ISSN 1392 – 124X INFORMATION TECHNOLOGY AND CONTROL, 2006, Vol. 35, No. 2

HUMAN-COMPUTER AUDIOVISUAL INTERFACE

Jonas Kaukėnas, Gediminas Navickas, Laimutis Telksnys

Institute of Mathematics and Informatics, Recognition Processes Department
Gostauto 12, LT-01108 Vilnius, Lithuania

Abstract. The concept of audiovisual interface between human and stochastic process modeling and analysis software is investigated. Examples revealing the advantages of audiovisual interface over audio-only interface are given.

Keywords: Audiovisual interface, multimodal interface, Coupled hidden Markov model.

1. Introduction

Computers are widely used for analyzing stochastic process characteristics, for solving the problems of process identification, clustering and recognition. Computers are also widely used for modeling the signals generated by stochastic dynamic systems.

While solving these problems, investigators have to work in loosely defined situations. They have to come to the solution not having or having very little information on the investigated processes. In such situations investigators refer to their previous experience, knowledge, intuition and step by step come to the solution. Computers are used in such an investigation process. They can calculate very quickly, they need no rest and they can present the computation results in a practical form. But if the investigator wants a computer to calculate anything, he has to give strictly formulated tasks or commands. During the whole process, the investigator has to deal with lots of different tasks and commands of this kind. Usually, these commands are given using keyboard and mouse. Such an interface sometimes distracts the investigator and makes the intellectual work less productive. It would be much more convenient if we could work with a computer not giving commands but just having a conversation and exchanging information – just like working with human laboratory assistant.

Such an interface can be realized by using speech recognizers and speech synthesizers. But when one has to work in the noisy environment, a lot of recognition errors make the situation complicated. This defect can be eliminated using the audiovisual interface for human-computer dialogue. In this case, beside the audio signal, visual information of user's articulatory tract (and/or body gestures) is used.

We present a concept of audiovisual interface between human and stochastic process analysis and modeling software. Some examples demonstrating the advantages of audiovisual interface over the audio-only interface are presented. Also, in this article we give an example of human-computer dialogue between investigator and stochastic process analysis and modeling software.

2. Situation overview

Multimodal interfaces (MMI) aim at integrating several communication means in an harmonious way and thus make computer behavior closer to human communication paradigms, and therefore easier to learn and use. This has been possible with the advent of multimedia systems that can sample, store and produce complex types of information in real time [3].

People usually use the language for the communication. Most often the speech and body language (different body movements – usually hand gestures and gaze) is used.

Communication psychologists say that about 60-80% of overall information during the conversation between two people is transmitted by non-verbal means of communication and only 20-50% in a verbal way [2].

The idea of multimodal interface is based on the belief that communication with a computer must become more natural – similar to the way we communicate to each other. We can say that the multimodal interface is trying to move to communication with machines rather than operation of machines [8].

Multimodal interface channels can be grouped in such a way:

1. **Tactile** – when different devices, which require a physical contact for input, are used: keyboard, mouse, touch screens, special pens, etc.

2. **Acoustic** – when the speech and other human produced sounds are detected.

3. **Visual** – when different movements of human body parts are detected, for example lip movements, hand gestures, head or eyes position.

Having in mind that most often people use speech, gaze (eye movements) and a wide spectrum of body language movements (usually hand gestures and the gaze) [11], we can say that tactile interfaces are not natural but determined by technical limitations which were relevant to the time when first computers appeared. Multimodal interface researchers aim at using audio and visual channels to make human-computer interaction more human-centered and anthropomorphic [10].

Multimodal interface enthusiasts do not try to eliminate the mouse and keyboard, at least in the nearest future. In some cases these input devices are hardly replaceable by anything better. But there are situations when traditional means of control are uncomfortable and restrictive. For example, when the user can not contact with control devices physically (when the user is at some distance from the machine, when the hands are busy, or when the user has to move around), when computer is intended to accept commands from several users, or when the user is not skilled at using the mouse ant the keyboard.

At the moment multimodal interfaces are used in such areas: control of CAD programs [6], robot control and interface [25], medical instrument control [1], GIS control, for collaborative knowledge work [13] and other areas.

