## Contents

*List of figures*  
page xi  
*List of tables*  
xvii  
*List of contributors*  
xviii  
*Preface*  
xxxiii  
*Acknowledgments*  
xxxvi  

**Introduction**  
1 Three puzzles of multimodal speech perception  

R. E. REMEZ  
1.1 Introduction  
4  
1.2 Organization  
5  
1.3 Event perception and speech perception  
10  
1.4 Experience  
15  
1.5 A conclusion  
20  
1.6 Acknowledgments  
20  

2 Visual speech perception  

L. E. BERNESTEIN  
2.1 Introduction  
21  
2.2 Evaluation of visemes and word homopheny  
27  
2.3 Phonetic distinctiveness of English words  
32  
2.4 Research strategies  
36  
2.5 General conclusions  
39  
2.6 Acknowledgments  
39  

3 Dynamic information for face perception  

K. LANDER AND V. BRUCE  
3.1 Introduction  
40  
3.2 Motion information for expression perception  
42  
3.3 Motion information for visual speech perception  
44  
3.4 Dynamic information for familiar face recognition  
47  
3.5 Dynamic information for unfamiliar face learning  
51  
3.6 Practical considerations  
54  
3.7 Theoretical interpretations  
55  
3.8 Future research and conclusions  
60
4 Investigating auditory-visual speech perception development 62
   D. BURNHAM AND K. SEKIYAMA
   4.1 Speech perception is auditory-visual 62
   4.2 Auditory-visual speech perception 63
   4.3 Methods for investigating development 64
   4.4 The ontogenetic development method 65
   4.5 The cross-language development method 69
   4.6 Combined methods 71
   4.7 Conclusions and an application: automatic speech recognition 73
   4.8 Acknowledgments 75
5 Brain bases for seeing speech: fMRI studies of speechreading 76
   R. CAMPBELL AND M. MACSWEENEY
   5.1 Introduction 76
   5.2 Route maps and guidelines 77
   5.3 Silent speechreading and auditory cortex 83
   5.4 Audiovisual integration: timing 92
   5.5 Speechreading: other cortical regions 94
   5.6 Speechreading in people born deaf 95
   5.7 Conclusions, directions 98
   5.8 Acknowledgments 99
   5.9 Appendix: glossary of acronyms and terms 100
6 Temporal organization of Cued Speech production 104
   D. BEAUTEMP, M.-A. CATHIARD, V. ATTINA,
   AND C. SAVARIAUX
   6.1 Introduction 104
   6.2 Overview on manual cueing 105
   6.3 First results on Cued Speech production 110
   6.4 General discussion 118
   6.5 Acknowledgments 120
7 Bimodal perception within the natural time-course of speech
   production 121
   M.-A. CATHIARD, A. VILAIN, R. LABOISSIÈRE,
   H. LOEVENBRUCK, C. SAVARIAUX, AND J.-L. SCHWARTZ
   7.1 Introduction 121
   7.2 The 2-Component-Vowel model 123
   7.3 The 2-Comp-Vowel model and visible speech 135
   7.4 The perceptual benefit of the model 146
   7.5 Conclusion and perspectives 155
   7.6 Post-scriptum 158
   7.7 Acknowledgments 158
8 Visual and audiovisual synthesis and recognition of speech by
   computers 159
   N. M. BROOKE AND S. D. SCOTT
   8.1 Overview 159
<table>
<thead>
<tr>
<th>Contents</th>
<th>ix</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.2 The historical perspective</td>
<td>161</td>
</tr>
<tr>
<td>8.3 Heads, faces, and visible speech signals</td>
<td>168</td>
</tr>
<tr>
<td>8.4 Automatic audiovisual speech processing</td>
<td>175</td>
</tr>
<tr>
<td>8.5 Assessing and perceiving audiovisual speech</td>
<td>184</td>
</tr>
<tr>
<td>8.6 Current prospects</td>
<td>189</td>
</tr>
<tr>
<td>9  Audiovisual automatic speech recognition</td>
<td>193</td>
</tr>
<tr>
<td>G. Potamianos, C. Neti, J. Luettin, and I. Matthews</td>
<td></td>
</tr>
<tr>
<td>9.1 Introduction</td>
<td>193</td>
</tr>
<tr>
<td>9.2 Visual front ends</td>
<td>197</td>
</tr>
<tr>
<td>9.3 Audiovisual integration</td>
<td>213</td>
</tr>
<tr>
<td>9.4 Audiovisual databases</td>
<td>229</td>
</tr>
<tr>
<td>9.5 Audiovisual ASR experiments</td>
<td>234</td>
</tr>
<tr>
<td>9.6 Summary and discussion</td>
<td>244</td>
</tr>
