International Conference on Auditory-Visual Speech Processing — AVSP'98 (Terrigal, Australia)
D. Burnham, J. Robert-Ribes & E.Vatikiotis-Bateson, Eds. - ISBN 0-643-063951

1118

Talking Heads

Philip Rubin (1) *and Eric Vatikiotis-Bateson* (2)

(1) Haskins Laboratories and Yale University School of Medicine

(2) ATR Human Information Processing Research Laboratories, Kyoto, Japan

ABSTRACT

This paper describes an interactive presentation that introduces the *Talking Heads* website, which was originally proposed at the AVSP'97 meeting in Rhodes, Greece. *Talking Heads* is an effort to bring together information from a wide range of sources. The site provides interactive access to multimodal material in both its original form and as summarized by us. In addition, the authors have provided historical information, supporting essays and tutorials, interviews, etc., that try to contextualize and make coherent this rapidly developing area. Both the website and the interactive presentation are described.

1. INTRODUCTION

The present project was proposed during the discussion at the AVSP'97 meeting in Rhodes, Greece. This paper describes the interactive presentation that formally introduces the *Talking Heads* website. As opposed to a more conventional form of collation such as mere print, the site provides interactive access to multimodal material in both its original form and as summarized by us. In the interactive presentation, issues relevant to a number of domains are presented and discussed. For example, in speech communication, some of the important research issues are auditory-visual (AV) integration, the interrelation of facial motion, acoustics, and vocal-tract behavior, and the relations between AV speech production and perception. In telecommunications, issues of transmission bandwidth (facial coding), realism (cosmetic and communicative), and utility (Do people really want videophones?) are important. The entertainment industry has had a somewhat different set of constraints and goals, and has in many cases worked out practical solutions to the problems being addressed by speech and telecommunications researchers. Yet, this industry is poised to reap the benefits of and to support the research and development achieved in these other areas. Finally, based on the assessment of the different domains, we summarize what these areas have to offer one another and present our view of what issues, pertinent to all areas, need to be addressed in the immediate future.

2. BEING BIOLOGICAL

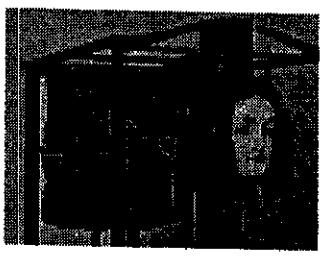
The desire to create talking human heads stems back at least several hundred years. This quest presently combines approaches that are computational, cognitive, and biological, and cuts across a wide variety of domains and interests. The *Talking Heads* website provides a brief overview of these areas and attempts to convey some of the excitement that has recently spurred so much interest and international research collaboration.

Of considerable importance to a variety of researchers is the need to model aspects of talking heads based upon an understanding of human physiology. This approach entails research in articulatory synthesis; models of speech motor control; the simulation and modeling of different articulator systems, including the lips, jaw, and tongue; and facial animation. Additional input comes from considerations of human-machine interface design. The applications of this work are also diverse, ranging from research tools for working with the hearing impaired to the creation of and improvements in avatars that are beginning to populate cyberspace, and virtual humans that are already beginning to appear in film, TV, and video games.

The Talking Heads website can be found at:

`www.haskins.yale.edu/haskins/heads.html`

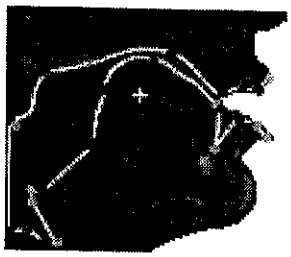
3. SIMULACRA: The Early History of Talking Machines



"The earliest speaking machines were perceived as the heretical works of magicians and thus as attempts to defy god. In the thirteenth century the philosopher Albertus Magnus is said to have created a head that could talk, only to see it destroyed by St. Thomas Aquinas, a former student of his, as an abomination. The English scientist-monk Roger Bacon seems to have produced one as well. That fakes were appearing in Europe in the late sixteenth and early seventeenth centuries is shown by Miguel de Cervantes' description of a head that spoke to Don Quixote -- with the help of a tube that led to the floor below. Like Magnus, this fictitious inventor also feared the judgement of religious authorities, though in his case he took it upon himslef to destroy the heresy. By the eighteenth century, science had started to shed its connection to magic, and the problem of artificial speech was taken up by inventors of a more mechanical bent." David Lindsay, "Talking Head," *Invention & Technology*, Summer 1997, 57-63. (1)

