ASSOCIATING CHILDREN'S NON-VERBAL AND VERBAL BEHAVIOUR: BODY MOVEMENTS, EMOTIONS, AND LAUGHTER IN A HUMAN-ROBOT INTERACTION

A. Batliner[1], *S. Steidl*[1,2], *E. Nöth*[1]

[1]Pattern Recognition Lab, Friedrich-Alexander Universität Erlangen-Nürnberg, Erlangen, Germany
[2]International Computer Science Institute (ICSI), Berkeley, CA, U. S. A.

ABSTRACT

In this article, we associate different types of vocal behaviour denoting emotional user states and laughter with different types of body movements such as gestures, forward bends, or liveliness. Our subjects are German children giving commands to Sony's Aibo robot; the data are fully realistic. The analysis reveals characteristic and significant co-occurrences of body movements and vocal events.

Index Terms— emotion, body movements, laughter, children, personality

1. INTRODUCTION

Within speech science, the focus of interest has broadened during the last years, encompassing all kinds of non-verbal/paralinguistic phenomena such as emotional user states. In the same vein, uni-modal processing (speech/video, etc.) has been supplemented by multi-modal processing. An overview of behavioural/social signal processing which is in the focus of all these new approaches is given in [1]. An ever-lasting problem for all these approaches is the sparseness of naturalistic/realistic data. In this paper, we report for the first time on associating children's body movements with their vocal behaviour expressing emotional user states in the fully naturalistic FAU Aibo Emotion Corpus. We do not know of many papers addressing this topic; [2] report on associating affective states with naturally occurring postures in a child-computer interaction.

2. THE DATABASE

The database used is a German corpus of children communicating with Sony's pet robot Aibo, the *FAU Aibo Emotion Corpus*, cf. [3, 4]. The children were led to believe that the Aibo was responding to their commands, whereas it was actually controlled by a human operator (Wizard-of-Oz, WoZ) using the 'Aibo Navigator' software over a wireless LAN (the existing Aibo speech recognition module was not used). The WoZ caused the Aibo to perform a fixed, predetermined sequence of actions; sometimes the Aibo behaved disobediently, thus provoking emotional reactions. The children believed that the Aibo was reacting to their orders – albeit often not immediately. In fact, the scenario was the opposite: the Aibo always strictly followed the same screen-plot, and the children had to align their orders to its actions. By these means, it was possible to examine different children's reactions to the very same sequence of Aibo's actions. In the so-called 'parcours' task, the children had to direct the Aibo from START to GOAL; on the way, the Aibo had to fulfil some tasks and had to sit down in front of three cups. This constituted the longest sub-task. In each of the other five tasks of the experiment, the children were instructed to direct the Aibo towards one of several cups standing on the carpet. The data were collected at two different schools from 51 children (age 10-13, 21 male, 30 female). Speech was transmitted via a wireless head set (UT 14/20 TP SHURE UHF-series with microphone WH20TQG) and recorded with a DAT-recorder (sampling rate 48 kHz, quantisation 16 bit, down-sampled to 16 kHz). Each recording session took some 30 minutes. The audio-stream was segmented automatically with a pause threshold of 1 sec. into interpausal units which can be conceived of as *turns*; this resulted in some 8.9 hours of speech. Moreover, the experiments were videotaped. The children were allowed to move freely as long as they remained on the small carpet the girl in Fig. 1 is kneeling on. The simple pragmatic reason for this restriction was to avoid the children pushing the Aibo or interfering with it in any other way but giving commands from some distance. At least two different conceptualisations could be observed: in the first, the Aibo was treated as a sort of remote-control toy (commands like *"turn left", "straight on", "to the right"*); in the second, the Aibo was addressed the same way as a pet dog (commands like *"Little Aibo doggy, now please turn left – well done, great!"* or *"Get up, you stupid tin box!"*), cf. [3]. Detailed information on the database is given in [4].[1]

The research leading to these results has received funding from the European Community under grant No. IST-2001-37599 (PF-STAR) and grant No. IST-2002-50742 (HUMAINE). The responsibility lies with the authors.

[1]The book can be downloaded from the web: `http://www5.informatik.uni-erlangen.de/Forschung/Publikationen/2009/Steidl09-ACO.pdf`.

