Private emotions versus social interaction: a data-driven approach towards analysing emotion in speech

Anton Batliner, Stefan Steidl, Christian Hacker, Elmar Nöth

Private Emotions vs. Social Interaction – a Data-driven Approach towards Analysing Emotion in Speech

Anton Batliner, Stefan Steidl, Christian Hacker, Elmar Nöth
Lehrstuhl für Mustererkennung, Martensstr. 3, 91058 Erlangen, F.R. of Germany

September 7, 2007

Abstract. The 'traditional' first two dimensions in emotion research are VALENCE and AROUSAL. Normally, they are obtained by using elicited, acted data. In this paper, we use realistic, spontaneous speech data from our 'AIBO' corpus (human-robot communication, children interacting with Sony's AIBO robot). The recordings were done in a Wizard-of-Oz scenario: the children believed that AIBO obeys their commands; in fact, AIBO followed a fixed script and often disobeyed. Five labellers annotated each word as belonging to one of eleven emotion–related states; seven of these states which occurred frequently enough are dealt with in this paper. The confusion matrices of these labels were used in a Non-Metrical Multi-dimensional Scaling to display two dimensions; the first we interpret as VALENCE, the second, however, not as AROUSAL but as INTERACTION, i.e., addressing oneself *(angry, joyful)* or the communication partner *(motherese, reprimanding)*. We show that it depends on the specifity of the scenario and on the subjects' conceptualizations whether this new dimension can be observed, and discuss impacts on the practice of labelling and processing emotional data. Two-dimensional solutions based on acoustic and linguistic features that were used for automatic classification of these emotional states are interpreted along the same lines.

Keywords: emotion, speech, dimensions, categories, annotation, data-driven, non-metrical multi-dimensional scaling

1.   Introduction

Most of the research on emotion in general and on emotion in speech in particular conducted in the last decades has been on elicited, acted, and by that rather full-blown emotional states. Of course, this means that the data obtained display specific traits: trivially but most importantly, the subjects only displayed those states that they have been told to display. The set of labels is thus **pre-defined**. The better actors the subjects were, the more pronounced and by that, easier to tell apart, these emotions were. The models and theories based on such data are normally not called 'data-driven' – however, in fact they are because they were founded and further developed with the help of these – pre-defined – data.

In linguistics and phonetics, the state of affairs had been similar: for decades, tightly controlled (and by that, pre-defined as well) and/or **'interesting'** data were objects of investigation - 'interesting' not be-

Manuscript

cause they were representative but because they were distinct and at the same time, well-suited to help deciding between competing theories, models, or explanations. However, when all these models had to be put into real practice, i.e., when real-life, spontaneous speech had to be processed, researchers learned that 'all of a sudden', their data looked pretty much different, and that their models could not be used any longer as such (Müller and Kasper, 2000). In the same vein, in the last decade, non-acted data were considered to be more and more important in research on emotion as well (Campbell, 2006).

1.1. Emotions and related states

An overview of emotional phenomena that are encoded in speech is given in (Cowie and Cornelius, 2003). We will address both 'emotions' in a narrow sense and 'emotion-related, affective states' in a broader sense, cf. (Scherer, 2003), p. 243, who lists the following types of affective states: emotion, mood, interpersonal stances, attitudes, and personality traits. Interpersonal stances are specified as "affective stance taken towards another person in a specific interaction, colouring the interpersonal exchange in that situation". (Schröder, 2004) gives a short overview of the multiple meanings of the word "emotion" and of the theories these different meanings are based on, such as the Darwinian, the Jamesian, the cognitive (with the central concept of appraisal), and the social constructivist perspective.

1.2. Categories vs. Dimensions

Broadly speaking, there are two different conceptualizations of emotion phenomena that are mirrored in the type of annotation performed for databases: dimensions and categories. Dimensions have been established by (Wundt, 1896), and for the first time used for judgments on emotional facial expressions by (Schlosberg, 1941; Schlosberg, 1952; Schlosberg, 1954). In the dimensional approach, emotion dimensions such as AROUSAL/ACTIVATION (high/low or active/passive), VALENCE/EVALUATION (negative/positive), and CONTROL/POWER (high/low) are assumed; emotional phenomena are annotated on continuous scales. Normally, only the two most important dimensions are used (Picard, 1997) which we henceforth will address as AROUSAL and VALENCE.[1] In contrast, a discontinuous, categorical conceptualization uses categories like the big n emotions (*anger, fear, sadness, disgust,*

---

[1] CONTROL would be necessary to tell apart, for instance, *angry* (high CONTROL) from *desperate* (low CONTROL), cf. Fig. 1. (Kehrein, 2002), p. 111, lists several other dimensions that have been proposed but are rather marginal nowadays such as: *attention–rejection, interest–lack of interest, yielding–*

Manuscript

etc.) or, within a broader concept, terms denoting different types of emotion-related states. Categories can be integrated into a theoretical, hierarchical system as in (Ortony et al., 1988), p. 191, who define emotions as "[...] *valenced reactions to events, agents, or objects, with their particular nature being determined by the way in which the eliciting situation is construed.*"; these authors argue against any dimensional representation: "It seems to us that the distinct emotion types cannot be arranged informatively into any single space of reasonably low dimensionality." (Ortony et al., 1988), p. 15.

In practice, categories are annotated as such, by using the term that describes best the phenomenon. The two conceptualizations are mapped onto each other by placing category labels onto appropriate positions within the two-dimensional emotional space with VALENCE and AROUSAL as dimensions, cf. (Cowie and Cornelius, 2003). Normally, this has been achieved by similarity judgment experiments using, e.g., the semantic differential (Osgood et al., 1957). Here, the position in the multidimensional space is obtained empirically; the dimensional terms themselves are pre-defined. Fig. 1 is a graphical representation of the two emotion dimensions VALENCE and AROUSAL (Cowie et al., 2000) with some prototypical emotions arranged in this space. These 'traditional' dimensions VALENCE and AROUSAL have been developed by looking at prototypical, acted emotions, be it for speech or for facial gestures. This holds for the 'traditional' category labels as well. Matters are different if we go over to real-life data: full-blown emotions are getting less important. As it turns out, interpersonal relations are coming to the fore instead. The alternative benefits and disadvantages of categorical vs. dimensional descriptions are summarized in (Cowie and Schröder, 2004), p.312: '[...] categorical and logical descriptions raise difficult statistical problems when there is a substantial range of emotions to deal with, dimensional descriptions are more tractable but fail to make important distinctions.'

1.3. Concepts, Data, Annotation and the Automatic Recognition of Emotional Speech

A dimension is rather a 'higher level', theoretical concept, encompassing several different categories, and more closely attached to models than categories. The latter ones can, of course, be 'higher level' as well, and can be used in a multi-layered, hierarchical description system (Ortony et al., 1988) but they can also be used in pre-theoretical, everyday language. In this section we will give a short account of the state of

---

resisting, *destruction–protection*, *reproduction-deprivation*, *incorporation-rejection*, *orientation-exploration*, or *relatedness*.
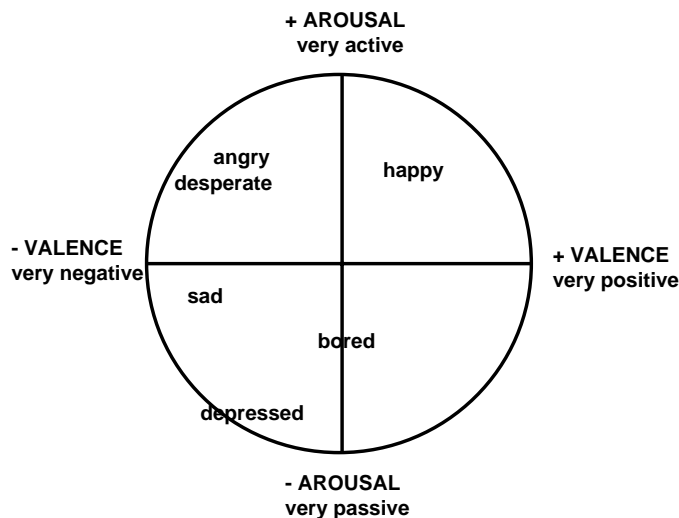
Manuscript

Figure 1. Graphical representation of the two emotion dimensions considered to be most important with some prototypical categories

the art in the automatic recognition of emotional, realistic speech; we will concentrate on the operationalisations of these different concepts utilized in this field.