There exist systems, which use such modalities: keyboard, mouse, visual information (of the face, eyes, lips), auditory information, gestures and special communication devices (for example special sensory gloves). The modalities are combined in different ways to achieve synergies that transcend the benefit of a single modality.

The „Put-That-There“ system [4], created in Massachusetts Institute of Technology (MIT), is the first example of the multimodal interface. In the ninth decade of the last century when computers didn't even have a graphical user's interface, R. A. Bolt demonstrated a system, which was controlled by speech and gestures. Later, based on these experiments, the book was released which became the first solid issue about multimodal interfaces [5]. Researchers at MIT keep on doing their work and develop the systems that use speech, gestures and gaze for the human-machine interface [29]. While the computers become more and more powerful, the multimodal interface becomes more realizable.

The modalities can be combined in the following ways:

1. **Data streams from different modalities make up a composite command**. For example, the user says "put the red object here" and shows the place by hand where the object has to be put [29].

2. **Data streams from different modalities are analyzed in parallel and complement each other** making the command more robust and reliable. For example, the command is given by voice and gesture. Both streams are analyzed and finally one command is comprised. In this case, if the voice command is not recognized because of the environmental noise, it is corrected using visual information, and vise versa.

The multimodal interface means not only using more input channels, but also new interface concepts: sound icons, smart interface ideology, dialogue versus operation of the computer.

One of the main ideas is that we need to escape from the WIMP (Windows, Icons, Menus, Pointers) paradigm thus seeking to make the interface more natural, intuitive, flexible and expressive [24].

While constructing MMI, we have to keep in mind that software must be adapted to such an interface if we want to benefit from it. At first we have to answer the questions: why are we creating MMI, why a traditional interface is not good in a special case, and is it possible, that MMI would limit user's behavior instead of helping?

The modalities must be combined so that the synergy would be achieved and MMI would overperform the single modality interface [13]. Thus, there appear new interface concepts [4, 5, 8], new interface design principles [16], metaphors (sound icons, smart interface), the criteria of interface efficiency [12] and even myths [24].

In respect of the research object, MMI is the cross-discipline area where the researchers with different backgrounds (computer science, mathematics, psychology, ergonomics, medicine and others) are working.

More information on the various aspects of MMI can be found in the works of scientists from MIT [5], Swedish Royal Institute of Technology [7], Chinese Academy of Sciences [16], Oregon Health & Science University [23, 16], Rutgers University (New Jersey) [13, 12], ATR Spoken Language Translation Research Laboratories (Japan) [19].

3. The problem of process analysis and the concept of human-computer interface

The functioning of mechanisms and organisms is related with random processes of a different nature. So there arises a necessity to analyze these random processes, study their features, and to develop their mathematical models. In solving this kind of problems, at first the investigator has no information on the processes. In order to describe the features of such random processes, an experimenter poses hypotheses and gives commands to a computer to verify them. The investigator obtains the calculation results from the computer and, basing on them, either accepts the posed hypotheses, or poses new ones. Thus he has to work in a dialogue mode.

To realize these jobs, a dialogue system STADIA3 for the statistical analysis of random processes was developed. When employing the dialogue system STADIA3, the researcher gives tasks to the computer by using the keyboard or/and mouse. However, this prevents him from concentrating attention on the analysis of the results obtained by computer and from formulating new hypotheses on the features of the process.

This kind of defect can be diminished using the audiovisual interface between the user and the statistical software of random process analysis.

To this end, a statistical analysis system of random processes STADIA4 is in progress by means of which the dialogue is realized through the audiovisual interface. The interface between the investigator and the process analysis system STADIA4 is illustrated in Figure 1.