Fig. 1. Caption of the video recording, parcours setting

3. ANNOTATION

Five labellers (advanced students of linguistics, 4 females, 1 male) listened to the turns in sequential order and annotated independently from each other each word as neutral (default) or as belonging to one of ten other **emotional states**, which were obtained by inspection of the data: joyful, surprised, emphatic, helpless, touchy, angry, motherese, bored, reprimanding, and rest. This procedure was iterative and supervised by an expert. The sequential order of labelling does not distort the linguistic and paralinguistic message; video information was not given. We resort to majority voting (henceforth MV): the most frequent label is attributed to the word. Exact descriptions and figures for the labels chosen and for different types of selections and mappings can be found in [4], p. 94ff. Some of the states were very sparse, e. g. *bored* (16 cases) or *joyful* (109 cases). In this work, we therefore use only one positive state, i. e. ***motherese*** (1 300 cases), and one negative main class, i. e. ***angry*** (1 718 cases) which is assembled from three raw labels: *reprimanding*, *touchy* (i. e., irritated), and *angry*. In total, 48 401 words were produced.

Two main types of **laughter** have been annotated, fully independently from the emotion annotation, by an experienced annotator and, in a sub-sequent pass, corrected by the first author: 100 instances of ***speech-laugh***, i. e. laughter modulated onto words, and 176 instances of 'normal' ***laughter*** between words or in isolated position; as for sub-types of laughter (strong vs. weak, and voiced vs. unvoiced), cf. [5]. This relatively low number – overall duration of laughters constituting only some 0. 4% of the total duration of all vocal events in the database – demonstrates that the children concentrated on fulfilling their tasks.

Fig. 1 illustrates the point of view of the video camera from the right side of the scene. We decided to resort to a coarse, binary type of labelling the children's body movements. For a few children, some of the first or the last subtasks were not videotaped due to technical problems. In order to keep annotations comparable across children, we therefore used only the central, longest, and most complicated task, the parcours task, which has been videotaped for all 51 children, for the annotation of **body features**. This annotation was done by the first author; only video information was used, audio was turned off fully during the annotation pass. In all, five binary body features were annotated, again obtained by iterative inspection of the data. The number of children displaying positive or negative characteristics is given as well; e. g., 22 were standing, 29 were in bent-knee posture:

posture: standing vs. bent-knee (sitting/kneeling) posture: [± STANDING] = 22/29
position: moving parallel to Aibo's position on the carpet, vs. same position throughout: [± MOVING] = 22/29
gesture: sometimes pointing towards the goal Aibo should go to, vs. never pointing: [± POINTING] = 15/36
bends: sometimes forward bending towards the Aibo vs. upright position only: [± BENDING] = 14/37
activity: lively (idle) movements vs. immobile position: [± LIVELY] = 21/30

Four children were standing at the beginning of the session but changed soon into a bent-knee posture; they were attributed to [− STANDING]. Six children did not use the pointing gesture throughout the session but only in the beginning; nevertheless, they were attributed to [+ POINTING]. With these specifications, posture, gesture, and bends are clearly binary phenomena. Activity is of course a continuum, pronounced at the edges but there might be some cases in between that are less clear; however, the decision was relatively easy to make. The same holds for position.

As we are not aligning the body features to verbal events (i.e. to the orthographic transcription) on the time axis, they have to be taken as personality features. Posture might simply be due to habits and/or comforting reasons; lowering oneself into the bent-knee posture, however, could be a sign of trying to be on the same level as the pet robot, i. e. trying to establish a closer relationship. Moving parallel to Aibo's position on the carpet and, by that, being as close to Aibo as possible – remember that forward movements, towards the Aibo, were not allowed becaused this would have meant to step onto Aibo's carpet – can be caused both by trying to closely observe and by being as close/intimate as possible to the Aibo. Choosing to point might be due to cognitive concepts (Aibo is intelligent, it can hear and thus, it can see as well) but at the same time, it could be as well a sign of establishing a closer, more intimate relationship. Forward bends are interpreted as an indication of engagement, and as trying to establish a closer, more intimate relationship. Activity is foremost a personality trait; it is well-known that some children are more lively and active, and that other ones are more calm and withdrawn.

The same way as for manual annotation, automatic annotation of these five body features should be rather easy to accomplish, even under less favourable recording and light conditions.