We distinguish between acted, prompted speech and non-acted, non-prompted speech; the latter will be called 'spontaneous speech' as well. Of course, there are different degrees of spontaneity and different degrees of realism in the recording scenario which are, however, not necessarily co-varying: spontaneity of speech goes together with informality of the situation; realism of emotions felt and/or expressed can be different for volunteering subjects behaving 'as if' they were in a specific situation, and for subjects being in real-life situations. (Note, however, that volunteering subjects pretending to be for instance interested in specific flight connections are not necessarily pretending getting angry if the system fails repeatedly to understand; normally, they really are.) The most important distinction is that between prompted vs. non-prompted speech.

The first paper on automatic emotion recognition using non-prompted, spontaneous speech was maybe (Slaney and McRoberts, 1998) on parents talking to their infants. At the turn of the century, studies concentrated on scenarios modelling human-machine communication: in (Batliner et al., 2000a; Batliner et al., 2000b), volunteering subjects

Manuscript

were recorded communicating with a so called Wizard-of-Oz (WoZ) system, i.e. a human operator pretending to be a system (appointment scheduling dialogues). (Ang et al., 2002) used volunteers calling an automatic system, (Lee et al., 2001) data from real users of a call-center application. All these studies were restricted to modelling a mapping onto a two-way distinction *negative* (encompassing user states such as *anger*, *annoyance*, or *frustration*) vs. the complement, i.e. *neutral*, even if at the beginning, more classes were annotated such as in (Ang et al., 2002) *neutral*, *annoyed*, *frustrated*, *tired*, *amused*, *other*, *not-applicable*. The minor reason for this mapping onto negative VALENCE vs. neutral/positive VALENCE was that in the intended application, it is most important to detect 'trouble in communication' (Batliner et al., 2003a). The major reason is simply that for statistical modelling, enough items per class are needed: the relation of non-marked / marked emotional user states is at best Pareto-distributed, i.e., 80% / 20%, but normally much more biased, up to >95% non-marked cases.

(Devillers et al., 2005) give a survey of these emotion detection studies and the labels used; the situation has not changed much recently: (Neiberg et al., 2006) model, label and recognize a three-way distinction *neutral*, *emphatic* and *negative* for one database (voice controlled telephone service), and for another (multi-party meetings), a three-way emotional VALENCE *negative*, *neutral*, and *positive*. (Devillers and Vidrascu, 2006) established an annotation scheme with a coarse level (8 classes) and a fine-grained level (20 classes) plus neutral for annotation; a coarse label is, for example, *anger* with the fine-grained sub-classes *anger*, *annoyance*, *impatience*, *cold anger*, and *hot anger*. For processing and classifying their real-life database (medical emergency call center), they use the four classes *anger*, *fear*, *relief*, and *sadness*. (Ai et al., 2006) use a three-way distinction for student emotion in spoken tutoring dialogs: *mixed/uncertain*, *certain*, and *neutral*. (D'Mello et al., 2008) model and classify five classes (*boredom*, *confusion*, *flow*, *frustration*, and *neutral*) in a tutoring scenario. In some few studies, up to seven different emotional user states are classified, cf. (Batliner et al., 2003c) (volunteers interacting with an information kiosk in a multi–modal setting) and the present paper; however, this 7–class problem cannot be used for real applications because classification performance is simply too low.

Even if some of these studies refer to (the possibility of) a (not yet existing) principled and fine-grained framework of annotation, in fact, all use eventually a data-driven, condensed annotation system with only a few categories.[2] As mentioned above, this is foremost simply

---

[2] Note that this is not confined to studies on automatic processing of emotions but might be characteristic for studies on real-life data in general. (Scherer and

Manuscript

due to the necessity of generating a representative set for training the classifiers with enough items (tokens) per class (type); of course, such a set is scenario-specific. Note that there is no exact definition of 'enough'; this depends on the number of features used for classification, on the variability within categories, on classifier performance, and on importance for intended applications.[3]

Basically, there are two different approaches towards annotations: an expert-based one, and a – more or less – 'naive' one. In the expert-based approach, chances are that the labelling is consistent but only corroborates the theory (Batliner and Möbius, 2005), yielding reliability but not validity; in the 'naive' approach, chances are that labelling is not consistent. Here validity is of course only granted if the reference, i.e., the classes that have to be annotated, is meaningful. For the experiments presented in this paper, we opted for a compromise, i.e., we instructed students iteratively – by that, they got experts without any theoretical bias – and relied on intersubjective correspondence. 10 labellers might have been an ideal number but this is normally too much effort; three labellers are the minimum for majority decision, five are a good compromise for telling apart weak from strong correspondences.

So far, studies on automatic emotion recognition have not really incorporated theoretical approaches towards emotion — and vice versa: emotion recognition is data-driven and application-oriented, emotion theories are model-driven and generic. In end-to-end systems, an 'up-link' to a theoretical model has to be mediated by more practical system requirements. This has been implemented in the SmartKom system (Streit et al., 2006); however, the complexity of this task resulted in several constraints: in order to obtain good and stable multi-modal recognition performance, the system had to be re-trained with acted data (Zeißler et al., 2006); the spontaneous speech data available (Batliner et al., 2003c) could not be used for this demonstration system. This implementation of the OCC model (Ortony et al., 1988) was restricted

---

Ceschi, 2000), p. 330 ff. use in the same vein for their rating of own or other's feeling states five combined categories: *angry/irritated, resigned/sad, indifferent, worried/stressed, in good humor.*

[3] As far as we can see, frequency as edge condition is not really discussed frequently in theoretical approaches towards emotion which heavily rely on example-based reasoning. Thus frequencies might not be constitutive in theory building but can, however, be of pivotal importance in social relationships, cf. the stereotypical male-female interaction: if a husband tells his wife once a year that he loves her, this constitutes a marital use case but might not prevent her from leaving him because for her, once a week or once a day would be the preferred frequency. It might be no coincidence that in our data, girls used markedly more *motherese* than *angry* than boys did (Batliner et al., 2005b); note that these labels are described below in section 3.

Manuscript

to some few so-called use cases; thus this module could be shown to be functional on a principled basis but had to await much more systematic testing and more robust recognition modules to be functional in any practical application.

