Laughter in Child-Robot Interaction

Anton Batliner*, Stefan Steidl*, Florian Eyben§, Björn Schuller§

* Chair for Pattern Recognition, Friedrich-Alexander-University Erlangen-Nuremberg, Erlangen, Germany
§ Institute for Human-Machine Communication, Technische Universität München, Munich, Germany

In this paper, we present a speech database with children's speech. We first describe the database, part of the emotional labelling conducted, and the distribution of `non-linguistic'/para-linguistic phenomena, esp. of `real' laughter vs. speech laughter, i.e. laughter modulated onto speech. Especially, we will characterize interactional attitude and speaker-specific behaviour. In a second, more technologically oriented part, we will address the automatic recognition of these phenomena.

The database used is a German corpus with recordings of children communicating with Sony's AIBO pet robot. 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. The children were led to believe that the AIBO was responding to their commands, whereas the robot was actually controlled by a human operator. The wizard caused the AIBO to perform a fixed, predetermined sequence of actions; sometimes the AIBO behaved disobediently, thereby provoking emotional reactions. The data was collected at two different schools, Mont and Ohm, from 51 children (age 10 - 13, 21 male, 30 female; about 9.2 hours of speech without pauses). Speech was transmitted with a wireless head set (UT 14/20 TP SHURE UHF-series with microphone WH20TQG) and recorded with a DAT-recorder (sampling rate 48 kHz, quantization 16bit, down-sampled to 16 kHz). The recordings were segmented automatically into `turns' using a pause threshold of 1 s.

Five labellers (advanced students of linguistics) listened to the turns in sequential order and annotated each word independently from each other as *neutral* (default) or as belonging to one of ten other classes. We resort to majority voting (henceforth MV). If three or more labellers agreed, the label was attributed to the word (majority voting MV). In the following, the number of cases with MV is given in parentheses: *joyful* (101), *surprised* (0), *emphatic* (2528), *helpless* (3), *touchy*, i.e., irritated (225), *angry* (84), *motherese* (1260), *bored* (11), *reprimanding* (310), *rest*, i.e. non-neutral, but not belonging to the other categories (3), *neutral* (39169). 4707 words had no MV; all in all, there were 48401 words. Note that for further processing, cf. Fig. 1, the negative classes *touchy, irritated,* and *angry*, were mapped onto a cover-class *angry*, to deal with the sparse data problem in automatic processing. In the orthographic transliteration of the data, the non-verbals displayed in Tab. 1 were annotated.

hesitation (vocal, nasal)	<"A>	20	1.1 %
hesitation (vocal)	<"a>	35	2.0 %
noise	<#>	809	46.2 %
breathing	<A>	570	32.5 %
human noise	<G>	151	8.6 %
cough	<H>	32	1.8 %
laughter	**<L>**	**110**	**6.3 %**
hesitation (nasal)	<m>	23	1.3 %
sum		1749	100.0 %

Tab. 1: non-verbals annotated in the database, with absolute and relative frequencies

Prosodic phenomena had been annotated as depicted in Fig. 1 displaying word based peculiarities, amongst them speech laughter [LAUGHTER] which co-occurs with *neutral* and esp. with *joyful* **-** note that the frequency of *joyful* only amounts to some 0.25 % of *neutral*, cf. above. It does not co-occur with *motherese.* Note, that it turned out to be necessary to re-annotate all utterances where laughter had been annotated; this was done by the first author. Eventually, 110 laughter instances <L> are found in 102 turns. For 94 words in 59 turns, speech laughter (adjacent to the resp. word in the transliteration) is annotated. Only in 22 turns, both <L> and [LAUGHTER] are found. This is an example of annotated laughter: *Mont_10_024: ja und jetzt iss Aibo [LAUGHTER] <L> komm [LAUGHTER] <L>. (yes, and now eat, Aibo, come on.)*

In a Mann-Whitney test, there are no significant differences between males and females as for the frequencies of *motherese*, <L>, or [LAUGHTER]. The correlations (Spearman r) between the frequencies of *motherese* per speaker and the frequencies of <L> (.24) and [LAUGHTER] (.09) per speaker are very low. These distributions make it most likely that both types of laughter in our data are a sort of `private' laughter denoting some amusement, and not social/phatic, i.e. communicative/conversational, laughter, both on a micro-level, i.e. at the co-occurrence of laughter with other phenomena on the word-level, and at the speaker level: speakers either tend towards a motheresing attitude towards the robot; in this case, they rather seldom display laughter. Or they conceive the AIBO rather as a sort of remote control toy; in that case, they might tend more towards laughter, if amused. We can speculate about the conceptualisation of pet robots and the difference between them and real pets or babies: the prosody of *motherese* in our data is similar to the one found for mother-child interaction. Moreover, there is no indication so far that children's laughter - at least at this age - displays other characteristics than adults' laughter does, cf. below. We do not know yet whether the distribution we found in our data for *motherese* vs. laughter is typical for human-pet robot communication in general, or for child-pet robot communication in particular.

