5-2 Spoken Communication in Multiple Modalities

Eric Vatikiotis-Bateson
Advanced Telecommunications Research Institute International

This article provides a brief overview of the recent activities of ATR International's Communication Dynamics Project. With the support of CRL, a multi-purpose research program has been established. The central theme of the project is to examine human communicative behavior as it occurs naturally: in multiple modalities and in complex environments. The forms of behavior being examined include both verbal communication and non-verbal gestures and expressions. A major research goal of the project is to achieve a better understanding of the link between the production and perception of these multi-modal behaviors. Computational models as they are developed will be applied to development of communicatively plausible systems for human-machine interaction.

Keywords
Auditory-visual speech processing, multi-modal communication, talking head animation, multi-modal perception and production, communicative gesture and expression

Introduction

Basic science and technology tend to have distinct interests, however they need not follow separate paths in achieving their goals. In our project with CRL, the ATR-I Communication Dynamics Project has developed a research program that serves these multiple purposes. The aim of the project is to understand how naturally occurring communicative events such as expressive speech, emotion, and gestures are produced, perceived, and processed neurally. Communication occurs in multiple modalities, principally the auditory and visual modalities, and in complex environments. Therefore, a major challenge is the analysis and synthesis of multi-modal speech behavior as it occurs naturally — integrated with emotional expression in multi-talker environments. Non-speech expressions of emotion and other gestural forms of communication are also being examined, as are basic visual processes such as the detection and representation of three-dimensional objects in interactive and changing environments. As computational models of communicative phenomena emerge, they will be applied to developing communicatively plausible human-machine systems.

1 Auditory-Visual Speech Processing

Although traditionally examined separately, the production and perception of communicative behavior co-constrain each other and therefore should be examined together. In order to link production and perception, we have devised a three-stage research methodology consisting of (1) empirical studies that afford a basic understanding of how multi-modal communicative behavior is produced and structured, (2) a talking head animation system whose parameters are under experimenter control and derive directly from time-varying and static production data and their