1.4. Overview

In the introduction, we shortly described the key concepts **dimensions** vs. **categories** in emotion research and sketched their relevance for the processing of real-life data. An overview of annotation practice for automatic recognition of realistic, spontaneous emotional speech was given. In the following chapter 2, we will present material and experimental design. Chapter 3 describes our annotations with emotion-related labels, conducted by five annotators. In chapter 4, we introduce Non-Metrical Multi-Dimensional Scaling (NMDS). As we employed several labellers, it is possible to compute confusion (similarity) matrices between each pair of labellers and/or average them across all labellers. These matrices were then fed into an NMDS analysis resulting in a two-dimensional representation of similarities and by that, of meaningful dimensions. This procedure was applied first to our German AIBO corpus (chapter 5), then to a parallel English corpus and another corpus with call-center data (chapter 7). In chapter 6 we interpret confusion matrices and dimensional solutions and relate them to theoretical approaches towards the social aspect of emotions. The labels chosen and annotated represent the 'ground truth' (reference) for automatic classification: the *significatum*. Automatic classification is done with the help of acoustic and linguistic features which can be called the *significans*. Result is again a confusion matrix for our labels, but this time not based on manual annotation but on automatic classification. In chapter 8, we present two-dimensional representations based on classifications using different types of features and discuss differences w.r.t. the solutions put forth in chapter 5. Assessment of solutions, less clear cases and different conceptualizations, user modelling, as well as consequences for annotation principles and 'presence' or 'absence' of emotion dimensions are discussed in chapter 9.

2. Material

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 (pet dog-like) Sony's AIBO robot. The

Manuscript

basic idea is to combine a new type of corpus (children's speech) with 'natural' emotional speech within a WoZ 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. The emotions and emotion-related states expressed by the children are 'realistic' in the above mentioned sense: they are not only acting 'as if' they were giving commands. In the 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). There were two different scenarios. The first was an 'object localisation' task, in which the children were told that they should direct the AIBO towards one of several cups standing on a carpet. The second was a 'parcours' task, in which the children had to direct the AIBO through a simple map towards a predefined goal. En route the AIBO had to fulfil several tasks such as sitting down in front of a cup, or dancing. The wizard caused 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. In this paper, we mainly want to deal with the German recordings; the parallel English data recorded at the University of Birmingham are described in more detail in (Batliner et al., 2004a) and below, in section 7. The German data were collected from 51 children (age 10 - 13, 21 male, 30 female); the children were from two different schools. 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. Based on pause information, the data were segmented automatically into 'utterances' or 'turns'; average number of words per turn is 3.5.

3. Annotation

The labellers listened to the utterances (no video information was given) of each child in sequential (not randomized) order. Five labellers annotated independently from each other each word[4] as neutral (default) or as belonging to one of ten other classes which were obtained by inspection of the data, cf. above.

The labellers first listened to the whole interaction in order to 'fine-tune' to the children's baseline: some children sound bored throughout, some other ones were lively from the very beginning. We did not want to annotate the children's general manner of speaking but only deviations from this general manner which obviously were triggered by AIBO's actions. In the following list, we describe shortly the annotation strategy for each label:

joyful: the child enjoys AIBO's action and/or notices that something is funny.

surprised: the child is (positively) surprised because obviously, he/she did not expect AIBO to react that way.

motherese: the child addressed AIBO in the way mothers/parents address their babies (also called 'infant-directed speech') — either because AIBO is well-behaving or because the child wants AIBO to obey; this is the positive equivalent to *reprimanding*.

neutral: default, not belonging to one of the other categories; not labelled explicitely.

rest: not neutral but not belonging to any of the other categories, i.e. some other spurious emotions.

bored: the child is (momentarily) not interested in the interaction with AIBO.

---

[4] The 'emotional domain' is most likely not the whole utterance and not the word but a unit in between: constituents (noun phrases, etc.) or clauses which, in turn, are highly correlated with prosodic pauses. If we label on the word level we do not exclude any of these alternatives. In a subsequent step, we therefore can perform and assess several different types of chunking. Moreover, the word is a well-established unit in speech processing. Our prosody module and other modules we use to extract acoustic features used for automatic classification, are integral part of an end-to-end system. Even if stand-alone extraction modules which are not based on word recognition can be meaningful for specific applications, in the long run, an integration into a whole speech processing system will be the right thing to do; such a system is described in (Batliner et al., 2000b).

Manuscript

emphatic: the child speaks in a pronounced, accentuated, sometimes hyper-articulated way but without 'showing any emotion'.

helpless: the child is hesitant, seems not to know what to tell AIBO next; can be marked by disfluencies and/or filled pauses.

touchy (=irritated): the child is slightly irritated; this is a pre-stage of anger.

reprimanding: the child is reproachful, reprimanding, 'wags the finger'; this is the negative equivalent to *motherese*.

angry: the child is clearly angry, annoyed, speaks in a loud voice.

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 resort to majority voting (henceforth MV): if three or more labellers agree on the same label, this very label is attributed to the word; if four or five labellers agree, we assume some sort of prototypes. Table I shows the labels used and the resp. number # and percent points % of MV cases for the German[5] and the English data. We will come back to the English figures below, in section 7.

We consider only labels with more than 50 MVs, resulting in seven classes.[6] *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 emotion 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, re-formulations, 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

---

[5] Due to a later check of the transliteration, these figures changed slightly as for the automatic classifications referred to below: *motherese*: 1260, *neutral*: 39169, and two 'new' words without emotion labels, resulting in a total of 48401.

[6] Note that for instance an MV of zero for *surprised* does not mean that this label was never given; it means that there was no agreement between the labellers. Moreover, it does not mean that the children displayed no surprise at all; it means, however, that this state cannot be modelled robustly enough.

Manuscript

Table I. Emotion labels used with # and % of majority voting (MV) cases for German (G) and English (E) data.

label	# **G**	% **G**	# **E**	% **E**
joyful	101	0.2	11	0.1
surprised	0	0.0	0	0.0
motherese	1261	2.6	55	0.6
neutral	39177	80.9	7171	84.6
rest (spurious emotions)	3	0.0	0	0.0
bored	11	0.0	0	0.0
emphatic	2528	5.2	631	7.4
helpless	3	0.0	20	0.2
touchy (irritated)	225	0.5	7	0.1
reprimanding	310	0.7	127	1.5
angry	84	0.2	23	0.3
no MV	4705	9.7	439	5.2
total	48408	100.0	8474	100.0

non-native, 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 (Batliner et al., 2003a).

From a methodological point of view, our **7-class problem** is most interesting. However, the distribution of classes is very unequal. Therefore, we down-sampled *neutral* and *emphatic* and mapped *touchy* and *reprimanding*, together with *angry*, onto **A***ngry*[7] as representing different but closely related kinds of negative VALENCE; this is a standard procedure for automatic recognition of emotions, cf. section 1.3. 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; this subset and different measures of interlabeller

---

[7] 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.

Manuscript

agreement are dealt with in (Steidl et al., 2005). Cases where less than three labellers agreed were omitted as well as those cases where other than these four main classes were labelled. We can see that there is a trade-off between 'interesting' and usable: our seven classes are more interesting, and our four classes are more equally distributed, and therefore better suited for automatic classification, cf. (Batliner et al., 2005b).

Some of our label names were chosen for purely practical reasons: we needed unique characters for processing. We chose *touchy* and not *irritated* because the letter 'I' has been reserved in our labelling system for *ironic*, cf. (Batliner et al., 2004b).[8] Instead of *motherese*, some people use 'child-directed speech'; this is, however, only feasible if there is in the respective database no negative counterpart such as *reprimanding* which is 'child-directed' as well. **A***ngry* was not named **N***egative* because we reserved **N** for **N***eutral*; of course, it stands for negative VALENCE.

