The Prosody of Pet Robot Directed Speech: Evidence from Children.

Anton Batliner[*] *& S. Biersack*[†] *& S. Steidl*[*]

[*] Lehrstuhl für Mustererkennung (Chair for Pattern Recognition),
University of Erlangen-Nuremberg, Erlangen, F.R. of Germany
[†] Department of Psychology, University of Stirling, United Kingdom
{batliner;steidl}@informatik.uni-erlangen.de,sb24@stir.ac.uk

Abstract

In this paper, we present a database with emotional children's speech in a human-robot scenario: the children were giving instructions to Sony's pet robot dog AIBO, with AIBO showing both obedient and disobedient behaviour. In such a scenario, a specific type of partner-centered interaction can be observed. We aimed at finding prosodic correlates of children's emotional speech and were interested to see which speech registers children use when talking to AIBO. For interpretation, we left the weighting and categorization of prosodic features to a statistic classifier. The parameters found to be most important were word duration, average energy, variation in pitch and energy, and harmonics-to-noise ratio. The data moreover suggests that the children used a register that resembled mostly child-directed and pet-directed speech and to some extent computer-directed speech.

1. Introduction

To date, research on the vocal expressions of emotions has mainly focused on adults. Moreover, the vast majority of this research has been based on acted emotions ([1] or, at best, on emotions evoked by explicit mood-induction techniques (for an overview see [2]). In the present study we made use of a database of emotional speech produced by children, with emotions solely resulting from the experimental context. The general framework for the database was human-machine – or to be more precise, human-robot – communication. The interaction was embedded in a Wizard-of-Oz task and the robot is Sony's (dog-like) AIBO robot. The speech in this study was spontaneous in that the children were not told to use specific instructions but to talk to AIBO as they would talk to a friend. The different emotional states occured when a child failed or succeeded in instructing AIBO; while the child was led to believe that AIBO was responding to his or her commands, the robot was actually being controlled by a human wizard, thus showing not only obedient but also disobedient behaviour. On the basis of these emotions that came as a reaction to AIBO's 'behaviour' we were interested to see how emotional speech is prosodically coded in children and in the subsequent recognition of these emotional user states. Note that we used a rather broad concept of 'emotion-related' user states, cf. below 2.1, and that in this paper, we were especially interested in 'partner-oriented', interactive speech, not in emotional speech in general.[1]

[1] The fact that non-interactive emotional speech has been by far more investigated than interactive speech is a scientifi c artifact caused by researchers choosing clean, but mostly solipsistic speech as object of investigation. Opinio communis is that speech has originated in and is mostly used in interaction and not in monologue.

The design of the task also allowed to examine which speech registers children were using when interacting with the pet robot dog, and to what extent these speech registers resembled adult speech registers described in the literature. With a pet robot dog as an interlocutor, the types of possible speech registers ranged from computer-directed speech over pet-directed speech, to possibly even child-directed speech. This potential variation of speech registers lay within the nature of AIBO as an interlocutor: AIBO is essentially a computer. Informal observation of the young participants in between recordings indicated, however, that they much more saw AIBO as a toy or even a pet than a machine; and while AIBO looks like a dog, the task led the children to believe that AIBO understands language to an extent that makes it more similar to a child than a pet [3]. Thus our scenario provided data for a specific combination of speech registers: pet robot directed speech produced by children.

The characteristics of child-directed speech and, to some extent, robot- and pet-directed speech are described in the literature. Computer- or robot-directed speech is probably the speech register with the largest inherent variation. Previous research emphasizes on differences between experienced and unexperienced users [4]. Particularly relatively unexperienced users are often found to speak slower, and to use many repetitions and shorter sentences. Depending on the type of robot used as an interlocutor, there is also evidence for robot-directed speech to show similarities to child-directed speech [3]. The unique features of child-directed speech comprise of elevated pitch, a wider pitch range, slower speech rate and hyperarticulation [5]. Pet-directed speech, finally, is often described as being a secondary or extended form of child-directed speech with relatively similar prosodic features, but features such as hyperarticulation, which mainly promote the acquisition of language, missing [6]. We were interested to see how the prosodic features found in our data compared to findings from the literature.