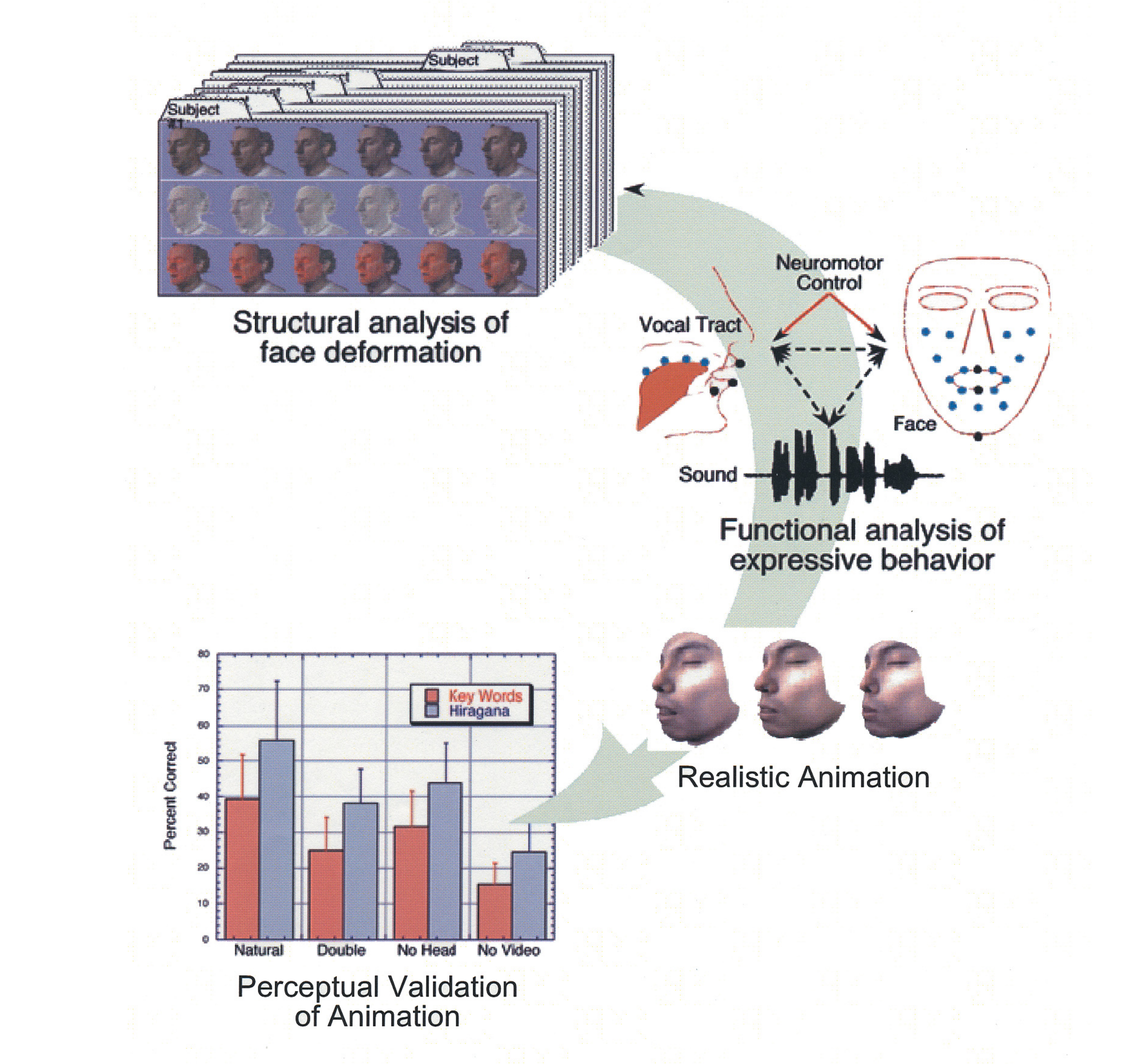

Fig.1 *Structural and functional data for many speakers are used to generate realistic animations of speaking and expressive faces. The animations are then used to generate stimuli for studies of multi-modal perception, neural processing (e.g., fMRI), and the links between production and perception.*

analyses, and (3) perception studies that use the talking head animations as stimuli to validate the realism of the animation system and to assess the relevance of the analytic production results for perception. The figure shows the general scheme of our research paradigm.

2 Measurement and Analysis of Multi-modal Speech Behavior

Time-varying production data are measured and analyzed at as many levels of observation as possible. These include the speech acoustics, motions of the vocal tract, face, and head, and the EMG (electromyographic) activity of associated muscles of the face and vocal tract. Having shown that simple aspects of the experimental protocol can severely effect the quality of the results, new recording and measurement techniques have been developed that enable more natural recordings of spontaneous communicative interactions between two or more people. An important consequence of this is the development of a non-invasive video recording and image analysis technique that provides reliable measures of the motions of the face and head most relevant to communication [Kroos et al., in press].

Analyses consist of multi-linear and nonlinear techniques for identifying and characterizing the patterns of time-varying events within and across the different measurement domains. For example, face motion and speech acoustics can both be reliably estimated from measures of the vocal tract or from the muscle activity that controls speech articulation. Such estimation has been used to create a computational model of multi-modal speech production that provides a plausible and fairly comprehensive characterization of the neuromotor control system and the physical structures [Vatikiotis-Bateson et al., 2000a]. More specific to speech coding, face and head motion can be recovered from the spectral and source properties of the acoustics [Yehia et al, 1998; 1999; in press]. Taken together, these results have shown that linguistic events are redundantly specified in the visual and auditory domains and that the two domains share a common control source for speech production. These findings provide a causal basis for understanding the long-standing psychophysical finding that being able to see a speaker's face can enhance speech intelligibility [e.g., Sumby & Pollack, 1954]. They also provide a starting point for examining the interaction and integration of speech and non-speech events, such as expressions of emotion and discourse markers (e.g., for 'turn-taking'). These are events that occur simultaneously on a speaker's face and must be processed in parallel by the perceiver.

