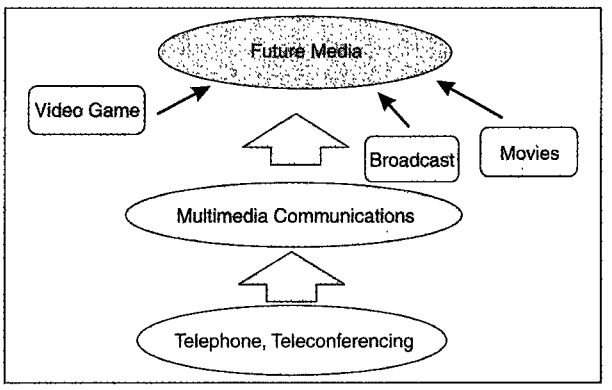
phone. Although we see teleconferencing gradually becoming a regular part of the business environment, videophones have yet to take off. The fact that such services are so intimately tied to activities in our everyday lives makes it extremely difficult to predict what services will be provided in the future based strictly on technologies. As such, we cannot focus on the telecommunications industry alone, but we must look at a broad range of other media in considering telecommunications services of the future.

Communications, entertainment, and other fields have recently undergone some rather significant changes. In essence, we are seeing the emergence of fields such as communications and entertainment in cyberspace. A good example of this in the telecommunications industry is the emergence of the Internet as a new communications venue. The Internet connects people all around the world in a way that could well be thought of as cyberspace on a massive scale. In this cyberspace, people communicate with others, shop, and seek out new information.

Let's look at the huge movie market that makes up the entertainment industry. Movies today incorporate digital and computer graphics technologies that are creating a whole new generation of movies. Digital and computer graphics technologies have given us ultra-realistic worlds never before possible in conventional movies. In other words, they give us the capacity to create new cyberspaces. On the other hand, video games, especially role playing games (RPGs), allow people to enjoy a good story as heroes in a cyberspace or virtual world.

Such trends must be taken into account when considering the future image of multimedia networks, and they are the same trends found in technological forecasts that point toward the integration of telecommunications and broadcasting. From this standpoint, therefore, we have no choice when looking at new services for the future but to consider images that include a variety of media outside of telecommunications, such as broadcast media like television, entertainment media like movies, and video-game media like those games so popular among children. Figure 2 shows the connection between conventional media and media of the future.

Let's look at the state of telecommunications in the future based on the trends described above. The general view on telecommunications of the future is that we will communicate across distance and time as well as cultures, and that we will communicate naturally with computers in cyberspace. This concept includes the original goal of telecommunications: communications with anyone, anywhere, and anytime. If we accept this broad concept, we can offer an extremely broad image for the future that includes telecommunications as well as other media. Therefore, it is important, I believe, for MMSP researchers to select research directions based on this future view of telecommunications. Although I do not list each research theme, the research areas connected to the following technologies will become important.



2. Toward the realization of future media.

- Technologies for generating any kind of cyberspace
- Technologies for warping into cyberspace
- Technologies for manipulating objects in cyberspace
- Technologies for communicating with residents of cyberspace

References

1. Mark Knapp, *Nonverbal Communication in Human Interaction*, New York: Holt, Rinehart and Winston. 1972.

2. P. Ekman, "Facial Expression and Emotion," *American Psychologist* 48, pp. 384-392. 1993.

3. Marshall McLuhan, *Understanding Media*, New York: McGraw-Hill, 1964.

4. Byron Reeves and Clifford Nass, *The Media Equation*, Cambridge, CSLI Publications, 1996.

Audio-Visual Interaction in Multimodal Communication

Rama Chellappa, University or Maryland;
Tsuhan Chen, AT&T Labs; Aggelos Katsaggelos,
Northwestern University

Multimedia signal processing is more than simply "putting together" text, audio, images, and video. It is the integration and interaction among these different media that creates new systems and new research challenges and opportunities. It is being realized that unimodal analysis using audio or video can deliver acceptable performance levels only in benign situations; the performance decreases rapidly when countermeasures are taken. For example, person authentication systems useful in security, access control, and surveillance applications do not perform well when subjects age, wear masks and disguises, or when the resolution is not good or poor lighting conditions are present. Many of these difficulties can be overcome by adding an audio signature along with video.

In multimodal communication where human speech is involved, audio-visual interaction is particularly significant. The most interesting phenomenon pertaining to this interaction is the "McGurk Effect" [1]. It shows that hu-

man perception of speech is bimodal in that acoustic speech can be affected by visual cues from lip movements. For example, one experiment shows that when a person "sees" a speaker saying /ga/, but "hears" the sound /ba/, the person perceives neither /ga/ nor /ba/, but something close to /da/.