2. Database and Processing

Our data was collected from 51 German children (age 10 - 13, 21 male, 30 female) from two different schools. Recordings took place in classrooms. The only persons present in the room were the child, the supervisor, who gave the instructions, the wizard (behind the children, pretending to be doing the recordings) and a third assistant. Each recording session took some 30 minutes. Due to the experimental setup, these recordings contained a huge amount of silence (due to the reaction time of AIBO), which caused a noticeable reduction of recorded speech after raw segmentation; we eventually obtained about 9.2 hours of speech. The basic idea of this study was to combine a new type of corpus (children's speech) with 'natural' emotional speech within a Wizard-of-Oz task. The 'behaviour' of

AIBO was hence carefully planned to evoke a maximum range of emotions while still allowing comparisons across children. The wizard caused AIBO to perform a pre-determined sequence of actions, which took no account of what the child said. 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 at the same time we did not want to run the risk of them breaking off the experiment. The children believed that AIBO was reacting to their orders - albeit often not immediately. It was, in fact, the other way round: AIBO always strictly followed the same screen-plot, and the children had to align their orders to its actions.

2.1. Labelling Emotions

Five labellers annotated independently from each other each word as neutral (default) or as belonging to one of ten other classes which were obtained by inspection of the data. Our choice of labels was strictly data driven; at the beginning, data of different children were annotated and discussed iteratively, until we found our final set of labels. Then, the whole database was processed. We do not claim that our labels represent children's emotions in general, only that they are adequate for the modelling of these children's behaviour in this specific scenario. We resorted to majority voting (henceforth MV): if three or more labellers agreed, the label was attributed to the word; if four or five labellers agreed, we assumed some sort of prototypes. The following raw labels were used; in parentheses, the number of cases with MV is given: *joyful* (101), *surprised* (0), *emphatic* (2528), *helpless* (3), *touchy*, i.e. irritated (225), *angry* (84), *motherese* (1261), *bored* (11), *reprimanding* (310), *rest*, i.e. non-neutral, but not belonging to the other categories (3), *neutral* (39177); 4705 words had no MV, all in all, there were 48408 words. This does not mean that user states such as *surprised*, *helpless*, or *bored* were hardly labelled at all – it only means that there was no clear agreement between the labellers when to use these labels. More details can be found in [7].

In [8] we considered for classification only labels with more than 50 MVs, resulting in a 7-class problem. *Joyful* and *angry* belong to the 'big' emotions, the other ones belong to 'emotion-related/emotion-prone' user states. The state *emphatic* has to be commented on separately: based on our experience with other emotional databases [9], 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 part of a particular speech register – 'computer talk' – which some people use while speaking to a computer, and which often resembles foreigner-directed speech, child-directed speech or even elderspeak, speech to elderly people hard of hearing.

In this paper, we concentrated on those labels which clearly denote **interactive** speech, namely *motherese* and *reprimanding*; these two constitute, in a NonMetrical MultiDimensional Scaling solution [7] the items with positive values on a dimension that we call 'interaction'. *Neutral* constitutes a baseline. By that we do not want to say that the other user states labelled are not interactive – they are, simply by being used in a communicative setting – but they are less interactive. An indication is, for instance, that verbs denoting these two user states are, at least in German, more transitive, having more slots to fill, than the other ones.

2.2. Prosodic Features

For spontaneous speech it is still an open question which prosodic features are relevant for the different classification problems, and how the different features are interrelated. We tried therefore to be as exhaustive as possible and used a highly redundant feature set leaving it to the statistic classifier to find out the relevant features and to do the optimal weighting of them. For the computation of the prosodic features, a fixed reference point had to be chosen. We decided in favour of the end of a word because the word is a well-defined unit in speech recognition, and because this point can be more easily defined than, for example, the middle of the syllable nucleus in word accent position. 95 relevant prosodic features modelling duration, energy and F0, were extracted from different context windows. The context was chosen from two words before, and two words after, around a word; by that, we used a sort of 'prosodic fivegram' and were able to model a speaker- or at least utterance-specific baseline. However, in our experience context features are - most probably because of sparse data - not that easy to interpret; in this paper we therefore confined our analyses to features computed for the single words.

