"You stupid tin box" - children interacting with the AIBO robot: A cross-linguistic emotional speech corpus.

A. Batliner*, C. Hacker*, S. Steidl*, E. Nöth*, S. D'Arcy‡, M. Russell‡, M. Wong‡

*Chair for Pattern Recognition
University of Erlangen-Nuremberg, Martensstr. 3, 91058 Erlangen, F.R.G.
{batliner,hacker,steidl,noeth}@informatik.uni-erlangen.de

‡School of Engineering
University of Birmingham, Edgbaston, Birmingham B15 2TT, Great Britain
{s.darcy,m.j.russell,l.p.wong}@bham.ac.uk

Abstract

This paper deals with databases that combine different aspects: children's speech, emotional speech, human-robot communication, cross-linguistics, and read vs. spontaneous speech: in a Wizard-of-Oz scenario, German and English children had to instruct Sony's AIBO robot to fulfil specific tasks. In one experimental condition, strictly parallel for German and English, the AIBO behaved 'disobedient' by following it's own script irrespective of the child's commands. By that, reactions of different children to the same sequence of AIBO's actions could be obtained. In addition, both the German and the English children were recorded reading texts. The data are transliterated orthographically; emotional user states and some other phenomena will be annotated. We report preliminary word recognition rates and classification results.

1. Introduction[1]

Desiderata for the automatic processing of realistic speech are corpora with children's speech, corpora with emotional speech, and - of course - corpora with emotional children's speech. It is well known that word recognition for children's speech is much more difficult than for adult's speech; this is at least partly due to the lack of training corpora, but also to physiological differences between children and adults and increased intra- and inter-speaker variability. Little is known about spontaneous emotional speech in general, and children's emotional speech in particular. These topics are not only interesting per se but also of great importance for applications, for example in the areas of 'edutainment', entertainment and human-robot-communication.

The modelling, generation and recognition of emotion has attracted more and more attention during recent years. Researchers have typically dealt with prototypical, 'full-blown' emotions and with elicited, prompted, acted speech (Cowie and Cornelius, 2003). Real life data differ, however, considerably from acted speech – not only because of different acoustic characteristics but because in a real life setting, much more means are available to signal and express, for instance, reactions to unsatisfactory system behaviour (Batliner et al., 2003a; Batliner et al., 2003c; Batliner et al., 2003b; Campbell, 2003). Of course, Labov's observer's paradox (Labov, 1970) holds for recordings of 'emotional' speech as well; thus we are not able to record 'real' real life data, but we can try to make them as real as possible.

[1]This work was funded by the EU in the project PF-STAR (*http://pfstar.itc.it/*) under grant IST-2001-37599 and by the German Federal Ministry of Education, Science, Research and Technology (*BMBF*) in the SmartKom project under Grant 01 IL 905 K7. The responsibility for the contents of this study lies with the authors. We want to thank the Ohm-Gymnasium and the Montessori school in Erlangen, and K. Fischer and T. Tenbrink, University of Bremen, for their kind co-operation.

2. Methodology

The general frame for the databases reported on in this paper is human-machine – to be more precise, human-robot – communication, children's speech, and the elicitation and subsequent recognition of emotional user states; moreover, we wanted to have at least one 'cross-linguistic' subsample, i.e., one identical experimental design for recordings in German and in English. The robot is the (dog-like) Sony's AIBO robot. The basic idea is to combine a new type of corpus (children's speech) with 'natural' emotional speech within a Wizard-of-Oz task.[2] The speech is intended to be 'natural' because children do not disguise their emotions to the same extent as adults do. However, it is of course not fully 'natural', as it might be in a non-supervised setting. Furthermore the speech is spontaneous, because the children were not told to use specific instructions but to talk to the AIBO like they would talk to a friend. In this experimental design, the child is led to believe that the AIBO is responding to his or her commands, but the robot is actually being controlled by a human operator, using the 'AIBO Navigator' software over a wireless LAN (the existing AIBO speech recognition module is not used). Three different versions of the experiment[3] were conducted:

In the experiment **E1** ('Parcours': AIBO obeys child's commands, **English** recordings), the child is asked to guide

[2]In (Tato et al., 2002), commands directed towards the AIBO were read by non-professional acting subjects.