4. Non-Metrical Multi-Dimensional Scaling

Input into Non-Metrical Multi-Dimensional Scaling (NMDS) (Kruskal and Wish, 1978) is normally a matrix indicating relationships amongst a set of objects. The goal is a visual representation of the patterns of proximities (i.e., similarities or distances) amongst these objects. The scaling is non-metrical if we do not assume distances based on a metric (interval) scale but on an ordinal or on a nominal scale; this is certainly appropriate for our annotations.[9] The diagonal (correspondence) is not taken into account; the matrices are either symmetric or are – as is the case for our data – made symmetric, via averaging. The computation encompasses the following steps: with a random configuration of points, the distances between the points are calculated. The task is to find the optimal monotonic transformation of proximities (i.e., of the distances), in order to obtain optimally scaled data (disparities); the so-called stress-value between the optimally scaled data and the distances has to be optimized by finding a new configuration of points.

---

[8] Note that our labellers were native speakers of German; they annotated according to the definitions given in the list and did not pay attention to the specific semantics of the English words.

[9] For instance, distances between cities are clearly metrical; human judgments such as school grades are ordinal. Categorical labels as such are originally nominal but can be interpreted as belonging to a higher scale of measurement if mapped onto a dimension axis, cf. below Fig. 4. Here we should not interpret exact distances but can make statements such as 'cluster together', 'are far away from each other' etc.

Manuscript

This step is iterated until a criterion is met. The output of NMDS is an n-dimensional visual representation; one normally aims at two dimensions, one dimension being often not interesting enough, and three or more dimensions often being difficult to interpret and/or not stable because of sparse data. The criteria for the goodness of the solution are the two measures of fit: Kruskal's stress and the squared correlation RSQ; a third one is interpretation quality – this is admittedly a rather vague but at the same time, very important criterion. The axes are meaningless, the orientation is arbitrary. Clusters and/or dimensions can be interpreted and, by that, more general concepts can be found than the single items (categories, labels) that were input into NMDS. Note that it is not the exact distance between items that should be interpreted and replicated but the basic configuration. Most useful is NMDS for exploration of new (types of) data. We will use the ALSCAL procedure from the statistical package SPSS.

5. NMDS solutions for our data: labels

We will call the MV cases described above **absolute majority (AM)** cases; in addition, we define as **relative majority (RM)** those cases where a relative majority or no majority at all (i.e., equal distribution) is given. RM is used to sort of **pre-emphasize** the non-MV cases.[10] Table II shows the number of cases per constellation, and Table III shows the combined confusion matrix for all labels, i.e., for AM and RM cases in percent.[11] To give two examples: For an AM case with a majority of 3/5 for **A**ngry, we enter 3 cases in the reference line into the cell for **A**ngry and the other two as 'confused with' into the cells for the resp. other labels in the same line. For an RM case with 1+1+1+1+1, i.e., equal distribution, we enter five times in turn each of the five different labels as reference and the other four as 'confused with' into the cells for the resp. other labels.

Fig. 2 shows the 2-dimensional NMDS solution for Table III. As mentioned above, axes and orientation are arbitrary; the underlying dimensions are thus not identical with the axes, and they are not

---

[10] Pre-emphasis increases in audio signals the magnitude of higher frequencies w.r.t. lower frequencies. If we 'pre-emphasise' our RM cases, we assign these rare but interesting cases higher weight by using the same case several times as reference. Another analogy is the logarithmic presentation of frequencies in a diagram if some classes have many tokens, some other only a few: here the bars for higher frequencies are lowered w.r.t. the bars for lower frequencies.

[11] In the tables, percent values per line sum up to 100%, modulo rounding errors. The labels are given recte, with boldfaced initials (row); for the columns, only the (unique) initials are given.

Manuscript

Table II. Emotion labels used with # of majority voting MV.

absolute majority AM	#
3/5	13671
4/5	17281
5/5	12751

relative majority RM	#
2+1+1+1	1554
2+2+1	3070
1+1+1+1+1	81
total	48408

Table III. confusion matrix for AM and RM in percent.

label	A	T	R	J	M	E	N
Angry	43.3	13.0	12.9	0.0	0.1	12.1	18.0
Touchy	0.5	42.9	11.6	0.0	0.9	13.6	23.5
Reprim.	3.7	15.6	45.7	0.0	1.2	14.0	18.1
Joyful	0.1	0.5	1.0	54.2	2.0	7.3	32.4
Mother.	0.0	0.7	1.4	0.8	61.0	4.8	30.3
Emphatic	1.3	5.7	6.7	0.5	1.2	53.6	29.8
Neutral	0.3	2.1	1.4	0.4	2.7	13.9	77.8

necessarily orthogonal to each other. 3- or higher-dimensional solutions would require much more items; they are therefore not stable enough in our case. [12] On the other hand, a comparison of stress and RSQ values between the 2-dimensional solutions and and the one with only one

---

[12] It is easy to use much more items in dimensional judgment studies (Scherer, 2001), p. 386, although these studies normally only interpret the two well-known dimensions AROUSAL and VALENCE — an outcome that has been characterized by (Ortony et al., 1988), p. 7, as "[...] as uninformative as it is surprising." In our approach, the items were not selected out of a pre-defined emotion dictionary but obtained in a data-driven way and filtered with frequency criteria; they can thus be considered being representative and ecologically more valid — not in a generic way but for this special application scenario. Actually, we doubt that in any specific scenario — which has to be modelled as such for automatic processing — there are more than a good few different emotional states that can be observed and modelled reliably.
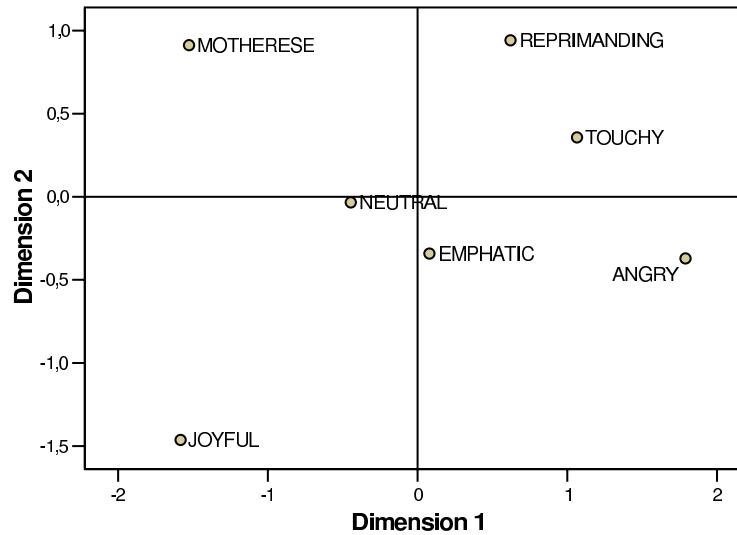
Manuscript

Figure 2. Original NMDS solution for MV data with # > 50, 2 dimensions; stress: .23, RSQ = .82

dimension displayed in Fig. 4 shows that the 2-dimensional solution is most adequate.