In pilot experiments using the present database, it turned out that formant-based features as for instance described in [10] could not be computed robustly enough, as adding them to the feature vector never improved classification performance (cf. section 4). In addition to our usual feature vector, we thus decided only to use five harmonics-to-noise-ratio features which were computed frame-wise for each voiced frame and all voiced parts of a word, as well as features modelling jitter and shimmer. For the computation of our features, we assumed 100% correct word recognition and used forced alignment for the spoken word chain. A full account of the strategy for the feature selection or for the choice of a word-based computation is beyond the scope of this paper; details are given in [9]. Here, we wanted to concentrate on acoustic, mostly prosodic features; thus we did not take into account our usual part-of-speech features, cf. [9]. This is a short account of the features used:

- length of filled/unfilled pauses before and after the word
- for energy, duration, and F0: a reference feature based on average values for all words in a turn
- for energy: maximum, mean, absolute value, normalized value, and regression coefficient with mean square error
- for duration: absolute and normalized
- for F0: minimum, maximum, mean, and regression coefficient with mean square error
- harmonics-to-noise-ratio (HNR) features: mean, standard deviation, minimum, maximum, and range
- for jitter and shimmer: mean and variance

3. Classification and Interpretation

We did not use the acoustic features as such as predictors, but principal components (PCs) based on these features. By that, we could reduce the number of predictors even more, and these predictors are orthogonal to each other; this facilitates interpretation. To use all features yields sometimes less classification

performance than selecting those features that are at the same time relevant and not too much correlated with other features. This loss is, however, not too severe. If we reduce the number of predictors to a considerable extent, classification performance goes down. However, by that it is possible to boil down a large number of predictors to a small one which can be interpreted more easily. There is thus always a certain trade-off: the clarity of interpretation is negatively correlated with classification performance. In this paper, we were not interested in optimizing classification performance. We confined ourselves to the reduction of the number of predictors and their interpretation: PC analyses were computed yielding 11 PCs with an eigenvalue greater than 1.0 which were used as predictors in an LDA; the 'semantics' of these PCs will be described below. We only put the three classes *motherese*, *neutral*, and *reprimanding* into the classifier. For a realistic classification, this was contrafactual as we blinded out the other labels. For interpretation, this enabled us, however, to concentrate on those user states we were interested in for the current study. In Table 1, the confusion matrix for our three classes is given; class-wise computed recognition rate (mean of diagonal) CW is 65.9.[2]

Table 1: *Confusion Matrix for the Three Classes in Percent Correctly Classified, Leave-one-out, with resp. Frequencies #*

label	*moth.*	*neut.*	*repr.*	#
motherese	52.6	33.2	14.2	1261
neutral	22.7	67.5	9.8	39177
reprimanding	8.4	13.5	78.1	310

For *n* classes, *n-1* functions had to be computed; functions at group centroid are given in Table 2. The higher the absolute value, the more important is this function for the respective class. The algebraic sign indicates whether higher or lower values are more important. Obviously, *neutral* is really in between the two other classes, with values close to zero. *Motherese* and *reprimanding* display opposite values, for function 1 low for *motherese* and high for *reprimanding*, and for function 2, negative for *motherese* and positive for *reprimanding*.

Table 2: *Functions at Group Centroids*

label	1	2
motherese	.467	-.838
neutral	.032	.021
reprimanding	2.141	.782

In Table 3, the correlation of the eleven PCs with both functions is given; we only interpreted values with a 'reasonable' value > 0.1, i.e. the first three PCs given in Table 3 are important for the first function, the fourth to the eights PCs are important for function 2, the rest is rather less important. The names given for the PCs are shorthand for their semantics. For interpretation, we only took into consideration features with a factor loading > 0.5 in the rotated component matrix.

Function 1 reveals that *reprimanding* displays longer duration and higher average energy: *PC1* is characterized by many features that model duration, *PC8* is composed of the average

[2]If we use all acoustic features, together with our part-of-speech features, for the same constellation, CW is 73.3. Classification performance will most probably be even higher, with feature evaluation, more sophisticated classifiers and the use of linguistic information. But as pointed out earlier, in this paper, we were not interested in tuning classification performance.

Table 3: *Correlation Coefficients between Variables and Functions*

principal components	1	2
PC1 $_{duration}$	.591	-.086
PC10 $_{pause\ before,\ energy\ regression}$	.439	-.104
PC8 $_{average\ duration\ \&\ energy}$	.437	-.258
PC5 $_{energy:\ absolut,\ maximum,\ mean}$	.054	.781
PC3 $_{HNR:\ std.\ dev.,\ maximum,\ range}$	.380	.443
PC6 $_{HNR:\ mean,\ maximum,\ minimum}$	-.035	.218
PC9 $_{mean\ square\ error:\ energy,\ F0}$	.056	-.191
PC11 $_{filled\ pauses}$	.049	-.163
PC7 $_{F0\ regression}$	-.012	.054
PC2 $_{F0\ maximum,\ minimum,\ mean,\ etc.}$	-.028	.039
PC4 $_{jitter,\ shimmer}$	.016	.033

energy and duration value; the second PC *PC10* is composed of higher energy regression coefficient and longer pauses before the respective word – it might indicate that the children typically react reprimanding by producing one-word chunks as, for example, *stop!* pause or *stop!* in a loud voice.

