Issues related to the functions of visual displays in multimodal communication

Loredana Cerrato

Abstract

Multimodal communication has locked my research interest since I attended the Elsnet Summer School MILASS in 1999[1] organised by the department of Speech Musing and Hearing of KTH.

Since the field of multimodal communication is very wide, in this review paper I will just introduce the concept of multimodal communication and I will then focus my attention on multimodal conversational computer interfaces.

I will briefly review some methodologies used in the development of computer animated faces, and I will discuss in particular some articles dealing with issues related to the display of facial expressions in multimodal conversational computer interfaces.

Index

[1] http://www.speech.kth.se/milass/

1.Multimodal communication

When humans communicate with each other signals from multiple channels are at work. We communicate not only through words, but also by intonation, gaze, hand and body gestures and facial expressions. These verbal and nonverbal signals have a role in the communicative process. They can add, modify, substitute information in discourse; they are highly linked with each other and continually in play complementing and supporting speech activity.

When we listen to a speaker we perceive not only semantic information conveyed by the words he/she is producing, but also regulative information about the conversational structure of the utterance. We also perceive evidential information marking the speaker identity and his or her affective state.

Communication between humans usually involves different modalities, these modalities are:

- **Verbal**: the uttered words and sentences.
- **Prosodic**: the speech rhythm, the pauses, the intensity, stress and intonation that characterise speech.
- **Gestural**: the movement of the hand and the arm that co-occur with speech.
- **Facial**: the movement of the head and eyes, the gaze, smiles and the facial expressions.
- **Bodily**: the posture and movement of the trunk and leg that co-occur in speech communication.

The ease and robustness of human-human communication is due to extremely high recognition accuracy (using multiple input channels) and the redundant and complimentary use of several modalities. For instance when speech is produced together with gesture and vision, it may result in a more robust, more natural and more efficient communication.

For people with hearing impairment the benefit of multi-modal transmission is very obvious: they use the visual information in support of the audio information they lack. Also in case of communication occurring in particular condition of noise the support from the visual channel can play an important role in conveying the message even for people with normal hearing.

Human computer interaction can benefit from modelling several modalities in analogous ways; therefore bringing this multimodal communication ability in the field of human-machine communication has become a big challenge.

The following ironic quotation of Descout [1992] summarise quite well the aim of multimodality, that is making man-machine interactions more robust more natural and more pleasant:

> *"the development of man-machine communication was accompanied by severe cuts in our perceptions which have transformed people into perceptually disabled persons.*
> *The celebrated extension of our senses by these technical surrogates has always been a limitation on dialogue.*
> *As a consequence, ingenious techniques for introducing intelligence into man-machine dialogue procedures have been proposed for replacing the dramatic lack of redundancy.*
> *In fact introducing naturalness into human or man-machine communication calls for a"multimodal dialogue".*

2.Multimodal conversational computer interfaces

Multimodal dialogue systems can facilitate communication since the interpretation of the communicative acts can be based on input from different modalities. At the same time devices and errors in one channel can be compensated by the information coming from another channel. Different channels can be audio, video and tactile.
For instance the prototype dialog system developed in the ESPIRIT MASK project-Multimodal-Multimedia Automated Service Kiosk-[2] enables interaction between the users and the system through the coordinated use of multimodal inputs (speech and touch) and multimedia output (sound, video, text, and graphics) for a ticket reservation task.
At Carnegie Mellon University a multimodal tourist assistant system has been developed. It consists of a multimodal interface (speech, gesture and handwriting) for an appointment-scheduling task [Yang et al. 1999].
The different modalities can be used to enter commands or to provide missing information or to solve an ambiguity, following a request from the system.
Anyway, since one of the most important forms of embodiment for social interaction is the face, most multimodal system display animated faces.
Demos of dialogue system with talking heads as agents have been developed at the department of speech music and hearing at KTH in the past years, with the main aim of testing if multimodality really increase the robustness and the performance of a dialogue system and to analyse various aspects of human-computer interaction in multimodal conversational dialogue systems.
The most recent multimodal system is called AdApt, reported in fig.1. Its domain is the real-estate. The design of the system allows multimodal interaction with its agent, Urban, that gives information about apartments in the Stockholm area, talking and showing illustration on the screen [Gustafson et al. 1999].

