Auditory Visual Speech Processing

Dominic W. Massaro

Department of Psychology, University of California
Santa Cruz, CA 95064 U.S.A.
massaro@fuzzy.ucsc.edu

Abstract

This paper provides an overview of the developments in Auditory Visual Speech Processing, a Special Interest Group within ISCA. I hope that this discussion will be informative and useful to readers in a variety of fields, including psychology, speech science, animation, psycholinguistics, human-machine interaction, hearing-impaired communication, and numerous other fields which also share in this fruitful intersection.

1. Introduction

We are celebrating Eurospeech 2001, the year in the title of director Stanley Kubrick's classic science fiction film 2001: A Space Odyssey, written by Arthur C. Clarke, and first released in 1968. Although most people are enthralled by this fable exploring their place in the universe, we AVSP-ers are thrilled by the computer Hal's ability to read the astronauts' lips, much to their demise. This implicitly acquired skill by an artificial system was the idea of Kubrick, the talented director. It has also been 25 years since Harry McGurk and MacDonald published their seminal paper on the McGurk effect in Nature [1]. Their serendipitous finding initiated a cottage industry in psychological inquiry and cognitive science more generally. One of the most exciting dimensions of this cottage enterprise is the mix of pure empirical research, theoretical development, important applications, and commercial enterprise.

Table 1: Illustration of the subfields of auditory visual speech processing (AVSP) with arbitrary distinctions between humans and machines and recognition and production.

	Recognizing	Producing
Human	AV Perception	AV Production
Machine	AV Recognition	AV Synthesis

The field of auditory visual speech processing can be delineated into four subfields shown in Table 1. We make arbitrary distinctions between humans and machines and recognition and production. It is particularly important that our SIG be understood and viewed as an AV one and not simply one with an interest in visible speech. Demonstrating the information value of visible speech is of value but most importantly it is necessary to understand how multimodal speech processing is the gold to be mined. The ability to integrate information from different sensory systems is a fundamental characteristic of human perception and holds great promise for intelligent machine recognition. Different sensory channels provide multiple "looks" at the stimulus situation and their perceptual integration or synthesis markedly enhances the detection, perception, and identification of the event. Both humans and machines have the ability to be excellent Bayesians [2]. The challenge for AVSP-ers is to determine the sources of information from different senses (or sensors) and how these can be combined to facilitate spoken language processing. We begin with human perception of multimodal spoken language.

2. Human Perception

Although both practitioners of speech therapy [3] and science [4] were both well aware of the potential richness of speech information in the face, the McGurk illusion (hearing *inappropriately* because of watching the face) captured the imagination of all persons who experienced it. The McGurk effect or some variant of it has been replicated and energetically studied across different languages (English, Japanese, Dutch, Spanish, French, German, Cantonese, and Thai) across the lifespan from infancy to chronological giftedness, from the perception of nonsense syllables to the understanding of prose [5][6]. Emerging from this impressive body of activity is the robustness of the phenomenon, holding up independently of the intention of the perceiver [5] and also existing in analogous fashion in other domains such as perceiving emotion from the face and the voice [7][8].

A persistent issue in audiovisual integration has been the time course of evaluation and integration of these modalities. Evidence based on the theoretical description of perceptual judgments has consistently shown that modality independent evaluation occurs for many of the phonetic properties of the speech input. Other investigators have been less willing to accept these conclusions partly because phenomenal experience seems to imply an early mixture of the auditory and visual inputs. Brain imaging results from fMRI during unimodal and audiovisual speech have been used to investigate the interaction between the two modalities [9]. Another possible source of evidence is the measure of event-related potentials (ERPs). In a recent study [10], stimuli were presented in an oddball paradigm to evoke the mismatch negativity effect (MMN). The MMN for incongruent audiovisual condition was much later than it was for the unimodal auditory MMN, suggesting that integration of the two sources occurs later than auditory-alone phonetic processing. We can expect an exponential increase in brain measures of auditory visual speech processing because it represents an important instance of sensory fusion in which

Paper presented at the 7th European Conference on Speech Communication and Technology - Eurospeech 2001 – Scandinavia, Aalborg, Denmark.

information from several senses is melded into a coherent percept.