This section of the website describes some of these mechanical talking machines and also some early electrical talking devices. Included in this survey are: the talking machine of Fredrick von Knaus that was demonstrated around 1770 before the Austrian emperor and the duke of Tuscany (1); Kratzenstein's resonators for the synthesis of vowel sounds (1779) (2); Von Kempelen's talking machine (1791) (2,3); the Euphonia: Joseph Faber's amazing talking machine (1830-40's) (1); Charles Wheatsone's refinements of Von Kempelen's talking machine (late 1800's) (4); Alexander and Melville Graham Bell's physical working model of the human vocal tract (5); Erasmus Darwin's speech synthesizer; R. R. Riesz's talking mechanism (1937) (2,3); J. Q. Stewart's electrical analog of the vocal organs (1922) (2); the Voder, which was introduced at the 1939 New York World's Fair (3); the Haskins Laboratories Pattern Playback (1950) (6); and Dunn's electrical vocal tract (1950) (7).

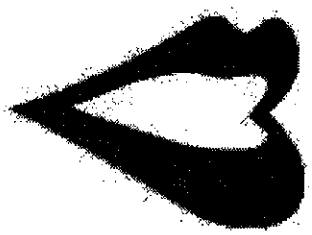
4. VIRTUAL VOCAL TRACTS



Mechanical talking devices were rapidly replaced by electronic speech simulators (known as speech synthesizers) as technological developments made experimentation more affordable. Models of the human vocal tract could now be created in the form of electrical circuits, leading to the creation of dedicated hardware speech synthesizers. A revolution occurred in speech technology when the digital computer permitted the simulation of electronic circuitry, the conversion of analog signals to digital form, and the creation of analog signals from digital information (in this case, sound in the form of speech). The advent of desktop computing in the 1980s and 1990s brought affordable speech synthesis and recognition within the reach of the average computer user.

This section provides additional information about the following topics: hardware synthesizers (OVE, etc.); digital synthesizers, formant and LPC-based software synthesis; inexpensive off-the-shelf, commercial synthesizers; synthesis-by-rule and text-to-speech synthesizers. In addition, there is an interactive demonstration of the Haskins articulatory synthesis system (*ASY*) in which speech acoustics are synthesized by controlling vocal-tract articulators such as the jaw, tongue, lips, and velum (8).

5. ARTICULATORS



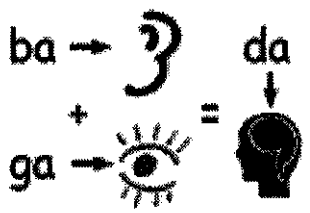
This section of the website provides details about state-of-the-art work going on around the world that attempts to model the complex physical and functional attributes of speech articulation (9-15). This section includes interviews with some of those who have worked in these areas including the late Christian Benoît, Lionel Revéret, Sumit Basu, David Ostry and others. In addition, there are links to sites containing more detailed information on the modeling of articulatory speech production.

6. MEASURING AND MODELING SPEECH PRODUCTION

In human communication, the speech system is specialized for the rapid transfer of information (16, 17). Significant events in the acoustic signal can occur in an overlapped or parallel fashion due to the physical coproduction of articulatory events that influence the acoustics. A result of this is that aspects of the signal corresponding to different linguistic units, such as consonants and vowels, often cannot be isolated in the acoustic stream. One way to help tease apart the components of the speech signal is to consider the physical system that gives rise to the acoustic information: the acoustic encoding of phonetic information is then viewed in light of the flexibility inherent in the production apparatus, particularly the human supralaryngeal vocal tract, in which individual articulators or groups of articulators can function semi-independently.

In this section of the site we review this approach (18). First, we show how the analysis of speech acoustics has benefited by treating the sound production system as one in which the contributions of physical acoustic sources and physiologically-determined filters are combined. We then discuss how acoustic diversity has resulted while achieving a high degree of articulatory simplicity. In the process, we review various methods that are used to examine articulatory activity, and also describe in detail the Haskins Laboratories lingustic/gestural approach to modeling the coordination of the speech articulators. We close by considering several recent efforts to explore the links between production, perception, and acoustics. Where possible, recent trends in the field have been exemplified by projects involving ourselves and our colleagues.