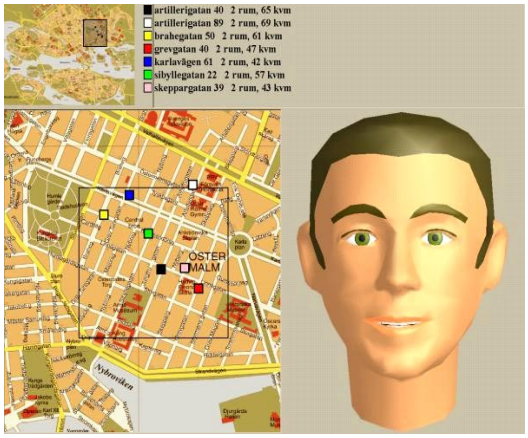

Fig. 1 Urban, the agent of the Multimodal dialogue system ADAPT .

2.1 Facial animation

The development of a computer animated face is not a simple task, but great effort has been devoted to this task and the actual results are quite good, two examples of extremely good and natural looking animated faces are given by ANANOVA,[3] the on-line new reader that resembles Kylie Minogue and the recent Japanese movie "Fantasy" which displays different realistic talking heads.

[2] http://m17.limsi.fr/Recherche/TLP/mask.html
[3] http://www.ananova.com/

In computer-based facial animation, three methods are mainly in use:

- concatenative,
- parametric models and
- muscle based models.

In **concatenative approach**, digitising and storing different faces create a library of expressions. Interpolating between two stored images, using key frame techniques, creates the animation. This method can be quite tedious and time-consuming since one needs to restart the whole digitization process each time a new model is animated.

The **parametric model** consists of a 3D structure of the face, which can be modified and deformed by the action of parameters [Parke 1983, Cohen & Massaro1993, Lundberg & Beskow 1999].
Generally, parametric models differentiate two sets of parameters: conformation parameters, that control the topology of the face and expression parameters, that control brow action, mouth shapes and so on.
Changing the values of control parameters and re-drawing the face by using the new values create facial animation. This approach has the advantage of being quite simple and efficient, as it requires low data storage.

In the **muscle based** approach, skin properties and muscle actions are simulated using an elastic spring mesh and forces. Waters proposes a parameterised facial muscle process to create more realistic facial animation [Waters 1987]. The principal drawback of these models is the high computational load.
The creation of facial animation control parameters is a tedious process at best.
Often several tens of parameters have to be specified and coordinated to create just a short animation sequence. For the face it is important to orchestrate the parameters carefully, as actions have to be timed and overlapped precisely to create believable facial expressions. For example, to create a believable expression of surprise, the onset and duration have to be correctly specified otherwise the motion will look more like a lazy yawn.

2.2 Audio-visual speech synthesis.
Animated faces are usually also able to automatically generate voice and facial animation from arbitrary text. This process is referred to as “audio-visual speech synthesis”.
The idea at the basis of this development is that visual information conveyed by the face can significantly improve intelligibility of synthetic speech, especially under degraded acoustic conditions because of noise, bandwidth filtering, or hearing-impairment [Summerfield 1979, Massaro 1987]
The strong influence of visible speech is not limited to situations with degraded auditory input, however. A perceiver's recognition of an auditory-visual syllable reflects the contribution of both sound and sight.
When an auditory syllable /ba/ is dubbed onto a videotape of a speaker saying /ga/, subjects perceive the speaker to be saying /da/. This effect, which is called “Mc Gurk” after the name of the researcher that detected it [Mc Gurk 1976] gives evidence that synthetic faces increase the intelligibility of synthetic speech. However this is true only under the condition that facial gestures and speech sounds are coherent. This means that in development of talking head it has been very important to synchronise the articulatory gestures with the speech production.