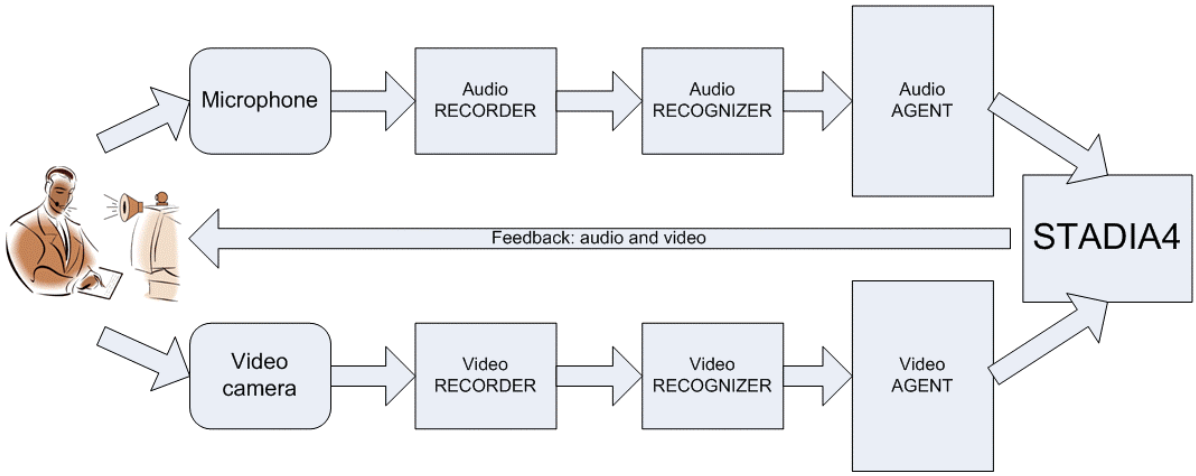

Figure 1. The scheme of interface between investigator and process analysis system STADIA4

The user communicates with the process analysis software STADIA4 using the voice and gestures. For that purpose the following tools are used:

1. Hardware: microphone, video camera.

2. Software that records audio and video. Both streams are passed to recognizers using the appropriate protocols.

3. Software that recognizes audio and video information. The audio software recognizes the given number of commands (phrases, words). Visual software recognizes the given number of micro or macro movements which we call gestures. The results of both programs are the control codes. The software can inform about recognition conditions, ask to complement or repeat the task or report that voice command is not synchronized with gestures.

4. Audio and video software agents. This software is responsible that control codes are formatted according to the appropriate communication protocols and passed to STADIA4.

After the STADIA4 software has got the command from the agent, it performs requested operations and reports success or failure.

The feedback, as well as the commands sent by user, is audiovisual. STADIA4 reports notifications by voice and displays them on the screen.

The concept of the dialogue between the investigator and computer is based on the finite set of the actions that program can perform.

The investigator gives commands to STADIA4 through the audiovisual interface, which consists of audio and visual recognizers and agents. STADIA4 reports the messages to the investigator through the audiovisual feedback channel.

In this situation, the following scenario of the dialogue between the investigator (marked as "I") and the system STADIA4 (marked as "S") can be an example:

I: Let‘s start.

S: Starting.

I: Generate autoregression sequence.

S: Please give the sequence parameters.

I: Sample size – five thousand values.

I: First equation coefficient – one point one.

I: Second equation coefficient – minus zero point nine.

I: Noise – one point three.

S: Invalid parameters. The model is not stable.

I: Change equation coefficients.

I: First equation coefficient – one point two.

I: Second equation coefficient – zero point nine.

S: Done (computer generates and draws the sequence graph).

I: Calculate spectral density.

S: Done (computer calculates spectral density function and draws its graph).

I: Increase the spectral density ordinate scale.

S: Done.

I: Let‘s finish.

S: Finishing.

4. The methods and tools used in the audiovisual human-computer interface

The audiovisual speech recognition system consists of audio and video detection and elementary processing, audio and visual feature extraction, and audio and visual feature integration (Figure 2).

At first, in the video signal the face and mouth are detected and then tracked. Further the visual features are extracted from the video signal, then the features from different streams are integrated and recognition decision is made. Our area of interest is the visual feature extraction and feature integration. The front-end for face and mouth detection and tracking is used as proposed by I. Shdaifat in 2005 [27].

In our concept of human-computer audiovisual interface, in the general case, audio and video streams are independent (Figure 1). Using Coupled hidden Markov model (CHMM), the scheme is as follows (Figure 3).