Table 1. Chi-square test for combination of body features; '*' indicates strong, 'significant' tendencies

(a) Chi-square test for combinations of body feaures

combination	p-value	phi
posture + gesture	.208	-.220
posture + bends	.770	-.085
posture + activity	.800	-.076
posture + position	* .000	-.600
gesture + position: P&M	* .012	.394
position + bends: M&B	* .028	.351
gesture + bends: P&B	* .003	.471
position + activity: M&L	* .002	.478
gesture + activity: P&L	* .001	.509
bends + activity: B&L	* .000	.735

(b) Cross-tabulations without posture, cf. explanation in text

P&M	−M	+M
−P	25	11
+P	4	11

P&B	−B	+B
−P	31	5
+P	6	9

P&L	−L	+L
−P	27	9
+P	3	12

M&B	−B	+B
−M	25	4
+M	12	10

M&L	−L	+L
−M	23	6
+M	7	15

B&L	−L	+L
−B	30	7
+B	0	14

abbreviations:
P: POINTING, M: MOVING, B: BENDING, L: LIVELY

4. RESULTS AND DISCUSSION

Some of the considerations on body features given above could have been formulated as hypotheses for a one-tailed testing of significance. However, as this work is rather a *'I wonder what will happen'* and not a *'I bet this will happen'* endeavour, we always choose the two-tailed test. For relating nominal data, we use the chi-square test (Yates' correction for 20-60 cases) and report effect size using phi. For the comparison of frequencies, which clearly are not normal-distributed, we use the non-parametric Mann-Whitney-U-test for two independent samples. In case of p-values above 0.05 and below 0.10, we may speak of weak tendencies, in case of 'significant' values below 0.05, we can assume strong tendencies. However, as we do not claim any 'significance' in its strict meaning, for reasons already discussed by [6], we do not adjust the level of significance for repeated measures; we rather use the p-values reported in a descriptive sense, the same way as the effect size measure phi, indicating strong tendencies that are worthwhile to be pursued further on.

Table 1(a) shows p-values (chi-square test) and phi values obtained for cross-tables when combining pairwise body features; here, we associate different types of body movements with each other. Table 1(b) shows the four-fold tables for those combinations we want to have a closer look at, cf. the lower part of Table 1(a). The only strong association for posture is the one with position; this might be due to the simple fact that it is easier to move around while being in upright, standing position. Thus, choosing between standing and bent-knee position might really mainly be due to comforting reasons. We therefore will not comment in more detail on associations with posture. Table 1(b) features a systematic pattern: highest values in the cells are for the combination with the same sign – in the case of negative signs (upper left cells), indicating that many of the children are rather withdrawn; however, most of the time, the combinations with positive signs (lower right cells) are second highest.

Table 2 displays p-values obtained in Mann-Whitney-U-tests for the five body features, and for overall frequencies of words (# words), *motherese* (# M), and *angry* (# A), and for their normalized frequencies in relation to overall frequency of words (# M_{norm} and # A_{norm}). Here, we associate different types of (non-verbal) body movements with verbal events, in this case, with the signalling of emotional user states. The association with normalized frequencies is a bit lower but the tendencies are the same. We do not display figures for laughter: there is only a (weak) tendency for speech laugh going together with [− STANDING]: p = .060 for frequency of speech laugh, and 0.049 for its normalized frequency. The missing association might be due to the fact that the children concentrated on their task, and that laughter in this data is not a sign of establishing any closer relationship – as is the case, e. g., in mother-baby interactions – but rather a sort of 'meta-comment': the children are not *smiling at* the Aibo but *laughing about* it, cf. [5]. Neither do we report figures for gender differences in Tables 1(a) and 2 as they are never significant for the associations with verbal events; between body movements, there is only a very weak tendency for male children to display more [+ LIVELY], [+ BENDING], and [+ MOVING] values.

