On the Attribution of Affective-Epistemic States to Communicative Behavior in Different Modes of Recording

Stefano Lanzini
SCCIIL (SSKKII)
Gothenburg University
lanzhbk@hotmail.it

Jens Allwood
SCCIIL (SSKKII)
Gothenburg University
jens.allwood@gu.se

Abstract

Face-to-face communication is multimodal with varying contributions from all sensory modalities; see e.g. Kopp (2013), Kendon (1980) and Allwood (1979). This paper reports a study of respondents interpreting vocal and gestural verbal and non-verbal, behaviour. 10 clips from 5 different short video + audio recordings of two persons meeting for the first time were used as stimulus in a perception/classification study. The respondents were divided in 3 different groups. The first group watched only the video part of the clips without any sound. The second group listened to the audio track without video. The third group was exposed to both the audio and video tracks of the clip. In order to collect the data, we used a crowdsourcing questionnaire. The study reports on how respondents classified clips containing 4 different types of behaviour (looking up, looking down, nodding and laughing) that were found to be frequent in a previous study (Lanzini 2013) according to which Affective Epistemic State (AES) the behaviours were perceived as expressing.

We grouped the linguistic terms for the affective epistemic states that the respondents used into 27 different semantic fields. In this paper we will focus on the 7 most common fields, i.e. the fields of Thinking, Nervousness, Happiness, Assertiveness, Embarrassment, Indifference and Interest. The aim of the study is to increase understanding of how exposure to video and/or audio modalities affect the interpretation of vocal and gestural verbal and non-verbal behaviour, when they are displayed uni-modally and multi-modally.

Keywords: Affective Epistemic States, Multimodality, Gesture, Speech, Verbal, Non-verbal Communication, vocal, auditory

1 Introduction

This paper explores the relative role of auditory and visual information for the attribution of affective-epistemic states to 4 different types of behaviour ("looking up", "looking down", "nodding" and "laughing") occurring in video clips taken from short video + audio recordings of two persons meeting for the first time.

By the term "Affective Epistemic State" we refer to internal human states that involve emotion, other aspects of cognition or perception (Allwood, 2012), e.g. Happiness, Sadness, Relaxation, Nervousness, alternatively described by Schroder (2011) "states which involve both knowledge and feeling" (Schroder, 2011).

We are considering both verbal and non-verbal behaviours expressed by vocal and gestural means, since many affective-epistemic and feedback functions are expressed simultaneously with all of these means, cf. Allwood, & Cerrato (2003) and Boholm (2011). Specifically, we are interested in investigating to what extent only visual, only auditory or both visual and auditory behaviour are involved.

K. Jokinen and M. Vels. 2015. Proceedings of The 2nd European and the 5th Nordic Symposium on Multimodal Communication.

2 Method

In this study we used 10 clips from 5 different recordings of pairs of 1:st language speakers of Swedish who are meeting for the first time, as stimulus in a crowd sourcing questionnaire study. The language used in the meetings is Swedish. The questionnaire was made with Google Drive and we employed random recruitment of respondents via social media. The duration of the clips varied from 7 sec to 20 sec, with an average length of 12.36 sec

There were 93 respondents, from different cultures. After having been exposed to the clips, they answered the questionnaire, in electronic form, available on the internet. We presented the subjects with recorded situations in three different conditions: video with audio (Video + Audio (30 persons)), video without audio (Video-only (35 persons)) and audio alone (Audio-only (28 persons)). The participants had to make an interpretation of which AES was expressed in a particular clip, in a particular presentation condition.

Each participant was exposed to 10 clips all in the same mode of presentation. The AESs had to be selected from a fixed list of options that were suggested by respondents in a previous study with Swedish stimulus material (Lanzini 2013). The AESs were given in English and were the following: Happiness, Sadness, Relaxation, Nervousness, Disinterestedness, Interest, Pride, Shyness, Confidence, Surprise, Sarcasticness, Aggressiveness, Thoughtfulness, Excitement, Unsureness and Playfulness

In addition, the participants could suggest other terms that according to them better described the AES they perceived. The participants also had to give a motivation for their answers.

3 Data

3.1 AESs grouped in Semantic fields

There are many words denoting different affective epistemic states in most languages. Some of the terms denote states that are closely related like “anger” and “wrath”. In our original study, we used free choice and consequently got very many different response terms for affective epistemic states. In order to make the data set more manageable we, in this study, grouped the terms used in the responses into semantic fields. A semantic field is a list of linguistic terms that share semantic characteristics. Below we present a list of the most frequent semantic fields made up of the linguistic terms for the AESs that we had obtained in a previous study (Lanzini 2013). For each clip, respondents were asked to write a term for only one Affective Epistemic State. The semantic fields were created after the data had been collected, and they were created on intuitive grounds by the researchers in order to group together AES terms with a similar semantic meaning. The following 7 semantic fields will be discussed below.