Not only asynchrony or incoherence between the two modalities doesn't increase speech intelligibility; but also it might even decrease it.
One of the primary goals of research in the field of audio-visual speech production has been to accurately measure the articulatory movement in humans and reproduce them in the talking head [Beskow 1995, Engwall 2001, Pelachaud 2001] then trying to synchronise these movement with speech synthesis in order to produce intelligible visual synthesis. Even if audiovisual speech synthesis is quite intelligible compared to natural speech, all the talking head still lack naturalness.

3 Facial displays

In order to make a talking head look more natural, pleasant and intelligible not only the articulatory movements of the lips, jaw and tongue should be carefully synthesised, but also a series of "non verbal behaviour" should be reproduced.
Much information related to phrasing, stress, intonation and emotions is expressed but a series of non-verbal behaviour conveyed by the face. These movements are referred to as "visual displays" and have been classified by Cassel [Cassel 2000] according to their placement with respect to the linguistic utterance and their significance in transmitting information. The most evident facial displays are:

- Eybrow position
- Expression of the mouth
- Movement of head
- Movement of the eyes

Facial displays carry out very important functions in communication, some of them have phonological functions (for instance the lip shapes that change according with the phoneme uttered) some fulfil a semantic function (for instance nodding instead of saying yes) some are used to cement social relationships (courtesy smile) and some correspond to grammatical functions, in fact much information related to phrasing, stress, intonation are expressed by nodding the head, raising and shaping of the eyebrow, eye movements and so on. This class of movements related to prosodical-intonational characteristics have been referred to as "visual prosody" [Granström1998]. There have been some attempts to display the most evident of these movements in connection with pitch movements, like raising the eyebrows at the end of a question or during the production of a stressed syllable [http://www.speech.kth.se/multimodal/].
Some other facial display are linked to speakers personality and remain constant across a life time, while some are linked to the emotional state of the speaker and may last as long as the emotion is felt. The facial displays related to the emotional state of the speaker are modelled according to the six universal basic emotions defined by Ekman [1979] and are reproduced on talking heads by modifying a series of parameters [Pelachaud 1996]
Many head movements have the special function to structure interactions, in particular to segment turn regulation and to give and elicit feedback expressions (head nods) [Godwin 980].
Other analyses have revealed that head movements in conversations have both cognitive and interactive functions, for instance the results of a microanalysis conducted at the California University State [Mc Clave 1998] have shown that a later sweep of the head can coincide with verbalizations such as "all, everything, whole", a

lateral shake of the head accompanying an affirmative statements can indicate some uncertainty of the speaker.
Some other facial displays are synchronised with the units of conversation and last only a very short time (eyebrow raise along with verbs that refer to special actions.)
All these movements related to non-verbal behaviours of communication are more difficult to model than the articulatory movements, so more effort is necessary to find a appropriate methodology to study them in human communication and reproduce them in the development of conversational speech interfaces.
As conversational speech interfaces become more advanced and human-computer dialogues appear more "natural", we may expect users of spoken dialogue systems to integrate a larger number of human discourse features into their speech. Therefore it is necessary to determine which visual displays are related to discourse features, like turn taking, feedback expressions, in order to model them and reproduce them in talking heads.

3.1 Visual correlates of emotions

Human facial expressions provide information about emotions; to enable also talking heads to display different moods, the six basic emotions: happiness, anger, fear, surprise, disgust, sadness- are currently implemented by modifying a series of parameters, on the base of a model proposed by Pelachaud 1996.
A good attempt to reproduce the basic emotions in talking heads is shown in the system August, developed at TMH-KTH. August has a library of communicative gestures of varying complexity and purposes, ranging from primitive punctuators, such as blinks and nods, to complex gestures tailored for particular meanings, like for instance: “eyebrow frowning when thinking or disagreeing. The various gestures are not only used to communicate emotions and attitudes, but also to signal turn-taking and to highlight information in speech, such as stressed syllables and phrase boundaries.
Each gesture is defined as a set of parameter tracks, which can be invoked in any point in time, either in-between or during the utterance.[Lundeberg & Beskow 1999][4].
The six basic emotions expressed by August are reported in fig. 1.

