad and single vs. dual focus environments. *Language and Speech*, 29, 233-251.
doi:10.1177/002383098602900304.
- Edwards, J., Beckman, M. E., & Fletcher, J. (1991). Articulatory kinematics of final lengthening. *Journal of the Acoustical Society of America*, 89, 369-382.
doi:10.1121/1.400674.
- Elman, J., & McClelland, J. (1986). Exploiting lawful variability in the speech wave. In J. S. Perkell & D. Klatt (Eds.), *Invariance and Variability in Speech Processes* (pp. 360-380). Hillsdale, NJ: Erlbaum.

- Engwall, O., & Beskow, J. (2003). Resynthesis of 3D tongue movements from facial data. *Proceedings of Eurospeech 2003*, pp. 2261-2264.
- Erickson, D. (2002). Articulation of extreme formant patterns for emphasized vowels. *Phonetica*, 59, 134-149.
- Erickson, D., Fujimura, O., & Pardo, B. (1998). Articulatory correlates of prosodic control: emotion and emphasis. *Language and Speech*, 41, 399-417.
doi:10.1177/002383099804100408.
- Fagel, S., & Madany, K. (2008). Guided non-linear model estimation (gnOME). *Proceedings of the International Conference on Auditory-Visual Speech Processing 2008 (AVSP'08)*, pp. 59-62.
- Fant, G. (1973). *Speech sounds and features*. Cambridge, Massachusetts: MIT Press.
- Fitzpatrick, M., Kim, J., & Davis, C. (2011). The effect of seeing the interlocutor on speech production in different noise types. *Interspeech 2011*, pp. 2829-2832.
- Flanagan, J. L., & Saslow, M. G. (1958). Pitch discrimination for synthetic vowels. *Journal of the Acoustical Society of America*, 30, 435-442.
doi:10.1121/1.1909640.
- Flecha-Garcia, M. L. (2006). Eyebrow raising in dialogue: discourse structure, utterance function, and pitch accents. PhD thesis, Theoretical & Applied Linguistics, University of Edinburgh.
- Flecha-García, M. L. (2010). Eyebrow raises in dialogue and their relation to discourse structure, utterance function and pitch accents in English. *Speech Communication*, 52, 542-554. doi:10.1016/j.specom.2009.12.003.

- Forster, K. I., & Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behavioral Research Methods: Instruments & Computers*, *35*, 116-124. doi:10.3758/BF03195503.
- Foxton, J. M., Riviere, L.-D., & Barone, P. (2010). Cross-modal facilitation in speech prosody. *Cognition*, *115*, 71-78. doi:10.1016/j.cognition.2009.11.009.
- Garnier, M., Henrich, N., & Dubois, D. (2010). Influence of sound immersion and communicative interaction on the Lombard effect. *Journal of Speech and Hearing Research*, *53*, 568-608. doi:10.1044/1092-4388(2009/08-0138).
- Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. *Trends in Cognitive Sciences*, *3*, 419-429. doi:10.1016/S1364-6613(99)01397-2.
- Graf, H. P., Cosatto, E., Strom, V., & Huang, F. J. (2002). Visual prosody: Facial movements accompanying speech. *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, pp.396-401.
- Granström, B., & House, D. (2005). Audiovisual representation of prosody in expressive speech communication. *Speech Communication*, *46*, 473-484. doi:10.1016/j.specom.2005.02.017.
- Granström, B., & House, D. (2007). Inside out: Acoustic and visual aspects of verbal and non-verbal communication. In *Proceedings of 16th International Congress of Phonetic Sciences*, pp. 11-18.
- Grant, K. W., & Seitz, P. -F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America*, *108*, 1197-1208. doi:10.1121/1.1288668.

- Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, genders, and sensory modality: Female faces and male voices in the McGurk effect. *Attention, Perception & Psychophysics*, *50*, 524-536. doi:10.3758/BF03207536.
- Guaitella, I., Santi, S., Lagrue, B., & Cavé, C. (2009). Are eyebrow movements linked to voice variations and turn-taking in dialogue? An experimental investigation. *Language and Speech*, *52*, 207-222. doi:10.1177/0023830909103167.
- Gussenhoven, C. (2007). Types of focus in English. In C. Lee, M. Gordon & D. Büring (Eds.), *Topic and Focus: Studies in Linguistics and Philosophy*, Vol. 82 (pp. 83-100). Dordrecht, The Netherlands: Springer.
- Hadar, U., Steiner, T.J., Grant, E. C., & Rose, F. C. (1983). Head movement correlates of juncture and stress at sentence level. *Language and Speech*, *26*, 117-129. doi:10.1177/002383098302600202.
- Hay, J. F., Sato, M., Coren, A. E., Moran, C. L., & Diehl, R. L. (2006). Enhanced contrasts for vowels in utterance focus: A cross-language study. *Journal of the Acoustical Society of America*, *119*, 3022-3033. doi:10.1121/1.2184226.
- Hill, H., & Pollick, F. E. (2000). Exaggerating temporal differences enhances recognition of individuals from point light displays. *Psychological Science*, *11*, 223-228. doi:10.1111/1467-9280.00245.
- Hirst, D. J. (2005). Form and function in the representation of speech prosody. *Speech Communication*, *46*, 334-347. doi:10.1016/j.specom.2005.02.020.

- Honda, K., Hirai, H., Masaki, S., & Shimada, Y. (1999). Role of vertical larynx movement and cervical lordosis in F0 control. *Language and Speech, 42*, 401-411. doi:10.1177/00238309990420040301.
- Horn, B. K. P. (1987). Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America A, 4*, 629-642. doi:10.1364/JOSAA.4.000629.
- Hostetter, A., Alibali, M., & Kita, S. (2007). I see it in my hands' eye: Representational gestures reflect conceptual demands. *Language and Cognitive Processes, 22*, 313-336. doi:10.1080/01690960600632812.
- House, D. (2002). Intonational and visual cues in the perception of interrogative mode in Swedish. *Proceedings of the International Conference on Spoken Language Processing (ICSLP-2002)*, pp. 1957-1960.
- Huber, J. E., & Chandrasekaran, B. (2006). Effects of increasing sound pressure level on lip and jaw movement parameters and consistency in young adults. *Journal of Speech, Language and Hearing Research, 49*, 1368-1379. doi:10.1044/1092-4388(2006/098).
- IEEE Subcommittee on Subjective Measurements (1969). IEEE recommended practices for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics, 17*, 227-246.
- Ishi, C. T., Haas, J., Wilbers, F. P., Ishiguro, H., & Hagita, N. (2007). Analysis of head motions and speech, and head motion control in an android. *Proceedings of the IEEE/RSJ International Conference. on Intelligent Robots & Systems*, pp. 548-553.

- Jefferson, G. (1974). Error correction as an interactional resource. *Language in Society*, 2, 181-199.
- Jiang, J., Alwan, A., Auer, E. T., & Bernstein, L. E. (2001). Predicting visual consonant perception from physical measures. *Proceedings of Eurospeech 2001*, pp. 179-182.
- Jiang, J., Alwan, A., Bernstein, L. E., Keating, P., & Auer, E. (2000). On the correlation between facial movements, tongue movements and speech acoustics. *Proceedings of the International Conference on Spoken Language Processing (ICSLP 2000)*, pp. 42-45.
- Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Attention, Perception and Psychophysics*, 14, 201-211.
doi:10.3758/BF03212378.
- de Jong, K. J. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *Journal of the Acoustical Society of America*, 97, 491-504. doi:10.1121/1.412275.
- de Jong, K. (2004). Stress, lexical focus, and segmental focus in English: patterns of variation in vowel duration. *Journal of Phonetics*, 32, 493-516.
doi:10.1016/j.wocn.2004.05.002.
- de Jong, K., Beckman, M. E., & Edwards, J. (1993). The interplay between prosodic structure and coarticulation. *Language and Speech*, 36, 197-212.
doi:10.1177/002383099303600305.
- Kendon, A. (1994). Do gestures communicate?: A review. *Research on Language and Social Interaction*, 27, 175-200. doi:10.1207/s15327973rlsi2703_2.

