ELSEVIER

Speech Communication 26 (1998) 1–4

SPEECH COMMUNICATION

Editorial

P. Rubin [a,b], E. Vatikiotis-Bateson [c]

[a] *Haskins Laboratories, 270 Crown St., New Haven, CT 06511, USA*
[b] *Yale University School of Medicine, Department of Surgery, Otolaryngology, New Haven, CT, USA*
[c] *ATR Human Information Processing Research Laboratories, 2-2 Hikaridai, Seika-cho, Soraka-gun, Kyoto 619-0288, Japan*

The interest in multimodal speech processes is by no means new. Sumby and Pollack (1954) provided the classic demonstration that visual information improves the intelligibility of speech degraded by acoustic noise. Two decades later, McGurk and MacDonald's description of the effect that now bears Harry McGurk's name intensified the interest in auditory-visual integration (McGurk and MacDonald, 1976).

The early phases of the work on audio-visual speech processes were summarized more than a decade ago by Summerfield (1979), Dodd and Campbell (1987) and Massaro (1987). Since then the volume and diversity of the research has grown exponentially. In the area of speech technology, machine recognition and speech coding have both made attempts to incorporate visual information into their task domains (e.g., Wolff et al., 1994).

In 1995 the diversity and vigor of research in audio-visual speech was acknowledged in a two-week NATO-ASI, organized by David Stork, Christian Benoît and N. Michael Brooke and held in Bonas, France. For perhaps the first time, perceptual psychologists, engineers, linguists and clinical researchers convened to discuss their work. One result of this meeting was a comprehensive volume (Stork and Hennecke, 1996) of conference proceedings. A second and perhaps more important result was the emergence of a group identity that effortlessly spanned the differences in specific research interest and venue of its members under the rubric of auditory-visual speech processing.

In order to nurture the fledgling bond among the group's members, while at the same time introducing auditory-visual speech processing to the larger speech community, Lynne Bernstein and Christian Benoît organized a special session of ICSLP'96 (International Conference on Spoken Language Processing, 1996). While extremely successful, it was clear that a half-day session could neither accommodate the volume of material being produced nor provide the time needed for discussion of the ideas stemming from so many research domains. Therefore, a two-day meeting on auditory-visual speech processing (AVSP'97) was organized by Christian Benoît and Ruth Campbell under the auspices of ESCA as a satellite of Eurospeech'97, held in Rhodes, Greece.

The expectation for AVSP'97 was that about 50 participants, most of them repeat customers from the NATO and ICSLP meetings, would attend. Surprisingly, more than 100 people joined the meeting. This surge in interest along with the high quality of the contributed papers and their subsequent discussion during the meeting inspired Christian Benoît to produce this special issue of *Speech Communication*. The 11 papers of this volume represent about a quarter of the papers presented in Rhodes.

In selecting these papers, Benoît's intent was, as usual, to show the diversity of topics in this burgeoning area of speech research. It was also his fervent wish to have the journal issue published in time for distribution at the next AVSP meeting,

PII: S0167-6393(98)00046-6

AVSP'98 in Sydney – a publication turnaround of just over 12 months. Unfortunately, our friend Christian Benoît did not live to see his effort bear fruit. Fortunately for all of us, this great lover of life, fun, people and bad beer, was extremely well-organized. It was amazingly easy, after his death in late April this year, for the two of us to pick up from where he left off and complete the editing process. Of course, this was not done without the yeoman efforts of people in Grenoble: his student Lionel Reveret, his best friend Pascal Perrier of the ICP, and Emmanuelle "Manu" Pechot-Gardia, the ESCA Assistant Secretary. We also thank Takaaki Kuratate and Erik McDermott at ATR for their assistance (Figs. 1 and 2).

Predictably, Christian chose his own paper for inclusion in the Special Issue, but how prophetic that the paper turned out to be perhaps the only comprehensive overview of his research activities since he joined the ICP as a CNRS Research Scientist in 1989. Far better than we could have done, his paper tells the story of his research goals and clearly indicates his vigorous mentoring of students.

Fig. 1. Christian Benoît at ATR, Japan (photograph by Takaaki Kuratate, January 1997).

We can only begin to describe the magnitude of the loss Benoît's death has caused to our research community. His great charm and quick-witted intelligence, his lousy pronunciation of English, questionable taste in clothes and friends were nothing compared to his passion for life and his profession. Benoît demonstrated constantly that it is not enough for us to study speech communication, we must also practice communication. It is rare these days to hear people described as wise, but those who knew Benoît recognized it as one of his foremost traits. In order to communicate with as many people as possible, he made it a point to be open-minded and supportive of the ideas and fancies of others. Imagine a smart guy who pays attention to our ideas and then encourages us to pursue them – is it any wonder that we all loved him?

Benoît did not leave matters at the level of interpersonal relations. For our benefit he worked wisely to improve lines of communication at organizational levels. Many of us are quite cynical when we see people actively seek positions of influence and power. Of course Benoît was neither stupid nor a saint, but perhaps more clearly than anyone we have known, he sought the power to dissolve the parochial barriers between researchers, institutions, and even nations. No wonder his colleagues were so happy to accommodate him!

