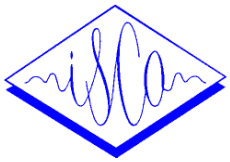
ISCA Archive
http://www.isca-speech.org/archive



AVSP 2003 – International Conference
on Audio-Visual Speech Processing
St. Jorioz, France
September 4–7, 2003

A method for the analysis and measurement of communicative head movements in human dialogues

Loredana Cerrato & Mustapha Skhiri*[¤]

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

ABSTRACT

The aim of this study is twofold: explore how people use specific gestures to serve important dialogue functions and show evidence that it is possible to measure and quantify their extent.

The paper focuses both on the presentation of the method used for the recording, coding and measurement of gestures and on the discussion on the obtained results.

Gestures, mainly head movements, were selected from naturally elicited dialogues *ad-hoc* recorded in lab-environment. The recordings consisted both of audio-video data and data measurements obtained with a motion tracking system.

Most of the analyzed head movements are produced to give feedback and the results show that it is possible to identify a specific pattern for a specific movement and that movements can be easily measured and their extent can be quantified. The results obtained with our method might eventually be implemented to improve the naturalness in animated talking heads.

1. INTRODUCTION

When human communicate with each other, signals from multple channels are at work. Communication takes place not only through words, tone of voice, stress given to words, but also by means of several gestures, such as facial expressions, gaze, head movements, hand movements, and body posture, which usually accompanies human speech. These accompanying gestures (as defined for instance by [14]"mouvements d´accompagnament") can be produced contemporarily with the production of utterances serving important syntactic, prosodic and dialogic functions [5] and can be produced to convey communicative intentions, feelings and attitudes.

In our investigation we focus our attention on a limited group of accompanying gestures, namely head movements.

The study of human communicative gestures is becoming more and more popular in the field of human-machine interfaces and speech technologies development. This is due to the fact that researchers, being aware of the important role that gestures play in communicative exchanges, are starting to integrate some of them in the development of dialogue systems endowed with embodied conversational agents, with the aim of enhancing their performance [5, 6, 7, 9, 12]. However the production of an accurate model of gesture realization is a time-consuming process, which requires extensive and detailed analysis of the gestures used in real communicative situation by human beings. Existing implementations of communicative gestures are therefore often based on observations that we might describe as intuitional. A consequence of the implementation of non-empirical results is for instance the arbitrary magnitude and unnaturalness of the reproduced movements. Since this stands in conflict with the reality demand, we propose a method for the recording, coding, measurement and quantification of specific gestures that presupposes the in-depth observation of how human beings use specific communicative gestures. 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.

Gestures were selected from naturally elicited dialogues ad-hoc recorded in lab-environment. The recordings consisted both of audio-video data and data measurements obtained with the Qualisys MacReflex motion tracking system [13].

For each identified gesture a 2D curve was plotted, each curve displays the amplitude of the gesture in millimeters on the Y-axis and the duration of the gesture in seconds on the X-axis.

By looking at the curves representing each gesture we tried to find plausible answers to the following questions:

Is there a one-to-one relationship between a specific verbal expression and its "accompanying" gesture?

Is there a one-to-one relationship between a specific gesture and a specific dialogic function?

Is it possible to notice inter-speaker and intra-speaker variability in the extent of the gesture?

The results show that it is possible to identify a specific pattern for a specific movement and that movements can be easily measured and their extent quantified.

A similar method of measurement and quantification has been applied with success to obtain data on articulatory gestures with the aim of reproducing them in talking heads [3,8].

2. EXPERIMENTAL DESIGN

2.1. Subjects and recording session

Four Swedish university students served as subjects (2 males, 2 females). Even if it is pretty difficult to predict when a particular gesture is going to occur in a spontaneous conversation, it is very likely that communicative gestures are produced in nearly all dialogue situations. For this reason the subjects were instructed to interact spontaneously with each other in a lab-environment.

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 foresee that our results might be implemented in animated conversational agents, we reproduced a communicative situation similar to the one that might arise between a user and an embodied conversational agent in a dialogue system, that is: "information seeking". Speaker A had the role of "information seeker" and speaker B had the role of "information giver". The information exchanged was related to movies: plots, schedules and so on. The focus of the recording was on speaker 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 the speaker B facing the interlocutor, the video camera and the four infra-red cameras. The people in figure 1 and 2 are not the subjects used in this study.