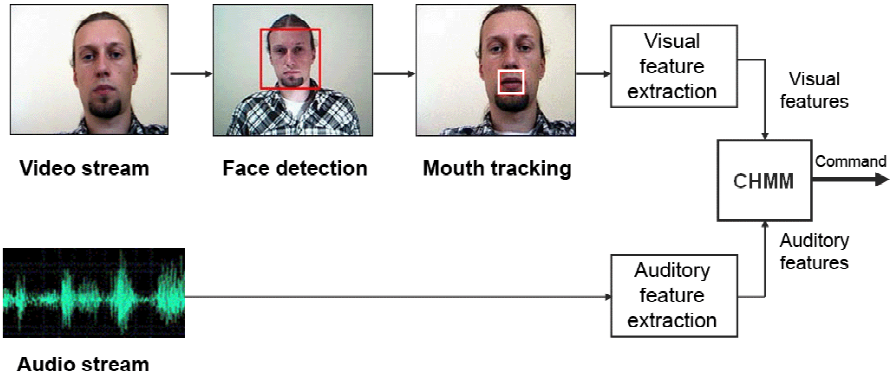

Figure 2. The components of the audiovisual speech recognition system

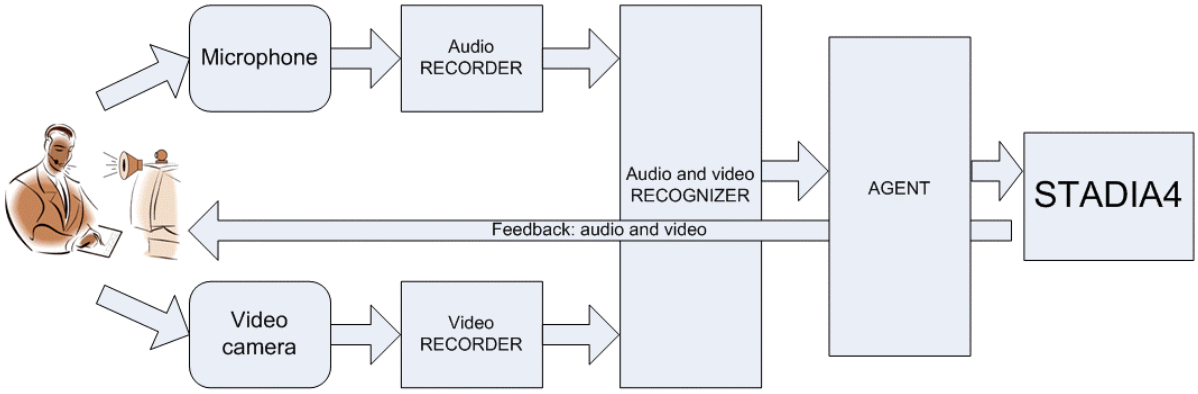

Figure 3. The scheme of interface between investigator and process analysis system STADIA4 when the CHMM is used for recognition

Further we present the visual feature extraction.

The 8 bit images of the user's mouth are analyzed: the pictures are called regions of interest (ROI). Then the visual features are extracted from a sequence of these regions by using the cascade feature extraction system similar to that described in [22, 17]: the sequence of images of ROI is normalized into the pictures of dimension 32x32 and then fed to the system illustrated in Figure 4. First of all, the primary ROI image is mapped into a 32-dimensional space of features by using the principal component analysis (PCA) functions. Afterwards, the digitized signal values are upsampled so that they correspond to the sequence of audio features. Next, the sequence is normalized by using feature mean normalization (FMN) according to the algorithms described in [22]. We obtain a description of the image that models interrelations of image observations. Finally, viseme-based linear discriminant analysis (LDA) is made. As a result, the visual observation vector is obtained.

Audiovisual integration or **feature fusion** is an operation when two streams (audio and visual) are "fused together" and the recognition decision is made. The recognition methods in the frame of audio-visual speech recognition are concentrated on the Hidden Markov model. Some alternative statistical classifications use artificial neural networks [27].

There are three strategies for audio-visual feature integration: the early, intermediate and late integration [21, 26, 9]. In our case we use **intermediate integration**.

In our system, the Coupled Hidden Markov model (CHMM) for audiovisual integration is used [20, 18] (Figure 5).