<tr>
<td>9.7 Acknowledgments</td>
<td>247</td>
</tr>
<tr>
<td>10 Image-based facial synthesis</td>
<td>248</td>
</tr>
<tr>
<td>M. Slaney and C. Bregler</td>
<td></td>
</tr>
<tr>
<td>10.1 Facial synthesis approaches</td>
<td>248</td>
</tr>
<tr>
<td>10.2 Image-based facial synthesis</td>
<td>250</td>
</tr>
<tr>
<td>10.3 Analyses and normalization</td>
<td>253</td>
</tr>
<tr>
<td>10.4 Synthesis</td>
<td>259</td>
</tr>
<tr>
<td>10.5 Alternative approaches</td>
<td>265</td>
</tr>
<tr>
<td>10.6 Conclusions</td>
<td>270</td>
</tr>
<tr>
<td>10.7 Acknowledgments</td>
<td>270</td>
</tr>
<tr>
<td>11 A trainable videorealistic speech animation system</td>
<td>271</td>
</tr>
<tr>
<td>T. Ezzat, G. Geiger, and T. Poggio</td>
<td></td>
</tr>
<tr>
<td>11.1 Overview</td>
<td>271</td>
</tr>
<tr>
<td>11.2 Background</td>
<td>272</td>
</tr>
<tr>
<td>11.3 System overview</td>
<td>275</td>
</tr>
<tr>
<td>11.4 Corpus</td>
<td>276</td>
</tr>
<tr>
<td>11.5 Pre-processing</td>
<td>277</td>
</tr>
<tr>
<td>11.6 Multidimensional morphable models</td>
<td>277</td>
</tr>
<tr>
<td>11.7 Trajectory synthesis</td>
<td>287</td>
</tr>
<tr>
<td>11.8 Post-processing</td>
<td>291</td>
</tr>
<tr>
<td>11.9 Computational issues</td>
<td>292</td>
</tr>
<tr>
<td>11.10 Evaluation</td>
<td>293</td>
</tr>
<tr>
<td>11.11 Further work</td>
<td>305</td>
</tr>
<tr>
<td>11.12 Acknowledgments</td>
<td>305</td>
</tr>
<tr>
<td>11.13 Appendix</td>
<td>306</td>
</tr>
<tr>
<td>12 Animated speech: research progress and applications</td>
<td>309</td>
</tr>
<tr>
<td>D. W. Massaro, M. M. Cohen, M. Tabain, J. Beskow, and R. Clark</td>
<td></td>
</tr>
<tr>
<td>12.1 Background</td>
<td>309</td>
</tr>
<tr>
<td>12.2 Visible speech synthesis</td>
<td>311</td>
</tr>
<tr>
<td>12.3 Illustrative experiment of evaluation testing</td>
<td>314</td>
</tr>
<tr>
<td>12.4 The use of synthetic speech and facial animation</td>
<td>317</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>12.5</td>
<td>New structures and their control</td>
</tr>
<tr>
<td>12.6</td>
<td>Reshaping the canonical head</td>
</tr>
<tr>
<td>12.7</td>
<td>Training speech articulation using dynamic 3D measurements</td>
</tr>
<tr>
<td>12.8</td>
<td>Some applications of electropalatography to speech therapy</td>
</tr>
<tr>
<td>12.9</td>
<td>Development of a speech tutor</td>
</tr>
<tr>
<td>12.10</td>
<td>Empirical studies</td>
</tr>
<tr>
<td>12.11</td>
<td>Additional potential applications</td>
</tr>
<tr>
<td>12.12</td>
<td>Acknowledgments</td>
</tr>
<tr>
<td>13</td>
<td>Empirical perceptual-motor linkage of multimodal speech</td>
</tr>
<tr>
<td>13.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>13.2</td>
<td>The perception of audiovisual speech</td>
</tr>
<tr>
<td>13.3</td>
<td>Bringing speech production to the face</td>
</tr>
<tr>
<td>13.4</td>
<td>Auditory-visual speech production</td>
</tr>
<tr>
<td>13.5</td>
<td>Correspondences of multimodal speech</td>
</tr>
<tr>
<td>13.6</td>
<td>Talking head animation</td>
</tr>
<tr>
<td>13.7</td>
<td>The importance of physical structure</td>
</tr>
<tr>
<td>13.8</td>
<td>Communicative versus cosmetic realism</td>
</tr>
<tr>
<td>13.9</td>
<td>Summary</td>
</tr>
<tr>
<td>13.10</td>
<td>Acknowledgments</td>
</tr>
<tr>
<td>14</td>
<td>Sensorimotor characteristics of speech production</td>
</tr>
<tr>
<td>14.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>14.2</td>
<td>Speech maps</td>
</tr>
<tr>
<td>14.3</td>
<td>Degrees-of-freedom in a speech task</td>
</tr>
<tr>
<td>14.4</td>
<td>Models of the underlying speech organs</td>
</tr>
<tr>
<td>14.5</td>
<td>Models of facial deformation</td>
</tr>
<tr>
<td>14.6</td>
<td>Linking articulatory degrees-of-freedom</td>
</tr>
<tr>
<td>14.7</td>
<td>Discussion</td>
</tr>
<tr>
<td>14.8</td>
<td>Conclusions</td>
</tr>
<tr>
<td>14.9</td>
<td>Acknowledgments</td>
</tr>
</tbody>
</table>