When looking at those six children who display [+ LIVELY], [+ BENDING], [+ MOVING], and [+ POINTING], all of them produce *motherese*, and three of them *angry* as well. In contrast, 22 children display [− LIVELY], [− BENDING], [− MOVING], and [− POINTING]; six of those do neither produce *motherese* nor *angry*. Thus out of 51 children, 12 represent a 'prototypical', all-or-nothing multi-modal behaviour, six employing either all the usual verbal and physical means indicating a closer relationship with the interaction partner that are available in this scenario, and six none of them. Note that only once, there is an empty cell in a cross-tabulation, i. e. [+ BENDING] & [− LIVELY]; in Table 1(a), this combination displays the highest phi value. This makes

Table 2. Mann-Whitney-U-Test for combining frequencies of verbal events with body features; p-values; '+' indicates weak, '*' strong, 'significant' tendencies

feature	# words	# M	# M_{norm}	# A	# A_{norm}	features going together
posture	.685	.924	.879	.821	.598	
position	.543	.238	.292	.669	.827	
gesture	.143	* .050	+ .067	.135	.294	[+ POINTING] & emotions
bends	.118	* .034	* .037	+ .063	+ .091	[+ BENDING] & emotions
activity	+ .088	* .005	* .005	* .036	+ .053	[+ LIVELY] & emotions

sense: withdrawn children do not go into closer contact with interaction partners.

Enhanced attention manifesting itself in [+ MOVING] **position** is obviously strongly associated with all other body movements, including posture, but not with emotional user states, cf. Table 2. All other three types of body movements are strongly associated with emotional user states, in this ranking: gesture, bend, activity; this can be seen when simply counting the stars and pluses in each of the last three lines of Table 2. As mentioned in section 3, gesture might be in between cognition and emotional user states. Obviously, [+ LIVELY] children display more freely their emotions, and the same holds for [+ BENDING], their combination displaying the highest phi-value in Table 1(a).

5. CONCLUDING REMARKS

With due caveat, the anecdotal observation of two different kinds of conceptualisations reported in section 2 can be extended, based on our empirical findings, onto body movements as well; different conceptualizations and establishments of relationships with the interaction partner go together both with using specific verbal means and displaying specific characteristics of body movements: we have seen prototypical combinations of verbal means and body movements, but of course, there is some degree of freedom to choose some – but not any – combination of them.

Laughter is often understood as simply indicating emotional user states; this has to be specified for our data: there is no association with the verbal expression of those 'interactive' user states, esp. not with *motherese*, we are dealing with in this paper, and any body movement we addressed. Laughter in our data is simply an expression of amusement [5].

A more fine-grained manual annotation of our data does not seem to be promising, due to the low quality of the video data. However, automatic video tracking, e. g. for liveliness, and automatic alignment of body movements with audio (via aligning audio recording from the video camera with the close-talk audio recording) seem to be feasible. Yet we have seen that even such a coarse annotation aiming at personality traits reveals systematic associations between body movements and speech characteristics.

An advantage of the corpus used in this study is its realism; a disadvantage is of course that it is still sparse data; thus, it was not possible to interpret closely any combination of phenomena that could be observed in these data.

In the long run, possible applications for such approaches might be the modelling of an adequate and consistent behaviour of virtual agents, the multi-modal recognition of emotional user states (e. g. adjusting the priors for audio classes, based on video information), the monitoring and screening of children with communicative problems, and generally, any application within the area of edutainment and entertainment.

6. REFERENCES

[1] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, pp. 1743–1759, 2009.

[2] S. Mota and R. W. Picard, "Automated Posture Analysis for Detecting Learner's Interest Level," in *Conference on Computer Vision and Pattern Recognition*, 2003, pp. 49–56.

[3] A. Batliner, S. Steidl, C. Hacker, and E. Nöth, "Private emotions vs. social interaction — a data-driven approach towards analysing emotions in speech," *User Modeling and User-Adapted Interaction*, vol. 18, pp. 175–206, 2008.

[4] S. Steidl, *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*, Logos Verlag, Berlin, 2009, (PhD thesis, FAU Erlangen-Nuremberg).

[5] A. Batliner, S. Steidl, F. Eyben, and B. Schuller, "On Laughter and Speech Laugh, Based on Observations of Child-Robot Interaction," in *The Phonetics of Laughing*, J Trouvain and N. Campbell, Eds. Mouton de Gruyter, Berlin, 2011, to appear.

[6] H.J. Eysenck, "The Concept of Statistcal Significance and the Controversy about One-Tailed Tests," *Psychological Review*, vol. 67, pp. 269–271, 1960.