Function 2 is negatively correlated with *motherese* and positively with *reprimanding*. Again, energy is higher for *reprimanding*, *PC5* being characterized by higher energy – absolute, maximum, and mean; HNR is more pronounced in *reprimanding*: *PC6* displays higher mean, maximum, and minimum, *PC3* higher standard deviation, maximum, and range, i.e. there is a higher proportion of non-periodicity and by that, breathiness, in *motherese* which would be in accordance with findings from the literature [11]. A caveat has to be made because HNR is rather dependent on segmental context; thus we do not know yet to which extent these findings will generalize. Mean square error for the regression coefficient of energy and F0 are higher for *motherese* (*PC9*) indicating that there is more 'variation' in *motherese*. Moreover, *motherese* displays some more/longer adjacent filled pauses (*PC11*) than *reprimanding*; this might seem counter-intuitive but it is only a slight tendency – possibly caused by some peculiarities in the data?

Less relevant with a correlation < 0.1 are F0 regression and F0 values (*PC7* and *PC2*), as well as jitter and shimmer (*PC4*). It might seem surprising that pitch is not relevant; this is, however, a result that has been observed throughout, in our experiments and in others, if (and only if) you deal with spontaneous speech and automatically extracted features. There are several – competing or corroborating – possible explanations for that, cf. [12]. *Neutral* might be in between the two other classes because of its frequency, and/or because it simply is the default state, cf. [7] where it has been shown in a two-dimensional representation that *neutral* indeed is close to the origin.

4. Discussion

The aim of this study was to describe the prosodic features of emotional speech of children addressing a pet robot and to examine which speech registers children are using in this set-up. We confined our study to interactive speech leaving aside more 'traditional' emotional user states such as *joyful*, *angry*, etc. The disadvantage of such an approach might be that we only considered a subsample, this has been discussed above; further, that class frequencies were very unequal, which is inevitable when looking at spontaneous speech. We do not know yet, for example, whether formant-based features – which we could not compute robustly enough – are simply not relevant or whether they

would have been if we had had a database greater by some order of magnitude. The advantages of our approach are a full coverage within a realistic scenario, spontaneous data, no selection of 'interesting' cases, and fully automatic processing (albeit, for this purpose, with the spoken word chain, i.e. not based on word hypotheses graphs).

We found that speech that was classified as *motherese* was mainly characterized by lower HNR, shorter duration, lower energy and, at the same time, more variation in energy and F0. Lower energy points towards the soothing features of child-directed speech, the energy and pitch variations towards the attention eliciting features of child-directed speech. The use of breathiness – as indicated by the lower HNR values – is a general strategy towards establishing a more intimate relationship, which is not confined to interaction with children [13]. Features that are described for child-directed speech and are thought to facilitate language acquisition – increased segment duration and with that, most probably hyperarticulation – were, however, not represented in the *motherese* user state of our sample. This seems certainly appropriate for the type of interlocutor the children were interacting with in this study. While AIBO might well have elicited a degree of affect in the children that caused them to produce child-directed speech, and while they did believe that AIBO was listening to what they were saying, there was no apparent reason for the children to believe that AIBO was trying to learn speaking. With the language acquisition component missing, the *motherese* class has in fact just as much in common with pet-directed speech as with child-directed speech. Contrary to what could be expected from a pet, however, the children generally seemed to presume AIBO to have a rather advanced linguistic capability; only if communication broke down, i.e. when AIBO disobeyed, they resorted to some other strategies. One boy, for instance, tried out hyperarticulated spelling: *Stop! Stop! Es Ti Ou Pi!* in such a case. Such strategies were part of the *reprimanding* user state which was mainly characterized by longer duration, higher (average) energy, with longer pauses in between words. Its characteristics were in direct opposition to those of *motherese*. Surprisingly, hardly anything can be found in the literature in view of features of reprimanding child-directed speech or reprimanding pet-directed speech, but a use of shorter utterances or one-word-chunks and louder voice is certainly in accordance with what we know from everyday experience. These features of the *reprimanding* user state could, of course, also be related to the features of computer-directed speech. The features of *motherese* on the other hand seem not to be part of what is normally understood to be computer-directed speech.

