Analysis and measurement of head movements signalling feedback in face-to-face human dialogues

Loredana Cerrato* & Mustapha Skhiri¤

*TMH/CTT, Dep. Speech, Music and Hearing, Royal Institute of Technology (KTH), Stockholm,
¤Dep. of Computer and Information Science (IDA), Linköping.
*¤Swedish Graduate School of Language Technology (GSLT)

Abstract

The aim of our study is twofold: explore how people use specific head movements to signal feedback and show evidence that it is possible to measure and quantify the extent of these movements with the Qualisys MacReflex motion tracking system.

Naturally elicited dialogues between humans have been recorded and analysed with focus on those gestures that occur contemporarily with the production of short verbal feedback expressions.

The results show that head nods and jerks are typical movements involved during the production and elicitation of feedback. These movements produce specific patterns and can be easily measured and quantified and eventually implemented in animated talking heads.

1. Introduction

When human communicate with each other, signals from multiple channels are at work. Communication takes place not only through speech, but also by means of several gestures, such as facial expressions, gaze, head movements, hand movements and body posture. These gestures do not take direct part in the articulation of speech, but they are produced to provide complementary information to speech, therefore they are referred to as "accompanying gestures" [Teston 1998]. Accompanying gestures can be produced to serve important syntactic and prosodic functions. For instance eye brow rises often accompany F0 rises [Cavè et al. 1996] [Graf et al. 2002], head movements are often produced to punctuate a stressed syllable, different gestures are also produced to convey communicative intentions, feelings and attitudes, consciously or unconsciously [Bolinger 1985].

For our investigation we had originally planned to observe gestures related to different dialogic functions, such as turn-taking and feedback. Turn-taking gestures consist usually of sequences of quick head movements coupled with changes of gaze direction and eyebrow rising. For instance a speaker who wishes to yield the turn can produce first a quick movement of the head towards his/her interlocutor, directing his/her gaze there; once facing his/her interlocutor, another slightly forward movement of the head is typically produced, often accompanied by eyebrow rising.

The sequence of head movements and gazes produced during turn-taking is to be considered more as

a continuum of movements than a sequence of distinct movements, for this reason it is quite difficult to isolate the single movements on the recording data and analyse them in details.

As a consequence we decided to focus our attention on the gestures related to feedback. These gestures consist mainly[1] of head movements, which are easy to isolate because they are mostly produced contemporarily with the production of short verbal feedback expressions such as: yes, no, mh, certainly and so on.

Feedback is here intended in accordance to [Alwood et al. 1993] as a mechanism "which enables the participants of a conversation to unobtrusively exchange information about four basic communicative functions: contact, perception, understanding and attitudinal reactions". This information can be exchanged by means of verbal and vocal expressions and also by means of gestures.

Researchers in the field of human-machine interfaces, aware of the important role that feedback gestures can play in communicative exchanges, are starting to integrate some of them in the design and development of dialog systems endowed with embodied conversational agents [Beskow et al.2002, Thorrrison 2002, Poggi & Pelachaud 2000]. We believe that the production of an accurate model of feedback gestures realization is a time-consuming process, which requires first an extensive and detailed analysis of feedback gestures production in real communicative situation by human beings. For this reason we recorded subjects interacting with each-other in a lab-environment in order to get some examples of spontaneous feedback gestures production.

2. Experimental Design

2.1. Subjects and recording session

Among the many difficulties that the analysis of gestures implies, probably the most crucial one is that of elicitation.

While different eliciting techniques have been developed to induce subjects to produce more or less spontaneous speech in somewhat controlled situation [Gybbon et al 1997], it is still very hard to find a good method of eliciting spontaneous gestures in a controlled recording session. This is primarily due to the fact that communicative gestures have not been studied and coded as much as speech, so it is quite difficult to predict exactly when they are going to occur in a spontaneous conversation.

For our data acquisition we used a recording set-up, which had been previously used with success to record and measure articulatory gestures for reproduction in visual synthesis [Granström et al 2002, Magno Caldognetto & Zmarich 1999, Hällgren & Lyberg 1998].

The set-up allows the recording of audio and visual data; audio data is recorded on a Dat-tape and visual data is recorded both by means of a video camera and with the optical motion tracking system Qualysis[2] .

