

MULTIMODAL INPUT FUSION IN HUMAN-COMPUTER INTERACTION

On the Example of the NICE Project

A. Corradini (1), M. Mehta (1), N.O. Bernsen (1), J.-C. Martin (2,3), S. Abrilian (2)

(1) Natural Interactive Systems Laboratory (NISLab), University of Southern Denmark, DK-Odense M, Denmark

(2) Laboratory of Computer Science for Mechanical and Engineering Sciences, LIMSI-CNRS, F-91403 Orsay Cedex, France

(3) Montreuil Computer Science Institute (LINC-IUT), University Paris 8, F-93100 Montreuil, France

Abstract: In this paper, we address the modality integration issue on the example of a system that aims at enabling users to combine their speech and 2D gestures when interacting with life-like characters in an educative game context. In a preliminary limited fashion, we investigate and present the use of combined input speech, 2D gesture and environment entities for user system interaction.

Key words: human-computer interaction, input fusion, gesture, speech

“..I feel that as a modern civilization we may have become intoxicated by technology, and find ourselves involved in enterprises that push technology and build stuff just because we can do it. At the same time we are confronted with a world that is increasing needful of vision and solutions for global problems relating to the environment, food, crime, terrorism and an aging population. In this information technology milieu, I find myself being an advocate for the humans and working to make computing and information technology tools that extend our capabilities, unlock our intelligence and link our minds to solve these pervasive problems...” (Thomas A. Furness III [1])

1. INTRODUCTION

Human-Computer Interaction (HCI) is a research area aiming at making the interaction with computer systems more effective, easier, safer and more seamless for the users.

Desktop-based interfaces also referred to as WIMP-based (Windows, Icons, Menus and Pointers) Graphical User Interfaces (GUIs), have been the dominant style of interaction since their introduction in the 80s when they replaced command line interfaces. WIMP interfaces enabled access to computers for more people by providing the user with a look and feel, visual representation and direct control using mouse and keyboard. Nevertheless, they have some intrinsic deficiencies: they passively wait for the user to carry out tasks by means of mouse or keyboard and often restrict input to single non-overlapping events. As the way we use computers is becoming more pervasive, it is not clear how GUI-WIMP interfaces will accommodate for and scale to a broader range of applications. Therefore, post-WIMP interaction techniques that go beyond the traditional desktop metaphor need to be considered.

In the scientific community, a shared belief is that the next step in the advancement of computing devices and user interfaces is not to simply make applications faster but also to add more interactivity, responsiveness and transparency to them. In the last decade much more efforts have been directed towards building multi-modal, multi-media, multi-sensor user interfaces that emulate human-human communication with the overall long-term goal to transfer to computer interfaces natural means and expressive models of communication [2]. Cross-disciplinary approaches have begun developing user-oriented interfaces that support non-GUI interaction by synergistically combining several simultaneous input and/or output modalities, thus referred to as multimodal user interfaces. In particular multimodal Perceptual User Interfaces (PUI) [3] have emerged as potential candidates for being the next interaction paradigm. On one hand, these kinds of interfaces can make use of machine perception techniques to sense the environment allowing the user to use input modalities such as speech, gesture, gaze, facial expression and emotion [4]; on the other they can leverage human perception by offering information and context through more meaningful output channels [5]. As benefits, PUIs will provide their users with reduced learning times, performance increase, an increased retention and a more satisfying usage experience.

So far, such interfaces have not yet reached widespread deployment. As a consequence this technology is not mature and most of these interfaces are still functional rather than social, thus far from being intuitive and natural.

The rigid syntax and rules over the individual modalities along with the lack of understanding of how to integrate them are the two main open issues.

In this paper, we will address the modality integration issue on the example of the NICE (Natural Interactive Communication for Edutainment) [6] project we are currently working on. We begin by giving an overview of multimodal fusion input in the next section. Section 3 presents related work while Section 4 describes the on-going NICE project. We conclude with discussion on other possible applications and future directions for development.