- Kim, J., & Davis, C. (2011). Auditory speech processing is affected by visual speech in the periphery. *Interspeech 2011*, pp. 2465-2468.
- Kim, J., Davis, C., & Krins, P. (2004). Amodal processing of visual speech as revealed by priming. *Cognition*, *93*, B39-B47.
doi:10.1016/j.cognition.2003.11.003.
- Kim, J., Sironic, A., & Davis, C. (2011). Hearing speech in noise: Seeing a loud talker is better. *Perception*, *40*, 853-862. doi:10.1068/p6941.
- Kita, S. (2000). How representational gestures help speaking. In D. McNeill (ed.), *Language and gesture: Window into thought and action* (pp. 162-185). Cambridge, United Kingdom: Cambridge University Press.
- Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness predicts prominence: fundamental frequency lends little. *Journal of the Acoustical Society of America*, *118*, 1038-1054. doi:10.1121/1.1923349.
- Krahmer, E., & Swerts, M. (2001). On the alleged existence of contrastive accents. *Speech Communication*, *34*, 391-405. doi:10.1016/S0167-6393(00)00058-3.
- Krahmer, E., & Swerts, M. (2004). More about brows: A cross-linguistic study via analysis-by-synthesis. In Ruttkay, Z., Pelachaud, C. (Eds.), *From Brows to Trust: Evaluating Embodied Conversational Agents* (pp.191-216). Dordrecht, The Netherlands: Kluwer Academic Press.
- Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, *57*, 396-414.
doi:10.1016/j.jml.2007.06.005.

- Krauss, R. M. (1998). Why do we gesture when we speak? *Current Directions in Psychological Science*, 7, 54-59.
- Krauss, R. M., Chen, Y., & Chawla, P. (1996). Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? In M. Zanna (Ed.), *Advances in experimental social psychology* (pp. 389-450). San Diego, California, USA: Academic Press.
- Krauss, R. M., Dushay, R. A., Chen, Y., & Rauscher, F. (1995). The communicative value of conversational hand gesture. *Journal of Experimental Social Psychology*, 31, 533-552. doi:10.1006/jesp.1995.1024.
- Krauss, R. M., Morrel-Samuels, P., & Colasante, C. (1991). Do conversational hand gestures communicate? *Journal of Personality and Social Psychology*, 61, 743-754. doi 10.1037/0022-3514.61.5.743.
- Kroos, C., Kuratate, T., & Vatikiotis-Bateson, E. (2002). Video-based face motion measurement. *Journal of Phonetics*, 30, 569-590.
doi:10.1006/jpho.2002.0164.
- Ladd, D. R. (1980). *The structure of intonational meaning*. Bloomington, Indiana, USA: Indiana United Press.
- Ladd, D. R., Faulkner, D., Faulkner, H., & Schepman, A. (1999). Constant “segmental anchoring” of F_0 movements under changes in speech rate. *Journal of the Acoustical Society of America*, 106, 1543-1554.
doi:10.1121/1.427151.
- Ladd, D. R., & Morton, R. (1997). The perception of intonational emphasis: continuous or categorical. *Journal of Phonetics*, 25, 313-342.
doi:10.1006/jpho.1997.0046.