The relevance of the stimulating research being carried out in multisensory fusion is readily apparent. There have been several important international conferences and this forum offers tremendous promise in linking the behavioral research and machine applications with the neuroscience of how multiple senses work together to provide coherent action in the environmental niche of the organism. Although not specifically concerned with spoken language, the International Multisensory Research Forum (http://www.wfubmc.edu/nba/IMRF/), brings together a broad range of scientists and scholars concerned with multisensory fusion. Their annual conferences highlights multisensory research from every corner of the field, including neuroscience, development, and perception. AVSP-ers will certainly benefit by learning how central integration of the senses is in natural perception and action. There has been tremendous progress in neural information processing, with many studies showing that "multisensory" integration is provided by neurons that receive convergent input from two or more sensory modalities [11]. Preliminary studies have even demonstrated that these neurons can be understood in terms of Bayesian processing [12].

3. Machine Recognition

Given the synergistic use of information from the face and the voice in humans, perhaps spoken language recognition by machines should be based on several sensors, not just a microphone. About 10 years after the McGurk effect, Eric Petajan [13] published his dissertation on a recognition system for automatically tracking the face and using information from it to aid in speech recognition by machine. Since Petajan's seminal study, several other researchers have advanced speechreading systems so that they might eventually be used effectively in spoken language understanding systems [14]. A challenge for speechreading systems is that the face is always moving about and therefore must be tracked in order to maintain it in view. Auditory speech, on the other hand, fills a room and the location of the microphone is not so critical. Of course, the robustness of auditory speech recognition systems is improved with a microphone that remains a fairly fixed distance from the mouth, and speechreading systems might follow this lead by mounting a camera on the user to move with movements of the head [15][16]. If the application requires an unencumbered head, a camera on the desktop computer can be used to track the head. There are several working applications of this technology [17][18].

One approach, particularly in accord with my bias that both people and machines can benefit from multiple sources of information, exploits many different cues to identify, locate, and track the face. Color, shape, and feature information have somewhat complementary properties. For example, color is very robust in locating the face but is not useful to give detail information about the exact pose. On the other hand, the use feature information can give very detailed information about the face but can be easily confused with changes in lighting, rapid movement, and other common changes. One strategy is to schedule these different operators intelligently, with the most precise operators applied when possible, but defaulting to more robust operators when necessary [14] [18].

4. Animation and Visible Speech Synthesis

At about the same time of the McGurk studies, Frederic Parke [19] carried out his seminal dissertation on facial animation. A synthetic talking head must be considered crucial to AV speech processing because it enables the scientist to have a level of control over the stimulus not possible by natural means--for example, in building continua of precisely defined steps between pairs of syllables, or systematically manipulating some property of the speech signal. Development and refinement of talking heads has thus been as important a part of research as has been use of it in psychological experiments; one ultimate goal of this development being to create a synthetic speaking face that is as realistic as a natural face in the movements of speech.

Although some research questions can be answered in part with natural speech stimuli, our overall progress in understanding speech perception has been critically dependent on the use of synthetic speech. Extending this approach to the visual dimension of speech, goals for this technology include gaining an understanding of the visual information that is used in speechreading, and how this information is combined with auditory information in language perception and understanding; I also foresee its use as an improved channel for human/machine communication, as a useful aid in education, and as a synthetic actor in entertainment [20].

In the last few years, there have been a plethora of agents serving as newscasters, helpers on desktops, messengers with email, and simply personal friends. For example, Microsoft® Agent can be included as part of Web pages or conventional applications at http://www.microsoft.com/msagent/. This user interface element can be used on Windows® platforms and it enables you to display and animate an interactive character. It has a good deal of flexibility and the sophisticated user can even compile original character animations using the Microsoft Agent Character Editor. The character's name and description, the way it looks, and even its methods of output can be controlled. Microsoft Agent Ring is worldwide organization of websites featuring Microsoft Agent [21]. However, it is fair to say that all of these characters have terrible visible speech and are actually detrimental rather than facilitative in communicating spoken language.