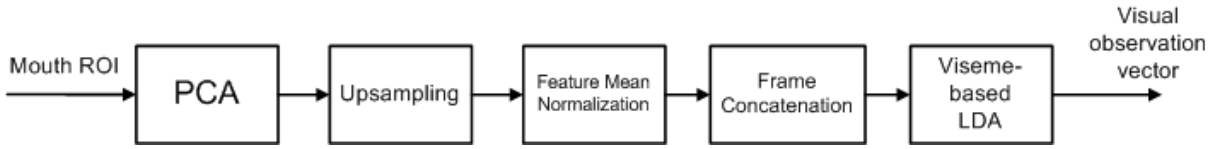

Figure 4. Flow Chart of visual feature extraction

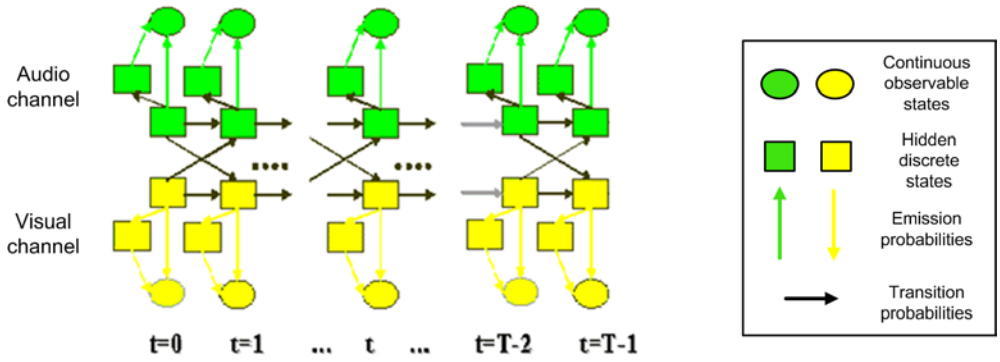

Figure 5. CHMM – Coupled Hidden Markov Model

In Figure 5, squares represent hidden discrete states (audio and video), circles represent the continuous observable states (audio and video), dark arrows represent transition probabilities and light arrows represent emission (observation) probabilities.

The CHMM can model the audio-visual state asynchrony and preserve the natural audio visual dependencies over time through the transition probabilities between the hidden states [18].

In the CHMM, the transition probability of either an audio or video state at time *t* is dependent on the previous audio and video state at time *t*-1. The emission probability of each stream is independent of each other.

Transition probability from state *j* to state *i* in CHMM:

$$a(i \mid j) = P(i_a = q_a^t \mid j_a = q_a^{t-1}, j_v = q_v^{t-1}) \times$$
$$P(i_v = q_a^t \mid j_a = q_a^{t-1}, j_v = q_v^{t-1}).$$

Emission probability in CHMM:

$$b_t(i) = b_t^a(i_a = q_a) b_t^v(i_v = q_v).$$

Where

q_a^t – the current hidden audio state at time t,

q_v^t – the current hidden visual state at time t.

The audio stream in our system can be evaluated using different parameters, for example, mel frequency cepstal coefficients (MFCCs) [28].

Audiovisual interface advantage over the audio-only interface reveals itself in noisy environments. The visual signal is resistant to acoustic noise and in this way it increases the recognition rate. Using the visual stream together with audio, we can avoid some problems specific to audio-only systems. For example, the consonants **m** and **n** are difficult to recognize from the acoustic signal, but using the visual signal, they are easily distinguished (Figure 6). One more example: consonants **p** and **t** (Figure 7).