- Ladd, D. R., & Schepman, A. (2003). "Sagging transitions" between high pitch accents in English: experimental evidence. *Journal of Phonetics*, 31, 81-112. doi:10.1016/S0095-4470(02)00073-6.
- Lansing, C., & McConkie, G. (1999). Attention to facial regions in segmental and prosodic visual speech perception tasks. *Journal of Speech, Language and Hearing Research*, 42, 529-539.
- Lappin, J. S., Tadin, D., Nyquist, J. B., & Corn, A. L. (2009). Spatial and temporal limits of motion perception across variations in speed, eccentricity, and low vision. *Journal of Vision*, 9, 1-14. doi: 10.1167/9.1.30.
- Lee, A., (2008). Virtual Dub (Version 1.8.6) [Software]. Available from <<http://www.virtualdub.org>>
- Lieberman, P. (1960). Some acoustic correlates of word stress in American-English. *Journal of the Acoustical Society of America*, 32, 451-454. doi:10.1121/1.1908095.
- Lieberman, P. (1967). *Intonation, perception and languages*. Cambridge, MA: MIT Press.
- Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H Theory. In W. J. Hardcastle, & A. Marchal (Eds.), *Speech Production and Speech Modeling* (pp. 403-439). Dordrecht, The Netherlands: Kluwer Academic.
- Lindblom, B. (1996). Role of articulation in speech perception: Clues from production. *Journal of the Acoustical Society of America*, 99, 1683-1692. doi:10.1121/1.414691.

- Linnankoski, I., Leinonen, L., Vihla, M., Laakso, M. –J., & Carlson, S. (2005).
Conveyance of emotional connotations by a single word in English. *Speech
Communication, 45*, 27-39. doi:10.1016/j.specom.2004.09.007.
- Løevenbruck, H., Dohen, M., & Vilain, C. E. (2009). Pointing is ‘special’. In S.
Fuchs, H. Løevenbruck, D. Pape & P. Perrier (Eds.), *Some Aspects of Speech
and the Brain* (pp. 211-258). Oxford, United Kingdom: Peter Lang.
- Lucero, J. C., Maciel, S. T. R., Johns, D. A., & Munhall, K. G. (2005). Empirical
modeling of human face kinematics during speech using motion clustering.
Journal of the Acoustical Society of America, 118, 405-409.
doi:10.1121/1.1928807.
- Maeda, S. (2005). Face models based on a guided PCA of motion capture data:
Speaker dependant variability in /s/ - /z/ contrast production. *ZAS Papers in
Linguistics, 40*, 95-108.
- Malestky, L. P., Sun, J., & Morton, N. A. (2007). Accuracy of an optical active-
marker system to track the relative motion of rigid bodies. *Journal of
Biomechanics, 40*, 682-685. doi:10.1016/j.jbiomech.2006.01.017.
- Massaro, D. W., & Beskow, J. (2002). Multimodal speech perception: A paradigm
for speech science. In B. Granström, D. House, & I. Karlsson (Eds.),
Multimodality in language and speech systems (pp. 45-71). Dordrecht, The
Netherlands: Kluwer Academic Publishers.
- McClave, E. (1998). Pitch and manual gestures. *Journal of Psycholinguistic
Research, 27*, 69-89. doi: 10.1023/A:1023274823974.
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*,
746-748. doi:10.1038/264746a0.

- McKee, S. P., & Nakayama, K. (1984). The detection of motion in the peripheral visual field. *Vision Research*, *24*, 25-32. doi:10.1016/0042-6989(84)90140-8.
- McMurray, B., Cole, J. S., & Munson, C. (2011). Features as an emergent product of computing perceptual cues relative to expectations. In G. N. Clements and R. Ridouane. (Eds.), *Where Do Phonological Features Come From?: Cognitive, Physical and Developmental Bases of Distinctive Speech Categories* (pp. 197-236). The Netherlands: John Benjamins Publishing Company.
- McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, *118*, 219–246. doi:10.1037/a0022325.
- Mol, L., Krahmer, E., Maes, A., & Swerts, M. (2012). Adaptation in gesture: Converging hands or converging minds? *Journal of Memory and Language*, *66*, 249-264. doi:10.1016/j.jml.2011.07.004.
- Munhall, K.G., Jones, J.A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science*, *15*, 133-137. doi: 10.1111/j.0963-7214.2004.01502010.x.
- Munhall, K. G., MacDonald, E. N., Byrne, S. K., and Johnsrude, I. (2009). Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate. *Journal of the Acoustical Society of America*, *125*, 384-390. doi:10.1121/1.3035829.