3 Structural Analysis of the Face and Head

In conjunction with the time-varying behavioral data, a large structural database is under construction. Sets of facial postures are being recorded with three-dimensional (3D) cylindrical scans of the face and head. The posture sets are intended to cover the extremes of face configuration during production of speech and other expressive behaviors, and eventually the face database will contain posture sets for 300 subjects (Japanese and non-Japanese). Analysis of the posture sets provides the face deformation parameters used in the talking head animation system. In addition, the database is being used to compute multi-dimensional scales for comparing structural (basic 3D morphology) and functional (postural) typing of faces. One use of the database is to examine the co-dependence of static structure on time-varying function. This is done by generating talking head animations where the animated face of one subject is contrasted with animations of other subjects whose face structure is of known dissimilarity (along some dimension of variability). Alternatively, animations can be constructed and compared psychophysically when the structural and functional data components are from different subjects of known dissimilarity. Finally, the posture sets include both speech and non-speech configurations. The deformation coefficients derived from these sets can be used to animate and evaluate speech produced with and without accompanying nonspeech expressions (of emotion et cetera).

4 Talking Head Animation System

Two types of talking head animation have been developed for synthesizing multi-modal behavior from production parameters. One is a kinematics-based system that combines time-varying face motion and face deformation parameters derived from multi-posture 3D scans of the subject's face. The face motion data can be either measured directly or estimated from muscle EMG or acoustic signals. Once parameterized, this system can generate video-realistic sequences of faces having good spatial and temporal resolution at near real time [Kuratate et al., 1998, submitted]. The second system was developed by collaborators in Canada [Lucero & Munhall, 1999], and is a physical model containing musculo-skeletal structures and a multi-layer model (mass-spring) of the facial skin. This system is based on the muscle-based face animation model developed by Waters [Waters, 1987] in collaboration with Terzopoulos [Terzopoulos & Waters, 1990]. It was modified to generate animations from time-varying EMG signals for 7 orofacial muscles. The physical model is more computation-intensive and difficult to control with fine temporal detail, but its physical skin provides more realistic rendering of skin deformations, particularly the areas surrounding the mouth, than the affine skin mesh used in the kinematics-based method developed at ATR [for details, see Vatikiotis-Bateson et al., 2000b]. Therefore, the kinematics-based method is being modified to incorporate a physical skin model.

5 Perceptual Evaluation of Multi-Modal Speech Behavior

Before production-based talking head animations can be used to assess the linkage between the perception and production of communicative expressions, they must be validated perceptually and kinematically. It must be shown that the talking heads make the same motion and convey the same sort of linguistic information as real faces. Both animation models used in our work are ‘feed-forward’, so their movement accuracy had to be verified independently [Vatikiotis-Bateson et al, 2000b]. The perceptual validation process has revealed interesting facts about the conditions under which perceivers’ are willing to suspend their disbelief and the effect this has on perception. The basic finding is that imperfections in video-realistic faces have more profound effects on perceived realism than do animated line-drawings or animal caricatures [Kuratate et al., submitted].

Numerous perception studies have now been conducted with Japanese and English speaking subjects in which the animations are evaluated under various auditory and visual conditions. The face animations have been compared to point-light displays, video of natural speaking faces, with and without head motion, and under various conditions where production control parameters are manipulated systematically (e.g., scaling the amount of head motion). Overall, the animations have been shown to elicit the same patterns of response as normal video presentations, despite the absence of eyes, teeth, and tongue in the animated faces. Furthermore, the strength of the response is sufficient to allow the effects of manipulating the model control parameters to be evaluated reliably. Thus, use of production-based stimuli to test perception has confirmed that head motion, observed to be correlated with the voice source [Yehia et al., in press], can convey important visual information about linguistic prosody when the voice source is masked with acoustic noise [Munhall et al., in preparation]. This test also exemplifies our intent to use our control of the production parameters to evaluate their role in perception.

