Tales of Tuning – Prototyping for Automatic Classification of Emotional User States

Anton Batliner, Stefan Steidl, Christian Hacker, Elmar Nöth, Heinrich Niemann

Lehrstuhl für Mustererkennung (Chair for Pattern Recognition),
University of Erlangen-Nuremberg, Erlangen, F.R. of Germany
batliner@informatik.uni-erlangen.de

Abstract

Classification performance for emotional user states found in the few realistic, spontaneous databases available is as yet not very high. We present a database with emotional children's speech in a human-robot scenario. Baseline classification performance for seven classes is 44.5%, for four classes 59.2%. We discuss possible strategies for tuning, e.g., using only prototypes (based on annotation correspondence or classification scores), or taking into account requirements and feasibility in possible applications (weighting of false alarms or speaker-specific overall frequencies).

1. Introduction

"Siobhan also says that if you close your mouth and breathe out loudly through your nose it can mean that you are relaxed, or that you are bored, or that you are angry and it all depends on how much air comes out of your nose and how fast and what shape your mouth is when you do it and how you are sitting and what you said just before and hundreds of other things which are too complicated to work out in a few seconds." (The Curious Incident of the Dog in the Night Time, by Mark Haddon, 2003) Shiobhan is right: it is not always easy for human beings to recognize emotions, and for machines, the same is true – if we are speaking of spontaneous emotions, not of acted ones: for acted speech and for four (or even more) classes, recognition rates can be above 90%. For spontaneous speech, however, recognition rates for a two-class problem are as yet normally below 80%, and for a four-class problem, below 60%.

This is a very rough picture of the processing chain: the 'inner circle' consists of the **event** – the phenomenon we want to classify, the necessary ingredients (**annotation** and **extracted features**), and the outcome – **classified data**. Recording **context** and **application** aimed at should be as close to each other as possible. Frequently used strategies for improving classification are to collect more data or to employ highly sophisticated classifiers. The problem with *"there's no data like more data"* is that sparse data prevail: in our experience, only some 10% or even less than 5% of spontaneous data are 'interesting'; the effort needed might thus be greater by some order of magnitude than the one needed for speech recognition. The problem with classifiers is that till now, no clear picture has emerged: one of the very few studies on benchmarking [1] concludes that highly sophisticated procedures such as Support Vector Machines are not necessarily superior to more traditional ones. Moreover, it is a dilemma that frequent benchmarking would tend to violate statistic validity [2] by not taking into account the multiplicity effect. More frequent is, however, to employ a bunch of different classifiers within a specific study and simply to report performance differences; we can, however, never be sure whether these results really generalize; the only tenable criterion – cumulative evidence – is not met.

Thus it might be worth while to have a look at other strategies in between; in this paper, we want to concentrate on annotations and subsequent prototyping, and on the eventual requirements put forth in possible applications; other promising 'building sites' are dealt with in the final section.

2. Material, annotation, and features

The general frame for the database 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. 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. 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).

The wizard causes the AIBO to perform a fixed, predetermined 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.

In this paper, we want to concentrate on the German recordings; parallel English data recorded at University of Birmingham are described in [3]. The data was collected from 51 children (age 10 - 13, 21 male, 30 female). The children are from two different schools; the recordings took place in two classrooms. Each recording session took some 30 minutes. Because of the experimental setup, these recordings 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.

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; 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. We resort to majority voting (henceforth MV): if three or more labellers agree, the label is attributed to the word; if four or five labellers agree, we assume 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. For classification, we consider only labels with more than 50 MVs, resulting in a **7-class problem**. *joyful* and *angry* belong to the 'big' emotions, the other ones rather to 'emotion-related/emotion-prone' user states. The state *emphatic* has to be commented on especially: based on our experience with other emotional databases [4], 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. 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 communication.

From a methodological point of view, the 7-class problem is most interesting, cf. the new dimensional representation of these seven categorical labels in [6]. However, the distribution of classes is very unequal. Therefore, we downsampled *neutral* and *emphatic* and mapped *touchy* and *reprimanding*, together with *angry*, onto **A***ngry*[1] as representing different but closely related kinds of negative attitude. For this more balanced **4-class problem** 'AMEN', 1557 words for **A***ngry*, 1224 words for **M***otherese*, and 1645 words each for **E***mphatic* and for **N***eutral* are used, cf. [5]. Cases where less than three labellers agreed were omitted as well as those cases where other than these four main classes were labelled.