- Nooteboom, S. (1997). The prosody of speech: Melody and rhythm. In W. J. Hardcastle & J. Laver (Eds.), *The Handbook of Phonetic Science* (pp. 640-673). London: Blackwell.
- Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, *85*, 172–191. doi:10.1037/0033-295X.85.3.172.
- Oviatt, S., Levow, G.-A., Moreton, E., & MacEachern, M. (1998). Modeling global and focal hyperarticulation during human-computer error resolution. *Journal of the Acoustical Society of America*, *104*, 3080-3098. doi:10.1121/1.423888.
- Paré, M., Richler, R. C., ten Hove, M., & Munhall, K. G. (2003). Gaze behavior in audiovisual speech perception: The influence of ocular fixations on the McGurk effect. *Attention, Perception and Psychophysics*, *65*, 553-567. doi:10.3758/BF03194582.
- Pell, M. D. (2001). Influence of emotion and focus location on prosody in matched statements and questions. *Journal of the Acoustical Society of America*, *109*, 1668-1680. doi: 10.1121/1.1352088.
- Peppé, S., Maxim, J., & Wells, B. (2000). Prosodic variation in Southern British English. *Language and Speech*, *43*, 309-334. doi:10.1177/00238309000430030501.
- Popat, H., Henley, E., Richmond, S., Benedikt, L., Msrhall, D., & Rosin, P. L. (2010). A comparison of the reproducibility of verbal and nonverbal facial gestures using three-dimensional motion analysis. *Otolaryngology- Head and Neck Surgery*, *142*, 867-872. doi:10.1016/j.otohns.2010.03.003.

- Popat, H., Richmond, S., Benedikt, L., Marshall, D., & Rosin, P. L. (2009). Quantitative analysis of facial movement- A review of three-dimensional imaging techniques. *Computerized Medical Imaging and Graphics*, *33*, 377-383. doi:10.1016/j.compmedimag.2009.03.003.
- Popat, H., Richmond, S., Marshall, D., & Rosin, P. L. (2011). Facial movement in 3 dimensions: Average templates of lip movements in adults. *Otolaryngology-Head and Neck Surgery*, *145*, 24-29. doi:10.1177/0194599811401701.
- Ramsay, J. O., Munhall, K. G., Gracco, V. L., & Ostry, D. J. (1996). Functional data analyses of lip motion. *Journal of the Acoustical Society of America*, *99*, 3718-3727. doi:10.1121/1.414986.
- Repp, B. H., & Crowder, R. G. (1990). Stimulus order effects in vowel discrimination. *Journal of the Acoustical Society of America*, *88*, 2080-2090. doi: 10.1121/1.400105.
- Rosenblum, L. D., Johnson, J. A., & Saldana, H. M. (1996). Visual kinematic information for embellishing speech in noise. *Journal of Speech and Hearing Research*, *39*, 1159-1170.
- Scarborough, R., Keating, P., Mattys, S., Cho, T., & Alwan, A. (2009). Optical phonetics and visual perception of lexical and phrasal stress in English. *Language and Speech*, *52*, 135-175. doi: 10.1177/0023830909103165.
- Schmidt, J., Berg, D. R., Ploeg, H.- L., & Ploeg, L. (2009). Precision, repeatability and accuracy of Optotrak® optical motion tracking systems. *International Journal of Experimental and Computational Biomechanics*, *1*, 114 – 127. doi:10.1504/IJECB.2009.022862.