7. THE MCGURK EFFECT



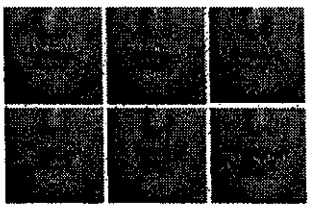
The McGurk effect has played an important role in audio-visual speech integration and speech reading. A recent paper by Massaro and Stork (19) describes the effect:

"The most striking demonstration of the combined (bimodal) nature of speech understanding appeared by accident. Harry McGurk, a senior developmental psychologist at the University of Surrey in England, and his research assistant John MacDonald were studying how infants perceive speech during different periods of development. For example, they placed a videotape of a mother talking in one location while the sound of her voice played in another. For some reason, they asked their recording technician to create a videotape with the audio syllable "ba" dubbed onto a visual "ga." When they played the tape, McGurk and McDonald perceived "da." Confusion reigned until they realized that "da" resulted from a quirk in human perception, not an error on the technician's part. After testing children and adults with the dubbed tape, the psychologists reported this phenomenon in a 1976 paper humorously titled 'Hearing Lips and Seeing Voices,' (20) a landmark in the field of human sensory integration. This audio-visual illusion has become known as the McGurk effect or McGurk illusion."

D. W. Massaro & D.G. Stork, "Speech recognition and sensory integration." *American Scientist*, 86, 1998, 236-244.

Related links are provided in this section of the website.

8. SPEECHREADING



Pointers to research and references on speechreading (21, 22, 23) can be found at the pioneering UCSC PSL Speechreading (lipreading) webpage:

`mambo.ucsc.edu/psl/lipr.html`.

Additional information about speechreading is available at a number of sites, including:

• Chris Bregler's Computer Lipreading page:

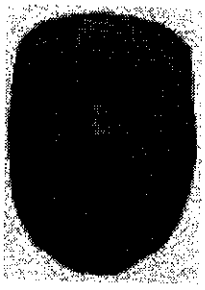
`www.cs.berkeley.edu/~bregler/vaspeech.html`

•NATO Advanced Study Institute: Speechreading by Man and Machine: `www.crc.ricoh.com/asi/` ; and

• Juergen Luettin's Speechreading Demo (Machine Vision Group at IDIAP):

`www.idiap.ch/cgi-bin/w3-msql/vision/visionNF.html`

9. FACIAL ANIMATION



This section features an interactive version of a recent research paper entitled *Kinematics-Based Synthesis of Realistic Talking Faces* by Eric Vatikiotis-Bateson, Takaaki Kuratate, Mark Tiede, and Hani Yehia, of the ATR Human Information Processing Research Laboratories (24). This paper, and other subsequent work, presents a method for animating talking faces that reaches towards cosmetic and communicative realism. The animations are derived from multilinear analyses of 3D facial motion data comprising a small set of time-varying positions measured on the face. Frame rates can be set at the video field rate or at much slower rates using via point analysis to extract key frame configurations for subsequent interpolation. The resulting animations are dynamically and cosmetically realistic to the extent that texture maps specific to the speaker are used in the animation and the audio and video components are perfectly synchronized.

Related analyses have shown that the facial motion parameters used in the animations are highly correlated with the measurable acoustic, vocal-tract, and neuromuscular events associated with speech production (25). Because of this strong interrelation of facial and other speech production parameters, the easily-obtained kinematic control parameters can be used to examine the nature and distribution of phonetically relevant facial information relative to changes in vocal tract configuration (e.g., tongue and jaw position) and the acoustics.

The animation system can also be used to examine the relations between speech production and perception and, along the way, to assess the communicative realism of the synthesized talking faces. A brief overview is given of our current use of the animations in studies of perception and correlated hemodynamic activity in the brain using functional magnetic resonances imaging (fMRI).