5. Speech Production

Students of speech production necessarily realize the intricacies of talking heads, but have been mainly concerned with how their movements occur and how they determine the acoustic output. The visibility of these movements has not been a central issue. Needless to say, however, is that this area of inquiry is central to AVSP. Several laboratories have independently collected measurements of real faces and tongues to better understand the visible and articulatory characteristics of speech, describe their relationship to the acoustic characteristics of speech, and to use these measures to guide accurate visible speech synthesis.

Speech scientists have collected 3D facial measurements and tongue movements from both Japanese and English talkers, and have correlated these measures with acoustic properties of the speech (specifically line spectral pairs—LSPs). They have found an impressive correlation between the facial and tongue measures and between these measures and LSPs. [22] These correlations have been recently replicated [23]. Kinematic signals predicted each other and are welll-predicted from LSPs. LSPs are not as well predicted from kinematics, however.

6. Data Bases of AVSP

In each of the four areas of AVSP, audio-visual speech databases from real talkers are essential. There are a number of data sources about speech production - both static and dynamic - that have been accummulated. The data sources include: 1) ultrasound and electropalatography (EPG) data for the tongue from a single talker [24], 2) X-Ray microbeam [25] and 3) cineradiographic recordings [26] of the vocal tract during articulation, 4) 3D Cyberware laser scans of static facial gestures and visemes, and 5) systematic measurements of visible speech articulation using various tracking systems of marked-up talking heads [27].

For facial animation and visible speech synthesis, an important goal is to enhance the quality of the synthetic visible speech and to make talking heads as realistic as possible. Measures of facial, lip, and tongue movements during speech production are required to optimize accuracy of speech production modeling. The goal is to minimize the error between the empirical observations of real human speech and the speech produced by a synthetic talker. In one case, a highly realistic palate, teeth, and tongue were modeled using 3D ultrasound data and electropalatography (EPG) [28].

7. Applications of AV Speech Processing

No one can deny the excitement generated by the challenges of AVSP applications. Thinking laterally, one distant goal of speech science and facial animation technology is to automatically translate language as it is being spoken and to produce it by a computer-animated talking head. Several researchers have developed several component technologies that bring us closer to this goal. One is texture mapping of a person's face onto the computer-animated talking head and another is the ability to drive talking heads directly from the symbolic input. The source language would be translated and this translation would drive a computer-animated talking head. Therefore, the speaker of the source language can be seen producing the target language [28][29]. The scenario would be that my Japanese colleague in Japan is speaking to me in Japanese. On my computer screen, I see him speaking to me in English. And he sees me speaking to him in Japanese even though my actual vocabulary tops out at "arigato".

Given the high correlation between speech acoustics and measures of visible speech production, we might then expect that visible speech could be driven directly from the auditory signal. Given that synthetic auditory speech still is unacceptable to most listeners, driving the visible speech directly from the auditory speech and simultaneously aligning the two modalities provides an ideal solution to many potential applications. Auditory speech, particularly in MP3 format, requires very little bandwidth. An animation application can exist locally on a user's desktop or laptop and auditory speech can be streamed in real time over the internet [30][31]. Texture mapping of an image of the person speaking is also possible. Thus, if I'm a fan of Mojo Nixon a disc jockey broadcasting from Ohio, I can see him as well as hear him at my desktop in Aalborg.

The development of a computer-animated talking head, Baldi, serendipitously showed promise for language tutoring. Baldi is fully integrated into a speech toolkit and is currently providing language training with profoundly deaf children [32], autistic children, and children with reading delays. Many different kinds of evaluation including assessment tests, children's reports and time spent on task, and the teachers and parents feedback indicate great promise for AVSP applications.

