The emotional and communicative significance of head nods and shakes in a naturalistic database

R. Cowie [1], H. Gunes[2] , G. McKeown[1] , L. Vaclavu-Schneider[1] , J. Armstrong[1] , and E. Douglas-Cowie [1]

[1.] Queen's University, Belfast
[2.] Imperial College, London
E-mail: r.cowie@qub.ac.uk, g.mckeown@qub.ac.uk, e.douglas-cowie@qub.ac.uk

Abstract

Head nods and shakes have been extracted from the SAL audiovisual database of spontaneous emotionally coloured dialogue. The dataset contains 154 nods and 104 shakes. Two trained observers rated them on multiple dimensions derived from linguistics on one hand, and the psychology of emotion on the other. One used audiovisual presentation, the other visual only. There was agreement on affective, but not linguistic significance – suggesting that the latter depends on speech context rather than the manner of movement per se. A few seem to form discrete types, but for the most part classical dimensional models of emotion captured the affective variation well.

1. Introduction

The phrase 'emotion in action and interaction' was a repeated refrain in the HUMAINE project. Research is gradually coming to grips with the specifics of the way emotion enters into interactions. This paper is based on work being done in the successor project SEMAINE, whose aim is to create agents capable of engaging in sustained emotionally coloured interactions.

One of the effects of work in that area is to challenge standard divisions. This paper focuses on an area where a particularly complex set of divisions comes into play. The general domain that it deals with is backchannelling, which is generally thought of as a linguistic function. But whereas spoken language is usually thought of as a primarily acoustic phenomenon, a large part of backchannelling is visual. The paper considers the most obvious visible components of backchannelling, that is, head nods and shakes. Although the framework in which these gestures are usually analysed is (in a broad sense) linguistic, subjectively, at least part of their significance would seem to be emotional. Dictionaries typically give one of the meanings of 'shake' along the lines of 'To brandish or wave, especially in anger', and it is hard to believe that there is no relationship between that and shaking the head to signify a negative reaction. The upshot is that the area brings together multiple modalities and multiple types of significance in an intriguing package.

Although the issues are complex, the end product is straightforward: a database of over 250 head movements extracted from spontaneous interactions, with associated labels that capture those attributes of each movement that seem to be most salient to human observers. It provides a basis for research on either recognition or synthesis of appropriate types of head movement during interaction.

2. Analyses of nods and shakes

Head movements have been viewed in various ways among the computational community and related disciplines. Historically, the usual practice been to treat head movements during conversation as noise to be ignored as best one can. Until recently, speaking avatars did not generally move their heads; and if people speaking to them made head movements (which was discouraged), the main response was to look for ways of recovering facial expression in spite of the complication produced by head movement.

An alternative which has become widely recognised is to regard head movement during speech as a default whose presence is not particularly informative, but whose absence is. The background to that position was provided by the motor theories of investigators such as Hadar (1984a,b) who argued that large scale movements during speech create a favourable environment for the subtle, co-ordinated actions required to produce speech as such. The idea was highlighted by evidence that suppression of default accompanying movements was associated with deliberate (and deceptive) manipulation of the communication process (Cohn et al 2004).

A second major alternative is linguistic. Head movements have been regarded as an integral part of the concept of backchannelling since the concept emerged (Yngve 1970, Duncan 1972). Nods in particular were seen as an integral part of the mechanism by which speakers manage control and exchange of the 'floor'. That approach was elaborated in an influential paper by McClave (2000), who distinguished nine types of linguistic function for nods. These are described in the method section.

That conception has influenced computational research in general, and the SEMAINE project in particular (Heylen et al, 2007). SEMAINE aims to synthesise agents that can hold a sustained, emotionally coloured conversation with a user. One of the key ways in which it creates a sense of interaction is by having the agents make head movements, and respond to the user's head movements. The database described here was created to support the development of that aspect of SEMAINE.

Although the linguistic perspective influenced SEMAINE directly, it is clearly not the only possible option. Two others will be mentioned here.

Psychology has a long-standing interest in interpersonal behaviours whose function seems to be to show 'convergence' between the interactants. There are famous examples involving posture (Beattie & Beattie 1981), but there has been a recent surge of interest in behaviours that show temporal alignment (often described as entrainment) (Varni et al 2009). That approach invites the idea that head movements support synchronisation, serving both to achieve and to display temporal coherence.