Notes: 397
References: 403
Index: 469
Figures

Figure 1.1 The relation between auditory and visual speech perception in a classic study by Sumby and Pollack (1954).

Figure 1.2 A video frame of the gamine used in a study of audiovisual intelligibility by Schwippert and Benoit (1997).

Figure 2.1 Results from the target identification task.

Figure 3.1 A functional model for face recognition (Bruce and Young 1986).

Figure 5.1 A schematic view of the left hemisphere, showing its major folds (sulci) and convolutions (gyri).

Figure 5.2 Functional organization of the cortex, lateral view, adapted from Luria (1973).

Figure 5.3 This schematic view “opens out” the superior surface of the temporal lobe.

Figure 5.4 Lateral views of the left and right hemispheres, with areas of activation indicated schematically as bounded ellipses.

Figure 5.5 Schematic showing differential activation in the lateral temporal lobes when watching non-speech actions (white-bordered grey ellipse) and watching speech actions (black-bordered ellipse).

Figure 5.6 Schematic lateral view of left hemisphere highlighting superior temporal regions.

Figure 5.7 Audiovisual binding: A role for STS.

Figure 6.1 Visible cues for English consonants, vowels, and diphthongs (from Cornett 1967).

Figure 6.2 Hand placements and hand shapes used in French.

Figure 6.3 Speech vs. lips and hand motion for the [pupøpu] sequence.

Figure 6.4 Cues for the [mabuma] sequence.

Figure 6.5 Speech vs. lips and hand motion for the [mabuma] sequence.
Figure 7.1 Sagittal contours for the three center phases of [u], [b], and [u], in the production of [ubu] (speaker J1X).

Figure 7.2 Contributions of the four main command parameters for the three center phases of [u], [b], and [u] in the production of [ubu] (speaker J1X).

Figure 7.3 Sagittal contours for the three center phases of [a], [b], and [a], in the production of [aba] (speaker J1X).

Figure 7.4 Contributions of the four main command parameters (below) for the three center phases for [a], [b], and [a], in the production of [aba] (speaker J1X).

Figure 7.5 Comparison between the modeled contour regenerated from the original [b] configuration (solid) and a simulated configuration without the compensating activity of the tongue (dashed).

Figure 7.6 Sagittal contours for [ʃ] in all combinations with [i, y, u, a].

Figure 7.7 Vertical displacement of upper lip (top traces in each figure) and lower lip (bottom traces) as a function of time during repetitive production of [bababa] by (a) an 8-month-old girl and (b) an adult (from Munhall and Jones 1998).

Figure 7.8 Acoustic signal (above) and time-course (below) of upper lip protrusion and lip area for the sentence “Tu dis: UHI ise?”

Figure 7.9 Acoustic signal (below; in abscissa: video frame numbers) and time-course of lip area (above; in cm²) for the sentence “Tu dis RUHI ise?”

Figure 7.10 Front images extracted from the sequence RUHI in: “Tu dis: RUHI ise?” with lip area measurements.

Figure 7.11 Front images extracted from the sequence “Tu dis”: “Tu dis: RUHI ise?” with lip area measurements.