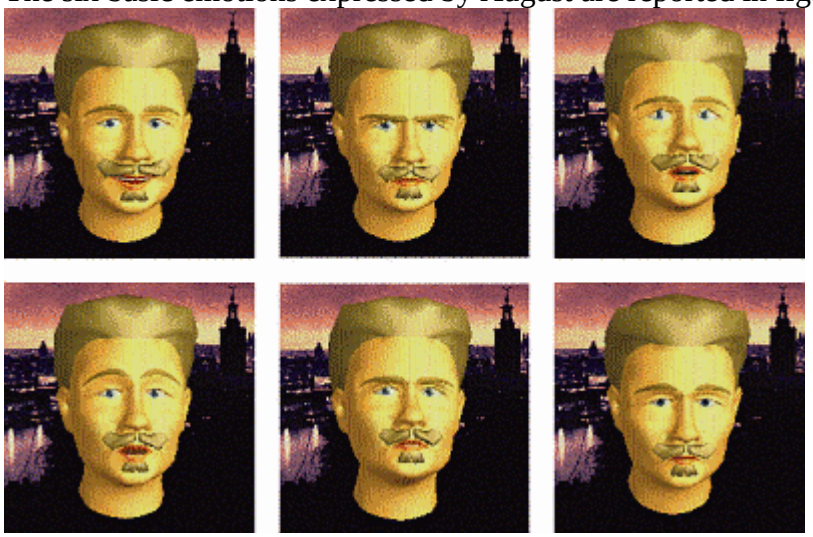

Fig. 1 Basic emotions in August:
Happiness anger fear
Surprise disgust sadness

[4] A nice demo of the basic emotions reproduced in talking can be seen at http://www.speech.kth.se/multimodal/emotions/.

The same parameters can be modified in any of the talking heads developed at TMH. In figure 2a is reported a talking head displaying "fear". The same emotion displayed by the talking head developed in Italy [Pelachaud 2001] is reported in figure 2 b.
The two emotions have been implemented following the same model, but if we compare the two expressions of fear we can understand that the ways in which some non-verbal behaviours are produced do certainly differ from culture to culture.
Maybe it is just not an impression that Italian facial and bodily gestures are more evident than the gestures produced by Northen European people!

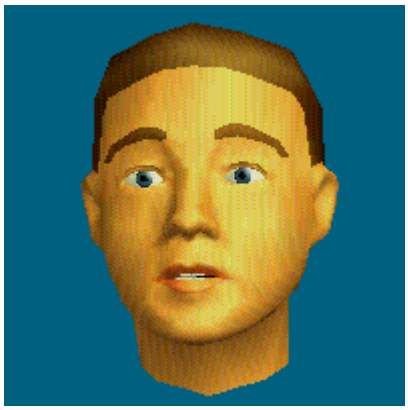

Fig. 2 a "Fear" displayed by a Swedish talking head

Fig. 2b "Fear" displayed by an Italian talking head

3.2"Visual prosody"

Facial movements have been proven to have strong relationship with speech, in particular with the prosodical-intonational characteristics of utterances: emphasis, structure of information (given-new, topic-comment) discourse segmentation.
In a recent article Julia Hirshberg [2000] provides a useful review of the functions of prosodic variation in human to human communication which could be used as a starting point for application in the design of spoken dialogue systems.
She believes that prosody could be used more effectively to improve naturalness in dialogue systems.
Prosodical cues can improve also the naturalness of multimodal conversational agents.
It has been proven for instance that some movements -like raising the eybrows or nodding the head- occurring simultaneously with stressed syllables, are powerful visual cues for prominence [House 2001]. These movements not only transmit important non-verbal information related to prosody, but they also make the face look alive and more natural since they can convey certainty, uncertainty and questioning behaviour.
In order to make conversational agents look more natural it is very important that the reproduced facial displays are appropriate and at the right time, in fact a wrong facial expression at the wrong time can be a source of misunderstanding and can convey the wrong message.