Figure 1: Reproduction of the recording session.

Figure 2 reports the position of the 13 hemispherical markers which were glued to the subjects' face.

Figure 2: The positions of the reflecting markers during the recording

The movements of the markers in 3D were stored together with the recorded acoustical and video signals.

The markers are ca 5 millimeters wide and passively reflect infra-red light. This way they are visible in the dark and easily traceable by infrared 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. In order to recover the rigid 3D motion of the head, the subjects wore special glasses, with 5 markers on.

Each recording session lasted about 15 minutes. During the first couple of minutes the subjects 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.

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 taken, the system was calibrated by using a calibration frame, to determine the geometrical relation between the image planes of the cameras and the coordinate system of the volume to be measured.

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 used to record the speech. The microphone was placed in front of the speaker B to accurately capture the speech of subject B. The voice of subject A was however recorded.

Transcriptions and labeling: orthographic transcription of the recorded dialogues was done with the support of the Multitool package software [2]. Multitool is a research tool for the analysis of digitized audiovisual data, which 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 has a score-based visual representation, which allows a multi-tier annotation of the phenomena under investigation, moreover it gives the possibility to manually code temporal alignment with speech. The time alignment is necessary in order to obtain perfect synchronization with the 3D recording.

Plotting of the curves: The 3D coordinates for each marker, were in each frame identified. The coordinates of the marker in the middle of the glasses was computed in order to get the movements of the head (3D rotation and translation, horizontal (x), vertical (y) and the depth (z)). To visualize the obtained data we used the dataplot function of the software package Wavesurfer [11].

3. ANALYSIS

3.1. Labeling of gestures

The gestures under analysis were labeled so that they could be coupled to the 3D data in order to produce clusters of patterns of movements. 98% of the labelled gestures occurred simultaneously with the production of speech, in particular with short expressions having specific dialogic functions. The labeling of gestures takes into account:

- the type of gesture,
- the communicative function of the related speech
- the relationship between the function of speech and the specific function of the accompanying gesture.

3.1.1. Type of gesture

To categorize the type of gestures the following labels were used:

Nod: a forward movement of the head, 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

3.1.2. Communicative function of the related speech

To label the communicative function of the related speech the following four main categories, which had been identified a priory, were used:

S: statement, positive or negative

FB: feedback,

A: answers, positive or negative,

D: disfluencies

Each category has some subcategories, however since 50% of the analyzed gestures accompany short feedback expressions, only the subcategories used for feedback, are reported in table1. These subcategories are to be interpreted as reaction to the previous communicative act, following [1, 4].

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

Function	Label	Comment
Continuation	**FBCPUi**	I want to go on
Continuation	**FCPUy**	you go on
Acknowledgement	**FBA**	acceptance
Refusal	**FBR**	refusal
Expressive	**FBE**	expression of attitude

3.1.3. Relationship between the function of speech and the specific function of the accompanying gesture

Gestures co-occurring with speech can either have a non-marked/neutral function, labeled as **N**, which does not modify the meaning of speech, or can be produced to modify the meaning [10] in one of the ways reported in Table 2.

Table 2: Schema of the labels used to code the relationship between the function of speech and the specific function of the accompanying gesture

Function	Label	Comment
Addition	**A**	the gesture adds some more info to speech
Emphasis	**E**	the gesture indicates a positive reinforcing attitude
De-Emphasis	**D**	the gesture weakens what has been said vocally
Contradiction	**C**	the gesture contradicts what has been said vocally

A schema of the relationships between gestures, speech and their functions are reported in Table 3

Table 3: Schema of the relationships between gestures, speech and their functions.

Type of gestures	Related short expressions	Communicative Functions	Relation gestures/ speech
Nod	*mh, ja, ah*	FBCPUy, FBCPUi, A, S	N,E,A
Shake	*nej*, negative statement	A{negative} S{negative} FBR	E
Jerk	*jaha, jusste*,	FBA, FBE{surprise}	A
Waggle	*Ehm*	D	A

Nods and jerks often accompany short verbal expressions having the main function of feedback (FBCPUy, FBCPUi) answers (A, usually positive answers). Sometimes nods are produced also during the production of statement (S), to emphasize what is being said. Jerks are produced together with feedback, mainly FBE when this conveys an attitudinal reaction of surprise. Shakes are produced always together with short negative answers and with negative statement (of the kind: I do not think so, I do not agree with you). Waggles are produced to express hesitation, doubt, mainly accompanying short disfluencies. The relation between the gesture and the meaning of the verbal expression is either N (neutral), usually when the feedback expression is a m-like sound produced with FBCPUy function, in a non-intrusive way or E when it emphasizes what has been said.