2. MULTIMODAL INPUT FUSION: AN OVERVIEW

In multimodal systems, complementary input modalities provide the system with non-redundant information whereas redundant input modalities allow increasing both the accuracy of the fused information by reducing overall uncertainty and the reliability of the system in the case of noisy information coming from a single modality. Information in one modality may be used to disambiguate information in the other ones. The enhancement of precision and reliability is the potential result of integrating modalities and/or measurements sensed by multiple sensors [7].

In order to effectively use multiple input modalities there must be some technique to integrate the information provided by them into the operation of the system. In the literature, two main approaches have been proposed. The first one integrates signals at the feature level whereas the second one fuses information at a semantic level. The feature fusion strategy is generally preferred for closely coupled and synchronized modalities, such as speech and lip movements. However, it tends not to scale up, requires a large amount of data for training and has high computational costs. Semantic fusion is mostly applied to modalities that differ in the time scale characteristics of their features. In this latter approach, timing plays an important role and hence all fragments of the modalities involved are time-stamped and further integrated in conformity with some temporal neighborhood condition. Semantic fusion offers several advantages over feature fusion. First, the recognizers for each single modality are used separately and therefore can be both trained separately and integrated without retraining. Furthermore, off-the-shelf recognizers can be utilized for standard modalities e.g. speech. An additional advantage is simplicity: modalities integration does not add any extra parameters beyond those used for the recognizers of each single mode allowing for generalization over number and kind of modalities.

Typically, the multimodal fusion problem is either formulated in a maximum likelihood estimation (MLE) framework or deferred to the decision level when most of the joint statistical properties have been lost. To make the fusion issue tractable within the MLE framework, the individual modalities are usually assumed independent of each other. This simplification allows the use of simple parametric models (e.g. Gaussian functions) for the joint distributions that cannot capture the complex modalities' relationships.

Very few alternatives to these classical approaches have proposed to make use of non-parametrical techniques or finite-state devices. [8] put forward a non-parametrical approach based on mutual information and entropy for audio-video fusion of speech and camera-based lip-reading modalities at signal level. Such a method does not make any strong assumptions about the joint measurement statistics of the modes being fused, nor does it make use of any training data. Nevertheless, it has been demonstrated over a small set of data while its robustness has not been addressed yet. In [9] multimodal parsing and understanding was achieved using a weighted finite-state machine. Modality integration is carried out by merging and encoding into a finite-state device both semantic and syntactic content from multiple streams. In this way, the structure and the interpretation of multimodal utterances can be captured declaratively in a context-free multimodal grammar. Whereas the system has been shown to improve speech recognition by dynamically incorporating gestural information, it has not been shown to provide superior performance, either in terms of error rate reductions, or in terms of processing speed, over common integration mechanisms. More importantly, it does not support mutual disambiguation (MD), i.e., using speech recognition information to inform the gestural recognition processing, or the processing of any other modality.

The kind of fusion strategy to choose may not depend upon the input modalities only. There is an empirical evidence [10] that distinct individual groups (e.g. children and adults) adopt different multimodal integration behaviors. At the same time, multimodal fusion patterns may depend upon the particular task at hand. A comprehensive analysis of experimental data may therefore help gather insights and knowledge about the integration patterns thus leading to the choice of the best fusion approach for the application, modalities, users and task at hand.

The use of distributed agent architectures, such as the Open Agent Architecture (OAA) [11], in which dedicated agents communicate with each other by means of a central blackboard, is also common practice in multimodal systems.

Besides architectures aiming at emulating the way human beings communicate with each other in their everyday lives, a variety of other multimodal systems have been proposed for recognition and identification of individuals based on their physiological and/or behavioral characteristics. These biometric systems address security issues with the purpose to ensure that only legitimate users access a certain set of services, e.g. secure access to buildings, computer systems and ATMs. Biometric systems typically make use of either fingerprints or iris or face or voice or hand geometry to assess the identity of a person. Because of issues related to non-universality of some single traits, spoof attacks, intra-class variability, and noisy, data architectures that integrate multiple biometric traits have shown substantial improvement in efficiency and recognition performance [12, 13, 14, 15]. Being a non issue for such systems, user traits temporal synchronization makes signal integration less complex than in HCI architectures and can be seen as a decision problem within a pattern recognition framework. Techniques employed for combining biometric traits range from the weighted sum rule [16], Fisher discriminant analysis [16], decision trees [15], to a decision fusion scheme [17].