It also is natural to assume that head movements may express affective content. There is an obvious connection with approaches that describe affect in terms of two dimensions, valence and arousal. To a first approximation, it would seem likely that a nod expresses positive affect towards the other party, whereas a shake expresses negative affect; and there is a relationship between the energy of the gesture and the arousal level.

Last, but not least, it should be noted that cultural factors are a major unknown. It is certainly true that some head movements take on some specific meanings in certain cultures (Brodsky & Griffin 2009). What is not clear is how deep and wide the influence of cultural factors is. This study does not try to answer that question, though the techniques that it describes might be relevant to doing so.

3. The study

The material to be considered was extracted from recordings of interactions using the SAL paradigm (Douglas-Cowie et al 2008). SAL is short for 'sensitive artificial listener'. The technique generates emotionally coloured conversation between a user and “characters” whose responses are stock phrases keyed to the user’s emotional state rather than the content of what he/she says. The model is a style of interaction observed in chat shows and parties, which aroused interest because it seemed possible that a machine with some basic emotion-detection capabilities could achieve it. In earlier versions of SAL, designed to provide training data, a person emulated the SAL characters. There are now versions where the characters’ speech and visible gestures are generated automatically. Recordings are available via http://www.semaine-db.eu/. SAL

Nods and shakes were extracted from interactions between a user and the person emulating the characters. The core task was to provide a description of each item that captured its functional significance as perceived by humans. In order to do that, it was necessary to address conceptual questions about the kind of framework that best captures the meaning that people attach to these movements. Two main levels of question are addressed. The first is whether distinctions between head movements are best captured in terms of linguistic categories (using McClave's system as the best developed) or affective descriptions. The second is whether the distinctions are best expressed in terms of categories (i.e. nods fall into n types) or dimensions (i.e. they lie at different points on n continua).

3.1 Method

3.1.1 Rating scales

The rating scales involved two parts. The first part used selected components of the system that has been developed for SEMAINE. It covers a range of descriptors, from the classical dimensions used to describe pure affect (arousal and valence) to terms that are purely cognitive (understanding and agreement). Between these are terms with both social and affective implications – 'solidarity' and 'antagonism', drawn from the categories developed by Bales (2000), which relate mainly to the valence dimension; and 'at ease', which relates mainly to arousal.

The second part used the linguistic categories proposed by McClave, i.e. inclusivity; intensification; uncertainty; direct quotes; expression of mental images of character; deixis and referential use of space; lists or alternatives; lexical repairs; backchanneling requests.

3.1.2 Rating procedure

All items were rated by two observers. They were students working on a year-long project, who were given prior training in the meanings of the categories as a preparation for the exercise.

Since there is a question over the role of linguistic information, we adopted the simple solution: one rater used audiovisual presentation, the other used visual only. Order of presentation was randomised, using different orders for the two raters.

The SEMAINE-derived components were used for both nods and shakes, the McClave components for shakes only.

4. Analysis

Ratings for nods and shakes were analysed separately, The same basic issues were covered in both. For each category, agreement between the two raters was examined. Note that since one rater had linguistic information and the other did not, what agreement indicates is that ability to assign that category does not depend radically on the presence of linguistic information. The two ratings were then averaged, and a second level of analysis was applied to the resulting averages. The first step was cluster analysis. Two step cluster analysis was used, as the most straightforward option. The results of that analysis were then studied graphically to establish whether some clusters might be better regarded as portions of a continuum (typically the upper and lower extremes). Where there seemed to be evidence of a continuum, factor analysis was used to gauge the number of dimensions present and the proportion of the variance that they accounted for.