If we want to refer to the dimensions we interpret for our solution, we will use the terms which refer to the compass rose: *west to east* thus means more or less along the x-axis, *south-west to north-east* means bottom left to upper right. Note that by that, we do not indicate any precise direction but only a rough orientation. *neutral* and *emphatic* cluster together, close to the origin; this means that they are rather neutral as for both dimensions. The first, most important dimension can clearly be interpreted as VALENCE *(south-)west* to *(north-)east*: from positive (*joyful* and *motherese*) over *neutral* and *emphatic* to negative (*reprimanding*, *touchy*, and *angry*). The second dimension (from *south(-east)* to *north(-west)*) cannot, however, be interpreted as something like the 'traditional' dimension AROUSAL; even if at first sight, *angry* and *joyful* could be interpreted as high AROUSAL, *emphatic* as medium AROUSAL, and *neutral* as no AROUSAL, it makes no sense to interpret *motherese* and *reprimanding* as having lower AROUSAL than *neutral*. Moreover, by listening to instances of *angry* and *joyful* we can say that *joyful* in our scenario definitely denotes not more pronounced AROUSAL than *angry* — rather the opposite. (We will come back to possible residuals of AROUSAL in section 6.) Another aspect that is partly entailed in our second dimension is interpersonal INTIMACY: *motherese* and *reprimanding* characterize a more intimate
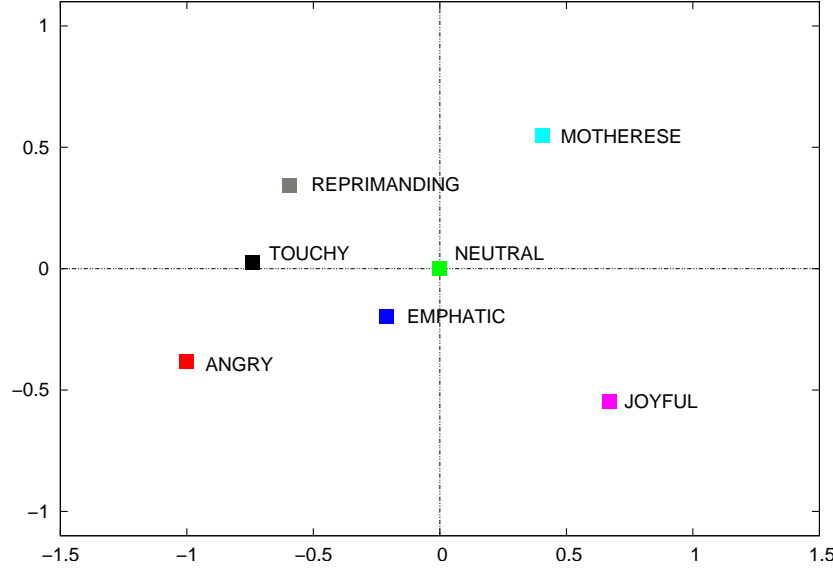
Manuscript

Figure 3. NMDS solution for MV data with # > 50, 2 dimensions; stress: .23, RSQ = .82

speech register (Batliner et al., 2006a) than *neutral* and *emphatic*. However, it makes no sense to interpret *angry* and *joyful* as being less intimate than *neutral*. Instead, we interpret the second dimension in more general terms as ORIENTATION towards the subject him/herself or towards the partner (in this case, the AIBO), as DIALOGUE aspect (MONOLOGUE vs. DIALOGUE), as SOCIAL aspect, or as [+/- INTERACTION]. In the following, we will use INTERACTION as term to describe this dimension.[13] User states like *angry*, i.e., [- VALENCE], and *joyful*, i.e., [+ VALENCE], represent [- INTERACTION]; subjects can be in such states even if they are alone; user states like *reprimanding*, i.e., [- VALENCE], and *motherese*, i.e., [+ VALENCE], represent [+ INTERACTION]; in these cases, some partner has to be present and addressed.

For a more intuitive graphical representation of our dimensions and the positions of our categories, we processed the co-ordinates of Figures 3 and 5 to 12 along the following lines: first, all points are moved in

---

[13] Actually, the other names might be, in other contexts, even more adequate depending on the specific theoretical and empirical background: if communication is restricted to speech (for instance, via telephone), we might prefer dialogue vs. monologue (i.e., speaking aside). At least in German, verbs with this type of [+ INTERACTION] tend to be more transitive, i.e., having more valence slots than verbs with [- INTERACTION]. Note that there are other, 'non-dimensional' terms to describe these phenomena such as 'speech register' or 'infant/child/pet-directed speech'.
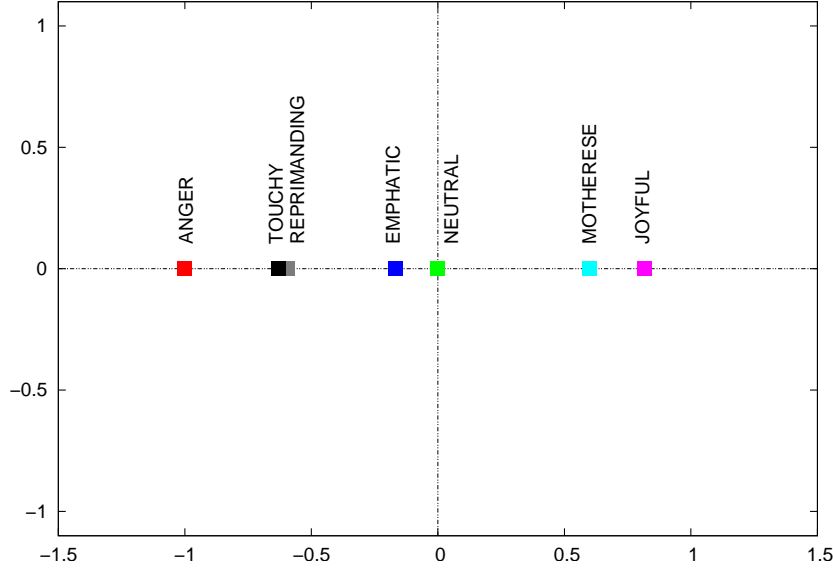
Manuscript

Figure 4. NMDS solution for MV data with # > 50, 1 dimension; stress: .32, RSQ = .73

such a way that *neutral* is in the origin. Second, all points are rotated in such a way that *motherese* is on the positive x-axis. Third, if needed, all point are flipped horizontally resulting in *joyful* having positive x-co-ordinates. Fourth, we rotated by 45 degree $\pm$ a heuristic angle to ensure that *motherese* is in the first quadrant (*north-east*), *joyful* in the fourth (*south-east*), and *angry* in the third (*south-west*); this results automatically in *reprimanding* being in the second quadrant (*north-west*). Fifth, all data points are scaled in such a way that they are in the region [-1,+1][-1,+1], i.e., the same factor for both axes is used. By that, Fig. 2 is transformed into Fig. 3; for both dimensions that we interpret, negative is now bottom and/or left, and positive is top and/or right.

The first, most important dimension is VALENCE. Fig. 4 displays the one-dimensional solution which clearly shows that the classes are not equidistant on this axis: *touchy* and *reprimanding* cluster together midway between *angry* and *emphatic*, *emphatic* is very close to *neutral*, and *motherese* clearly denotes positive VALENCE albeit *joyful* is most pronounced as for positive VALENCE, the same way as *angry* is for negative VALENCE. This one-dimensional solution has, however, markedly higher stress and lower RSQ values; thus, the second dimension clearly contributes to interpretation.