3. RELATED WORK

Several multimodal systems have been proposed after Bolt's pioneering system [18]. Speech and lip movements have been merged using histogram techniques [19], multivariate Gaussians [19], artificial neural networks (ANNs) [20, 21] or hidden Markov models (HMMs) [19]. In all these systems, the probabilistic outputs of modalities have been combined assuming conditional independence by using either Bayes' rule or a weighted linear combination over the mode probabilities for which the weights were adaptively determined.

While time synchrony is inherently taken care of (at least partially) in the ANN-based systems described in [20, 21], this cannot be adequately addressed in the other systems. To address temporal integration of distinct modalities, a generic framework has been put forward in [22]. It is characterized by three steps and makes use of a particular data structure named melting pot. The first step, referred to as microtemporal fusion, combines information that is produced either in parallel or over overlapping time intervals. Further, macrotemporal fusion takes care of either sequential inputs or time intervals that do not overlap but belong to the same temporal time window. Eventually, contextual fusion serves to combine input according to contextual constraints without attention to temporal constraints.

In speech and gesture systems it is common to have separate recognizers for each modality. The outcome of the single recognizers may be used for further monomodal processing at a higher level (e.g. a natural language understanding module to deal with the spoken input representation from the speech recognizer) and/or followed by the late fusion module. QuickSet [23] is a multimodal pen-gestures and spoken input system for map-based applications. A multi-dimensional chart parser semantically combines the statistically ranked set of input representations using a declarative unification-based grammar [24]. Temporal fusion relies on time proximity: time-stamped features from different input channels are merged if they occur within a 3 to 4 second time window.

In [25], two statistical integration techniques have been presented: an estimate and a learning approach. The estimate approach makes use of a multimodal associative map to express, for each multimodal command, the meaningful relations that exist between the set of single constituents. During multimodal recognition, the posterior probabilities are linearly combined with mode-conditional recognition probabilities that can be calculated from the associative map. Mode-conditional recognition probabilities are used as an approximation of the mode-conditional input feature densities. In the learning approach, called Members to Teams to Committee (MTC), multiple teams are built to reduce fusion uncertainty. Teams are trained to coordinate and weight the output from the different recognizers while their outputs are passed on to a committee that establishes the N-best ranking.

The EMBASSI system [26] combines speech, pointing gesture and the input from a graphical GUI into a pipelined architecture. The Smartkom [27] is multimodal dialogue system that merges gesture, speech and facial expressions for both input and output via an anthropomorphic and affective user interface. In both systems, input signals are assigned a confidence score that is used by the fusion module to generate a list of interpretations ranked according to the combined score.

4. THE NICE PROJECT

4.1 The NICE Project and Its Multimodal Scenario

The NICE PC-based system aims at enabling users to combine their speech and 2D gestures when interacting with characters in an educative game context. It addresses the following scenario. 3D animated life-like fairy tale author Hans Christian Andersen (HCA) is in his 19th Century

study surrounded by artifacts. At the back of the study is a door which is slightly ajar and which leads out into the fairy tale games world. This world is populated by some of his fairy tale characters and their entourage, including, among others, the Naked Emperor and the Snow Queen. When someone talks to HCA, this user becomes an avatar that walks into HCA's study. In the study, the user can have spoken conversation with HCA, including the use of gesture input to support interaction by e.g. indicating artifacts during conversation. At some point, the user may wish to visit the fairy tale world and is invited by HCA to go through the door at the back of the study. Once in the fairy tale world, the user may interact with the characters populating the fairy world using speech and 2D gesture. The intended users are primarily kids and youngsters and, secondarily, everyone else. The primary scenario of use is in technology and other museums in which, expectedly, the duration of individual conversations will be 5-30 minutes. Secondarily, we investigate the feasibility of prototyping the world's first spoken computer game for home use with its average of 30 hours of user interaction time.