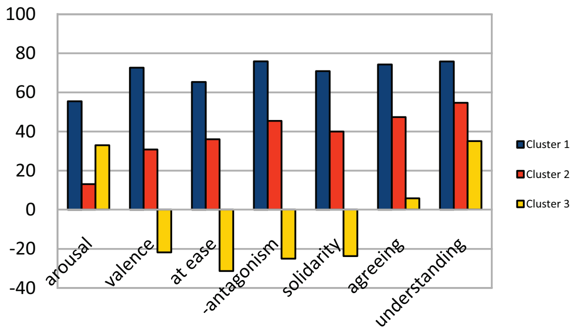

Figure 1: coordinates of cluster centroids for nods

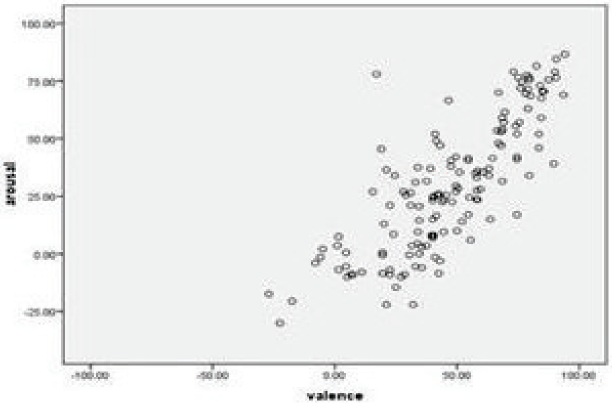

Figure 2: valence and arousal in nod clusters 2 & 3

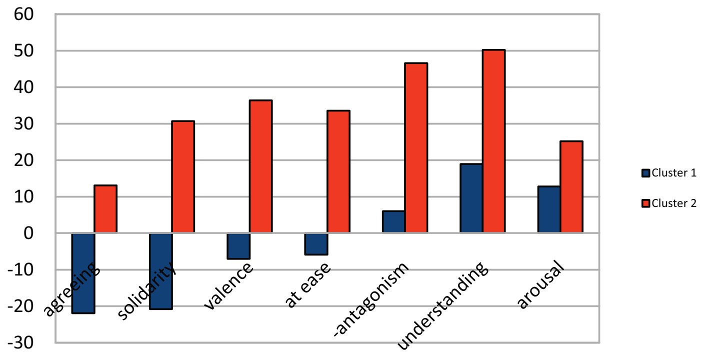

Figure 3: coordinates of cluster centroids for shakes

Variable	factor 1	factor 2
valence	0.86	0.24
arousal	0.31	0.91
agreeing	0.68	-0.34
at ease	0.83	0.1
solidarity	0.87	-0.27
antagonism	-0.8	0.2
understanding	0.8	0.08

Table 1 loadings of shake factors on the items

4.1 Nods

Inter-rater agreement was significant for most of the SEMAINE variables, though the strength of the relationship varied. It was highest for arousal (r =0.585), and valence (r=0.393), and low (but still significant) for solidarity (r=0.232), antagonism (r=0.190) and agreement (r= 0.213). Agreement was non-significant for understanding and at ease. The natural interpretation is that seeing a nod provides good information about affect, but relatively little about the more interpersonal and cognitive issues.Cluster analysis identifies three clusters. Figure 1 shows the coordinates of the cluster centroids. The profiles of clusters 1 and 2 are almost parallel, which is what would be expected if the clusters actually represented upper and lower ends of a single continuum. Cluster 3 is very different. It is marked by arousal and understanding, along with a range of negative evaluations – in other words, the nods convey that a message is understood and rejected.

To clarify the meaning of clusters 1 and 2, the 'understand and reject' nods were removed and factor analysis was applied to the remaining points. It recovered one factor, which corresponds to a well-established concept in emotion research, 'positive activation' (Watson & Tellegen 1985): that is to say, there is a continuum from low activation and neutral to high activation and positive. Figure 2 shows the distribution with respect to the two affective variables. It seems much more natural to regard these nods as a single continuum than as two clusters.

4.2 Shakes

The pattern of inter-rater agreement for shakes was broadly similar to the pattern for nods, with clearly significant agreement for arousal (r=0.605), intermediate for valence (r=0.264), solidarity (r=0.324), and antagonism (r=0.307); and marginal or non-significant relationships for agreement, understanding, and at ease.

Again, the natural reading is that what the appearance of shakes signals is mainly affective. An interesting additional point can be made because raters were asked to give not only their rating, but also their confidence in it. The items of which the rater with audiovisual information was most confident were 'understanding' and 'at ease', suggesting that the reason for inter-rater differences on these items is not that the information is poor, but that the audio provides very good information for them.

Cluster analysis identifies two clusters. Figure 3 shows the coordinates of the cluster centroids. However, if we ignore arousal, again, the profiles are almost parallel, suggesting that they may represent upper and lower ends of a single continuum. If so, arousal follows a different pattern. Factor analysis confirms that reading. It finds two dimensions. The loadings, shown in Table 1, show that they correspond admirably to the classical affect dimensions of valence and arousal. That suggests that the information in shakes is even more straightforwardly affective than that information in nods.

Analysis of the linguistically motivated categories reinforces that point. Most of the categories were not applied with any consistency at all. Specifically, for seven of the nine categories, the number of clips where both raters agreed that the category applied was two or less. The two exceptions were intensification and uncertainty, which it is reasonable to regard as the most affective of the categories. The lack of agreement on the others, which are more straightforwardly linguistic, has a straightforward interpretation: it seems very likely to mean that what marks these functions is not the appearance of the shake per se, but its relationship to speech.