This section of the website also contains links to other facial animation sites. Of particular note is the UCSC PSL Facial Animation webpage:

`mambo.ucsc.edu/psl/lipr.html`

which provides an extensive overview of facial animation and includes links related to the pioneering work of Parke (26, 27, 28), Waters (28, 29), and Cohen and Massaro (30, 31, 32). Other sites include:

- KTH Multimodal Synthesis in Stockholm:

  `www.speech.kth.se/multimodal/` ;

- ICP Facial Synthesis (33, 34) in Grenoble:

  `ophale.icp.inpg.fr/2.6.html` ;

- MikeTalk (Ezzat and Poggio) (35) at MIT:

  `http://cuneus.ai.mit.edu:8000/research/miketalk/miketalk.html` ;

- Video Rewrite: Driving Visual Speech with Audio (Bregler, Covell & Slaney, Interval Research Corp.) (36):

  `web.interval.com/papers/1997-012/` ;

and many others.

10. AVATARS

The recent advances in animation, computer graphics, fast microprocessors, human-machine interface design, motion capture, physiological modeling, and the growth of the world wide web are all factors that have influenced the growing interest in developing avatars/virtual humans: computer-generated representations of entities with personalities and the ability to communicate (37, 38, 39, 40, 41). In the future, will interaction and communication with virtual humans be commonplace in the workplace, community and home? Will virtual personalities fundamentally change entertainment media? The exciting international work in the nascent research area of audio-visual speech will have a profound impact on this future.

This section of the site includes links to just a portion of this burgeoning volume of work and, as with the other sections of the site, is being constantly expanded as new developments are reported and made available.

11. REFERENCES

1. Lindsay, D., Talking Head. (1997). *Invention & Technology*, Summer 1997, 57-63.
2. Cater, John P. (1983). *Electronically Speaking: Computer Speech Generation*. Howard W.Sams & Co., Inc., Indianapolis, Indiana.
3. Flanagan, J.L. (1965). *Speech Analysis Synthesis and Perception*. Springer-Verlag, New York.
4. Wheatstone, Charles. (1879). *Scientific Papers of Sir Charles Wheatstone*. Physical Society of London, London.
5. Bell, A. G. (1922). Prehistoric Telephone Days. *National Geographic Magazine*, 41, 223-242.
6. Cooper, F.S., Liberman, A.M. & Borst, J.M. (1951). The interconversion of audible and visible patterns as a basis for research in the perception of speech. *Proceedings of the National Academy of Science*, 37, 318-325.
7. Dunn, H. K. (1950). The calculation of vowel resonances, and an electrical vocal tract. *Journal of the Acoustical Society of America*, 22, 740-753.
8. Rubin, P. E., Baer, T., & Mermelstein, P. (1981) An articulatory synthesizer for perceptual research, *Journal of the Acoustical Society of America*, 70, 321-328.
9. Basu, S., Oliver, N., & Pentland, A. (1998). 3D modeling and tracking of human lips. *Proc. Internat. Conf. on Computer Vision*, Bombay, 4-7 Jan. 1998 (IEEE Computer Society Press, Los Alamitos), 337-343.
10. Ostry, D.J., Vatikiotis-Bateson, E., & Gribble, P.L. (1997). An examination of the degrees of freedom of human jaw motion in speech and mastication. *Journal of Speech and Hearing Research*, 40, 1341-1351.
11. Ostry, D. J., & Munhall, K. G. (1994). Control of jaw orientation and position inmastication and speech. *Journal of Neurophysiology*, 71, 1515-1532.
12. Perkell, J., Cohen, M., Svirsky, M., Mathies, M., Garabieta, I., & Jackson, M. (1992). Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *Journal of the Acoustical Society of America*, 92, 3078-3096.
13. Perrier, P., Ostry, D.J., and Laboissiere, R. (1996). The equilibrium-point hypothesis and its application to speech motor control. *Journal of Speech and Hearing Research*, 39, 365-377.
14. Revéret, L. (1997). From raw images of the lips to articulatory parameters : A viseme-based prediction. Proceedings of the Fifth EUROSPEECH Conference, Rhodes, Greece, Sept. 22-25, 1997, vol. 4, 2011-2014.
15. Wilhelms-Tricarico, R (1995). Physiological modeling of speech production: methods for modeling soft-tissue articulators. *Journal of the Acoustical Society of America* 97, 3085-3098
16. Liberman, A.M. & Mattingly, I.G. (1985). The motor theory of speech perception revised. *Cognition* 21:1-36.
17. Mattingly, I.G. & Liberman, A.M. (1988). Specialized perceiving systems for speech and other biologically significant sounds. In: G. M. Edelman, W. E. Gall, & M.E. Cowan (Eds.), *Auditory function: the neurobiological bases of hearing*. Wiley, New York, 775-793.
18. Rubin, P. E. & Vatikiotis-Bateson, E. (1998). Measuring and modeling speech production. In : S.L. Hopp, M.J. Owren, & C.S. Evans (Eds.), *Animal Acoustic Communication*. Springer-Verlag.
19. Massaro, D. W. & Stork, D. G. (1998). Speech recognition and sensory integration. *American Scientist*, 86, 236-244.
20. McGurk, H. & MacDonald, J. W. (1976). Hearing lips and seeing voices. *Nature*, 264, 746-748.
21. Summerfield, A. Q. (1979). Use of visual information in phonetic perception. *Phonetica*, 36, 314-331.
22. Stork, D.G. & Hennecke, M.E. (Eds.). (1996). *Speechreading by Humans and Machines: Models, Systems, and Applications*, Springer-Verlag, New York.
23. Brooke, N. M. (1996). Talking Heads and Speech Recognisers That Can See: The Computer Processing of Visual Speech Signals. In D. G. Stork & M. E. Hennecke (Eds.), *Speechreading by Humans and Machines: Models, Systems, and Applications*, Springer-Verlag, New York, 351-371.
24. Vatikiotis-Bateson, E., Kuratate, T., Tiede, M. K., & Yehia, H C. (1998). Kinematics-based synthesis of realistic talking faces. Technical Report TR-H-237, ATR-HIP.
25. Yehia, H., Rubin, P. , & Vatikiotis-Bateson. (In press). Quantitative association of vocal-tract and facial behavior. *Speech Communication*.
26. Parke, F. I. (1975). A model for human faces that allows speech synchronized animation. *Journal of Computers and Graphics*, 1(1), 1-4.
27. Parke, F. I. (1982). Parameterized models for facial animation, *IEEE Computer Graphics*, 2(9), 61-68.
28. Parke, F. I. & Waters, K. (1994). *Computer Facial Animation*, AK Peters. ISBN 1-56881-014-8.
29. Waters, K. (1986). Expressive three-dimensional facial animation. *Computer Animation* (CG86), October, 49-56.
30. Massaro, D. W. (1987). *Speech perception by ear and by eye: A paradigm for psychological enquiry*. Hillsdale, NJ: Lawrence Erlbaum Associates.