[3]The 'Wizard-of-Oz' methodology in which the child gives spoken instructions to a Sony AIBO robot and the experimental design were developed at the University of Erlangen. The purpose was to elicit as many emotional utterances as possible without making the child unwilling to fulfil the task. The main motivation for the English recordings was a bit different, namely to record spontaneous, not necessarily only emotional children's speech. Thus only the design of experiment **E2** was kept identical for both languages.

AIBO around a map, starting at a square labelled 'start' and ending at a square labelled 'goal'. The map is printed on a floor carpet measuring approximately 2 x 3 meters. A diagram of the map is shown in Figure 1. A number of cups are placed in pre-determined positions on the map, and the child is instructed to make AIBO look into the cups. In addition, special squares on the map are labelled with instructions, such as *'dance'*, and the child is asked to make AIBO obey the instruction when it reaches the square. The 'wizard' is able to listen to the child and to watch the child and AIBO through one or more video cameras. In **E1** the wizard tries to make AIBO follow the child's instructions[4]. Thus **E1** is representative of a child controlling AIBO using a very high performance spoken language understanding system.

In the experiment **E2** ('Parcours': AIBO's actions pre-determined, **German** and **English** recordings), the child is given exactly the same task and instructions as in **E1**. However, the wizard causes the AIBO to perform a fixed, pre-determined sequence of actions, which takes no account of what the child says. For the sequence of AIBO's actions, we tried to find a good compromise between obedient and disobedient behaviour: we wanted to provoke the children in order to elicit emotional behaviour but of course we did not want to run the risk that they break off the experiment. The children believed that the AIBO was reacting to their orders - albeit often not immediately. In fact, it was the other way round: the AIBO always strictly followed the same screen-plot, and the children had to align their orders to it's actions. By this means, it is possible to examine different children's reactions to the very same sequence of AIBO's actions. **E2** simulates spoken language control of AIBO using a rather poor spoken language processing system. A short sequence of – mostly disobedient – actions performed by AIBO is shown in the following:

POSITION: START	
AIBO addresses child :	gesture "Hi"
CHILD:	tells AIBO what to do
+ co-operative:	gets up
+ co-operative:	goes forward
....	
POSITION C, 4th crossing	
- co-operative:	stops
- co-operative:	lays down
+ co-operative:	stands up
- co-operative:	lays down
+ co-operative:	stands up
- co-operative:	lays down
+ co-operative:	stands up
+ co-operative:	turns left
AIBO addresses child:	turns head towards child
+ co-operative:	goes forward
....	

In each of the five tasks OL A - OL E of the experiment **E3** (**O**bject **L**ocalisation, **German** recordings), the children were instructed to direct the AIBO towards one out of several cups standing on the carpet. One of these cup was 'poisoned'. The children applied different strategies to direct the AIBO. However, similar to **E2**, all actions of AIBO were pre-determined. The AIBO was fully controlled by the wizard. In OL A AIBO was 'obedient' in order to make the children believe that AIBO would understand their commands. In the other tasks AIBO was 'disobedient'. In some tasks AIBO went directly toward the 'poisoned' cup in order to evoke emotional speech from the children.

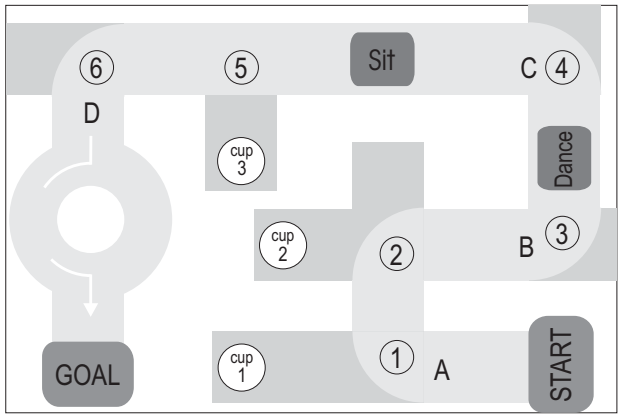

Figure 1: Map of the Parcours task; numbers 1-6: crossings; A-D: AIBO behaves disobediently; *Dance* and *Sit*: tasks to fulfil; goals: cups 1-3 and GOAL