3.3 Visual displays for turn regulation

Many studies have been carried out in the past 30 years to try to understand how human face-to-face dialogue is organised [Douglas 1995, Grice 1975].
One of the most important structures that emerged from these studies is the organisation around the "turn taking system". The purpose of this system is to manage the flow of interaction, minimizing overlapping speech and pauses. Turn taking has

the properties of requiring real-time responsiveness and concurrent input and output. The inputs and outputs are multimodal and they include speech gestures and other visible behaviours, like eye gaze, blinks and also hand and body movements; for instance it has been observed that to give the turn speakers can gaze at addressee, without moving hands and arms, while to take the turn speakers gaze at addressee, gesticulating at the same time. These movements of the heads and gaze that in human to human communication are used to regulate the flux of conversation, could be reproduced also in dialogue systems to signal when the agent is keeping the floor and when it wishes to relinquish it. Much confusion in spoken language systems arises in fact when users become confused about when they are expected supply input and when the system is still processing their prior utterance and not listening for a new one.

3.4 "Visual feedback"

In dialogues (not only face to face dialogues) interaction quickly breaks down if communication only happens at or above the turn taking level. There needs to be a two-way incremental exchange of information. Part of the task for a listener is to make sure that the other person taking part in the conversation knows that he-she is paying attention, and indicate that he-she is at the same state in the conversation. This is mainly done by means of feedback expressions: an exchange of information that support the interaction. Feedback can be expressed vocally, verballly, gesturally or by any combination of the three.

A few studies have already been carried out to find out which are the basic ways of expressing verbal and bodily feedback during a conversation [Alwood & Nivre 1993, Cerrato 2000], even if there are cultural and speaker-dependent variables, there seem to be general ways of expressing feedback:

- by means of short verbal expressions like: *yes, no* together with some phonological phenomena of vocalic lengthening
- by means of short non lexical items like: ah, ah ah, mhmh
- by repeating either the last word uttered by the interlocutor or by repeating the core words of the last sentence by other types of reformulation of the meaning of the received message.
- by means of head nods, by rising the eyebrow or by making specific hand movements.

In current multimodal dialogue systems elicit visual feedback expressions are not produced by the agents, while in human-human conversation, dialogue participants continuously give each other positive and negative feedback as a way of showing attention, recognizing the intention what the other conversant is saying or to signal non-understanding or misunderstanding.

3.5 "Visual correlates of speech acts"

Some recent studies ave been concerned with the process of understanding how the communicative intention of the speaker in performing communicative acts[5] is

[5] Austin was the first one to draw attention to the many functions performed by utterances as a part of interpersonal communication. In particular he pointed out that humans do not just perform actions randomly, but they plan their actions to achieve various goals, and in the case of communicative actions, those goals include changes to the mental states of the listeners. When for instance a priest or a justice of peace says to a couple " I now declare you to be Husband and Wife" the utterance not only communicates the intention to "join them in the holy sacrament", but is equivalent to the action of "joining them in the holy sacrament" and therefore it conveys a new psychological or social reality.

communicated through facial expressions and how can these expressions be simulated in animated faces system.
For instance in the ADAPT system the agent called "Urban" has an associated library of gestures representing communicative functions that can be used in the dialogue. Actions are triggered by the state of the agent in such a way that appropriate gestures are automatically selected when he enters, exits or remains in a particular state (like speaking or attending etc.)
In a recent article Poggi & Pelachaud [2000] propose a meaning-to face approach aiming at a face simulation automatically driven by semantic data.
Following Austin's and Searle's theory of speech acts [Austin1962, Searle 1969], they believe that utterances are not simply strings of words, but rather are the observable performance of communicative actions, or speech acts, such as requesting, informing, warning, suggesting, and confirming. Moreover humans do not just perform actions randomly, but plan their actions to achieve various goals, and in the case of communicative actions, those goals include changes to the mental states of listeners. For instance speakers' requests are planned to alter the intentions of their addressees. Every sentence has a literal goal, explicitly stated by its literal meaning, but beyond its goal it might have one or more super-goals. A super-goal is not explicitly stated by the literal meaning, but is to be drawn by the addressee through inferential work.
For instance the sentence "*can you pass me the salt*" is a request having the literal meaning and goal to get the salt. But a super-goal could be inferred by the context and shared knowledge. So for instance if we insert the sentence in the context of a familiar dinner, where the wife is talking to her husband, the sentence could be interpreted as *" stop putting salt in your dish, watch your blood pressure!"*
Every goal can be expressed by a performative[6]. Since performatives are semantically richer than the single goals, they can be distinguished in term of specific features and can be representable in terms of cognitive units.
As an instance of feature that distinguish performatives from each other, we can consider the feature of "interest", in other words we can ask whose goal does the requested action serve? In whose interest?
If a speaker says to a person: "can you get me the newspaper" we can well think that he is pursuing his own goal since it is in his interest to receive the newspaper and read it. While if a speaker says to a person "take your umbrella with you when you go out" the speaker is suggesting that taking the umbrella would prevent the person to get wet, which is not a goal of the speaker, since he is speaking in the interest of the receiver of the message. So Poggi and Pelachaud state that every communicative act has two faces:

- a signal, or the visual realisation, which consists in the muscular actions performed or the morphological features displayed-said, the vocal articulation for speech acts and facial,
- a meaning, which consists in the set of goals and beliefs that the sender wants to transfer to the receiver of the message[7].

---

[6] Performative: Austin stressed that every sentence has a performative aspect, since it performs some action. In fact, any sentence performs a locutionary, an illocutionary and a perlocutionary act: it is an act of doing something physically, but in being uttered (in locution) it also performs a social action, and through this (per locution) it may also have some effect on the receiver of the message. The illocutionary force of the sentence (the type of action it performs) is its performative, and it can be made explicit verbally by performative verbs like promise, command, order.

The signal part of a communicative act may be represented formally by facial actions, phonetic articulation, intonation contours and the meaning may be presented in terms of logical propositions that Poggi and Pelachaud call for “cognitive units”.
Their hypothesis is then that to each cognitive unit, which defines a performative, is associated one or more non-verbal behaviour. For example performatives whose general type of goal is to request are signalled by “head kept right”.
The cognitive structure of the communicative act is combined with the contextual information.
On the basis of this hypothesis they are developing a system endowed of a “performative library”. The system should be able to generate a complete cognitive structure for every performative, thanks to the support of the semantic analysis and of the conextual information. Once identified the performative, the system should be able to combine to it all non-verbal signals specified for each cognitive unit related to the given performative.
Even if this system is still a project rather than an applicaqtion, it seems to be quie interesting, since among the available multimodal dialogue systems, there is no agent able to display visual correlates of communciative acts.

4. Conclusions

Research and development in the field of multimodal communication is a very large domain and there are many aspects that still need further investigations.
In the past 10 years very good results have been achieved in the development of audio-visual speech synthesis, and several talking heads have been inserted in multimodal dialogue systems. Unfortunately these talking head still lack naturalness and this is due to the fact that researchers have mostly focused their attention on the study and reproduction of the articulatory movements of the lips jaw and tongue, without making too much effort in reproducing a series of facial expression, like eyebrow position, expression of the mouth and movement of head and eyes that absolve very important functions in communication and make the face look more natural and pleasant.

Only few recent studies have been focusing on the observation of facial movements related to speech and the results have shown the existence of strong relationships between facial gestures and the prosodical-intonational characteristics of utterances: such as emphasis, structure of information (given-new, topic-comment) discourse segmentation, turn regulation. Moreover some results have also shown that some facial expressions can convey information about the communicative intention of the speakers.

[7] This description finds its root in the classical linguistic descriptions of the structuralisms, in particular Ferdinande de Saussure and his description of a sign, which is a two faced entity composed by a sequence of sounds*:* signifier or *signifiant* in the original French and a meaning: signified or *signifié* in French.