The computation of the confusion matrices might affect the dimensional solution. Thus for Table IV, another computation was chosen:

Manuscript

Table IV. confusion matrix for 'probability in percent (cf. explanation in text).

label	A	T	R	J	M	E	N
Angry	15.4	16.7	12.8	0.1	0.1	17.6	36.7
Touchy	3.6	12.8	11.1	0.1	1.2	19.9	49.2
Repr.	3.4	14.1	17.8	0.2	2.2	24.5	37.1
Joyful	0.1	0.6	0.7	17.6	4.7	9.4	64.3
Mother.	0.0	0.9	1.2	0.7	32.8	5.8	58.1
Emphatic	0.7	3.5	3.4	0.3	1.5	21.2	68.7
Neutral	0.3	2.2	1.3	0.6	3.6	17.0	73.9

each cell represents the probability for a word to be labelled with one emotion (line) by one labeller and with the same or another emotion (row) by another labeller, averaged across all 10 possible combinations of labellers {A,B}: $P(A \leftrightarrow B)$; the values of all cells in the triangular matrix sum up to 100. This raw matrix, however, does not yield any meaningful dimensional solution because distribution in the cells is very unequal. Therefore, we normalized each line; by that, the values in percent of each line sum up to 100%. Thus for Table III we sort of 'pre-emphasised' the unclear, mixed cases, for Table IV we sort of 'pre-emphasised' the rare cases.

Fig. 5 displays the 2-dimensional solution for the matrix of Table IV. The general picture remains the same: *neutral* and *emphatic* cluster together close to the origin, *joyful* and *motherese* are positive, i.e., [+ VALENCE] and [-/+ INTERACTION], *angry* is like *joyful* but negative, i.e., [- VALENCE]. In Fig. 3, *touchy* is in between *angry* and *reprimanding*, in Fig. 5, it is on the INTERACTION dimension at the same height as *reprimanding*.

As mentioned in section 3, for automatic classification, cf. (Steidl et al., 2005; Batliner et al., 2005b), we mapped our labels onto a 4-class problem with > 1000 tokens in each class. Table V displays the confusion matrix for these four labels, computed the same way as in Table III. In Fig. 6, the 2-dimensional NMDS solution for the confusion matrix of Table V is shown. There are only four items; this 2-dimensional solution is therefore not stable. The first dimension seems to be VALENCE again: from **A**ngry to **E**mphatic to **N**eutral to **M**otherese. However, a second dimension is not easy to interpret; it rather looks as if the relevant classes at the top left and bottom right edges are missing – which in fact is true: there is no *reprimanding* or *joyful*. *reprimanding*
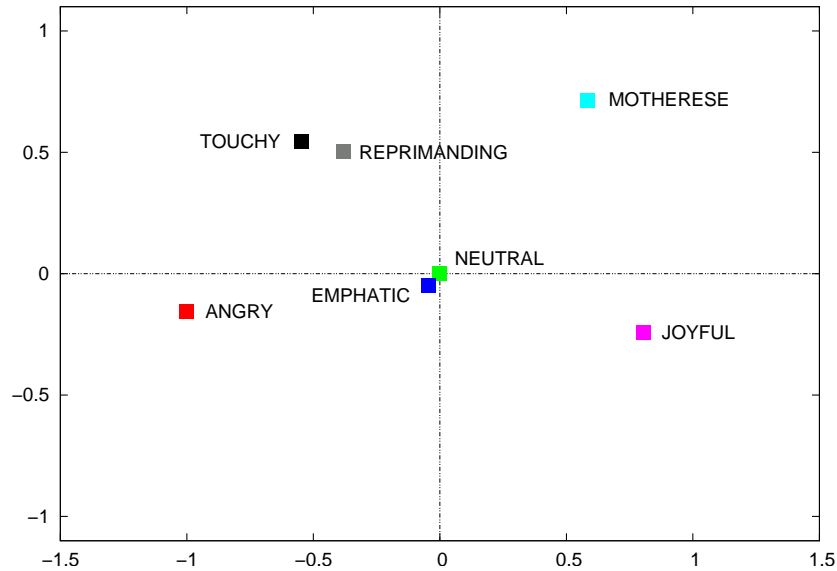
Manuscript

Figure 5. NMDS solution for 'probability' data with # > 50, 2 dimensions; stress: .21, RSQ: .85

has been mapped onto **A**ngry, and *joyful* has been discarded altogether because of sparse data (101 tokens).

Table V. confusion matrix for AMEN.

label	**A**	**M**	**E**	**N**
Angry	70.6	0.4	10.7	18.2
Motherese	0.4	68.8	1.5	29.3
Emphatic	5.7	0.2	65.5	28.5
Neutral	2.1	2.6	13.3	82.0

As usual in research on realistic emotions, we are facing a sparse data problem: with less representative data, we can find interesting dimensions but of course, automatic classification performance is not high, cf. (Batliner et al., 2005b). With (statistically) representative data – obtained via mapping onto cover classes/dimensions — classification performance is higher but our interesting dimension INTERACTION is gone, i.e., no longer visible.
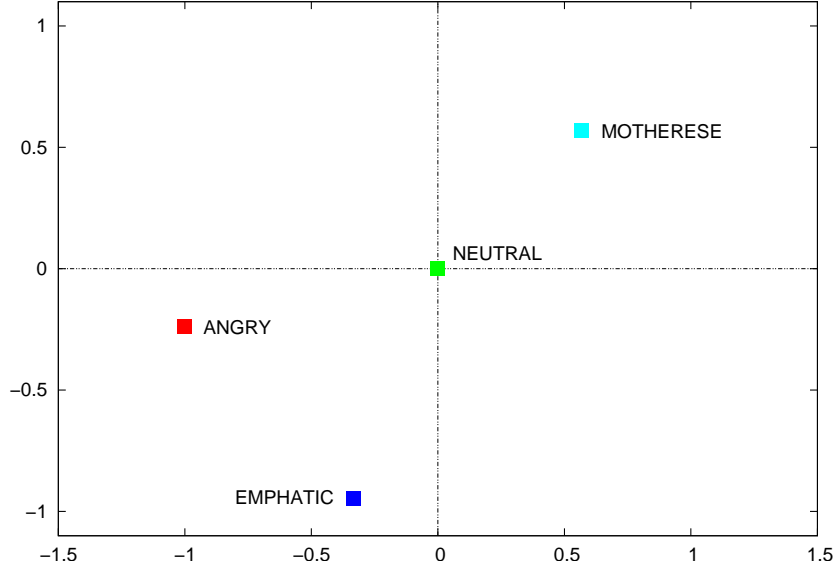
Manuscript

Figure 6. NMDS solution for the 4-class problem AMEN, 2 dimensions; stress: .19, RSQ: .90

6. Interpretation

The clusters and the localisation in the 2-dimensional space find their counterpart in the confusions displayed in Tables III and IV: most confusion takes place between all other labels and *neutral*, and to a somewhat lesser degree, with *emphatic*, cf. the last and the second-last columns. Therefore, *neutral* and *emphatic* are close to the origin in the original dimensional solution in Fig. 3. This illustrates at the same time the difficulty of telling apart the neutral baseline from any marked state. *motherese* and *joyful* are almost never confused with the labels denoting negative VALENCE, i.e., with *angry*, *touchy*, or *reprimanding*; therefore they are localized at the opposite end, i.e. at positive VALENCE. (An interesting exception is discussed below in section 9.3.) The three negative labels are less stable, more often confused with each other, and can change place, according to different computations of the confusion matrices; this holds mostly for *touchy* which obviously is in between *angry* and *reprimanding*. Actually, it has been defined as a sort of pre-stage of *angry*.