31. Massaro, D. W. (1996). Bimodal Speech Perception: A Progress Report. In D. G. Stork & M. E. Hennecke (Eds.), *Speechreading by Humans and Machines: Models, Systems, and Applications*, Springer-Verlag, New York, 79-101.

32. Cohen, M. & Massaro, D., (1990). Synthesis of visible speech. *Behavior Research Methods: Instruments & Computers*, 22, 260-263.

33. Bailly, G. & C. Benoît, C., (Eds.). (1992) *Talking machines: Theories, synthesis. Theories, Models, and Designs*, Amsterdam: North Holland.

34. Benoît, C., Lallouache, T., Mohamadi, T., & Abry, C. (1992). A set of French visemes for visual speech synthesis. In G. Bailly & C. Benoît, (Eds.), *Talking machines: Theories, Models, and Designs*, Amsterdam: North Holland, 485-504.

35. Ezzat, T. & Poggio, T. (1998). MikeTalk: A Talking Facial Display Based on Morphing Visemes. *Proceedings of the Computer Animation Conference*, Philadelphia, PA, June 1998.

36. Bregler, C., Covell, M. & Slaney, M. Video Rewrite: Driving Visual Speech with Audio.

37. Damer, Bruce. *Avatars!* (1998). Peachpit Press, Berkeley, CA.

38. Magnenat-Thalmann, N. & Thalmann, D. (1995). Digital actors for interactive television, *Proceedings of the IEEE*, August 1995.

39. Noser, H., Renault, O., Thalmann, D. & Magnenat Thalmann, N. (1995). *Navigation for Digital Actors based on Synthetic Vision,Memory and Learning, Computers and Graphics*, Pergamon Press.

40. Thalmann, N. & Thalmann, D. (1990) Synthetic Actors in Computer-Generated 3D Films. Tokyo: Springer-Verlag.

41. Thalmann, N. & Thalmann, D. (1991). Complex models for animating synthetic actors, *IEEE Computer Graphics and Applications*, 11, 5, 32-44.