- Schröder, M., & Trouvain, J. (2003). The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology*, 6, 365-377. doi: 10.1023/A:1025708916924.
- Schulman, R. (1989). Articulatory dynamics of loud and normal speech. *Journal of the Acoustical Society of America*, 85, 295-312. doi: 10.1121/1.397737.
- Schwartz, J. –L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition*, 93, B69-B78. doi:10.1016/j.cognition.2004.01.006.
- Selkirk, E. (1995). Sentence prosody: Intonation, stress and phrasing. In J. Goldsmith (Ed.), *The Handbook of Phonological Theory* (pp. 550-569). Oxford: Blackwell.
- Schriberg, E. (1993). Four new speech and prosody-voice measures for genetics research and other studies in developmental phonological disorders. *Journal of Speech and Hearing Research*, 36, 105-140.
- Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A., & Stolke, A. (2005). Modeling prosodic features for speaker recognition. *Speech Communication*, 46, 455-472. doi:10.1016/j.specom.2005.02.018.
- Silbert, N., & de Jong, K. (2008). Focus, prosodic context, and phonological feature specification: Patterns of variation in fricative production. *Journal of the Acoustical Society of America*, 123, 2769-2779. doi:10.1121/1.2890736.
- Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C. Price, P., Pierrehumbert, J., & Hirschberg, J. (1992). TOBI: A standard for labelling English prosody. *Proceedings of the International Conference on Spoken Language Processing (ICSLP-1992)*. pp. 867-870.

- Srinivasan, R. J., & Massaro, D. W. (2003). Perceiving prosody from the face and voice: Distinguishing statements from echoic questions in English. *Language and Speech, 46*, 1-22. doi:10.1177/00238309030460010201.
- States, R. A., & Pappas, E. (2006). Precision and reliability of the Optotrak 3020 motion measurement system. *Journal of Medical Engineering and Technology, 30*, 11-16. doi:10.1080/03091900512331304556.
- Streeck, J. (1993). Gesture as communication I: Its coordination with gaze and speech. *Communication Monographs, 60*, 275-299. doi:10.1080/03637759309376314.
- Stone, M. (1997). Laboratory techniques for investigating speech articulation. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The Handbook of Phonetic Sciences, Second Edition* (pp. 9-38). Oxford, United Kingdom: Blackwell Publishing. doi:10.1002/9781444317251.ch1.
- Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America, 26*, 212-215. doi:10.1121/1.1907309.
- Summerfield, Q. (1979). Use of visual information for phonetic perception. *Phonetica, 36*, 314-331. doi: 10.1159/000259969.
- Summerfield, Q. (1992). Lipreading and audio-visual speech perception. *Philosophical Transactions: Biological Sciences, 335*, 71-78. doi:10.1098/rstb.1992.0009.
- Summers, W. V. (1987). Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analyses. *Journal of the Acoustical Society of America, 82*, 847-863. doi:10.1121/1.395284.

- Swerts, M., & Krahmer, E. (2008). Facial expression and prosodic prominence: Effects of modality and facial area. *Journal of Phonetics*, *36*, 219-238. doi:10.1016/j.wocn.2007.05.001.
- Swerts, M., & Krahmer, E. (2010). Visual prosody of newsreaders: Effects of information structure, emotional content and intended audience on facial expressions. *Journal of Phonetics*, *38*, 197-206. doi:10.1016/j.wocn.2009.10.002.
- Thomas, S. M., & Jordan, T. R. (2004). Contributions of oral and extraoral facial movement to visual and audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *30*, 873-888. doi: 10.1037/0096-1523.30.5.873.
- Tseng, C. -Y., Pin, S. -H., Lee, Y., Wang, H. -M., & Chen, Y. -C. (2005). Fluent speech prosody: Framework and modelling. *Speech Communication*, *46*, 284-309. doi:10.1121/1.423888.
- Vaissiere, J. (2004). Perception of Intonation. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 236-263). Oxford: Blackwell.
- van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, *45*, 598-607. doi:10.1016/j.neuropsychologia.2006.01.001.
- Vatikiotis-Bateson, E., Eigsti, I., Yano, S., & Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics*, *60*, 926-940. doi:10.3758/BF03211929.