Figure 7.12 Acoustic signal (above) and time-course (below) of upper lip protrusion (line with black boxes) and lip area (continuous line) for the sentence (a) “T’as dit: Hue hisse?” and (b) “T’as dit: Huis?”.

Figure 7.13 “Hue hisse” identification percentages.

Figure 7.14 Identification curves obtained by contrasting subjects (group 1) compared with non-contrasting subjects (group 2).

Figure 7.15 Identification curves obtained for “Hue hisse” (HH) and “Huis” (Huis) stimuli, with static (Stat) and dynamic (Dyn) instructions.

Figure 9.1 The main processing blocks of an audiovisual automatic speech recognizer.
List of figures

Figure 9.2 Region-of-interest extraction examples. 200
Figure 9.3 Examples of lip contour estimation by means of active shape models (Luetten et al. 1996). 201
Figure 9.4 Geometric feature approach. 206
Figure 9.5 Statistical shape model. 207
Figure 9.6 Combined shape and appearance statistical model. 209
Figure 9.7 DCT- versus AAM-based visual feature extraction for automatic speechreading. 210
Figure 9.8 Three types of feature fusion considered in this section. 217
Figure 9.9 Left: Phone-synchronous (state-asynchronous) multi-stream HMM with three states per phone in each modality. 224
Figure 9.10 Example video frames of 10 subjects from the IBM ViaVoice™ audiovisual database. 233
Figure 9.11 The audiovisual ASR system. 236
Figure 9.12 Comparison of audio-only and audiovisual ASR. 242
Figure 10.1 The range of options on the knowledge- to data-based axis of facial synthesis methods. 249
Figure 10.2 The Video Rewrite synthesis system. 251
Figure 10.3 The effects of coarticulation. 252
Figure 10.4 The masked portion of the face shown at the top is a reference image used to find the head pose. 255
Figure 10.5 EigenPoints is a linear transform that maps image brightness into control point locations. 258
Figure 10.6 The Video Rewrite synthesis process. 263
Figure 10.7 Images synthesized by Video Rewrite showing John F. Kennedy speaking (from Bregler et al. 1997b). 265
Figure 10.8 Ten of the sixteen static visemes used by the MikeTalk system. 266
Figure 10.9 Output from the Voice Puppetry system. 268
Figure 11.1 Some of the synthetic facial configurations output by the Mary101 system. 272
Figure 11.2 An overview of our videorealistic speech animation system. 276
Figure 11.3 The head, mouth, eye, and background masks used in the pre-processing and post-processing steps. 278
Figure 11.4 Twenty-four of the 46 image prototypes included in the MMM. 281
Figure 11.5 The flow reorientation process. 283
Figure 11.6 Top: Original images from our corpus. 285
Figure 11.7 Top: Analyzed α, flow parameters computed for one image. 286
Figure 11.8  Histograms for the $\alpha_1$ parameter for the /w/, /m/, /aa/, and /ow/ phones. 288
Figure 11.9  Top: The analyzed trajectory for $\alpha_{12}$ (in solid), compared with the synthesized trajectory for $\alpha_{12}$ before training (in dots) and after training (in crosses). 291
Figure 11.10  The background compositing process. 293
Figure 11.11  BACKWARD WARP algorithm. 307
Figure 11.12  FORWARD WARP algorithm. 307
Figure 12.1  Top panel shows dominance functions for lip protrusion for the phonemes in the word “stew.” 313
Figure 12.2  Viseme accuracy and confusions for natural and synthetic visual speech. 316
Figure 12.3  New palate and tongue embedded in the talking head. 320
Figure 12.4  Half of palate with velum in three different states of opening. 321
Figure 12.5  Tongue development system (see text for description). 321
Figure 12.6  Teeth and palate, showing regular quadrilateral mesh liner. 323
Figure 12.7  Voxel Space around the left jaw region, with the anterior end to the right in the picture. 324
Figure 12.8  Sagittal curve fitting. 324
Figure 12.9  Four typical ultrasound-measured tongue surfaces (for segments /a, i, N, T/) with synthetic palate and teeth, and EPG points (data from Stone and Lundberg 1996). 325
Figure 12.10  3D fit of tongue to ultrasound data. 326
Figure 12.11  EPG points on the synthetic palate. 327
Figure 12.12  Face with new palate and teeth with natural (top left) and synthetic (bottom left) EPG displays for /N/ closure. 328
Figure 12.13  Original canonical head (left), a target head (center), and the morphed canonical head (right) derived from our morphing software. 330
Figure 12.14  Speaker DWM with OPTOTRAK measurement points. 331
Figure 12.15  Illustrates placement of the points for the new model of WM, which corresponds to Baldi’s wireframe morphed into the shape of DWM. 331
Figure 13.1  Schematic representation of the four measurement domains used in our research. 350
Figure 13.2  Results of within and across domain analysis for a speaker of English (left) and Japanese (right) show the small number of correlates required to characterize multimodal speech data. 351
List of figures