In the **German** recordings, the order of the tasks was OL A, OL B, Parcours, OL D, OL E, OL C. No child broke off the experiment, although it could be clearly seen towards the end that many of them were bored and wanted to put an end to the experiment - a reaction that we wanted to provoke. Interestingly, in a post-experimental questionnaire, all the children reported that they had much fun and liked it very much. 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!*).[5]

The majority of the **English** children completed two recording sessions of **E1** and **E2** described above, in this order. The children were told that they were using two alternative systems in the two conditions. Under the second condition in **E2**, similar emotional responses were observed to those in the German experiments. However, it is likely that the successful experience of the first task may have influenced the children's reaction to the second task, making them less willing to accept the apparent poor performance of the system.

3. Recordings

The **German** data was collected from 51 children (age 10 - 13, 21 male, 30 female). The children are from two different schools ('Mont' and 'Ohm'). The recordings took

[4]In some cases this was not possible. For example one child used the command "paw" to ask AIBO to shake hands.

[5]In addition, the same children read different German texts which were previously unknown to them (approx. 9.3 hours of speech, with a vocabulary of approx. 7800 words). These audio files will be processed in parallel.

place in two class-rooms, one in school 'Mont' and one in school 'Ohm'. The only persons in the room were the child, the supervisor, who gives the instructions, the wizard (behind the children, pretending to be doing the recordings) and a third assistant.[6] Originally, each recording session took some 35 minutes. Because of the experimental setup these 25.5 hours contain a huge amount of silence (reaction time of the AIBO), which caused a noticeable reduction of recorded speech after raw segmentation. Finally we obtained about 9.2 hours of speech.

Thirty **English** children, between the ages of 4 and 14, took part in **E1** and **E2**[7]. Recordings were made in a special multi-media studio in CETADL (the Centre for Educational Technology and Distance Learning) in the Department of Electronic, Electrical and Computer Engineering (EECE). The total duration of the recordings is approximately 8.5 hours, which corresponds to just over 1.5 hours of speech once silences, pauses and 'babble' have been removed.[8] In both the German and English experiments, video recordings were also made. These are only for internal use, due to privacy restrictions.

4. Annotation

Both the German and English data have been transliterated orthographically. In addition to the spoken word chain, other verbals (filled pauses etc.) and non-verbals (microphone noise etc.) have been annotated.

4.1. Prosodic peculiarities

One experienced labeller annotated the German data basically along the same lines as had been done for another database (Batliner et al., 2003b) with the following phenomena: very long pauses (child waits for AIBO to fulfil a command: [PAUSE_LONG]; unusual pauses between phrases: [PAUSE_WORD]; pauses within a word, between syllables: [PAUSE_SYLL]; lengthening of syllables: [LENGTH_SYLL]; insertion of syllables: [INS_SYLL], for instance /stop/ `['StO:|hOp]`; marked emphasis: [EMPHASIS]; shouting: [SHOUTING]; shift of accent position: [ACC_SHIFT], for instance /Aibo/ `[aI|'bo:]`; very clear articulation: [CLEAR_ART]; laughter: [LAUGHTER]; vocative: [VOCATIVE] (only for the word *Aibo*); A word can have more than one label. This annotation has been finished. There are 51.393 word tokens. To give some examples, the frequency of the 11 combinations with more than 100 tokens are: 114 LENGTH_SYLL & CLEAR_ART, 122 PAUSE_LONG, 186 PAUSE_WORD, 242 EMPHASIS & CLEAR_ART, 254 VOCATIVE, 287 SHOUTING, 616 LENGTH_SYLL & EMPHASIS, 2117 CLEAR_ART, 2901 LENGTH_SYLL, 4328 EMPHASIS, and 39669 NEUTRAL.