5. Discussion

The concrete outcome of the research is a database containing substantial numbers of nods and shakes from spontaneous, emotionally coloured interactions, and a variety of labellings. That provides a resource for research interested in either learning or synthesising head movements during interaction.

The conceptual outcome can be expressed in terms of the way the various labellings included in the database can be understood. There is a strong tendency to assume that gestures like nods and shakes should be understood in terms of categories. Membership of clusters is given, and it does seem to be a useful descriptor for one, rather small group of nods, those that convey “message understood and rejected”. However, in most cases, the natural descriptor follows one of the classical patterns described by the psychology of emotion – positive activation for nods, and valence/activation space for shakes.

Describing these patterns highlights an issue which, to the best of our knowledge, has not been brought into focus before. It is the role of statistical reduction techniques in labelling. It is a standard procedure in psychology to translate responses on a number of raw rating scales into a smaller number of scores on (ex hypothesi) more basic dimensions. Derived measures of that kind are generated by the factor analyses described here, and the results are included in the databases. It would be consistent with practice elsewhere in psychology to think of that kind of variable as a more natural source of information than responses to individual items.

Linked to doubts about classification are doubts about the way paradigms from linguistics apply to non-verbal communication. The database includes descriptions using McClave's categories. It is not in dispute that the categories are useful. However, there clearly is reason to question the basis on which they are assigned. It has not been emphasised, but it is possible, that the categories simply cannot be assigned with much consistency. However, it seems more likely that the reason for the inter-rater differences is that assigning these categories is not a matter of classifying the head movement as such, but of gauging its relationship to what is being said.

A second level of relationality has not been addressed directly, but the results raise questions about it. There are reasons to predict that the meaning of a head movement will only be apparent in the context of the other party's movements. That appears to be at most part of the picture. It may be that relative timing can change the perception of a head movement, but there seem to be conclusions that people can draw with high confidence without considering that context.

Acknowledgements

Preparation of this paper was supported by the FP7 projects SEMAINE and SSPnet.

References

Bales, R.F. (2000) Social Interaction Systems: Theory and Measurement. New Brunswick, NJ: Transaction.

Beattie, G. and Beattie, C. (1981) Postural congruence in a naturalistic setting Semiotica 35 (1-2), 41–56

Brodsky, S.L & Griffin, M.P. (2009) When jurors nod The Jury Expert 31(6) 38-40

Cohn, J.F. Reed, L.I. Ambadar, Z. Xiao, J. & Moriyama, T. (2004) Automatic Analysis and Recognition of Brow Actions and Head Motion in Spontaneous Facial Behavior, Proc. IEEE Int'l Conf. Systems, Man, and Cybernetics (SMC '04), vol. 1, pp. 610-616.

Douglas-Cowie, E. Cowie, R. Cox, C. Amir, N. & Heylen, D. (2008) The sensitive artificial listener: an induction technique for generating emotionally coloured conversation. Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC ’08), pp. 1–4, Marrakech, Morocco, May 2008.

Duncan, S. (1972) Some signals and rules for taking speaking turns in conversations. Journal of Personality and Social Psychology 23: 283-292.

Hadar, U., Steiner, T.J. Grant E.C. and Rose, F. C.

(1984a). The timing of shifts of head postures during conversation. Human Movement Science 3: 237-245.

Hadar, U., Steiner, T.J. and Rose, F. C. (1984b) The relationship between head movements and speech dysfluencies. Language and Speech 27: 333-342.

Heylen, D., Nijholt, A., & Poel, M. (2007) Generating Nonverbal Signals for a Sensitive Artificial Listener. Verbal and Nonverbal Communication Behaviours, pp. 264-274.

McClave, E.Z. (2000) Linguistic functions of head movements in the context of speech Journal of Pragmatics 32 (7), 855-878

Varni, G., Camurri, A., Coletta, P. Volpe, G. (2009) Toward a Real-Time Automated Measure of Empathy and Dominance. International Conference on Computational Science and Engineering, 2009 vol. 4, pp. 843-848,

Watson, D., Tellegen, A. (1985). Toward a consensual structure of mood. Psychological Bulletin, 98, 219–235

Yngve, V. 1970. On getting a word in edgewise. Papers from the 6th regional meeting of the Chicago Linguistic Society, 567-578.

.