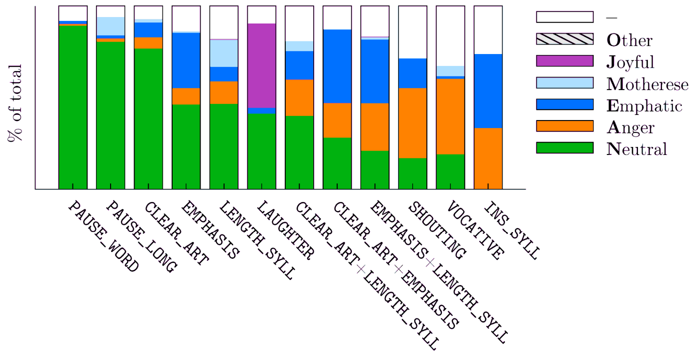

Fig.1: Prosodic peculiarities vs. emotional user states; frequencies in percent of total

Phonetically, speech laughter in our data is often characterized as either tremolo with *2-n* pulses/cycles, or as more or less breathy. Isolated laughter can be some variety of (repeated) [h@], sometimes with laryngealization as delimiter between pulses, or voiceless expiration. In a few cases, inspiration can be observed as well. The following figure displays an example with most characteristic traits: *[na:|hI|?n]* , with tremolo, i.e. poly-syllabification CV →CVCVCV with insertion of "voicelessness". Generally, laughter is characterized by one of more of these traits: tremolo without or with polysyllabification by inserting mostly voiceless segments, sometimes nasalization, and/or breathiness/voicelessness.

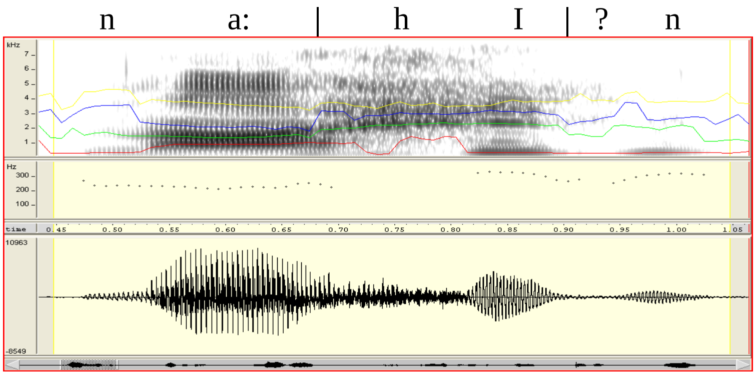

Fig. 2: Mont_10_086: *nein* (*no*); from top to bottom, Sampa transcription, spectrum with formants, pitch contour, and time signal

For automatic classification, we used a large state-of-the-art feature set consisting of 5,967 features representing statistical functionals of 39 low-level feature contours such as energy, pitch, Harmonics to Noise Ratio (HNR), Mel-Frequency Cepstral Coefficients (MFCC), and Spectral features. A single feature vector was computed for every word or word like unit in the corpus. To test automatic discrimination of Speech (S) vs. Speech Laughter (SL) and Laughter (L), 94 words without Laughter selected randomly from the whole corpus were added. As classifier, Support-Vector Machines (SVM) with a linear kernel function, trained using a Sequential Minimal Optimisation (SMO) algorithm, were used.

Mean % correct	**w. FS**	**w/o FS**
M/O	62.6	66.0
LOSO	64.1	**72.0**

Table 2. – Results with Feature Selection (w. FS) and without Feature Selection (w/o FS)

# classified as	**S**	**SL**	**L**
S	**27**	21	5
SL	6	**28**	4
L	2	12	**42**

Table 3. – Confusion Matrix for M/O

Results for two test scenarios are shown in Table 2. In the first test, M/O, the data from Mont is used for training, and the data from Ohm is used for testing; in the second test, a Leave-One-Speaker-Out (LOSO) evaluation is performed and the results are averaged over all 51 individual evaluations. In both tests, best results are achieved using the full feature set; the feature selection conducted seems to select features too specific to the training set. Table 3 displays a confusion matrix for the M/O test (w/o FS). As expected, most confusions occur with class SL, where S or L are mis-classified as SL. The confusion between S and L is very low.

In the next steps, we will investigate different pre-segmentation methods, in order to be able to spot the 3 classes in a continuous audio stream without word alignments; this means at the same time to move from classification to the more realistic task of spotting, using all data. Further, we will investigate the relevance of feature types, using additional databases and expert knowledge, aiming at an optimal feature selection and by that, a reduced set of most relevant features for laughter detection.