The primary research challenge addressed in NICE is to move from the existing paradigm of task-oriented spoken dialogue with computer systems to the next step which we call domain-oriented spoken dialogue. In domain-oriented spoken dialogue, there is no longer any user task to constrain the dialogue and help enormously in its design and implementation, but only the semi-open domain(s) of discourse which, in the case of HCA, are: his life, his fairy tales, his 3D physical presence, his modeling of the user, and his role as kind of gate-keeper for the virtual fairy tales world. In a limited fashion, however, we also investigate the use of combined input speech and 2D gesture for indicating objects and other entities of interest.

4.2 Requirements for Multimodal Input from Experimental Data

Early multimodal prototypes have been developed without much knowledge about how the potential final users would combine the distinct modes to interact with the system. This design approach has changed over the years and it is now considered important to collect behavioral data prior to and/or while the design phase via a simulation of the future system using a Wizard of Oz (WoZ) approach. In this kind of study, an unseen assistant plays the role of the computer, processing the user's input and responding, as the system is expected to.

In order to collect data on the multimodal behavior that our future system might expect from the users, we have built a simple 2D game application. In

this application the user can interact with several 2D characters located in different rooms to which he/she has to bring some objects back. The user can issue spoken input and/or pen-gesture to accomplish the desired task. In the following, we focus on how we are currently taking these observations into account for the specification and development of a first demonstrator of the NICE multimodal module.

The observed commands were classified into six sets: *getIn* where the user wants to get in a room from the corridor, *askWis* when the user asks the character for an object, *getOut* when the user wants to leave the room he/she is currently in, *takeObject* when the user wants to take an object in the current room and later hand it over to another character, *giveObject* when the user wants to give an object to the character in the current room and this is placed in a deposit area graphically visible from the interface, and finally *social dialogues* when the user utterance is not directly related to the task at hand.

By analyzing the way the user carried out these commands, we were able to detect a few common multimodal patterns useful for the design of the multimodal module. For example, we were able to find out that a few single commands are always issued unimodally (e.g. when the user utters “What do you want?” without any accompanying gesture) while others are issued indifferently either unimodally, with no dominant modality (e.g. in the case of the user either uttering “get into the red room” to express the wish to enter a red painted room or just circling the door of the red room), or multimodally (providing both spoken and gestural input to the system). In case of multimodal commands, we have seen that gesture always precedes speech and this is consistent with previous empirical evidence [28]. Other commands were noticed to use multiple gestures in sequence (e.g. to get into a room the user clicks on a door and then circles it). Also, gesture-only commands have at present a high semantic variability which can be resolved only if information about location of the gesture or the object is known (e.g. drawing a circle about an object in the room means *takeObject* whereas the same gesture referring to an object in the deposit area means *giveObject*). Eventually, few unexpected speech and gesture combinations were observed such as when the user utters “thank you” while for instance performing a *takeObject* gesture. The observed gestures were classified into the following shape categories: pointing (makes up for 66% of the data), circling (18,1%), line (5.4%), arrow (2.1%) and explorative gestures (8.5) i.e. those that occur when the user gestures without touching the screen. Accurate details on the experiment and its results can be found in [29].

4.3 Gesture Recognition Module

While both pointing and exploring categories observed in the corpus do not need any specific recognition algorithm, to recognize circling, line and arrow, a 2D gesture recognition module was developed using Ochre Neural Networks technology [30] trained with templates extracted from the experimental data corpus. The approach is easily extendable to more gestures and other patterns may be added later if it will turn out necessary.

An N-best hypotheses list results from the gesture classification task. The list is wrapped into an XML-like format that has been agreed upon to allow messages to be exchanged by the different modules.

4.4 The Speech Processing Module

In order to test the input fusion we developed a very simple speech processing module to provide input to the Input fusion module. So far, a fairly simple speech grammar has been manually specified out of the set of utterances in the corpus. 94 sentences were defined: 18 formulations of the *askWish* command, 15 for *giveObject*, 37 for *takeObject*, 16 for quitting and 8 for greetings. We used the off-the-shelf IBM ViaVoice [31] technology as speech recognizer. Currently, no natural language processing module is employed. In addition, the grammar being very limited no conversation dialogue is possible with the system. In the near future, we will be adding a natural language processing module to add partial dialogue conversation capabilities. Similarly to the gesture modules, the speech processing results in an XML-like message to be passed on to the input fusion component.