8. Appendix

AVISA (the Auditory-Visual Speech Association), which is the second ISCA Special Interest Group, started its official business on December 5, 1998, at the AVSP'98 (Auditory Visual Speech Processing) meeting in Terrigal, New South Wales, Australia. AVISA was first conceived by Christian Benoît, who wanted to see the auditory-visual speech community come together in a more formal way. Before his untimely death, Christian drafted a proposal to create AVISA. We are proud to dedicate this SIG to Christian's memory. The web site is http://www.haskins.yale.edu/AVISA/AVISA.html.

As part of the AVISA effort, the "Talking Heads" website [33] was created to provide an overview of the rapidly growing international effort to create talking heads (physiological / computational / cognitive models of audio-visual speech), the historical antecedents of this effort and related work. Links are provided (where possible) to the sites of many researchers and commercial entities working in this diverse and exciting area. The Talking Heads site was officially released on Dec. 6, 1998, at the AVSP '98 meeting in Terrigal, New South Wales, Australia. The URL is http://www.haskins.yale.edu/haskins/heads.html.

This overview is dedicated to Christian Benoit and Kerry Green, who were early and influencial students of AV speech processing. Christian almost single-handedly established visible speech as an important domain of research and application and Kerry grounded the study of AVSP in innovative and informative psycholinguistic experiments.

9. Acknowledgement

This writing of this paper was supported in part by NSF grants BCS-9905176 and IIS-0086107 and Public Health Service Grant PHS R01 DC00236.

10. References

[1] McGurk, H., & MacDonald, J. "Hearing lips and seeing voices", *Nature, 264*, 746-748, 1976.

[2] Massaro, D.W., & Stork, D. G. "Sensory integration and speechreading by humans and machines," *American Scientist, 86*, 236-244, 1998.