- Vatikiotis-Bateson, E., & Yehia, H. C. (2000). Estimation and generalization of multimodal speech production. *Proceedings of the 2000 IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing X, 1*, pp.23-32.
- Wagner, M., & Watson, D. G. (2010). Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes, 25*, 905-945. doi:10.1080/01690961003589492.
- Walden, B. E., Prosek, R. A., Montgomery, A. A., Scherr, C. K., & Jones, C. J. (1977). Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research, 20*, 130-145.
- Walker, M. B., & Trimboli, C. (1983). The expressive function of the eye flash. *Journal of Nonverbal Behavior, 8*, 3-13. doi:10.1007/BF00986326.
- Watson, D.G. (2010). The many roads to prominence: Understanding emphasis in conversation. In B. H. Ross (Ed.), *Psychology of Learning and Motivation, Vol. 52* (pp. 163-183). Dordrecht, The Netherlands: Academic Press.
- Wells, J. C. (2006). *English Intonation: An Introduction*. Cambridge: Cambridge University Press.
- Wightman, C. (2002). ToBI or not ToBI? *Proceedings of the International Conference on Speech Prosody 2002*. pp.25-29.
- Xu, Y. (2011). Speech prosody: A methodological review. *Journal of Speech Sciences, 1*, 85-115.
- Xu, Y., & Xu, C. X. (2005). Phonetic realization of focus in English declarative intonation. *Journal of Phonetics, 33*, 159-197. doi:10.1016/j.wocn.2004.11.001.

- Yehia, H. C., Kuratate, T., & Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. *Journal of Phonetics*, 30, 555-568. doi:10.1006/jpho.2002.0165.
- Yehia, H. C., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26, 23-43. doi:10.1016/S0167-6393(98)00048-X.

APPENDICES

Appendix A. IEEE Stimulus Sentence Properties

Table A.1. IEEE stimulus sentences and associated properties. The critical constituent is indicated in italics.

Sentence #	Segmental Content	Utterance Type*	Word Count	Number of Syllables				
				Pre-Critical	Critical	Post-Critical	Total	
1	It is a band of <i>steel</i> three inches wide	L/L	9	5	1	4	10	
2	The pipe ran almost the <i>length</i> of the ditch	L/L	9	6	1	3	10	
3	It was hidden from sight by a <i>mass</i> of leaves and shrubs	L/L	12	8	1	4	13	
4	The weight of the <i>package</i> was seen on the high scale	L/E	11	4	2	6	12	
5	Wake and rise, and <i>step</i> into the green outdoors	L/E	9	4	1	6	11	
6	The green light in the <i>brown</i> box flickered	S/L	8	5	1	3	9	
7	The brass <i>tube</i> circled the high wall	S/E	7	2	1	5	8	
8	The lobes of her ears were <i>pierced</i> to hold rings	L/L	10	6	2	3	11	
9	Hold the <i>hammer</i> near the end to drive the nail	L/E	10	2	2	7	11	
10	Next <i>Sunday</i> is the twelfth of the month	L/E	8	1	2	7	10	
11	The poor boy missed the <i>boat</i> again	S/L	7	6	1	2	9	
12	The big red <i>apple</i> fell to the ground	S/E	8	3	2	4	9	
13	A <i>pink</i> shell was found on the sandy beach	L/E	9	1	1	8	10	
14	The sheep were led <i>home</i> by a dog	S/L	8	4	1	3	8	

Sentence #	Segmental Content	Utterance Type*	Word Count	Number of Syllables			
				Pre-Critical	Critical	Post-Critical	Total
15	Feed the white <i>mouse</i> some flower seeds	S/E	7	3	1	4	8
16	Both <i>brothers</i> wear the same size	S/E	6	1	2	4	7
17	<i>Two</i> blue fish swam in the tank	S/E	7	0	1	6	7
18	Nine rows of <i>soldiers</i> stood in a line	S/E	8	3	2	4	9
19	Soap can wash <i>most</i> dirt away	S/L	6	3	1	3	7
20	Clams are round, <i>small</i> , soft and tasty	S/E	7	3	1	4	8
21	We talked of the <i>sideshow</i> in the circus	L/L	8	4	2	4	10
22	Use a <i>pencil</i> to write the first draft	S/E	8	2	2	5	9
23	He ran half way to the <i>hardware</i> store	S/L	7	6	2	1	9
24	The clock struck to mark the <i>third</i> period	L/L	8	6	1	3	10
25	A <i>small</i> creek cut across the field	S/E	7	1	1	7	9
26	Cars and <i>busses</i> stalled in snow drifts	S/E	7	2	2	4	8
27	The set of china hit the <i>floor</i> with a crash	L/L	10	7	1	3	11
28	This is a grand <i>season</i> for hikes on the road	L/E	10	4	2	5	11
29	The <i>dune</i> rose from the edge of the water	L/E	9	1	1	8	10
30	Those words were the cue for the <i>actor</i> to leave	L/L	10	7	2	2	11