4.5 Input Fusion

The input processing architecture of the NICE system has been specified as shown in Figure 1. The speech recognizer sends a word lattice including prioritized text string hypotheses about what the user said to the natural language understanding module (NLU), which parses the hypotheses and passes a set of semantic hypotheses to the input fusion module. In parallel, the gesture recognizer sends hypotheses about the recognized gesture shape to the gesture interpreter. The gesture interpreter (GI) consults with the simulation module (SM) to retrieve information on relevant objects visible to the user, interprets the gesture type, and forwards its semantic interpretations to the input fusion module. The input fusion module combines the information received and passes on its multimodal input interpretation to the dialogue manager (DM).

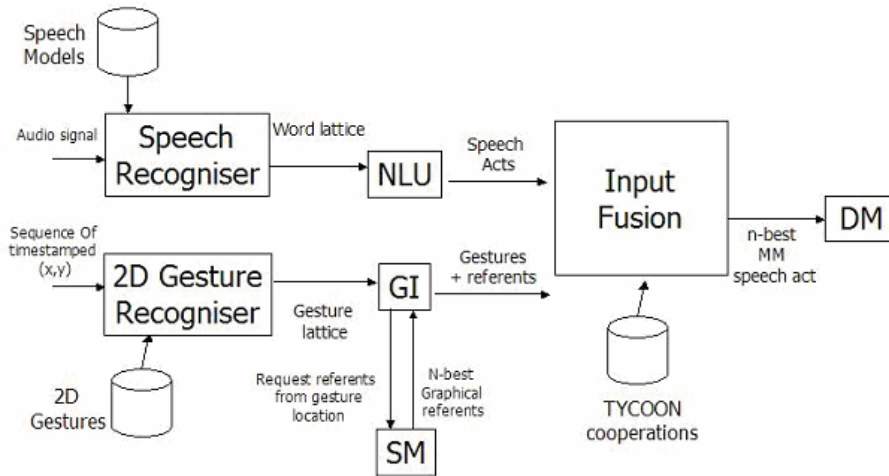


Figure 1. Sketch of the NICE input processing architecture

In previous work [32, 33] we have proposed a typology made of several types of cooperation between modalities for analyzing and annotating the user’s multimodal behavior and also for specifying interactive multimodal behaviors. Basic types of cooperation among modalities are: *equivalence* to specify modalities that occur interchangeably in the same unimodal command, *specialization* for commands that are always specified with the same modality, *redundancy* for modalities that either combined or taken separately produce the same command, and *complementarity* for modalities that need to be merged to result in a meaningful command. We have also included the notion of referenceable objects to specify entities the user can refer to using uni- or multimodally utterances.

We utilize a text file to contain the description of the expected modalities combination where the variables are defined and reused later by multimodal operators such as specialization, complementarity, etc.. For example, a *giveObject* command can be specified using the following text script

```

...
#- giveObject command
specialisation CC3 IS3
specialisation CC4 IG1

semantics CC4 position

complementarity temporalProximity 5000 CC5 CC3 CC4
endHypothesis CC5 giveObject
...

```

Here, IS3 stands for one of the possible utterances associated with a *giveObject* spoken command, IG1 stands for the detection of a gesture associated to the gestural part of the same command, and the CC# tags are contextual units which are activated by different multimodal patterns. For example CC5 gets activated if CC3 and CC4 are activated within a 5000ms time window. The multimodal module [33] parses this text file and makes use of the TYCOON symbolic-connectionist technique to classify multimodal behaviors. TYCOON was inspired by the Guided Propagation Networks [34] that are composed of processing units exchanging symbolic structures representing multimodal fusion hypotheses.

4.6 Input fusion and Message Passing: an Example

The following example illustrates the result of the fusion given the incoming messages from the distinct modes. The messages were generated when the user, after asking permission for picking up an object (a coffee machine), uttered “thanks” while pointing to the object.