References

Alwood & Nivre 1992 :Allwood J., Nivre J., 1992, On the semantics and pragmatics of Linguistic feedback Journal of Semantics

Austin 1962: Austin J., How to do things with words OUP

Beskow 1995 Beskow J, Rule-based visual speech synthesis Proceedings of Eurospeech 1995 Spain, 299-302

Cassel 2000: Cassel J., Sullivan J., Prevost S., Churchill E., Embodied conversational agents MIT press 2000

Cerrato 2000: Cerrato L. Il feedback verbale in dialoghi elicitati con la tecnica del map task, Atti delle X Giornate del Gruppo di Fonetica Sperimentale Napoli, 1999

Cohen & Massaro 1993: Cohen M., Massaro D., Modeling co-articulation in synthetic visual speech . In Thalmann N. and Thalman D (eds.)Models and techniques in Computer animation. Tokyo: Springer –Verlag 1993 p.139-156

Descaut 1992: Descaut R.,"Visual Speech Synthesis"in Talking Machines: Theories, Models And Design. Bailly G. Benoit C. eds , 1992 pp. 475-478

Douglas 1995: Douglas S., Conversational analysis and Human-computer interaction Design, in Thomas P.(ed.) The social and interactional dimensions of human-computer interfaces. CUP 1995

Ekman 1979: Ekman P. About brows: Emotional and conversational signals. In Cranach K. et al. (eds.) Human ethology : Claims and limits of a new discipline: Contributions to the Collooquium pp. 169-248 CUP 1979

Engwall 2001: Engwall O. Considerations in Intraoral Visual Speech Synthesis: Data and Modeling In Proc of 4th International Speech Motor Conference, 23-26.

Goodwin1980: Goodwin M.H. Processes of mutual monitoring implicated in the production of description significancy

Granström 1998: Granstrom B., Multi-modal Speech Synthesis with aplication Speech processingrecognition and artificial neural networks III International school on Neural nets Eduardo Canciello edited by Crollet G., Di Benedetto M. Esposito A., Springer 1988

Grice 1975: Grice H.,Logic and conversation In Cole P., Morgan J eds Syntax and semantics: Speech Acts NT Academic Press

Gustafson J. et al 1999: Gustafson J., Bell L., Beskow J., Boye J., Carlson R., Edlund J., Granström B., House D., Wirén M. AdApt- a Multimodal conversational dialogue system in a n apartament domain …

Hirshberg 2000: Hirshberg J., Communication and prosody: functional Aspects of prosody – To appear in Speech Comm

House 2001: House D., Interaction of visual cues for prominence Lund University Working Papers 49, 2001

Lundberg & Beskow 1999: Lundberg M., Beskow J., Developing a 3D-Agent for the august dialogue system AVSP 1999

Massaro 1997: Massaro D. Perceiving Talking facesFrom Speech perception to a behavioural Principle Cambridge Massachussets

Mc Gurk 1976: Mc Gurk H MacDonald j 1976 Hearing lips and seeing voices Nature 264, p. 746-748

McClave 1998: McClave E., Cognitive and Interactional Functions of Head Movements in Conversation Oralité et Gestualité ed. By S. Santi et al. 1998 p. 366-370

Park 1982: Park F., Parametrized models for facial animation. IEEE Computer Graphics 2 (9) 1982 pp. 61-68

Pelachaud 1996: Pelachaud C. Badler N. Steedman M. 1996 Generating facial expressions in speech cognitive Science 28, p. 1-46

Pelachaud 2001: Pelachaud C., Magno Caldognetto E., Zmarich C., Cosi P., An approach tyo an Italian talking head proceedings of Eurospeech 20001 p. 1035-1039

Poggi & Pelachaud 2000: Poggi I., Pelachaud C., Performative Facial Expressions in Animated faces, in Cassel J., Sullivan J., Prevost S., Churchill E., Embodied conversational agents MIT press 2000 pp. 155-187

Searle 1969: Searle J.R Speech Acts, CUP

Summerfield 1979: Summerfield A.Q., Use of visual information in phonetic perception. Phonetica, 36, 314-331

Waters 1987: waters K. A muscle model for animating 3-dimensional facial expressions, Computer Graphics, 21, p.17-24

Yang 1999: Yang J., Yang W. Denecke M., Waibel, Smart sight: a tourist assistant system, ISWC'99 San Francisco, California 1999