4.2. Emotional user states

The annotation of the emotional user states for the German data is still ongoing. Several labellers annotate independently from each other each word as neutral (default) or as belonging to one of the following classes: *joyful, surprised, emphatic, helpless, touchy* (=irritated), *angry, motherese, bored, reprimanding, rest* (non-neutral, but not belonging to the other categories). These classes were obtained by inspection of the data; we do not claim that they represent children's emotions in general, only that they are adequate for the modelling of these children's behaviour in this specific scenario. *joyful* and *angry* belong to the 'big' emotions, the other ones rather to 'emotion-like/emotion-prone' user states. The state *emphatic* has to be commented on especially: based on our experience with other emotional databases (Batliner et al., 2003a), any marked deviation from a neutral speaking style can (but need not) be taken as a possible indication of some (starting) trouble in communication. If a user gets the impression that the machine does not understand her, she tries different strategies – repetitions, reformulations, other wordings, or simply the use of a pronounced, marked speaking style. Such a style does thus not necessarily indicate any deviation from a neutral user state but it means a higher probability that the (neutral) user state will possibly be changing soon. Of course, it can be something else as well: a user idiosyncrasy, or a special style – 'computer talk' – that some people use while speaking to a computer, like speaking to a non-native, or to a child, or to an elderly person who is hard of hearing. Thus the fact that *emphatic* can be observed can only be interpreted meaningfully if other factors are considered. There are three further – practical – arguments for the annotation of *emphatic*: firstly, it is to a large extent a prosodic phenomenon, thus it can be modelled and classified with prosodic features. We do not know yet whether this holds for the other user states. Secondly, if the labellers are allowed to label *emphatic* it might be less likely that they confuse it with other user states. Thirdly, we can try and model emphasis as an indication of (arising) problems in the communication.

For labelling, there are basically two different strategies: either to use highly qualified experts (or make the labellers experts by training them thoroughly) or to use several less experienced labellers – who so to speak represent the 'man on the street' – and rely on a majority voting. With the 'expert' approach that for instance normally has been used by the ToBI community for the labelling of intonation, a higher interlabeller correspondence and by that, reliability, can be obtained. However, it is not clear yet, whether this means at the same time a high validity, i.e., whether it really models the object of investigation and not only the human ability to train hard to harmonize with each other.

[6] Speech was transmitted with a wireless head set (UT 14/20 TP SHURE UHF-series with microphone WH20TQG) and recorded with a DAT-recorder. The sampling rate of the signals is 48 kHz, quantisation is 16bit. However, the data is downsampled to 16 kHz.

[7] Recordings were also made in an audiometric booth of each of these children reading English texts.

[8] The recordings used two head-mounted wireless microphones: the UT 14/20 TP SHURE UHF-series with microphone WH20TQG and a Senheiser ew100 range lapel microphone (SK100 transmitter, EK100 receiver), which was clipped to the SHURE head-mount. The speech was also recorded using existing wall-mounted microphones in CETADL. Analogue to digital conversion used the Edirol UA-5 external sound card with USB interface. The sample rate was 44.1kHz.

In the past, we have had good experiences with the 'majority' approach (Kießling et al., 1994) for the annotatation of boundary and accent position.[9] Accentuation is as well as user states not an all or none but rather a continuous phenomenon, i.e., different thresholds are possible. Thus we decided to use the majority approach for the annotation of user states in the German data, i.e., a user state will be defined by the majority of all labellers. We plan to label the English data along the same lines, but possibly, due to time restrictions, rather with a sort of 'expert' approach.

4.3. Interaction alignment

With a pause detection algorithm, the German audio data have been segmented automatically into turns. To relate AIBO's actions to the child's verbal re-actions, we want to make an alignment: by looking at the video recording, AIBO's actions (an example is given in section 2.) will be attributed to the turn-ID of the utterance with which the child reacts to this action; to reduce effort, we refrain from an exact time alignment. With such a rough alignment, we get so to speak an interaction structure of the child-robot communication along the same lines as a dialogue structure and can try to establish some - even primitive - 'interaction act sequences'.