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 try therefore to be as exhaustive as possible, and we use a highly redundant feature set leaving it to the statistic classifier to find out the relevant features and the optimal weighting of them. For the computation of the prosodic features, a fixed reference point has to be chosen. We decided in favour of the end of a word because the word is a well-defined unit in word 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, are extracted from different context windows. The context could be chosen from two words before, and two words after, around a word; by that, we use so to speak a 'prosodic five-gram' and can model some sort of speaker- or at least utterance-specific baseline. 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 is beyond the scope of this paper; details are given in [4]. Additionally, we included some nine 'spectral' features modelling jitter, shimmer, and harmonicity-to-noise ratio.

A Part of Speech (POS) flag is assigned to each word in the lexicon. Six cover classes are used: AUX (auxiliaries), PAJ (particles, articles, and interjections), VERB (verbs), APN (adjectives and participles, not inflected), API (adjectives and participles, inflected), and NOUN (nouns, proper nouns), i.e., for the context of +/- two words, $6 \times 5 = 30$ features. An additional feature is a flag denoting whether the word is a fragment or not. All in all, our feature vector thus comprises 135 features.

3. Classification: baseline and prototyping

For the experiments reported on in the following, we always use a simple linear classifier (LDA) with the full feature vector.[2] By using two-fold cross-validation (children from one school for training, from the other school for test, and vice versa) we can reduce effort and at the same time, secure strict speaker independence. Reported is the **CL**ass-wise computed recognition rate CL as mean value of the two cross-validations.[3]

Table 1: *classification rates in percent, 7 and 4 classes*

	7 classes		4 classes	
constellation	#	CL	#	CL
all = MV {3,4,5}	43686	44.5	6071	59.2
MV {3}	13374	35.5	3271	50.1
MV {4,5}	30312	43.2	2800	70.0
prob. $\geq .8$	22990	53.9	4284	73.5
MV {4,5}+ prob. $\geq .8$	-	-	2067	77.5

In Table 1 we display number of items # and percent correctly classified CL for the following constellations:

- *all = MV {3,4,5}*: all cases with MV ≥ 3 used for training and testing
- *MV {3}*: only cases with MV $= 3$ used for training and testing

[1] If we refer to the resulting 4-class problem, the initial letter is given boldfaced and recte. Note that now, **A***ngry* can consist, for instance, of two *touchy* and one *reprimanding* label; thus the number of **A***ngry* cases is far higher than the sum of *touchy*, *reprimanding*, and *angry* MV cases.

[2] In pilot experiments, more sophisticated classifiers as NN were not much better; neither yielded a reduction of the number of features much better classification rates; as we focus in this paper on other aspects, we postpone optimization of classifiers and feature vectors, cf. below last section.

[3] Although LDA is relatively robust as for unequal distribution, for the unbalanced 7-class problem, *neutral* is always classified much better than the other classes. For in-depth interpretation, we therefore resort to the more balanced 4-class problem.

- *MV {4,5}*: only cases with MV ≥ 4, i.e., prototypes used for training and testing
- *only prob.* $\geq .8$: all cases with MV ≥ 3 used for training; only 22990 or 4284 cases, resp., with a class assignment with high probability $\geq .8$ are given in the table
- *MV {4,5}+ only prob.* $\geq .8$: only cases with MV ≥ 4, i.e., 2800 prototypes used for training; out of these cases, only those 2067 with a class assignment with high probability $\geq .8$ are given in the table

We can see that for the 7-class problem in Table 1, the prototypes *MV {4,5}* do not yield better recognition rates than *all* cases – most probably because for some classes, there were not enough items for a robust classification (remember that we use 135 features!). For the more balanced 4-class problem, however, prototypes are classified some 10 percent points better than the baseline. If training is done with all cases but performance is only computed for the cases with high probability, the classification rate is better, 53.9% for the 7-class and 73.5% for the 4-class problem. The best figure can be obtained if we combine the two selection criteria in the last line of Table 1 yielding 77.5% for the 4-class problem. Due to sparse data, no figure is here given for the 7-class problem.

The better performance of **prototypes** indicates that the emotional user states labelled are either a graded and/or a mixed phenomenon: obviously, there are more or less clear cases.[4] A classification performance of up to 77.5% for four classes and for real life data looks good; however, we achieved this only by leaving aside two third (for 77.5%) or one third of the items (for 73.5%) – those cases which did not meet our selection criteria. Note that due to the persistent sparse-data problem which holds for our data as well, researchers sometimes only deal with 'interesting' chunks cut out of longer passages; we have seen that by following comparable strategies, we really can improve classification performance. Such prototypes can be very valuable for modelling but we have to keep in mind that, at the same time, we sort of blind out reality up to some extent. In the next section, we will therefore go back to our 4-class problem with all cases (baseline, first line in Table 1) and sketch some applications.