6 Expanding the Horizon

Now that the research methodology is established for analysis, synthesis, and evaluation for speech behavior, we are expanding its scope in three directions: 1) to incorporate non-speech gestures and emotion; 2) to investigate brain responses to multi-modal stimuli [Callan et al., 2001] and compare them with psychophysical studies of perception [Munhall et al., submitted]; and 3) to examine our ability to adapt in navigating complex environments. These are but a few of the myriad possibilities for expansion.

7 Non-Speech Expression

Despite the overwhelming focus of previous research on static faces, be they scripted caricatures of emotional expression [e.g., Ekman et al., 1972] or familiar faces to be identified [Bruce and Young, 1986], faces provide information through time, seemingly just like everything else associated with human information processing. For example, simple animation of emotional extremes (based on morphing between neutral and extreme postures) at different transition speeds (velocity) suggests that emotions may inhabit different kinematic ranges. For example, smiles are quicker than frowns and are more reliably perceived when presented at the appropriate speed [Kamachi et al., 2001]. Our system allows these phenomena to be examined in detail and in more realistic situations where spontaneous, rather than scripted, expressions can co-occur with speech.

8 Brain Responses to Multi-Modal Stimuli

An emergent issue in studies of brain function (e.g., fMRI, MEG) is that brain responses to stimuli do not always match the results of behavior studies. For example, human infants show a developmental asynchrony in their ability to discriminate non-native sound categories and to register those differences in the activity of auditory centers of the brain. Specifically, in the first year of life, children's cognitive and neural processes develop at different rates, and it is only after about a year that auditory and categorical discrimination of speech sounds coincide [Rivera-Gaxiola et al., 2000]. In adults, we do not expect such startling differences between cognitive and neural behavior; however, multi-modal communication invokes a wide range of motor, auditory, visual, and even tactile processes. On the other hand, the redundancy across modalities observed behaviorally for audiovisual speech communication suggests that cognitive processes may be less affected by the loss of information in one or another modality. We are now conducting comparative cognitive and brain function studies that assess the degree of redundancy between neural and cognitive processes, and the mechanisms the rapid adaptation to accommodate sudden changes (e.g., in modality) of the perceived information stream.

9 Adaptation and Navigation of Complex Environments

Generally when humans interact with one another or their environment, they must retrieve information under less-than-ideal conditions because the information and/or its context are either novel or degraded in some way. This requires that information retrieval be flexibly adaptive and sensitive to redundancy. The work on auditory-visual speech processing already has a strong focus on characterizing the redundancy of communicative information over different modalities. However, a more basic and probably related issue is to understand how we retrieve the information used to characterize and navigate our inanimate environment. Humans see three-dimensional objects in the real world "actively". Thus, when the information obtained about some object is not sufficient for recognition, we change the parameters of the retrieval process by changing the viewpoint. Structural ambiguity is minimized through iterative manipulation of the of the camera pose using a viewpoint control system implemented into a vision-based 3D recognition system. Experiments with this system have shown that the total accuracy of object detection can be improved by as much as four times after only several iterations of the pose correction process [Kinoshita & Tonko, in press].

10 Summary

It is impossible in a few short pages to discuss adequately the range of possible issues and potential applications that can be addressed within the research framework of multi-modal communication described here. Indeed, the rapid rate of development of tools and applications makes predicting even the near future almost impossible. For example, we can now produce videophonic speech signals at low bit rates, and soon we should be able to implement control systems for simple interactive communication into physical robotic devices [Kozima & Vatikiotis-Bateson, 2001]. Our greatest hope, however, is that development of such systems can help us understand the neural and cognitive bases of multi-modal communication.

References

1 Bruce, V. and Young, A. W., (1986), "Understanding face recognition", British Journal of Psychology, 77, 305-327.

2 Callan, D. E., Callan, A. M., Kroos, C., and Vatikiotis-Bateson, E., (2001), "Multimodal contribution to speech perception revealed by independent component analysis: A single-sweep EEG case study", Cognitive Brain Research 349-353.