The interpretation of the second dimension as an interactional, social one is corroborated by other, theoretical approaches towards the social aspects of emotions: 'Social emotions' are addressed in (Ortony et al., 1988), for instance, *Reproach*, and in (Scherer, 2001), p. 381 who claims that "a large majority of emotion episodes are eminent social — both

Manuscript

with respect to the eliciting situation and the context of the emotion reaction [...]". (Gratch et al., 2006) define *anger* as social emotion per se whereas (Poggi et al., 2001) make a difference between *social anger* and *non-social anger*, depending on the addressee; thus, *anger* is claimed to be not intrinsically a social emotion while others such as *Reproach* are (note that *Reproach(ing)* is almost synonymous with our *reprimanding*). We will come back to different types of *anger* in section 9.6 below.

We want to stress that we do not intend to 'get rid' of AROUSAL as emotion dimension; we only claim that — in specific scenarios — it is not amongst the two most important ones. Due to our sparse data problem, we cannot say whether there is some 'residual' of AROUSAL encoded in our second dimension. However, this might be plausible if we consider that social control can prevent the signalling of 'too much emotion' up to the same degree as it favours social behaviour oriented towards the interaction partner. If we look at recent studies on human-human multi-party interaction we can see that even if the researchers started with the intention to annotate the two classic dimensions, they found out that something like AROUSAL is not really represented in their data: "[...] most of the changes in the mental state of participants that one can observe do not relate to the two emotional dimensions that are captured by the FeelTrace procedure [i.e., VALENCE and AROUSAL]. The major mental states that are identified relate to cognitive processing or expressions of propositional attitudes: 'concerned', 'interested', 'doubting', 'distracted', 'uncertain' are more relevant terms for this kind of data." (Reidsma et al., 2006). (Laskowski and Burger, 2006) note that "We chose not to annotate emotional activation, studied in the context of meetings [before] as there was not as much intra-speaker variability in our data relative to the seemingly larger differences between baselines for different speakers." Note that these human-human multi-party interactions are of course far more complex than those in our scenario where only one user interacts via speech while the addressee (the AIBO) is always silent and only (re-)acts.

7. Other Types of Data

If data are not pre-defined, i.e., if we only can label what we can find in realistic databases, then we will most likely find something different – even different categories and by that, different dimensions – for different types of databases. To illustrate this aspect, we first computed a 2-dimensional NMDS solution for our parallel English data, exactly along
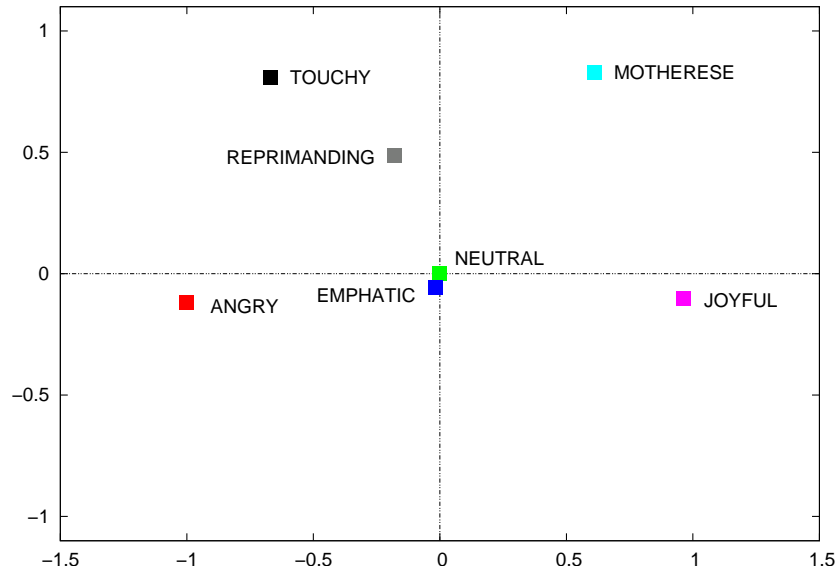
Manuscript

Figure 7. NMDS solution for English MV data, 2 dimensions; stress: .17, RSQ: .89

the same lines as for our German data: MV, 'pre-emphasis'. The English data do not only represent another language but differ in several aspects slightly from our German data: there were 30 English children who took part, with a wider range of age, namely between 4 and 14. There were two recordings, the second being parallel to one of our sub-designs, the so called 'parcours'; details can be found in (Batliner et al., 2004a). In the first recording, the same sub-design was used but the AIBO behaved obediently and followed the children's commands. The children were not told that they could communicate with the AIBO as with a friend. The data was annotated by three out of the five labellers who annotated our German data. MV therefore means that two out of three labellers agreed. This is a typical situation that we often face in daily practice: parallel does not mean strictly parallel – for our English data, there are, e.g., less subjects, age distribution is different, there are less labels and less labellers. Fig. 7 displays the 2-dimensional NMDS solution for the English data. For comparison, we take exactly the same labels as we did for our German data, even if MV frequency is now sometimes below 50 cases, cf. Table I. We can find our two dimensions, we can replicate the clustering found in Figures 3 and 5; the positions of *touchy* and *reprimanding* resemble those found in Fig. 5. If we consider that the sparse data problem for our English data is even more pronounced than for our German data, cf. Table I, this is a reassuring result.

Now we now want to have a look at the dimensions we can extract for data obtained within a totally different material, recorded within a call-

Manuscript

Table VI. SympaFly: Confusion matrix for emotional user states annotated per turn, two labellers.

L1 ↓	L2 →	J	N	S	I	C	E	A	P	H	T	Total
Joyful		12	5	-	3	-	-	-	-	-	-	20
Neutral		13	5355	3	31	18	110	1	6	31	72	5640
Surprised		-	1	3	1	-	1	-	-	1	-	7
Ironic		4	17	1	28	1	1	-	-	2	8	62
Compassionate		-	-	-	-	-	-	-	-	-	-	
Emphatic		2	340	-	8	11	218	2	8	7	54	650
Angry		-	2	-	-	-	-	-	-	2	4	8
Panic		-	1	-	-	-	-	-	7	-	-	8
Helpless		-	16	-	5	2	1	-	2	21	9	56
Touchy		2	39	-	1	-	21	1	-	3	76	143
Total		33	5776	7	77	32	352	4	23	67	223	6594

center scenario: the German SympaFly database was recorded using a fully automatic speech dialogue telephone system for flight reservation and booking. In the first, preliminary stage of this system which was achieved by rapid prototyping, performance was rather poor (approx. 30% dialogue success rate); in the last, third stage, performance was very good (above 90% dialogue success rate). In the second, intermediate stage, system performance was increased little by little, sometimes from one day to the other. Recordings were made with volunteering subjects (2. stage) and with employees of a usability lab (1. and 3. stage). A full description of the system and these recordings can be found in (Batliner et al., 2003b; Batliner et al., 2004b). We employed two labellers; as is the case for the AIBO labels, the labels were chosen in a pilot pass. The confusion matrix, this time with the absolute number of items in each cell in order to indicate the sparse data problem more clearly, is given in Table VI. Note that here, we annotated whole turns and not words. Each turn had 4.3 words on average.