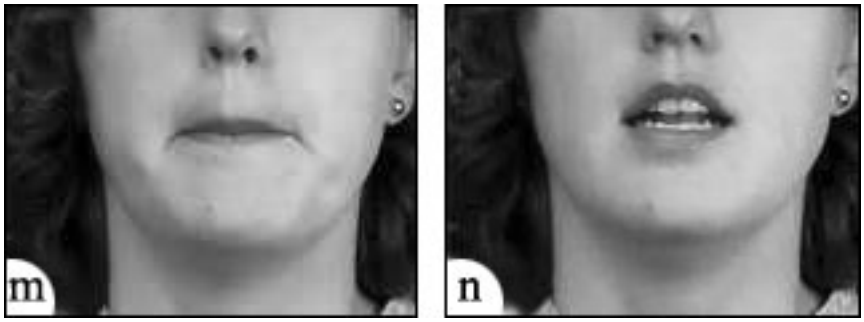

Figure 6. Articulatory image of the consonants *m* and *n*

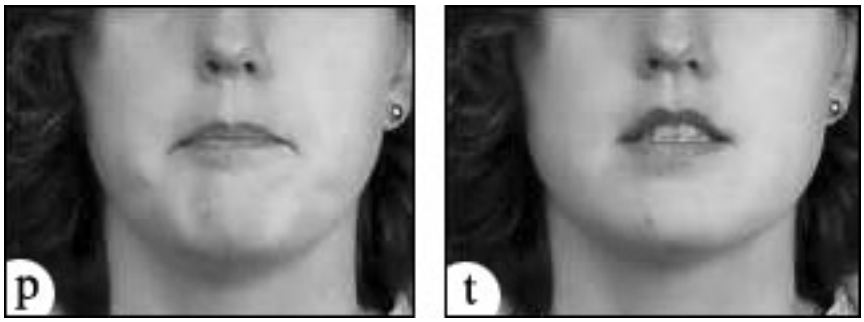

Figure 7. Articulatory image of the consonants *p* and *t*

6. Conclusions

1. To ensure efficient work with stochastic process analysis and modeling system, it is reasonable to invoke new ways of communication between the investigator and computer which realize the conversation between the user and machine.

2. The audiovisual interface gives an opportunity to make the communication between the investigator and special software more reliable.

3. The concept of the audiovisual interface between the investigator and software was presented.

4. The example of the dialogue between the investigator and stochastic process analysis and modeling system STADIA4 was presented.

References

[1] **M. Akay, I. Marsic, A. Medl, G. Bu.** A system for medical consultation and education using multimodal human/machine communication. *IEEE Transactions on Information Technology in Biomedicine*, 1998, 2(4), 282-291.

[2] **A. Ališauskas.** Sutrikusios klausos asmenų vizualinė komunikacija. *ŠPI, Šiauliai*, 1996, 8.

[3] **Y. Bellik**. Media Integration In Multimodal Interfaces. *IEEE First Workshop on Multimedia Signal Processing*, 1997, pages: 31 – 36

[4] **R. A. Bolt**. "Put-that-there": Voice and gesture at the graphics interface. *Proceedings of the 7th annual conference on Computer graphics and interactive techniques*, *ACM Press*, 1980, 262 – 270.

[5] **R. A. Bolt**. Human Interface: Where People and Computers Meet. *John Wiley & Sons Inc., New York, USA*, 1984.

[6] **M. Billinghurst**. Put That Where? Voice and Gesture at the Graphics Interface. *SIGGRAPH Computer Graphics Newsletter* [online], *Vol*.32 *No*.4, November 1998, [*cited* 2006-03-05]. *Available from WWW: <http://old.siggraph.org/publications/newsletter/v32n4/contributions/billinghurst.html>*.

[7] **L. Bretzner, I. Laptev, T. Lindeberg, S. Lenman, Y. Sundblad**. A Prototype System for Computer Vision Based Human Computer Interaction. *Technical report* ISRN KTH/NA/P-01/09-SE, KTH *(Royal Institute of Technology)* [Online], 2001 [*cited* 2006-03-05]. *Available from WWW: <http://www.nada.kth.se/cvap/abstracts/cvap251.html>*.

[8] **S. Card, T. Moran, A. Newell**. The Psychology of Human-Computer Interaction. *Erlbaum, Hillsdale, NJ*, 1983, 7.

[9] **C.C. Chibelushi, F. Deravi, J.S.D. Mason**. A review of speech-based bimodal recognition. *IEEE Transactions on Multimedia*, 2002, *Vol*.4, *Issue* 1, 23 – 37.

[10] **L. M. Encarnação, L. J. Hettinger**. Guest Editors' Introduction: Perceptual Multimodal Interfaces. *IEEE Computer Graphics and Applications, IEEE Computer Society Press Los Alamitos, CA, USA, Vol*.23, *Issue* 5, (*September* 2003), 24 – 25.

[11] **J. Flanagan.** Multimodal Communication for Collaborative Environments. [*Online*] [*cited* 2006-03-05]. *Available from WWW: <http://nsf-workshop.engr.ucf.edu/papers/Flanagan.asp>*.