3.2. Analysis of the 3D data

For each identified head movement (in vertical[1] (y) and in horizontal[2] (x)) 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 second on the X-axis.

By looking at the curves representing each gesture we tried to answer the following questions:

- Is there a one-to-one relationship between a specific verbal expression and its accompanying gesture?
- Is there a one-to-one relationship between a specific gesture and a specific dialogic function?
- Is it possible to notice inter-speaker and intra-speaker variability in the extent of the gesture?

4. RESULTS

In table 4 is reported the number of occurrences of the selected gestures per subject. In total it was possible to select 32 gestures in dialogue 1 and 25 gestures in dialogue 2.

Table 4: Occurrence of gestures per subject.

Gestures	subject1	subject2
Nod	15	14
Shake	11	6
Waggle	5	1
Jerk	1	2
Side way-turn	-	2
tot gestures	**32**	**25**

By looking at the curves we obtained for each selected head movement, we tried to identify some general patterns and understand the relationships between gestures, speech and the function carried out in the communicative situation.

The most frequent head movement is"nod", which can be produced either as single or as multiple. In the video recordings it is easy to detect a multiple nod since it is possible to observe the head go up and down more than once. This is reflected in the curves because to every nod corresponds a single arc/peak as shown in the exemple reported in figure 3..

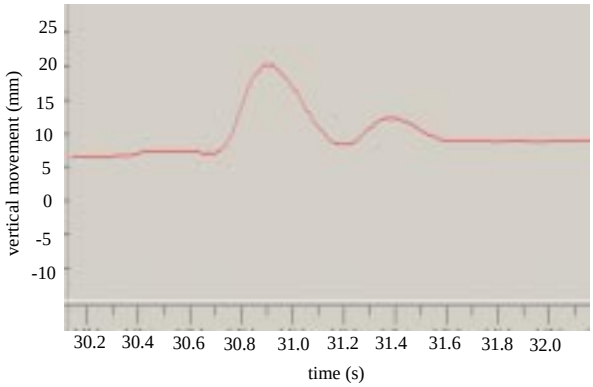


Figure 3: Curve of multiple nods accompanying the short expression "mh" by subject-1 with FBCPUy function.

The curves reported in figure 3 and 4 represent respectively a multiple nod and jerk which were produced by the same subject (1) with the same function: FBCPUy. The gestures accompany two different short expressions, respectively *mh* and *ja* -yes- however these expressions carry out the same communicative function. These 2 curves show that it is not possible to establish a one-to-one relationship between a specific gesture and a specific communicative function: different movements can in fact be produced to convey the same function/meaning.

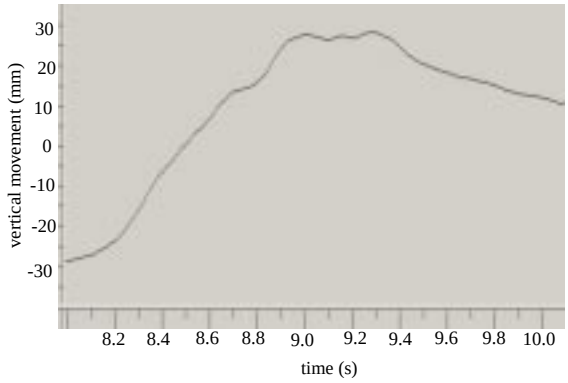


Figure 4: Curve of a jerk in vertical, accompanying the short expression "ja" by subject 1, with FBCPUy function.

The curves reported in figure 5 and 6 represents the gesture coded as"multiple nods" produced respectively by subject 1 and 2. These multiple nods were accompanying two different short verbal feedback expressions, respectively *javis*st and *jusste*, but they had the same communicative function: FBA and were used to emphasize speech. 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 pattern.

1. vertical (y) for nods and jerks
2. horizontal (x) for waggle and shakes

.