Figure 13.3 A frame taken from EMG-driven animation of the nonsense utterance ['upa] during production of the stressed vowel [u]. 354
Figure 13.4 Schematic overview of kinematics-based animation. 356
Figure 13.5 Percentages of correctly identified Hiragana (syllabic) characters for Japanese sentences. 357
Figure 13.6 Simple depiction using optical flow. 358
Figure 13.7 Shown are images band-pass filtered at the two lowest spatial resolutions. 359
Figure 13.8 The original 3D position data were recorded using OPTOTRAK. 361
Figure 13.9 The cumulative contribution of ranked components (PCA) to the variance of 2700 3D face scans (300 subjects × 9 postures). 363
Figure 13.10 Multiple discriminant analysis (MDA) computed for the entire 3D face database recovers 95 percent of the variability. 363
Figure 13.11 Intelligibility results for Japanese sentences animated from the same motion data, but with different sets of postures. 366
Figure 14.1 2D and 3D biomechanical models of the tongue. 373
Figure 14.2 Independent movements of the four articulators shaping the vocal tract geometry: jaw, lips, tongue, and larynx. 375
Figure 14.3 Illustration of a functional 3D model of the tongue (from Badin and Serrurier 2006). 377
Figure 14.4 The biomechanical model of the face developed by Nazari et al. (2010). 379
Figure 14.5 Example of video image for /a/. 381
Figure 14.6 Dispersion ellipses for each original (left) and residual (right) facial and lip point (± 1 standard deviation). 382
Figure 14.7 Articulatory dof of facial speech movements (from Revéret et al. 2000). 383
Figure 14.8 Comparison of midsagittal profiles extracted from X-ray images and fitted with the midsagittal vocal tract model (Beautemps et al. 2001), for two articulations. 386
Figure 14.9 Predicting vocalic vocal tract configurations from the face. 387
Figure 14.10 Failing to predict consonantal vocal tract constriction from the face. 388
List of figures

Figure 14.11  Comparing confusion trees for vowels (left) and consonants (right).  389
Figure 14.12  Original (left) compared to recovered (right) lingual constriction in a tentative face-to-vocal tract inversion procedure.  390
Tables

Table 2.1 Phoneme equivalence classes.  
Table 4.1 Ontogenetic (amount) and cross-language (type) methods for investigating linguistic development.  
Table 9.1 Taxonomy of the audiovisual integration methods considered in this section.  
Table 9.2 The forty-four phonemes to thirteen visemes mapping considered by Neti et al. (2000).  
Table 9.3 The IBM audiovisual databases.  
Table 9.4 Comparisons of recognition performance based on various visual features.  
Table 9.5 Test set speaker-independent LVCSR audio-only and audiovisual WER (%).  
Table 9.6 Adaptation results on the speech impaired data.  
Table 10.1 Comparing animation systems.  
Table 11.1 Results from Experiment 1 “Single presentations.”  
Table 11.2 Results from Experiment 2 “Fast single presentations.”  
Table 11.3 Results from Experiment 3 “Pair presentations.”  
Table 11.4 Numbers of subjects and stimuli, and mean numbers of words, syllables, and phonemes used in Experiment 4, “Visual Speech Recognition.”  
Table 11.5 Percentage of responses and percentage correct identification of words, syllables, and phonemes for Experiment 4, “Visual speech recognition.”  
Table 12.1 The 10 facial control parameters.  
Table 12.2 The views which best illustrate which views best suit each internal viseme (a category of different phonemes that have very similar internal visible speech).  
Table 12.3 Optimal view to be chosen when direct comparisons are being made between two visemes.  
Table 13.1 Eclectic summary of findings for the analysis of multimodal speech and their causes and/or implications.  
Table 14.1 Mapping visible \( \text{dof} \) to underlying articulatory \( \text{dof} \).