OUTPUT FROM SPEECH PROCESSING MODULE

```
<semanticRepresentation>
  <score>0.8</score>
  <function>thank</function>
</semanticRepresentation>
```

OUTPUT FROM GESTURE RECOGNITION MODULE

```
<recognisedGesture>
  <hyp n="1">
    <score>0.75</score>
    <shape>point</shape>
    <begin> ...</begin>
    <end>...</end>
    <2DboundingBox>...</2DboundingBox>
  </hyp>
  <hyp n="2">
    <score>0.2</score>
    <shape>line</shape>
    <begin> ...</begin>
    <end>...</end>
    <2DboundingBox>...</2DboundingBox>
    <direction>...</direction>
  </hyp>
</recognisedGesture>
```

OUTPUT FROM GESTURE INTERPRETER MODULE

```
<semanticRepresentation>  
<score>0.75</score>  
<function>takeObject</function>  
<object>coffeeMachine#1</object>  
</semanticRepresentation>
```

OUTPUT FROM INPUT FUSION

```
<semanticRepresentation>  
<score>0.9</score>  
<function>takeObject</function>  
<object>coffeeMachine#1</object>  
</semanticRepresentation>
```

In this example, the function and the object were not provided by speech but by gesture. Yet, the compatible fusion enables the increase of the score of the command after merging hypothesis from speech processing and gesture recognizer.

5. CONCLUSION AND FUTURE DIRECTIONS

There is evidence that people are polite to the computer they are using, treat them as member of the same team but also expect them to be able to understand their needs and be capable of natural interaction. In [35], for instance, is reported that when a computer asked a human being to evaluate how well the computer had been doing, the individual provides more positive responses than in the case of a different computer asking the same question. Likewise, it was shown that people tend to give computers higher performance ratings if the computer has recently praised the user. In light of these inclinations, systems making use of human-like modalities seem to be more likely to provide users with the most natural interface for many applications. Humans will benefit from this new interface paradigm as automatic systems will capitalize on the inherent capabilities of their operators, while minimizing or even eliminating the adverse consequences of human error or other human limitations.

The rigid syntax and rules over the individual modalities along with the lack of understanding of how to integrate them are the two main open issues in the development of multimodal systems. This paper provided an overview of techniques to deal with the latter issue and described the fusion in the ongoing NICE project. The current version of the input fusion module will have to be improved in the following directions: recognize more complex

and multi-stroke gestures, integrate with the other modules such as the NLU and the 3D environment, and add environment information to resolve input ambiguities.

To illustrate this latter issue, suppose, for instance, that the user says, “What is written here?” whilst roughly encircling an area on the display. Let’s assume the speech recognizer passes on hypotheses, such as “what is it gray here”, “what does it say here”, along with the correct one, while the gesture recognizer passes on hypotheses, such as that the user wrote the letter Q and that the user drew a circle. The simulation module would inform the gesture interpreter that the user could have referred to the following adjacent objects: a bottle up front on the display and a distant house. We would refer to these objects as environment content. Eventually, the input fusion module will have to combine the time-stamped information received from the natural language understanding and gesture interpretation modules, select the most probable multimodal interpretation, and pass it on to the dialogue manager. The selection of the most probable interpretation should allow ruling out inconsistent information by both binding the semantic attributes of different modalities and using environment content to disambiguate information from the single modalities [36].

Multimodal fusion can be adopted to deal with either multimodal sensors or multimodal inputs or a combination of the two. Several relevant families of applications could benefit from an accurate and reliable fusion integration strategy. Possible applications range from gesture-cum-speech systems for battlefield management [23, 37], biometric systems [15], remote sensing [38], crisis management [39], to aircraft and airspace applications [40].

ACKNOWLEDGMENTS

The support from the European Commission, IST Human Language Technologies Programme, IST-2001-35293 is gratefully acknowledged.