Since our study is based on the assumption that understanding how humans use gestures in dialogues can be very useful in the design of more natural-looking animated talking heads, and since we fore-

1. Smiles, laughter, gaze and shoulder shrug are also used to signal feedback, however for this investigation we only looked at head movements
2. Qualysis MacReflex Motion Tracking System: http://www.qualisys.se (July 2003)

see that our results might be implemented in animated conversational agents, we reproduced a communicative scenario similar to the one that might arise between a user and an embodied conversational agent in a dialogue system, that is: "information seeking".

In our scenario there are two dialogue participants: A, who has the role of "information seeker" and B who has the role of "information giver". The information exchanged is relative to movies: plots, schedules and so on.

The focus of the recording was on participant B, the "information giver", and only his movements were recorded. However the audio recordings included the production of both subjects.

In both dialogues the "information giver" was a male speaker; as a consequence we have data from 2 male speakers, which from now onward we will refer to as subject-1 and subject-2.

Figure 1 reports a reproduction of the recording session, with participant B facing participant A, the video camera and the four infra-red cameras[1] .

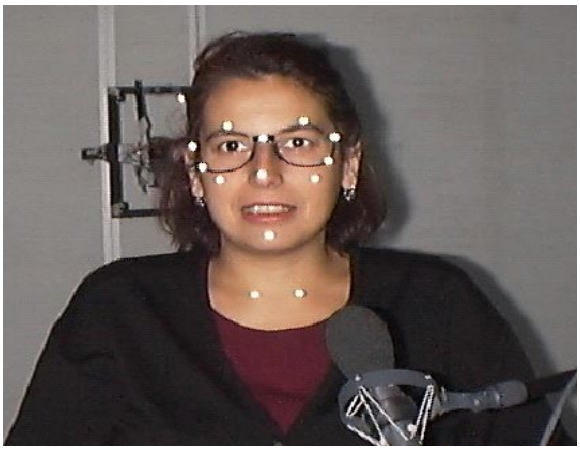

Figure 1: *the displacement of the reflecting markers on subject B during the recordings.*

1. The people reported in figure 1 and 2 are reproducing the experimental set-up, but are not the subjects used in this study

Figure 2: *subject B, whose head movements are to be recorded, sits facing four infrared cameras, a digital video recorder, a microphone and his/her interlocutor.*

A total of thirteen hemispherical markers were used for the 3-D recording. Six markers were attached on the subject's face, two on the chest. Moreover to recover the rigid 3D motion of the head, the subjects wore special glasses, with five markers on. The markers are ca. 5 millimetres wide and reflect infra-red light. This way they are visible in the dark and easily trackable by IR sensitive cameras.

The Qualisys system uses 4 infra-red cameras to recover the full 3D motion of each marker, operating at about 60 frames per second.

Each recording session lasted about 15 minutes. During the first couple of minutes of each session the participants got acquainted with each other and with the recording environment so to feel at ease when starting the actual task. None of the subject thought that wearing the markers and the glasses during the recording was uncomfortable.

The movements of the markers in three dimensions were stored together with the recorded acoustical and video signals.

2.2. Technical equipment

3D recordings: 3D recordings are obtained by calculating 3D coordinates from four cameras with different viewing angles.

Before any measurements were recorded, we calibrated the system to determine the geometrical relation between the image planes of the cameras and the coordinate system of the volume to be measured. The calibration is performed with the help of a calibration frame placed in the measurement volume prior to the measurements.

Video recordings: a Sony digital videocamera DCR-PC-115 E, focusing on speaker B was placed 2 meters away from the speaker in the recording studio. The video recording signal was then digitalized in order to be used for the detailed analysis.

Audio recording: a microphone SHURE Model 16A was placed at an appropriate distance from the speaker in focus in order to assure good quality of audio-recording.

3. Analysis and results

3.1. Transcriptions of dialogues

The recorded dialogues were orthographically transcribed with the support of the Multitool package software et [Allwood et al. 2002]. Multitool is a research tool for the analysis of digitized audiovisual data. Multitool has a partiture-based visual representation, which displays a multi-tier annotation of the phenomena under investigation. It simultaneously displays the video and the relative orthographic transcription of the dialogues, so that the operator can easily observe when gestures are produced together with speech. Multitool gives also the possibility to code gestures in temporal alignment with speech.The time alignment is necessary to obtain perfect synchronisation with the 3D recordings.