* S/E = Short Utterance, Early Critical Constituent

S/L = Short Utterance, Late Critical Constituent

L/E = Long Utterance, Early Critical Constituent

L/L = Long Utterance, Late Critical Constituent

Appendix B. Auditory Visual Speech Prosody Corpus

B.1. Corpus Instructions and File Naming Convention

The Auditory Visual Speech Prosody Corpus (AVSPC) contains a total of 2160 tokens, comprised of 30 sentences (Appendix A) recorded in three prosodic conditions (broad focus, narrow focus, echoic question) across two interactive conditions (face-to-face, auditory only) by six talkers, with two repetitions of each item.

For each token, several different files are provided (along with software that can be used for playback). The auditory files (Appendix B.2) for each token, normalised to a peak intensity of approximately 65 dB, are provided in .wav format and can be played back in VLC Media Player or Praat. The corresponding phonemic transcription files (Appendix B.3) are provided in .TextGrid format, and can be viewed in Praat independently, or in conjunction with the auditory file.

The shape normalised motion capture data has been provided in two different formats. The raw motion capture data (Appendix B.4) is provided in .n3d format, and can be played back by dragging and dropping the desired file into the OptoViewer program (optoviewer.exe). This program allows for changes to be made to the viewpoint of the motion capture data in three-dimensions by zooming and rotating the talkers face using mouse and keyboard commands (a readme file is included with the program). Alternatively, the motion capture files have been converted to point-light displays in .avi format (Appendix B.5), and can be played back using VLC Media Player.

Appendix B: Auditory Visual Speech Prosody Corpus

Finally, the processed version of each utterance (i.e., processed with guided principal components analysis and reprojected into component space) is included as Appendix B.8 in .avi format. These files contain both the auditory track and augmented point-light video display, along with the $F0$ contour and principal component amplitude curves over time. The segmental boundaries (i.e., pre-critical, critical and post-critical) have also been included. Note that to minimise the total file size of the corpus, these recordings have been compressed using the Cinepak Codec and down-sampled to 30 fps (from the original 60 fps).

The files share a structured naming convention consisting of a five digit number (ABCDE), followed by the file type. The naming convention is outlined in Table B.1. For example, the filename 52101.wav corresponds to the auditory token of Talker 5 producing the first repetition of sentence 10 in the narrow focused prosodic condition in the FTF interactive setting.

Appendix B: Auditory Visual Speech Prosody Corpus

Table B.1. File naming convention for the files in the Auditory Visual Speech Prosody Corpus

Filename Position	Corresponding Property	Values	
A	Talker	1	Talker 1 (MB)
		2	Talker 2 (MS)
		3	Talker 3 (WC)
		4	Talker 4 (RR)
		5	Talker 5 (EC)
		6	Talker 6 (TP)
B	Prosodic Condition and Interactive Setting	1	Broad Focus, FTF Setting
		2	Narrow Focus, FTF Setting
		3	Echoic Question, FTF Setting
		4	Broad Focus, AO Setting
		5	Narrow Focus, AO Setting
		6	Echoic Question, AO Setting
CD	Sentence	00 - 30	See Appendix A
E	Repetition	1	First Repetition
		2	Second Repetition
.filetype	File Type	.wav	Auditory Wave File
		.TextGrid	Transcription File
		.n3d	3D Motion Capture File
		.avi	Audio Video Interleaved File
		.jpeg	Image File