- **Thinking:** Thinking, Remembering, Reflective, Thoughtful, Giving Explanations
- **Nervousness**: Nervous, Uneasy, Unsure, Insecure, Uncomfortable, Hesitant, Reluctant, Uncertain, Unconfident
- **Happiness**: Happy, Good Mood, Amused, Joyous, Happy and Calm, Glad
- **Assertiveness**: Assertive, Sure, Proud, Confident, High Self Esteem, Assured Persistent, Insistent
- **Embarrassment**: Embarrassed, Self-conscious, Timid, Intimidated, Ashamed, Humbled, Shy, Reserved, Modest, Submissive
- **Indifference**: Indifferent, Apathetic, Lazy, Neutral, Evasive, Not Concentrated, Disinterested, Bored, not Interested
- **Interest**: Interested, Surprised, Participating, Engaged, Curious, Concerned, Hopeful, Motivated, Willing

3.2 Gestural behaviour

We will now present the most common interpretations of the following four gestural behaviours; “looking up” (2 clips from 2 videos), “looking down” (3 clips from 3 videos), “nodding” (3 clips from 3 videos) and “laughing” (2 clips from 2 videos). The 4 behaviours and their descriptive labels (“looking up” etc.) were chosen on the basis of being the most selected behavioural descriptive labels and, thus likely to be associated with easily perceived behaviour, in the previous study (Lanzini 2013). Every recorded behaviour was presented in the three presentation conditions (only audio, only video,

audio + video) introduced above. The word "whole" means that the whole body including feet is presented on the video, while in other cases only the upper part of the body is presented. The yellow fields indicate the AES attribution with the highest proportion of respondents for a particular clip in a particular condition of presentation. All AESs that turned out to be the most popular in any of the three conditions of presentation for any recording of the chosen 4 types of behaviour are included. Capital letters are used when referring to a semantic field, e.g. "Nervousness". All tables below show the most frequent semantic fields used by respondents. The percentages are generated by dividing the number of responses using a particular semantic field with the number of respondents for a particular condition of presentation. There were 28 respondents in the audio condition, 30 respondents in the video+audio condition and 35 respondents in the video condition.

3.3 Looking-up (2 clips)

Only Video (35 persons)	Nervousness	Thinking
Clip(4) looking up	26%	49%
Clip (5) whole body	20%	60%

Video+Audio (30 persons)	Nervousness	Thinking
Clip(4) looking up	43%	20%
Clip (5) whole body	27%	50%

Only Audio (28 persons)	Nervousness	Thinking
Clip(4) looking up	36%	14%
Clip (5) whole body	21%	21%

Table 1. Percentage of respondents for each condition of presentation using the 2 most common AES interpretations of "looking up" in the 3 conditions of presentation

In table 1, the two most common AES interpretations of "looking-up" behaviour are Nervousness and Thinking. The Thinking field interpretation is most popular when respondents have access to only video without sound, while the second most popular; Nervousness, is most frequent when they have access to both audio and video. The data also shows that an audio presentation does not strongly evoke a Thinking interpretation while it does evoke an interpretation of Nervousness. In the audio-only presentation condition, the speech in clip (4) was mostly perceived as a sign of Nervousness (36%), while in clip (5) "whole body" a smaller number of respondents (21%) perceived the speech as a sign of Thinking, which was the same percentage of respondents that interpreted it as an expression of Nervousness. According to respondents in the multimodal condition, the two clips of "looking up" are perceived differently mostly because of the verbal vocal behaviour which sounded more nervous in the audio-only presentation. In clip (4) the combination of speech and body movements increased the percentage of respondents that perceived Nervousness in comparison to both unimodal audio and unimodal video. In audio-only, Nervousness got a higher percentage of perceptions (36%) than Thinking (14%).

In contrast, in clip (5) "whole body", Thinking as an interpretation of "looking up", increases both in unimodal video and multimodal condition. This can be related to the fact that for clip (5), Nervousness and Thinking got the same number of interpretations (21%) in audio-only mode. Thus, Nervousness is most commonly attributed with multimodal data, less so with audio-only and least with video-only. So the combination of nervous speech and nervous body movement increases the perception of Nervousness.

Thoughtfulness and Thinking are most commonly attributed with video-only, less with multimodal data and least with audio-only. So the attribution of Thinking AESs decreases when the gestural behaviour of looking up is presented together with speech. It decreases a lot, so that if in audio-only, the speech is perceived as a sign of Nervousness.