REFERENCES

- [1] Furness, T.A. III, *Towards Tightly Coupled Human Interfaces*, In: *Frontiers of Human-Centred Computing, Online Communities and Virtual Environments*, Earnshaw, R., Guedj, R., van Dam, A., and Vince, J., eds, pp. 80-98, 2001
- [2] Bernsen, N.O., *Multimodality in language and speech systems. From theory to design support tool*, In: Granström, B., House, D., and Karlsson, I. (Eds.): *Multimodality in Language and Speech Systems*, Dordrecht: Kluwer Academic Publ., pp. 93-148, 2002
- [3] <http://www.cs.ucsb.edu/conferences/PUI/index.html>

- [4] Picard, R.W., *Affective Computing*, MIT Press, 1997
- [5] Turk, M., *Perceptual User Interfaces*, In: *Frontiers of Human-Centred Computing, Online Communities and Virtual Environments*, Earnshaw, R., Guedj, R., van Dam, A., and Vince, J., eds, pp. 39-51, 2001
- [6] www.niceproject.com
- [7] Kittler, J., On Combining Classifiers, *IEEE Transactions on PAMI*, vol. 20, no 3, pp. 226-239, 1998
- [8] Fisher, J.W., and Darrell, T., *Signal Level Fusion for Multimodal Perceptual Interface*, In: *Proceedings of the Conference on Perceptual User Interfaces*, Orlando, Florida, 2001
- [9] Johnston, M., and Bangalore, S., *Finite-state Multimodal Parsing and Understanding*, *Proceedings of the International Conference on Computational Linguistics*, Saarbruecken, Germany, 2000
- [10] Xiao, B., Girand, C., and Oviatt, S.L., *Multimodal Integration Patterns in Children*, *Proceedings of 7th International Conference on Spoken Language Processing*, pp. 629-632, Denver, Colorado, 2002
- [11] Cheyer, A., and Martin, D., *The Open Agent Architecture*, In: *Journal of Autonomous Agents and Multi-Agent Systems*, vol. 4, no. 1, pp. 143-148, March 2001
- [12] Dieckmann, U., Plankensteiner, P., and Wagner, T., *Sesam: A biometric person identification system using sensor fusion*, In: *Pattern Recognition Letters*, vol 18, no 9, pp. 827-833, 1997
- [13] Kittler, J., Li, Y., Matas, J., and Sanchez, M.U., *Combining evidence in multimodal personal identity recognition systems*, *Proceedings 1st International Conference on Audio-Video Personal Authentication*, Crans-Montana, Switzerland, pp. 327-334, 1997
- [14] Maes, S., and Beigi, H., *Open sesame! Speech, password or key to secure your door?*, *Proceedings 3rd Asian Conference on Computer Vision*, Hong-Kong, China, pp. 531-541, 1998
- [15] Ross, A., Jain, A., *Information Fusion in biometrics*, *Pattern Recognition Letters*, vol. 24, no. 13, pp. 2115-2121, 2003
- [16] Wang, Y., Tan, T., Jain, A.K., *Combining Face and Iris Biometrics for Identity Verification*, *Proceedings of 4th International Conference on Audio- and Video-Based Biometric Person Authentication*, Guildford, UK, 2003
- [17] Jain, A., Hong, L., and Kulkarni, Y., *A Multimodal Biometric System Using Fingerprint, Face and Speech*, *Proceedings of 2nd Int'l Conference on Audio- and Video-based Biometric Person Authentication*, Washington D.C., pp. 182-187, 1999
- [18] Bolt, R.A., *Put that there: Voice and gesture at the graphic interface*, *Computer Graphics*, vol. 14, no. 3, pp. 262-270, 1980
- [19] Nock, H. J., Iyengar, G., and Neti, C., *Assessing Face and Speech Consistency for Monologue Detection in Video*, *Proceedings of ACM Multimedia*, Juan-les-Pins, France, 2002
- [20] Meier, U., Stiefelhagen, R., Yang, J., and Weibel, A., *Towards Unrestricted Lip Reading*, *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 14, no. 5, pp. 571-585, 2000
- [21] Wolff, G.J., Prasad, K.V., Stork D.G., and Hennecke, M., *Lipreading by neural networks: visual processing, learning and sensory integration*, *Proc. of Neural Information Proc. Sys. NIPS-6*, Cowan, J., Tesauro, G., and Alspector, J., eds., pp. 1027-1034, 1994
- [22] Nigay L., and Coutaz, J., *A Generic Platform for Addressing the Multimodal Challenge*, *Proceedings of CHI'95, Human Factors in Computing Systems*, ACM Press, NY, pp. 98-105, 1995