3.2. Coding and Alignment of Gestures

To code gestures we adapted a coding schema originally proposed by [Alwood 2001] and Cerrato [2002]. We annotated the type of gesture and its function in the given context. This information was displayed, in alignment with the dialogue transcription, on two tiers displayed on the partiture of Multitool.

To categorize the type of head movements the following labels were used:

Nod: a forward movement of the head up and down, which can be single or multiple.

Jerk: a backward movement of the head, which is usually single

Shake: a left-right or right-left movement of the head, which can be single or multiple

Waggle: a movement of the head back and forth, left to right

Side-way turn: a single turn of the head left or right

To code the possible function of feedback gestures the labels reported in table 1 were used.

Function	Label
Continuation, Perception, Understanding	FBCPU
Acknowledgement	FBA
Refusal	FBR
Expressive	FBE

Table 1: Schema of the labels used to code the communicative function of feedback

3.3. Analysis of the 3D data

Thanks to the time alignment it was possible to isolate the coded gestures in the measurement data recorded with Qualisys. For each head movement a 2D curve was plotted, the curve displays the amplitude of the gesture in millimeters on the Y axis and the duration of the gesture in milliseconds on the X axis.

The curves were plotted by means of the software package Wavesurfer [Sjölander & Beskow 2000].

By looking at the curves representing each movement, we tried to see if there was a one-to-one relationship between a specific verbal expression and its "accompanying" gesture.

4. Results

The measurement data are full of information related to the different movements the subjects made, the thirteen markers can in fact measure both head movements and facial expressions (like eyebrow rising, smiles). However for this paper we restricted our analysis to the data related to the movements coupled with feedback expressions, data related to other gestures are published elsewhere [Cerrato & Skhiri 2003].

Table 2 reports a list of the occurrence of the gestures related to feedback that we observed and coded in the audio-video recordings

Gestures	Related expression	Function	Subject1	Subject2
single nod	*ja, mh, javisst*	FBCPU,FBA	1	2
multiple nod	*ja, mh, javisst, jusste*	FBCPU,FBA	5	3
single jerk	*ja, jaha*	FBCPU,FBA	3	2
multiple jerk			-	-
side way-turn		FBCPU,FBU	2	-

Table 2: Occurrences of the head movements accompanying feedback expressions

Head nods and jerks are movements accompanying the production of short feedback expressions such as "ja" -yes- and "mh" produced with the main function of showing continuation of contact, perception and understanding of the message (label: FBCPU). These movements have been coded as "single" or "multiple". In the case of a head nod for instance, this means that it is possible to observe the head go up and down once (single nod) or more than once (multiple nods).

The difference between single and multiple nods is evident to detect on the measurement data, since every nod is represented by a single arc/peak as shown in figure 3.

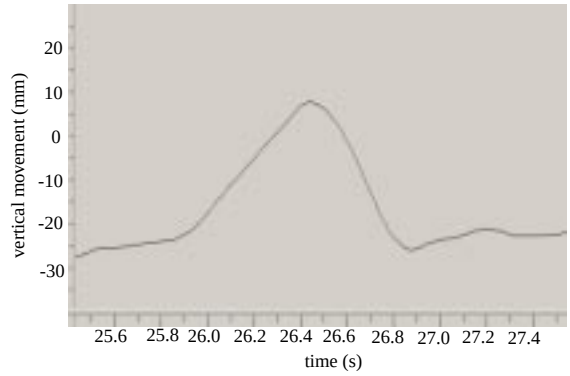

Figure 3: *Curve of a nod produced together with the feedback expression ”mh” by subject-1.*

Figure 4 reports another head movement produced by the same speaker: a jerk produced when saying "ja" with FBCPU function

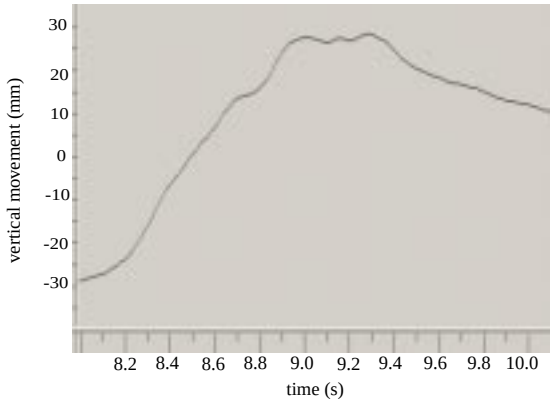

Figure 4: *Curve of a jerk, which is a backword movement of the head, produced together with the feedback word ”ja” yes by subject-1.*

As we can see the curves are different, but both the head movements (nod and jerk) were produced by subject-1 with the same function: "show continuation of contact perception and understanding".