3.4 Looking-down (3 clips)

Only Video (35 persons)	Nervousness	Embarrassment	Indifference
Clip (3) looking down	40%	11%	11%
Clip (7) looking down	31%	60%	0%
Clip (2) whole body	23%	31%	26%

Video+Audio (30 persons)	Nervousness	Embarrassment	Indifference
Clip (3) looking down	40%	3%	13%
Clip (7) looking down	43%	33%	0%
Clip (2) whole body	33%	13%	10%

Only Audio (28 persons)	Nervousness	Embarrassment	Indifference
Clip (3) looking down	25%	7%	50%
Clip (7) looking down	36%	14%	4%
Clip (2) whole body	29%	25%	4%

Table 2. Percentage of respondents for each condition of presentation using The 3 most common AES interpretations of "looking down", in 3 presentation conditions, and 3 clips.

In table 2, "Looking-down" is most strongly related to the 3 semantic fields of Nervousness, Embarrassment and Indifference. If we compare the multimodal mode with the unimodal conditions, we see that when the three clips (3), (7) and (2) "whole body" were presented with speech and gesture together, for clip (7) and clip (2) "whole body", the attribution of Nervousness increased, like it did in relation to "looking up". Clip (3) got the same number of attributions of Nervousness in the video-only and multimodal mode. Nervousness seems clearly noticeable in both speech and gesture, with speech cues possibly slightly more important.

If we consider the unimodal video condition, for "looking-down", the semantic field of Embarrassment has a higher number of attributions than it has for unimodal audio and multimodal audio+video, in all three clips.

In conclusion, it seems that speech has a negative effect on the attribution of Embarrassment-related AESs. This observation is supported by the fact that for these three clips, the semantic field of Nervousness got a much higher number of attributions, in the audio-only condition, than the field of Embarrassment.

It is also interesting to note that in the audio mode, 50% of respondents of clip (3), interpreted the speech as a sign of Indifference. The attribution of Embarrassment decreased in all three clips when presentation of body movement was combined with presentation of speech or given only in speech while it clearly increased the attribution of Nervousness. The attribution of Indifference shows a more varied picture, being most frequent for clip 3 when presented in audio-only.

3.5 Nodding (3 clips)

Only Video (35 persons)	Nervousness	Assertiveness	Interest
Clip (5) nodding	6%	14%	26%
Clip (6) nodding	31%	6%	29%
Clip (4) whole body	11%	11%	29%

Video+Audio (30 persons)	Nervousness	Assertiveness	Interest
Clip (5) nodding	7%	43%	10%
Clip (6) nodding	23%	0%	43%
Clip (4) whole body	3%	7%	50%

Only Audio (28 persons)	Nervousness	Assertiveness	Interest
Clip (5) nodding	0%	39%	32%
Clip (6)	4%	4%	54%
Clip (4) whole body	7%	7%	46%

Table 3. Percentage of respondents for each condition of presentation using The 3 most common AES interpretations of "nodding", in 3 presentation modes, and 3 clips.

Table 3 shows us that "Nodding" is most strongly related to the semantic field of Interest. For all clips this effect is strongest in the audio-only condition and lowest in the video-only condition. For clip (5), the Interest attribution is most infrequent in the multimodal condition. Thus, for Interest attributions, the vocal behaviour produced while people are nodding has equal or more influence than the nodding itself. Probably the video unimodal condition provides too little information for respondents to clearly attribute the AESs of Interest.

For the semantic field of Assertiveness, the case is less clear. For clip (5) speech plays an important role and this attribution decreases in the video-only presentation. However, the case is less clear for clip (6) and (4).

"Nodding" is also related Nervousness but here the relation to the video mode is stronger than in the case of "looking-up" and "looking-down

3.6 Laughing (2 clips)

Only Video (35 persons)	Nervousness	Happiness	Assertiveness	Embarrassment
Clip (2) laughing	23%	23%	6%	14%
Clip (7) whole body	29%	11%	3%	23%

Video+Audio (30 persons)	Nervousness	Happiness	Assertiveness	Embarrassment
Clip (2) laughing	23%	7%	3%	33%
Clip (7) whole body	23%	17%	7%	13%

Only Audio (28 persons)	Nervousness	Happiness	Assertive-ness	Embarrassment
Clip (2) laughing	32%	11%	18%	7%
Clip (7) whole body	21%	7%	21%	7%

Table 4. Percentage of respondents for each condition of presentation using The 4 most common AES interpretations of "laughing", in 3 presentation modes, and 2 clips.

Laughing involves both gestural and vocal behaviours. In table 4, we see that when laughing is presented multimodally with both sound and visible behaviour, in clips (2) and (7), it is mostly interpreted as a sign of Nervousness and/or Embarrassment. However, the two clips are quite different and the video participants laughed in very different ways.

If we consider the semantic field of Assertiveness we can note that respondents more frequently attributed AESs of this type when the laughter was presented in audio-only condition (clip (2), 18% and clip (7), 21%). The attributions of Assertiveness decrease in both unimodal video condition and in multimodal condition. So it seems that the properties providing Assertiveness in speech loose their effect when combined with gesture. In contrast, the semantic fields of Happiness and Embarrassment got a higher number of attributions in the video mode than in the audio mode, indicating that for these types of AES, visual cues seem to carry more influence than auditive cues.