[3] Hutton, C. "Combining auditory and visual stimuli in aural rehabilitation", *Volta Review, 61*, 316-319, 1959.
[4] Campbell, H. W. *Phoneme recognition by ear and by eye: A distinctive feature analysis*, Unpublished doctoral dissertation, University of Nijmegen, Holland, 1974.
[5] Massaro, D. W. *Perceiving talking faces: From speech perception to a behavioral principle*, MIT Press, Cambridge, Massachusetts, 1998.
[6] Campbell, R., Dodd, B., & Burnham, D. (Eds.), *Hearing by eye II : advances in the psychology of speechreading and auditory-visual speech*, Psychology Press, Hove, East Sussex, UK , 1998.
[7] Massaro, D. W., & Egan, P. B. "Perceiving affect from the voice and the face", *Psychonomic Bulletin & Review, 3*, 215-221, 1996.
[8] Vroomen, J. & de Gelder, B. "The perception of emotions by ear and by eye", *Cognition and Emotion, 14*, 289-311, 2000.
[9] Calvert, G. A.; Campbell, R.; Brammer, M. J. "Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex", *Current Biology*, 10 (11), 649-657, 2000.
[10] Bernstein, L. E., Ponton, C., Auer, E. T. "Is audiovisual speech integration an early perceptual effect? An event-related potential study of the McGurk effect", *Cognitive Neuroscience Society*, New York City, March 25-27, 2001.
[11] Stein, B.E. "Neural mechanisms for synthesizing sensory information and producing adaptive behaviors", *Exp. Brain Res*., 123(1-2):124-35, 2001.
[12] Anatasio, T. "Modelling Multisensory Enhancement in the Superior Colliculus: The second annual Multisensory Research Conference, held 6 - 7 October, 2000 in Tarrytown, New York."
[13] Petajan, E. D. "Automatic Lipreading to Enhance Speech Recognition: Proceedings of the Global Telecommunications Conference", *IEEE Communication Society*, 265-272, 1984.
[14] Toyama, K. "Prolegomena for Robust Face Tracking: Microsoft Research Technical Report", *MSR-TR-98-65*, 1998.
[15] Gagne, J., Le Monday, K., Desbiens, C. Lapanlme, M. & Ducas, L. "Evaluation of a visual –FM system to enhance speechreading. *Proceedings of Auditory Visual Speech Perception*, 201-206, Terrigal-Sydney Australia, December, 1998.
[16] Ganymedia, http://www.ganymedia.com/
[17] Yang, J., Stiefelhagen, R., Meier, U., & Waibel, A. "Visual tracking for multimodal human computer interaction", *Proceedings of SIGCHI*, 140-147, 1998.
[18] Shpungin, B, & Movellan,, J. R. " A multimodular approach to real-time face tracking", *Proceedings of the 7th symposium on Neural Computation*, California Institute of Technology, June 2000.
[19] Parke, F. I. "A model for human faces that allows speech synchronized animation", *Computers and Graphics Journal*, 1(1-4), 1975.
[20] Massaro, D.W., Cohen, M. M., Beskow, J., & Cole, R. A. "Developing and evaluating conversational agents", In J. Cassell, J. Sullivan, S. Prevost, & E. Churchill (Eds.) *Embodied conversational agents*, MIT Press, Cambridge, MA, 2000.
[21] http://www.msagentring.org/
[22] Yehia, H., Rubin. P., & Vatikiotis-Bateson, E. *Quantitative Association of Vocal-Tract and. Speech Communication*, 26(1-2):23-43, 1998.
[23] Jiang, J., Alwan, A., Bernstein, L. E., Keating, P., & Auer, E. T. "On the correlation between facial movements, tongue movements, and speech acoustics", *ICSLP2000, International Congress on Spoken Language Processing*, Beijing, China, 16-20 October, 1998.
[24] Munhall, K.G., Vatikiotis-Bateson, E. & Tohkura, Y. "X-ray Film database for speech research", *Journal of the Acoustical Society of America, 98*, 1222-1224, 1995.
[25] Westbury, J. R. "X-Ray Microbeam Speech Production Database User's Handbook", Madison, WI: Waisman Center on Mental Retardation and Human Development, 1994.
[26] Stone, M. & Lundberg, L. "Three-dimensional tongue surface shapes of English consonants and vowels", *Journal of the Acoustical Society of America, 99*, 1-10, 1996.
[27] Cohen, M. M., Beskow, J., & Massaro, D. W. "Recent developments in facial animation: An inside view", *Proceedings of Auditory Visual Speech Perception*, 201-206, Terrigal-Sydney Australia, December, 1998.
[28] Waibel, A. "Interactive translation of conversational speech", *Computer, 29*, 41-48, 1996.
[29] Ogata, S., Murai, K., Nakamura, S., & Morishima, S. "Model-based lip synchronization with automatic translated synthetic voice toward a multi-modal translation system", 2001.
[30] Eva Agelfors, Jonas Beskow, Bjrn Granstrm, Magnus Lundeberg, Giampiero Salvi, "Synthetic Visual Speech Driven From Auditory Speech", speech.kth.se/~bes...avsp99teleface.pdf
[31] Massaro, D.W., Beskow, J., Cohen, M. M., Fry, C. L., & Rodriguez, T. "Picture my voice: Audio to visual speech synthesis using artificial neural networks", In D. W. Massaro (Ed.), *Proceedings of AVSP'99: International Conference on Auditory-Visual Speech Processing*, 133-138, Santa Cruz, CA., August, 1999.
[32] Massaro, D.W., & Cole, R. "From 'Speech is special' to talking heads in language learning", In *Integrating Speech Technology in the Language Learning and Assistive Interface*, University of Abertay Dundee, Dundee, Scotland, August 29-30, 153-161, 2000.
[33] Rubin, P. & Vatikiotis-Bateson, E. "Talking heads", In D. Burnham, J. Robert-Ribes, & E. Vatikiotis-Bateson (Eds.), *International Conference on Auditory-Visual Speech Processing - AVSP'98*, 231-235, Terrigal, Australia, 1998.