5. Conclusions

Taken together, our findings suggest that children of this age, just like adults [14], are able to fine-tune their speech register to the addressee and to their own emotional state. What they produce when interacting with AIBO is probably most closely related to pet-directed speech, with overlaps with child-directed and computer-directed speech. These overlaps seem to be representative of the untypical addressee: AIBO is in its role essentially a mixture of pet and robot, with childlike traits.

6. Acknowledgements

This work was partly funded by the EU in the framework of the two projects PF-STAR under grant IST-2001-37599 and HUMAINE under grant IST-2002-507422, and by the German Federal Ministry of Education and Research (*BMBF*) in the framework of the two projects SmartKom (Grant 01IL905K7) and SmartWeb (Grant 01IMD01F). The responsibility for the contents of this study lies with the authors.

7. References

[1] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "Desperately Seeking Emotions: Actors, Wizards, and Human Beings," in *Proc. ISCA Workshop on Speech and Emotion*, R. Cowie, E. Douglas-Cowie, and M. Schröder, Eds., Newcastle, September 2000, pp. 195–200.

[2] K. Scherer, "Vocal Communication of Emotion: A Review of Research Paradigms," *Speech Communication*, vol. 40, pp. 227–256, 2003.

[3] C. Breazeal and L. Aryananda, "Recognition of Affective Communicative Intent in Robot-Directed Speech," *Autonomous robots*, vol. 12, pp. 83–104, 2002.

[4] L. Bell and J. Gustafson, "Utterance types in the August dialogues," in *Proc. IDS 99, ESCA workshop on Interactive Dialogue in Multi-Modal Systems*, 1999, pp. 81–84.

[5] A. Fernald, T. Taeschner, J. Dunn, M. Papoušek, B. Boysson-Bardies, and I. Fukui, "A Cross-Language Study of Prosodic Modifications in Mothers' and Fathers' Speech to Preverbal Infants," *Journal of Child Language*, vol. 16, pp. 477–501, 1989.

[6] D. Burnham, C. Kitamura, and U. Vollmer-Conna, "What's New Pussycat? On Talking to Babies and Animals," *Science*, vol. 296, p. 1435, 2002.

[7] A. Batliner, S. Steidl, C. Hacker, E. Nöth, and H. Niemann, "Private Emotions vs. Social Interaction - towards New Dimensions in Research on Emotion," in *Proc. Workshop on Adapting the Interaction Style to Affective Factors, 10th Int. Conf. on User Modelling*, Edinburgh, 2005, 8 pages, no numbering.

[8] ——, "Tales of Tuning – Prototyping for Automatic Classification of Emotional User States," in *Proc. 9th Eurospeech - Interspeech 2005*, Lisbon, 2005, pp. 489–492.

[9] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "How to Find Trouble in Communication," *Speech Communication*, vol. 40, pp. 117–143, 2003.

[10] R. Tato, R. Santos, R. Kompe, and J. Pardo, "Emotional space Improves Emotion Recognition," in *Proc. ICSLP 2002*, 2002, pp. 2029–2032.

[11] N. Campbell and P. Mokhtari, "Voice Quality: The 4th Prosodic Dimension," in *Proc. 15th ICPhS*, 2003, pp. 2417–2420.

[12] A. Batliner, J. Buckow, R. Huber, V. Warnke, E. Nöth, and H. Niemann, "Boiling down Prosody for the Classification of Boundaries and Accents in German and English," in *Proc. 7th Eurospeech*, Aalborg, 2001, pp. 2781–2784.

[13] J. Laver and P. Trudgill, *The Gift of Speech*. Edinburgh University Press, 1991, ch. Phonetic and Linguistic Markers in Speech, pp. 235–264.

[14] S. Biersack, V. Kempe, and L. Knapton, "Fine-Tuning Speech Registers: A Comparison of the Prosodic Features of Child-Directed and Foreigner-Directed Speech," in *Proc. 9th Eurospeech - Interspeech 2005*, Lisbon, 2005, pp. 2401–2404.