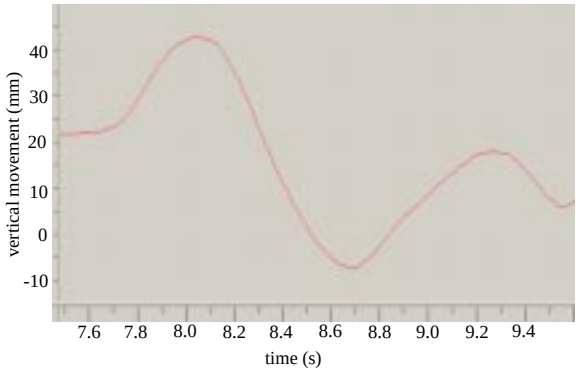


Figure 5: Curve of multiple nod produced with the feedback expression ”javisst” (certainly) produced by subject-2 with FBA function.

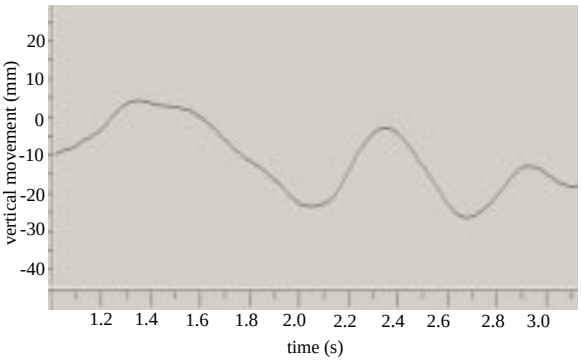


Figure 6: Curve of multiple nod in vertical, accompanying the expression ”jusste” (just that) produced by subject-1 with FBA function.

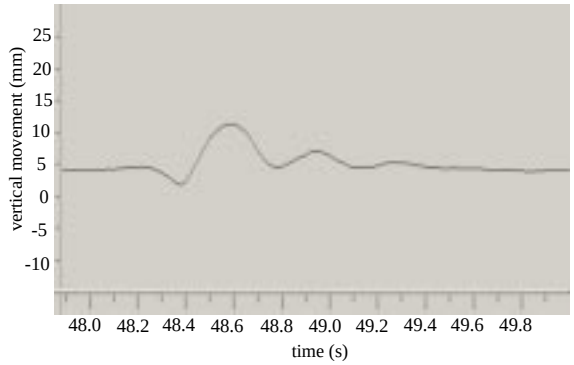


Figure 7: Curve of multiple nods accompanying the short expression “ja” by subject 1 with FBCPUi function.

The examples reported in figure 5 and 6 seem to support the idea that it is likely to link up some patterns of movements with categories of meaning. However the examples reported in fig 3 and fig 7 show instead that it is not possible to claim a one-to one relationship between a specific gesture and a specific verbal expression: figure 3 and figure 7 represent in fact two very similar curves, for the same gesture (multiple nods) produced by the same subject (1) to accompany two different expressions, *mh* and *ja*, with the same function FBCPUy.

Figure 8 reports an example of a shake. Shakes are represented on the curve as inverse peak.

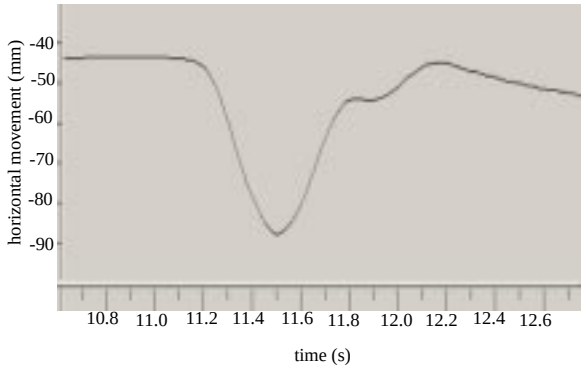


Figure 8: Curve representing a shake in horizontal, accompanying the short expression “nej” by subject 2, produced as a negative answer.