Due to the bimodality in speech perception, audio-visual interaction is an important design factor for multimodal communication systems such as video telephony and video conferencing. A prime example of this interaction is lip reading or speech reading. Lip reading is not only used by the hearing-impaired for speech understanding. In fact, everyone utilizes lip reading to some extent, in particular in a noisy environment such as at a cocktail party. Communication systems must be able to provide the full motion necessary for speech reading by the hearing-impaired. Researches have studied the importance of frame rates with impaired listeners [2] and analyzed the effects of frame rates on isolated viseme recognition [3]. Research in these areas will lead to multimedia systems that account for the perceptual boundaries of the hearing-impaired. Researchers have also tried to teach computers to lip-read [5]. Based on computer-vision techniques for tracking lip movements of a speaking person, a computer can be trained to understand visual speech. In addition, automatic lip reading has also been used to enhance acoustic speech recognition.

What can one do if the frame rate is not adequate for lip synchronization perception, which is a typical situation in video conferencing equipment due to the bandwidth constraint? One solution is to extract the information from the acoustic signal that determines the corresponding mouth movements, and then process the speaker's mouth image accordingly to achieve lip synchronization [4]. On the other hand, it is also possible to warp the acoustic signal to synchronize with the person's mouth movements. The latter approach is very useful in nonreal-time applications, such as dubbing in a studio.

One key issue in bimodal speech analysis and synthesis is the establishment of the mapping between acoustic parameters and the mouth shape parameters. In other words, given the acoustic parameters, such as the cepstral coefficients, one needs to estimate the corresponding mouth shape, and vice versa. A number of approaches have been proposed for this task that utilize vector quantization [7], neural networks [8], Gaussian mixtures, and hidden Markov models [9].

Audio-visual interaction can be exploited in many other ways. The correlation between audio and video can be utilized to achieve more efficient coding of both audio and video [6, 7]. Audio-visual interaction can also be used for person authentication and verification [10, 11, 12]. Other applications include dubbing of movies, segmentation of image sequences using video and audio signals [13], human-computer interfaces, and cartoon animation. All these clearly demonstrate that the joint processing of audio and video provides additional capabilities that are not possible when audio and video are studied separately. It is clear that once we break down the artificial boundary between audio/speech and image/video processing, many new research opportunities and innovative applications will arise.

References

1. H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, pp. 746-748, December 1976.

2. Frowein, et al., "Improved speech recognition through videotelephony: Experiments with the hard of hearing," *IEEE Journal on Selected Area in Communication*, vol. 9, no. 4, May 1991.

3. J. Williams, J. Rutledge, D. Garstecki, A. Katsaggelos, "Frame Rate and Viseme Analysis For Multimedia Applications," *Proc. of IEEE Multimedia and Signal Processing Conference*, Princeton, NJ, June 1997.

4. T. Chen, H.P. Graf, and K. Wang, "Lip-synchronization using speech-assisted video processing," *IEEE Signal Processing Letters*, vol. 2, no. 4, pp. 57-59, April 1995.

5. D. Stork, "'I could see your lips move': HAL and Speechreading," HAL's Legacy, The MIT Press, 1997.

6. D. Shah, and S. Marshall, "Multi-modality coding system for videophone application," WIASIC'94, Berlin, Germany, October 1994.

7. S. Morishima, K. Aizawa, and H. Harashima, "An intelligent facial image coding driven by speech and phoneme," ICASSP, p. 1795, Glasgow, UK, 1989.

8. F. Lavagetto, "Converting speech into lip movements: A multimedia telephone for hard of hearing people," *IEEE Trans. on Rehabilitation Engineering*, pp. 114, March 1995.

9. T. Chen and R. Rao, "Audio-Visual Interaction in Multimedia Communication," vol. 1, pp. 179-182, ICASSP, Munich, April 1997.

10. J. Luettin, N.A. Thacker, S.W. Beet, "Speaker identification by lipreading," ICSLP, October 1996.

11. M.R. Civanlar and T. Chen, "Password-free network security through joint use of audio and video," *SPIE Photonic East*, November 1996.

12. *Proceedings of the First International Conference on Audio and Video Biometric Person Authentication*, Crans-Montana, Switzerland, March 12-14, Springer-Verlag, Berlin, 1997.

13. J. Nam and A.H. Tewfik, "Combined Audio and Visual Streams Analysis for Video Sequence Segmentation," *ICASSP*, vol. 4, pp. 2665-2668, Munich, April 1997.

Modeling and Evaluation of Multimodal Perceptual Quality

Kim Tilgaard Petersen, Steffen Duus Hansen, John Aasted Sørensen, Tech. Univ. of Denmark

The increasing performance requirements of multimedia modalities, carrying speech, audio, video, image, and graphics, emphasize the need for assessment methods of the total quality of a multimedia system and methods for simultaneous analysis of the system components. It is important to take into account still more perceptual characteristics of the human auditory, visual, tactile systems, as well as combinations of these systems. It is also highly desirable to acquire methods for analyzing the main perceptual parameters, which constitute the input for the total quality assessment. Altogether, this is necessary for opti-