- [23] Cohen, P.R., Johnston, M., McGee, D.R., Oviatt S.L., Pittman, J., Smith, I., and Clow, J, *Quickset: Multimodal Interaction for Distributed Applications*, In: Proceedings of the 5th International Multimedia Conference, ACM Press, pp. 31-40, 1997
- [24] Johnston, M., *Unification-based multimodal parsing*, Proceedings of the 17th International Conference on Computational Linguistics, ACL Press, pp. 624-630, 1998
- [25] Wu, L., Oviatt, S.L., Cohen, P.R., *Multimodal Integration – A Statistical View*, IEEE Transactions on Multimedia, vol. 1, no. 4, pp. 334-341, December 1999
- [26] Elting, C., Strube, M., Moehler, G., Rapp, S., and Williams, J., *The Use of Multimodality within the EMBASSI System*, M&C2002 - Usability Engineering Multimodaler Interaktionsformen Workshop, Hamburg, Germany, 2002
- [27] Wahlster, W., Reithinger, N., and Blocher, A., *SmartKom: Multimodal Communication with a Life-Like Character*, Proceedings of Eurospeech, Aalborg, Denmark, 2001
- [28] Oviatt, S.L, DeAngeli, A., and Kuhn, K., *Integration and synchronization of input modes during multimodal human-computer interaction*, Proceedings of the Conference on Human Factors in Computing Systems (CHI '97), ACM Press, New York
- [29] Buisine, S., and Martin, J.-C., *Experimental Evaluation of Bi-directional Multimodal Interaction with Conversational Agents*, Proceedings of INTERACT, Zurich, Switzerland, pp. 168-175, 2003
- [30] <http://www.hhs.net/tiscione/applets/ochre.html>
- [31] <http://www.ibm.com/software/speech/>
- [32] Martin, J.C., Julia, L., and Cheyer, A., *A Theoretical Framework for Multimodal User Studies*. Proceedings of 2nd International Conference on Cooperative Multimodal Communication, Theory and Applications, Tilburg, The Netherlands, 1998
- [33] Martin, J.C., Veldman, R., and Beroule, D., *Developing multimodal interfaces: a theoretical framework and guided propagation networks*, In: Multimodal Human-Computer Communication. Bunt, H., Beun, R.J. & Borghuis, T. (Eds.), 1998
- [34] B eroule, D., *Management of time distortions through rough coincidence detection*, Proceedings of EuroSpeech, pp. 454-457, 1989
- [35] Reeves, B., and Nass, C., *The media equation: How People Treat Computers, Television, and New Media Like Real People and Places*, Cambridge University Press, Cambridge, 1996
- [36] Kaiser, E., Olwal, McGee, D.R., A., Benko, H., Corradini, A., Li, X., ., Cohen, P.R., and Feiner, S, *Mutual Disambiguation of 3D Multimodal Interaction in Augmented and Virtual Reality*, In: Proceeding of the International Conference on Multimodal Interfaces, Vancouver (BC), Canada, pp. 12-19, 2003
- [37] Corradini, A., *Collaborative Integration of Speech and 3D Gesture for Map-based Applications*, In: Proceedings of the International Conference on Computational Science, Workshop on Interactive Visualization and Interaction Technologies, Lecture Notes in Computer Science 3038, Springer Verlag, pp. 913-918, 2004
- [38] Aleotti, J, Bottazzi, S., Caselli, S., and Reggiani, M., *A multimodal user interface for remote object exploration in teleoperation systems*, IARP International Workshop on Human Robot Interfaces Technologies and Applications, Frascati, Italy, 2002
- [39] Kraahnstoever, N., Schapira, E., Kettebekov, S., and Sharma, R., *Multimodal Human-Computer Interaction for Crisis Management Systems*, IEEE Workshop on Applications of Computer Vision, Orlando, Florida, 2002
- [40] Blatt, M., Grossman, T., and Domany, E., *Maximal a-posteriori multi-sensor multi-target neural data fusion*, submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), can be downloaded from: <http://www.weizmann.ac.il/~fedomany/papers.html>