5. Some preliminary results

The **English** data, summed over **E1** and **E2**, comprises 5,822 words from a vocabulary of 247 words. The average number of words spoken per session for conditions **E1** and **E2** are 85.7 and 135.5, and the average session durations are 7.3 minutes and 10.3 minutes, respectively. Thus the average numbers of words per session and session durations increase by 58% and 41% respectively, between conditions **E1** and **E2**. The word frequencies, summed over all conditions, follow a Zipfian distribution. The 20 most frequent words (and their number of occurrences) are: *stop* (1533), *forward* (945), *turn* (884), *left* (771), *walk* (540), *right* (372), *up* (314), *forwards* (305), *stand* (224), *the* (216), *go* (187), *sit* (159), *around* (152), *dance* (150), *a* (50), *to* (137), *backwards* (122), *round* (115), *OK* (115) and *and* (113).

The **German** data, summed over **E2** and **E3**, comprises 51,393 words from a vocabulary of 1190 words (841 real words, 349 fragments/non-words; 580, i.e., 49%, hapax legomena). The 20 most frequent words (and their number of occurrences) with English translations are: *Aibo* Aibo (7466),[10] *nach* to (2960), *links* left (2561), *stopp* stop (1807), *geh* go (1756), *lauf* go (1586), *rechts* right (1443), *und* and (1354), *jetzt* now (1242), *dich* you (1138), *steh* stand (1110), *auf* up (1088), *komm* come (1010), *g'radeaus* straight on (899), *dreh* turn (883), *weiter* (go) on (874), *mal* – (840), *ja* yes/OK (778), *sitz* sit (722), and *aufstehen* get up (703).

For the German data, first word recognition experiments yielded a word accuracy of 76.7% with a bigram Language Model. We mapped the prosodic labels onto four cover classes: *neutral* without any labels, *laughter* and *vocative* with or without any other label, and *marked* for any other combination of prosodic peculiarities. We used a large prosodic feature vector with 95 prosodic and 30 part-of-speech features, and a Neural Network for classification; details of such a constellation are given in (Batliner et al., 2003b). For this four-class problem and for learn≠test, the overall recognition rate is 69.8%, the class-wise computed recognition rate (mean of recognition rates per class) 48.7% (without laughter 57.2%).

[9] Note that for rather complicated tasks as the labelling of intonation within a specific theoretical model, e.g., ToBI, only an expert approach is feasible.

[10] In the English data, *Aibo* is on rank 21 with 110 occurrences.

6. Future work

After completion of the annotations, amongst others the following topics will be addressed: classification of emotional user states (different granularities) with different classifiers (Neural Networks, Linear Discriminant Analysis, Decision Trees), without and with other linguistic information, e.g., word classes, Language Models; modelling of interaction sequences; optimization of word recognition; cross-linguistic differences; and a characterisation of read vs. spontaneous speech. Eventually, the databases will be made available for the research community.

7. References

Batliner, A., K. Fischer, R. Huber, J. Spilker, and E. Nöth, 2003a. How to Find Trouble in Communication. *Speech Communication*, 40:117–143.

Batliner, A., C. Hacker, S. Steidl, E. Nöth, and J. Haas, 2003b. User States, User Strategies, and System Performance: How to Match the One with the Other. In *Proc. of an ISCA tutorial and research workshop on error handling in spoken dialogue systems*. Chateau d'Oex: ISCA, pages 5–10.

Batliner, A., V. Zeissler, C. Frank, J. Adelhardt, R. P. Shi, and E. Nöth, 2003c. We are not amused - but how do you know? User states in a multi-modal dialogue system. In *Proc. Eurospeech 2003*, Geneva, Switzerland, pages 733–736.

Campbell, N., 2003. Towards Synthesising Expressive Speech: Designing and collecting Expressive Speech Data. In *Proc. Eurospeech 2003*, Geneva, Switzerland, pages 1637–1640.

Cowie, R. and R.C. Cornelius, 2003. Describing the emotional states that are expressed in speech. *Speech Communication*, 40:5–32.

Kießling, A., R. Kompe, A. Batliner, H. Niemann, and E. Nöth, 1994. Automatic Labeling of Phrase Accents in German. In *Proc. ICSLP 1994*, Yokohama, pages 115–118.

Labov, W., 1970. The Study of Language in its Social Context. *Studium Generale*, 3:30–87.

Tato, R., R. Santos, R. Kompe, and J.M. Pardo, 2002. Emotional space Improves Emotion Recognition. In *Proc. ICSLP 2002*, pages 2029–2032.