[12] **J. Flanagan, I. Marsic**. Issues in measuring the benefits of multimodal interfaces. *IEEE International Conference on Acoustics, Speech, and Signal Processing,* ICASSP-97, 1997, *Vol*.1, 163 – 166.

[13] **J. Flanagan, I. Marsic, A. Medl et al**. Multimodal Human/Machine Communication.

[14] **F. J. Huang, T. Chen**. Real-Time Lip-Synch Face Animation driven by human voice. *IEEE Workshop on Multimedia Signal Processing, Los Angeles, California*, 1998.

[15] **D.B. Koons, C.J. Sparrell, K.R. Thorisson**. Integrating simultaneous input from speech, gaze, and hand gestures. *Intelligent Multimedia Interfaces. M. Maybury, Ed. MIT Press*, Menlo Park, CA, 1993, 257–276.

[16] **M. Li, G. Zhang, G. Dai**. A Primitive-Based Architecture of Multimodal Interface (PBA_MMI). *IEEE International Conference on Intelligent Processing Systems, ICIPS*'97, 1997, *Vol*.1, 858 – 862.

[17] **L. H. Liang, X. X. Liu, Y. B. Zhao, X. Pi, A.V. Nefian.** Speaker Independent Audio-Visual Continuous Speech Recognition. *In Proc. of IEEE ICME, Lausanne, Switzerland,* 2002.

[18] **X. Liu, Y. Zhao, X. Pi, L. Liang, A. V. Nefian.** Audio-Visual Continuous Speech Recognition Using A Coupled Hidden Markov Model. *Proc. Internatinal Conference of Spoken Language Processing, Denver,* 2002, 213–216.

[19] **K. Murai, S. Nakamura**. Real Time Face Detection for Multimodal Speech Recognition. *Proc. IEEE International Conference on Multimedia and Expo* (ICME2002), 2002, *Vol*.2, 373-376.

[20] **A. V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, K. Murphy**. A Coupled HMM For Audio-Visual Speech Recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2002, 2013-2016.

[21] **A. V. Nefian, L. Liang, X. Pi, X. Liu, K. Murphy**. Dynamic Bayesian Networks for Audio-Visual Speech Recognition. *EURASIP, Journal of Applied Signal Processing* 11, 2002, 1–15.

[22] **C. Neti, G. Potamianos, J. Luettin, I. Matthews, D. Vergyri, J. Sison, A. Mashari, J. Zhou**. Audio visual speech recognition. *In Final Workshop* 2000 *Report*, 2000.

[23] **S. Oviatt, R. Coultson, R. Lunsford**. When Do We Interact Multimodally? Cognitive Load and Multimodal Communication Patterns. *Proceedings of the* 6*th international conference on Multimodal interfaces, State College, PA, USA*, 2004, 129 – 136.

[24] **S. Oviatt**. Ten Myths of Multimodal Interaction. *Communications of the ACM, ACM Press*, November 1999, *Vol*.42, *Issue* 11, 74 – 81.

[25] **D. Perzanowski, D. Brock, W. Adams, M.Bugajska, A.C. Schultz, J.G. Trafton, S. Blisard, M. Skubic**. Finding the FOO: a pilot study for a multimodal interface. *IEEE International Conference on Systems, Man and Cybernetics*, 2003, *Vol*.4, 3218 – 3223.

[26] **G. Potamianos, J. Luettin, C. Neti**. Hierarchical discriminant features for audio-visual LVCSR. *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP'01), 2001, 165-168.

[27] **I. Shdaifat**. Design of a Visual Front End for Audio-Visual Speech Recognition. *Ph.D. Disertation, Technische Universität Hamburg-Harburg, Hamburg*, 2005.

[28] **D.G. Stork, M.E. Hennecke**, eds. Speechreading by Humans and Machines. *Springer, Berlin*, 1996, 351-371.

[29] **K. Thorisson, D. Koons, R. Bolt**. Multi-Model Natural Dialogue. *CHI 92 Video Proceedings*, 1992, 653. *Also available from WWW: <http://www.open-video.org/details.php?videoid=8113>*.

Received April 2006.