As for the contents of the Special Issue, Benoît chose a total of 11 papers whose topical diversity gives a good indication of the breadth of the growing interest in auditory-visual speech processing. The papers deal (almost always non-exclusively) with the perception, production and processing of audio-visual speech behavior by humans and machines. Three papers report audio-visual perception studies using normal subject populations: Remez et al. examine the perception of audio-visual speech where portions of the acoustic information have been re-synthesized using sinewave synthesis. Sams et al. tested the

Fig. 2. Christian Benoît (left), Eric Vatikiotis-Bateson, Philip Rubin (right) in Lindos, Greece (photograph by Elisabeth Louveau, September 1997).

McGurk effect on Finnish subjects in order to address issues of auditory-visual integration. Iverson et al. argue that lexical as well as phonemic information is needed for perception of audio-visual speech stimuli. A fourth perception paper by de Gelder and Vroomen examines the dependence of speech-reading ability on a prosopagnosic patient's ability to process faces.

Four papers deal with aspects of machine recognition and synthesis of lip motion. These include Benoît and Le Goff's summary of the lip-modeling work within Benoît's group at the ICP and Basu et al.'s elegant approach to the automatic tracking and re-synthesis of 3D lip structures from the 2D video stream. Yamamoto et al. propose a text-to-audio-visual-speech system based on look-ahead HMMs, and Rogozan and Deléglise use HMMs to elucidate the adaptive fusion of audio and visual sources.

Two papers consider aspects of audio-visual speech production: Yehia et al. demonstrate strong correlations among the measured motions of the vocal tract and face and the spectral and segmental acoustics. Blokland and Anderson examine the effects of the visual medium on interactive speech performance. Finally, Poggi and Pelachaud consider higher level linguistic constraints and pragmatic factors in determining appropriate expressive facial behavior.

In completing the publication process since the time of Benoît's death in late April, we have tried to strike the balance we believe he intended to achieve between the presentation of serious research topics and the enjoyment of this new and exciting field of research. In successfully juggling the editing duties between three continents (Europe, Asia and North America), we owe a great debt to everyone involved. In addition to

Christian's associates at the ICP, the authors, reviewers, and senior editorial and publications staff of *Speech Communication*, particularly Sadaoki Furui, have all been exemplary in their co-operation and enthusiasm for seeing us and this project through a very sad time.

Philip Rubin, New Haven
Eric Vatikiotis-Bateson, Kyoto
July 1998

References

Dodd, B., Campbell, R. (Eds.), 1987. Hearing by eye: The psychology of lip-reading. Lawrence Erlbaum, London, Hillsdale, NJ.

Massaro, D.W., 1987. Speech perception by ear and eye: A paradigm for psychological inquiry. Lawrence Erlbaum London, Hillsdale, NJ.

McGurk, H., MacDonald, J., 1976. Hearing lips and seeing voices. 264, 746–748.

Stork, D., Hennecke, M. (Eds.), 1996. Speechreading by humans and machines. Vol. 150, Springer, Berlin.

Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Amer. 26, 212–215.

Summerfield, Q., 1979. Use of visual information for phonetic perception. 36, 314–331.

Wolff, G.J., Prasad, K.V., Stork, D.G., Hennecke, M., 1994. Lipreading by neural networks: Visual preprocessing, learning and sensory integration. In: Cowan, J.D., Tesauro, G., Alspector, J. (Eds.), Advances in neural information processing systems 6. Morgan Kaufmann, San Francisco, pp. 1027–1034.

Philip Rubin is the Chief Operating Officer and Vice President, Technology, of Haskins Laboratories in New Haven, Connecticut. He is also a Senior Scientist at Haskins Laboratories, a Research Affiliate in the Department of Psychology at Yale University, and a Professor (Adjunct) in the Department of Surgery (Otolaryngology) at the Yale University School of Medicine. Dr. Rubin's general research interests include speech perception, speech production, audio-visual speech, cognitive and ecological psychology, motor behaviour, and connectionist and nonlinear dynamical modeling. Dr. Rubin developed ASY, the Haskins Articulatory Synthesizer, which is presently being used by a large group of Haskins researchers as part of a computational model of speech production. In addition, he is directing the development of CASY, a configurable version of the articulatory synthesizer. He developed SWS, the Haskins Sine Wave Synthesizer and continues a long-standing collaboration with Robert Remez at Barnard College that uses SWS to explore temporal patterning in speech perception and the characteristics of normal conversational speech and speaker identity. The HADES system was also designed by Dr. Rubin. This is a set of programs for physiological and speech signal analysis and manipulation with an integrated programming language (SPIEL). Presently he is co-developing, with Eric Vatikiotis-Bateson of ATR, the Talking Heads website.

Eric Vatikiotis-Bateson is a member of Department 4 in the ATR Human Information Processing Research Laboratories in Kyoto, Japan. When left no alternative, he attempts to model the production of audio-visual speech behaviour computationally through examination of the neuromotor, biomechanical, and functional properties of the vocal tract and external orofacial structures. In pursuing his research interests, Dr. Vatikiotis-Bateson has been very fortunate to have fruitful collaborations with a broad range of talented researchers in Asia, North America and Europe.