4. Online vs. offline applications

In this section, we want to concentrate on some possible applications. For that, we will tell apart **online** application (the system reacts immediately to some emotional user state and by that, influences the interaction with the user) from **offline** application (the system does not interact with the user but draws conclusions based on the emotional user states found in the interaction). Online application can be more touchy: imagine a system monitoring user state in a car turning off the engine if the driver seems to be highly aroused – and he is not. (Even if he were, this is not a feasible application but only a striking example.) Thus we have to imagine an application where correct classification adds to system performance and pleases the user whereas some false alarms do not harm up to a large extent. If we take our scenario 'child playing with a pet robot' as an extension of the old scenario 'child playing with a teddy-bear': the one and only possible sound produced by the bear was interpreted sort of top-down differently by the child depending on the phase of interaction, i.e., the context played a major role. Thus the pet robot's actions do not have to be fully clear, it might suffice if they are consistent and not contradicting severely the expected behaviour. Let's assume that the AIBO reacts positively to **M***otherese* by for instance looking towards the child, waging it's tail etc., and behaves repentantly to **A***ngry* by for instance sitting down, whining, etc. As even a real pet dog's actions are not always fully understood by humans, a confusion with **N***eutral*/**E***mphatic* might not matter much; only a confusion of **M***otherese* with **A***ngry* and vice versa might puzzle the child. In the second to fifth column of Table 2, the confusion matrix is given for the first line *all*= *MV {3,4,5}* of Table 1. The severe false alarms **M***otherese* $\rightarrow$ **A***ngry* amount to 11.6%, and **A***ngry* $\rightarrow$ **M***otherese* to 4.8%. Remember that we downsampled esp. **N***eutral* by some order of magnitude. As we assume that a confusion with **N***eutral* does not matter much, a 'fatal' confusion might really occur very seldom.

Table 2: *confusion matrix for 4 classes in percent correctly classified, and correlation of # labels with # classified (Spearman/Pearson)*

label	**M**	**N**	**E**	**A**	Spearm./Pears.
Motherese	53.0	31.7	3.8	11.6	.83/.98
Neutral	12.4	59.0	14.4	14.3	.75/.69
Emphatic	1.8	17.8	64.4	16.0	.81/.80
Angry	4.8	17.1	18.2	60.1	.83/.83

Even less fatal is a single misrecognition in an offline application where frequencies matter and not single events. Imagine that we are interested in the evaluation and follow-up screening of children's linguistic and interactive social behaviour. There are many speech development tests available but not that many that deal with (linguistic) interaction – aiming at some alternative IQ, i.e., **I**nteraction **Q**uotient. We computed for our 4-class problem the sum of each of the labels for each subject and correlated this value with the sum of the correctly classified labels for each label and for each subject, again based on a two-fold cross-classification. In Table 2, Spearman's rho (non-parametric correlation) is .83 for **M***otherese*, .75 for **N***eutral*, .81 for **E***mphatic*, and .83 for **A***ngry*; for **M***otherese*, Pearson's parametric correlation coefficient is even higher, due to one outlier with many **M***otherese* items. Thus we might be able to use an automatic procedure for detecting overall frequencies, i.e., trends, for the marked user states **M***otherese* and **A***ngry* which both display a correlation above .8.

Figure 1 shows a scatterplot with the distribution of **M***otherese* vs. **A***ngry* per speaker (leaving aside the one outlier subject which displays very high frequencies for both). Spearman's rho for these two distributions is .47 (without the outlier) or .50 (with the outlier). There seem to be, however, two distinct trends in this plot: one type of children tends towards using **A***ngry* but not (much) **M***otherese*, another type uses both. Maybe we can even tell apart three different interaction types: one addresses the robot as a sort of remote control tool, without showing much emotions. The second one is sort of mixed, showing anger sometimes, and the third one addresses the AIBO really as an interaction partner, as a real pet: encouraging, if need be, and reprimanding, if need be. Here, the **target prototypes** are thus at the origin (no interactive behaviour at all, only commands), high on the y-axis and low on the x-axis (showing **A***ngry*), and high on both axes (showing both **M***otherese* and **A***ngry* which means a fully developed interactive behaviour). If children belong to the third type, we

[4] At the moment, we therefore run an additional annotation pass where the labellers can annotate more than one label for the same word and assign percentages for each label summing up to 100%.