4 Summary and discussion

The main conclusion concerning the four behaviours we have studied (looking up, looking down, nodding and laughing) is that no easy generalizations are available. What type of affective-epistemic state the behaviours are seen as expressing depends on the particular person expressing the AES and which sensory modality it is presented in. If we consider the four types of behaviour, some of the main results are the following with regard to mode of presentation:

(i) Looking-up

The most frequent semantic field to be associated with this behaviour is the field of Thinking and thoughtfulness. The association is strongest with the visible behaviour of "looking-up" and much weaker with the speech accompanying the visible behaviour. Perhaps this reflects that for Thinking the visual cue of looking-up is the strongest. For the second most common AES field, Nervousness, the opposite holds. Nervousness is most frequently associated with the speech accompanying "looking-up" behaviour, if respondents also hear the speech accompanying the bodily movement or do not see the bodily behaviour.

(ii) Looking-down

"Looking-down" is most strongly related to the semantic field of Nervousness followed by Embarrassment and Indifference. As is the case for "looking-up", Nervousness is most frequently attributed when both speech and gesture are available.

The semantic field of Embarrassment has a higher number of attributions for unimodal video than it has for unimodal audio and multimodal audio+video, in all three clips. Thus Embarrassment like Thinking seems to have a strong visual side.

50% of the respondents to clip (3), interpreted the speech accompanying the "looking-down" sequence as a sign of Indifference, when only presented with the audio condition. When presented in video-only or multimodally this decreased the attribution of Indifference, indicating that for this clip the important cue for Indifference was auditive rather than visual.

(iii) Nodding

The most common semantic field attributed to "nodding" is Interest, followed by Assertiveness and Nervousness ". The connection is strongest in the audio-only and multimodal presentation condition and slightly weaker in the video-only condition indicating that audio cues play an important role in making nodding an expression of Interest.

For Assertiveness, the case is less clear. Only one clip (5) shows a clear pattern of speech playing an important role, similar to what is the case for Interest, occurring in the audio-only presentation and in the multimodal presentation with a decrease in the video-only presentation.

Somewhat surprisingly for nodding, the relation of Nervousness to the video mode is stronger than in the case of "looking-up" and "looking-down" where the auditive cues were more important.

(iv) Laughing

The most common attribution to "laughter" was Nervousness followed by Happiness, Assertiveness and Embarrassment, all about equally common. Nervousness was attributed to laughter to roughly the same degree in all conditions of presentation. For Happiness and Embarrassment visual cues seemed slightly more important than auditive while for Assertiveness the opposite seemed to the case, making auditive cues the most important.

References

Allwood, J. 1979."Ickeverbal kommunikation - en översikt" in Stedje and af Trampe (Ed.) *Tvåspråkighet*. Stockholm, Akademilitteratur. Also in *Invandrare och Minoriteter* nr 3, 1979, pp. 16-24.

Allwood, J. and Cerrato, L. 2003 A Study of Gestural Feedback Expressions. *First Nordic Symposium on Multimodal Communication*. Paggio P. Jokinen, K. Jönsson, A. (eds). Copenhagen, 23-24, September 2003, pp. 7-22.

Boholm M. and Lindblad G. 2011. Head movements and prosody in multimodal feedback. *NEALT Proceedings Series: 3rd Nordic Symposium on Multimodal Communication,* 15, p. 25-32.

Chindamo, Massimo, Allwood, Jens & Ahlsén, Elisabeth. 2012. Some suggestions for the study of stance in communication. *Proceedings of IEEE SocialCom Amsterdam 2012*, 3-5.

Kopp Stefan. 2013. Giving interaction a hand: deep models of co-speech gesture in multimodal systems. *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 245-246.

Kendon,Adam. 1980. Gesticulation and speech: two aspects of the process of utterance. In M.R.Key (ed), *The Relationship of Verbal and Nonverbal Communication*, pp. 207-227. The Hague: Mouton and Co.

Lanzini, Stefano. 2013. *How do different modes contribute to the interpretation of affective epistemic states*. Published master's thesis for master's degree. University Gothenburg, Division of Communication and Cognition, Department of Applied IT.

Schroder Marc, Bevacqua Elisabetta, Cowie Roddy, Eyben Florian, Gunes Hatice, Heylen Dirk, ter Maat Mark, McKeown Gary, Pammi Sathish, Pantic Maja, Pelachaud Catherine, Schuller Bjorn, de Sevin Etienne, Valstar Michel, and Wollmer Martin. 2011. Building Autonomous Sensitive Artificial Listeners. *IEEE Trans. Affective Computing*. 9. (1). p. 1