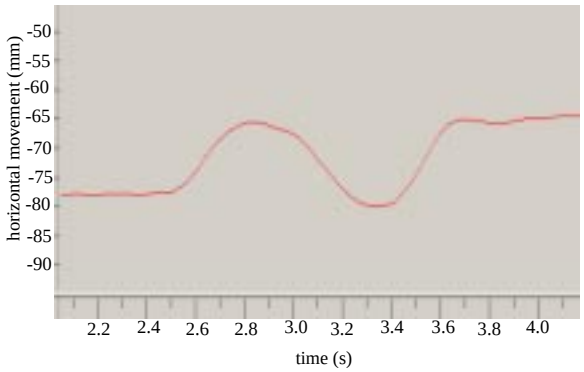


Figure 9: Curve of a waggle in horizontal, accompanying the short expression “ehm” by subject 1 produced as disfluency.

Waggles are produced to express hesitation, doubt and they are represented on the curve by peaks, which are wider than those representing nods. These results show that even if it is not possible to systematically establish a one-to-one correspondence between a specific gesture and a specific verbal expression, and even if it is not possible to establish a one-to-one correspondence between a specific gesture and a specific communicative function, it is possible to establish a one-to-one relationship between a specific gesture and a specific shape in the relative curves; moreover the following trends seem to be consistent within and across subjects:

- when the short expression *mh* produced to give FBCPUy is accompanied by a gesture, the gesture is always a nod (single or multiple);
- when short expressions like *ja* (yes), *preci*s (exactly), produced as a positive answer or as a FBCPUy are accompanied by a head movement, the movement is always a nod;
- when the short expression *ne*j -no-, produced as a negative answer or as FBR is accompanied by a head movement, the movement is always a shake. Shakes accompany also negative state-

ments;

- when the short expression *ehm* produced as D is accompanied by a head movement, the movement is always a waggle. Waggles are usually produced to show doubt and hesitation;
- when the short expressions *mh* and *ja* are used to give FBCUy in a non-intrusive way, the accompanying gesture, which 90% of the times is a nod, is minimal. This means that it has a short duration (about 100msec) and the peaks on the curve are not high;
- when expressions like *jaha, jusste javisst* are used to give FBA, or FBE with an attitudinal reaction of surprise, enthusiasm with the intention of emphasizing the message (E), the gesture accompanying them, which is usually a jerk or a nod, is quite extended, which means that it shows a longer duration and higher peaks.

5. CONCLUSIONS

The result of this study show evidence that it is possible to measure and quantify the extent of the selected gestures and it is possible to identify a general pattern for each specific movement, even if there are both intraspeaker and interspeaker variability in the duration and extent of the movements. In some cases it is possible to establish a one-to-one relationship between a specific verbal expression and its accompanying gesture, but it is not possible to establish a one-to-one relationship between a specific gesture and a specific dialogic function. This means that gestures are polisemic: they can carry out different functions/meaning depending on the context in which they are produced. 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 was 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 interlocutors mimicked each others gesture, how they exchanged gaze and so on. 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. However before being able to produce a model of human gestures, which could 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 attitudes.

6. REFERENCES

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

[2] Allwood J, Grönqvist L, Ahlsén E, Gunnarsson M, 2002:"Annotations and Tools for an Activity Based Spoken Language Corpus", Proc. of 2nd SIGdial Workshop on Discourse and Dialogue, Aalborg, Denmark.

[3] Beskow J, Granström B, and Spens KE, 2002 "Articulation strength - readability experiments with a synthetic talking face", in Proceedings of Fonetik 2002, Stockholm, Sweden.

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

[5] Elisei, F., Odisio, M., Bailly, G., and Badin, P. Creating and controlling video-realistic talking heads. In Auditory-Visual Speech Processing Workshop, Scheelsminde, Denmark, 2001.

[6] Mäkinen E., Patomäki S., and Raisamo R., Experiences on a Multimodal Information Kiosk with an Interactive Agent. Proceedings of NordiCHI 2002, The Second Nordic Conference on Human-Computer Interaction, ACM Press, 2002, 273-276.

[7] Hällgren Å, Lyberg B, 1998, Visual speech synthesis with concatenative speech. AVSP'98 Terrigal Australia.

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

[9] 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.

[10] Poyatos F., Non verbal communication across disciplines, John Benjamins Publishing Company, 2002.

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

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

[13]Qualisys MacReflex motion tracking system: http://www.qualisys.se.

[14]Teston B, 1998, L'observation et l'enregistrement des mouvements dans la parole: Problèmes et methods Oralitè et Gestualitè Santi S et al (eds) L' Hartmattan.