The curves reported in figure 5 and figure 6 represent the movement coded as multiple nod produced by respectively subject-1 and subject-2. These multiple nods were accompanying two different verbal feedback expressions, respectively "ja visst" and "jusste" which however were used with the same meaning (certainly) and with the same function, that is to accept the information received (label:FBA).

Even if the curve in figure 5 shows three peaks and the curve in figure 6 shows two peaks (each peak corresponding to a single nod) the two curves show a similar shape. This seems to supports the idea that it is possible to identify a general pattern for each specific movement.

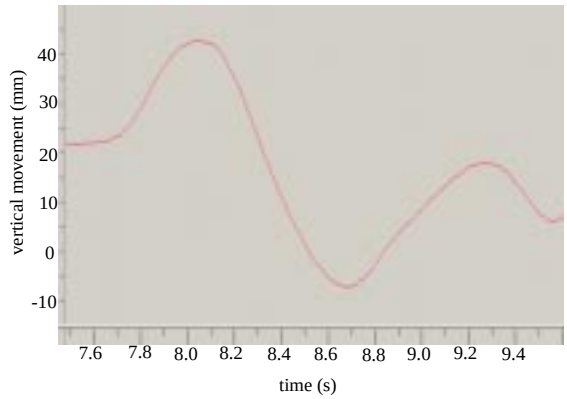

Figure 5: *Curve of multiple nod produced with the feedback expression ”javisst” (certainly) produced by subject-2.*

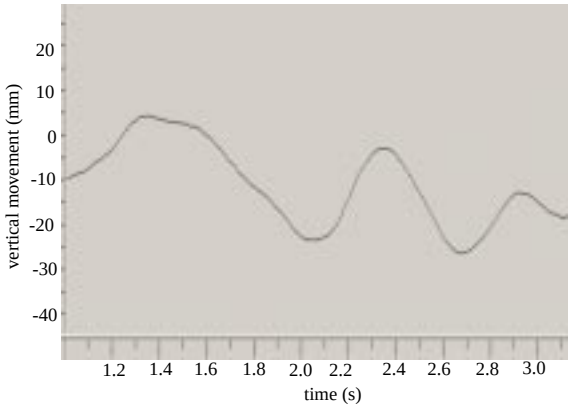

Figure 6: *Curve of multiple nod produced with the feedback expression ”jusste” (just that) produced by subject-1.*

However if we look at the examples reported in figure 3 and 4 we can observe that the same subject (1) produces two different movements, (a nod and a jerk), coupled to two different feedback expressions ("mh" and "ja") which nonetheless convey the same meaning: showing continuation of attention, perception and understanding (label: FBCPU).

These results show that it is not possible to establish a one-to-one correspondence between a movement and a function. In fact different movements can be produced with the same function.

5. Concluding remarks and future work

The aim of this study was to explore the potential of analysing how people use specific head movements that serve feedback functions, and propose a method for the recording, measurement and quantification of these movements which could be useful in the development of a model of human gestures to implement in virtual agents.

The result show evidence that it is possible to identify a general pattern for each specific movement even if it is not possible to identify a one-to-one relationship between a specific verbal expression and its "accompanying" gesture.

The method of analysis and measurement we tested in our experiment seems to be quite useful to extract data related to the extent and the duration of different gestures.

However our experimental set up has shown some limitations that can be improved in future collection of data. One limitation is for instance that only the subject with the markers on his face was video-recorded, as a consequence it was not possible to observe how dialogue participants mimicked each other's gesture, how they exchanged gaze and so on.

In the future we will improve the video recording by using two video-cameras to record both interlocutors and we will also synchronize the video recording camera with the tracking system.

Another limitation is that the subjects in our experiment belong to just one cultural community, and this might reflect culture-specific behaviour.