3 Ekman, P., Friesen, W. V., and Ellsworth, P., (1972), "Emotion in the human face: Guidelines for research and a review of findings", New York: Pergamon Press.

4 Kamachi, M., Bruce, V., Mukaida, S., Gyoba, J., Yoshikawa, S., and Akamatsu, S., (2001), "Dynamic properties influence the perception of facial expressions", Perception, 30, 875-887.

5 Kinoshita, K. and Tonko, M., (in press), "3D accuracy improvement using an uncalibrated image (in Japanese)", IEICE Transactions D-II.

6 Kozima, H. and Vatikiotis-Bateson, E., (2001), "Communicative criteria for processing time/space-varying information", In P. Coiffet (Ed), 10th IEEE International Workshop on Robot and Human Communication (ROMAN 2001), (pp. 377-382), Bordeaux-Paris.

7 Kroos, C., Kuratate, T., and Vatikiotis-Bateson, E., (in press), "Video-based face motion measurement", Journal of Phonetics.

8 Kuratate, T., Yehia, H., and Vatikiotis-Bateson, E., (1998), "Kinematics-based synthesis of realistic talking faces", In D. Burnham, J. Robert-Ribes and E. Vatikiotis-Bateson (Eds.), International Conference on Auditory-Visual Speech Processing (AVSP'98), (pp. 185-190), Terrigal-Sydney, Australia.

9 Kuratate, T., Vatikiotis-Bateson, E., and Yehia, H. C., (submitted), "Talking faces synthesized by facial motion mapping", Speech Communication.

10 Lucero, J. C. and Munhall, K. G., (1999), "A model of facial biomechanics for speech production", Journal of the Acoustical Society of America, 106, 2834-2842.

11 Munhall, K. G., Kroos, C., and Vatikiotis-Bateson, E., (submitted), "Spatial frequency requirements for audiovisual speech perception", Perception & Psychophysics.

12 Rivera-Gaxiola, M., Csibra, G., Johnson, M. H., and Karmiloff-Smith, A., (2000), "Electrophysiological correlates of cross-linguistic speech perception in native English speakers", Behavioural Brain Research, 111, 13-23.

13 Sumby, W. H. and Pollack, I., (1954), "Visual contribution to speech intelligibility in noise", Journal of the Acoustical Society of America, 26, 212-215.

14 Terzopoulos, D. and Waters, K., (1990), "Physically-based facial modeling, analysis, and animation", Visualization and Computer Animation, 1, 73-80.

15 Vatikiotis-Bateson, E., Kuratate, T., Munhall, K. G., and Yehia, H. C., (2000), "The production and perception of a realistic talking face", In O. Fujimura, B. D. Joseph and B. Palek (Eds.), Proceedings of LP'98, Item order in language and speech (pp. 439-460), Prague: Karolinum Press (Charles University).

16 Vatikiotis-Bateson, E., Kroos, C., Kuratate, T., Munhall, K. G., and Pitermann, M., (2000), "Task constraints on robot realism: The case of talking heads", In K. Kamejima (Ed), 9th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN 2000), (pp. 352-357), Osaka, Japan.

17 Waters, K., (1987), "A muscle model for animating three-dimensional facial expression", Computer Graphics, 22, 17-24.

18 Yehia, H. C., Rubin, P. E., and Vatikiotis-Bateson, E., (1998), "Quantitative association of vocal-tract and facial behavior", Speech Communication, 26, 23-44.

19 Yehia, H. C., Kuratate, T., and Vatikiotis-Bateson, E., (1999), "Using speech acoustics to drive facial motion", In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville and A. C. Bailey (Eds.), Proceedings of the 14th International Congress of Phonetic Sciences, (pp. 631-634), San Francisco, CA.

20 Yehia, H. C., Kuratate, T., and Vatikiotis-Bateson, E., (in press), "Linking facial animation, head motion, and speech acoustics", Journal of Phonetics.

***Eric Vatikiotis-Bateson**, Ph. D.*

Project Leader, Communication Dynamics Project, Information Sciences Division, ATR International