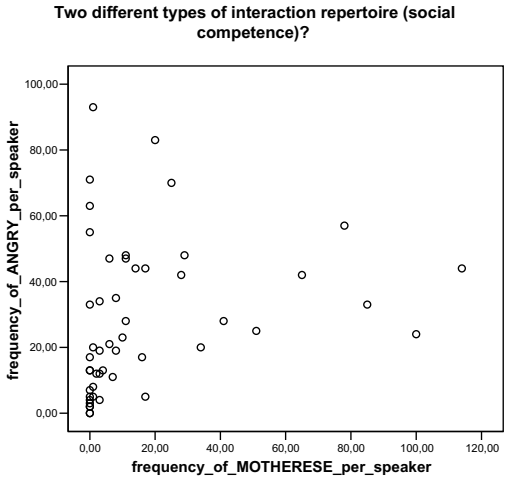

Figure 1: *Scatterplot: Distribution of Motherese and Angry per Speaker*

can conclude that they use a more elaborated linguistic and by that, interaction repertoire. It is an interesting question whether such an elaborated repertoire goes along with a higher social competence. Furthermore we can find out whether there are gender-specific differences: in our database, girls tend to use more **M***otherese* and less **A***ngry* than boys. This difference is, in a two-tailed t-test, not significant but in a one-tailed – as this difference was not formulated as alternative hypothesis, we had to use the two-tailed test. Anyway, it seems to be promising to use automatic classification procedures for such screening tests, especially if we consider that till now, we only took into account acoustic cues, disregarding all other possible cues which might contribute to a better performance. Some of these will be sketched in the following, final section.

5. Concluding remarks and future work

We have shown that it might be worth while to concentrate not only on (overall) classification performance but on the notion of prototypical examples (found in the annotations) and prototypical targets (envisaged in possible applications). For classification, we concentrated on acoustic, mostly prosodic, features, and standard procedures. In the following list, we display other possible, promising sources of information: optimized acoustic features, linguistic features (language model, lexicon, syntactic-sematic chunking) [4, 7], interaction background, multi-modal features [8], and personality traits. These sources will possibly be not additive in a simple sense; on the contrary, if one tries, e.g., to consider multi-modal information, intervening factors might – at least in the beginning – deteriorate performance. All of these factors are, together with those mentioned in the introduction, namely sparse data and classifier evaluation, worth to be addressed. We want to deal with some of them in an initiative within the Network of Excellence HUMAINE which is called CEICES, i.e., **C***ombining* **E***fforts for* **I***mproving automatic* **C***lassification of* **E***motional user* **S***tates, a 'forced cooperation' initiative* where different sites will take part: an old and yet unresolved problem is, for instance, the relevance of pitch vs. other types of features. Traditionally, pitch was considered to be most important; however, almost all studies which used automatically extracted feature values reported that pitch is less relevant than, e.g., duration or energy. We do not know whether this is because pitch is really less relevant or whether this is due to gross extraction errors which might deteriorate performance for several cases. Therefore, we have manually corrected F0 values in our database which enables us to have a look at the relevance of pitch vs. other feature types. Further, at different sites, different (types of) features have been implemented. If we pool all these features, we can try to aim at a sort of 'hyper'-vector encompassing all those features which showed best performance. By using different and highly sophisticated classifiers used at different sites for one and the same database, we at least meet one of the criteria for benchmarking, namely keeping databases constant across studies.[5]

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 German Federal Ministry of Education and Research (*BMBF*) in the framework of the two projects SmartKom (Grant 01 IL 905 K7) and SmartWeb (Grant 01IMD01F). The responsibility for the contents of this study lies with the authors.

7. References

[1] D. Meyer, F. Leisch, and K. Hornik, "Benchmarking Support Vector Machines," Report Series No. 78, Adaptive Informations Systems and Management in Economics and Management Science, 2002.

[2] S. Salzberg, "On comparing classifiers: Pitfalls to avoid and a recommended approach," *Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 317–328, 1997.

[3] A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D'Arcy, M. Russell, and M. Wong, ""You stupid tin box" - children interacting with the AIBO robot: A cross-linguistic emotional speech corpus," in *Proc. LREC 2004*, Lisbon, 2004, pp. 171–174.

[4] 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.

[5] S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann, ""Of All Things the Measure is Man": Automatic Classification of Emotions and Inter-Labeler Consistency," in *Proc. ICASSP 2005, Philadelphia, U. S. A.*, 2005.

[6] A. Batliner, S. Steidl, C. Hacker, and E. Nöth, "Private Emotions vs. Social Interaction - towards New Dimensions in Research on Emotion," in *Proc. Workshop on Adapting the Interaction Style to Affective Factors, User Modelling 2005*, Edinburgh, 2005, to appear.

[7] L. Devillers, L. Lamel, and I. Vasilescu, "Emotion Detection in Task-Oriented Spoken Dialogs," in *ICME*, Baltimore, July 2003.

[8] R. Cowie and R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication*, vol. 40, pp. 5–32, 2003.

[5] Other promising approaches as, e.g., personalized, speaker-dependent modelling and the exploitation of multi-modal cues are beyond the scope of this database.