A natural continuation of this study is to test a larger group of subjects, using two video-cameras, and taking into account a larger number of communicative gestures, in order to obtain empirically-based data which can eventually be implemented in a virtual agent.

Before being able to produce a model of human gestures, to be implemented in talking agents, further investigation is necessary to capture some more subtlety of human gestural communication and get more insight in how speech and gestures integrate each other to express different nuances of meanings.

6. References

[Allwood et al. 1993] Allwood J, Nivre J, Ahlsén E, (1993)"On the semantics and pragmatics of linguistic feedback". In Journal of Semantics. Vol. 9. No. 1.

[Allwod 2001] Allwood, J. (ed) (2001), Dialog Coding - Function and Grammar. Göteborg Coding Schemas. Gothenburg Papers in Theoretical Linguistics, 85. Department of Linguistics, Göteborg University

[Alwood et al. 2002]: Allwood J, Grönqvist L, Ahlsén E, Gunnarsson M.(2002)"Annotations and Tools for an Activity Based Spoken Language Corpus" (2002) Procedings of 2nd SIGdial Workshop on Discourse and Dialogue, Aalborg, Danmark

[Beskow et al. 2002]: Beskow J., Edlund J. & Nordstrand M., (2002) Description and realisation of multimodal output in dialogue systems. Proceedings of ICSLP'02 Denver, USA, (in press)

[Bolinger 1985] Bolinger D,L, Intonation and its parts Edward Arnolds, 1985

[Cavé et al 1996] Cavé R. et al. (1996) About the relationship between eyebrow movements and F0 variation. Proceedings of ICSLP'96 Philadelphia.

[Cerrato 2002] Cerrato L. (2002), Some characteristics of feedback expressions in Swedish. Proceedings of Fonetik '02. Speech, Music and Hearing Quarterly Progress and Status Report, vol 44, Stockholm, KTH , pp. 101-104..

[Cerrato & Skhiri 2003] Analysis and measurement of communicative gestures in human dialogues. Pro-

ceedings of AVSP ’03, St Jorioz, France (in press)

[Granström et al. 2002] Granström B; House D; Beskow J. (2002) "Speech and gestures for talking faces". In Granström B; House D; Karlsson (eds) Multimodality in Language and Speech Systems Kluwer Academic Publishers, p. 209-241

[Hällgren & Lyberg 1998]}Hällgren Å, Lyberg B, (1998) Visual speech synthesis with concatenive speech. Proceedings of AVSP’98 Terrigal Australia.

[Graf et al. 2002] Graf H.P. et al. (2002) Visual Prosody; Facial Movements Accompanying speech. Proceedings of the 5th IEEE International Conference on Automatic Face and Gesture Recognition, Washington

[Gybbon et al. 1997] Gybbon D.Moore R. and Winski R.. (1997) Handbook of standards and Resources for Spoken Language Systems Mouton de Croutier

[Magno Caldognetto & Zmarich 1999] Magno Caldognetto E., Zmarich C., (1999) Visual spatio-temporal characteristics of lip movements in defining Italian consonantal visemes Proceedings of ICPhS ’99 S. Francisco USA, I-/ August 1999, Vol 2 881, 884

[Poggi and Pelachaud 2000] Poggi I., Pelachaud C. (2000) "Performative Facial Expressions in Animated faces". In: Cassel J., Sullivan J., Prevost S., Churchill E., 2000, Embodied conversational agents, MIT press, 155-187.

[Sjölander and Beskow 2000] Sjölander, K. and Beskow, J. (2000). "WaveSurfer - an Open Source Speech Tool". Proceedings of ICSLP 2000, Bejing, China, October 2000

[Thorrisón 2002] Thorrisón K.(2002) Natural Turn-taking Needs no Manual: Computational Theory and Model, from Perception to Action. In: Granström B; House D; Karlsson (eds) Multimodality In Language and Speech Systems Kluwer Academic Publishers, 173-208.

[Teston 1998] Teston, B., (1998). L’observation et l’enregistrement des mouvements dans la parole : Problèmes et méthodes. In: Santi S., Guaitella I., Cavé C. et Konopczynski G. (eds), Oralité et Gestualité, ORAGE L’Harmattan, Paris, 41-58