Fig. 8 shows for those items with a frequency above 50 for each of the two labellers the 2-dimensional solution for the SympaFly data. With only two labellers, there is no MV. We therefore took each labeller in turn as reference (line), normalized each line summing up to 100%, and computed the mean percent value per cell for these two matrices. (Needless to say that this solution can only be taken as some indication because we only have two labellers, and because the distribution of our items is extremely unequal.) It is self-evident why we do not find the IN-
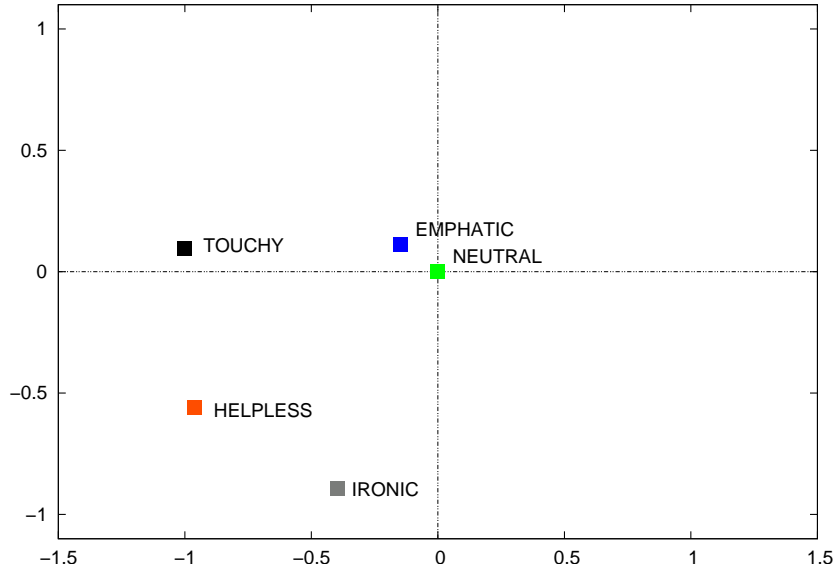
Manuscript

Figure 8. NMDS solution for SympaFly (call-center data) with # > 50; stress: .24, RSQ: .80

TERACTION dimension that is specific for our AIBO data: call-center clients do not use *motherese* or this specific type of *reprimanding* while communicating with a human operator, let alone with an automatic system. However, we do not find the clear-cut dimensions AROUSAL or VALENCE either. The first dimension could be some sort of EXPRES-SIVITY from *south-east* to *north-west* – related to but not necessarily identical with AROUSAL: it is typical for *ironic* that it lacks EXPRES-SIVITY the same way as *neutral* does – otherwise, it would no longer be irony. *touchy* on the other hand, displays EXPRESSIVITY. *helpless* is a very specific type marked by disfluencies, hesitations, and pauses. The second dimension might be another type of INTERACTION (related to CONTROL) from *north-east* to *south-west*: the normal one in the case of *neutral* and *emphatic*, and withdrawal from normal interaction, i.e., rather some sort of meta-communication, in the case of *helpless* and *ironic*.

The chunking of *neutral* and *emphatic* can be observed throughout in all figures and is consistent with our explanation in section 3 that *emphatic* does not necessarily indicate any (strong) deviation from a neutral state.

Manuscript

8.   NMDS solutions for our data: features

Instructions and data presented for annotation can be quite different: if we, for instance, were only interested in the relevance of pitch curves for the perception of emotional states, we could low-pass filter the signal and by that, devoid it of any linguistic content. We decided in favour of the opposite approach: the speech signals were presented without any distortion in natural order. By that, the labellers could establish speaker-specific baselines as well as notice and take into account changes of these speaker-specific baselines over time. They were told that for the actual word they had to label, they should pay attention to this word in relation to its immediate context. The question is now: which characteristic traits (types of features) did our labellers pay attention to — only acoustic, or linguistic, or both? Decoding this information is hopefully closely related to encoding by the speakers.

For automatic classification of word-based emotion, we extracted large feature vectors modelling acoustic and linguistic properties of the actual word and of its immediate context and used them subsequently in an automatic classification. The results of such an automatic classification is a confusion matrix and, based on that, recognition rates. In this paper, we use the following three feature vectors:

PROSODIC-SPECTRAL features: prosodic and harmonics-to-noise ratio HNR (Batliner et al., 2006a), prosody modelling duration, energy, F0, shimmer and jitter. We compute features for the actual word and other features modelling a context of two words before or two words after. In (Batliner et al., 2003a) a more detailed account of prosodic feature extraction is given. All in all, there were 124 prosodic-spectral features.

MFCC features: the mean values of the first 12 mel-frequency-cepstral-coefficients MFCC and their first derivatives computed per frame and averaged per word summing up to 24, for the actual word, and for the two words before and after. By that, we sort of model a 'MFCC five-gram'. MFCCs are standard features in speech recognition and model the segmental content of words; however, they proved to be very competitive for language identification and emotion recognition as well. All in all, there were 120 MFCC features.[14]

SEMANTIC features: the usual bag-of-word approach is not applicable for word-based processing. Thus we decided in favour

---

[14] Note that MFCCs model the spectrum but cannot easily be interpreted as such – we could say that they are 'implicit' spectral features – whereas a direct interpretation of our 'explicite' prosodic-spectral features is possible.
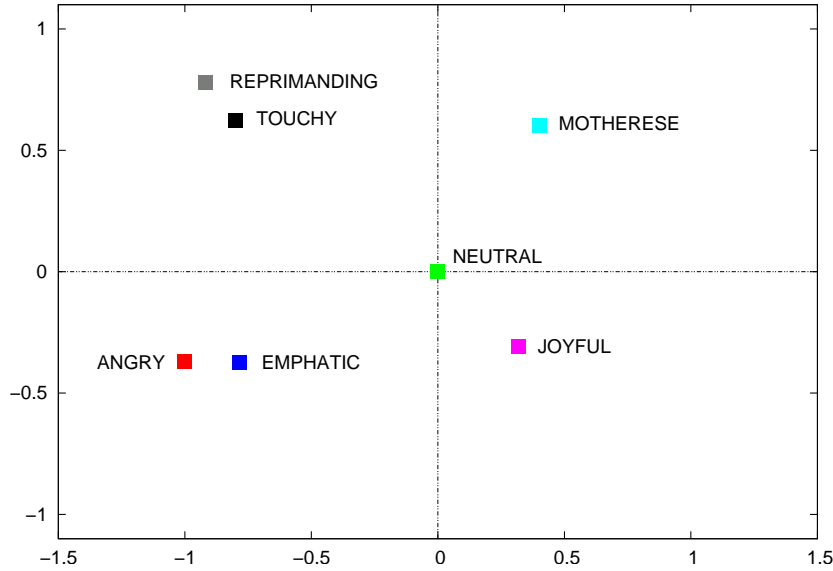
Manuscript

Figure 9. NMDS solution for prosodic-spectral features, RR = 64.4, CL = 48.5; 2 dimensions; stress: .25, RSQ = .78

of a scenario-specific mapping of lexicon-entries onto six semantically/pragmatically meaningful cover classes: *vocative*, *positive valence*, *negative valence*, *commands and directions*, *interjections*, and *rest*. Again, for each word, the two words before and the two words after are modelled as well, resulting in 30 'semantic' features.

As we want to model the strategies of our annotators who know all the speakers, we use leave-one-out and not leave-one-speaker-out. We employ LDA (Linear Discriminant analysis), a linear classifier which proved to be rather competitive in comparison with more sophisticated ones such as Random Forests or Support-Vector-Machines for our four-class AMEN problem (Batliner et al., 2006b). For computation of word boundaries, a forced alignment with the spoken word chain was used; by that, we simulate 100% correct word recognition. The three different classifications with prosodic-spectral, MFCC and semantic features resulted in three different confusion matrices which were put into NMDS yielding the 2-dimensional solutions given in Figures 9 to 11. Besides Kruskal's stress and the squared correlation RSQ, the captions display overall recognition rate RR (number of correctly classified cases divided by total number of cases, also known as weighted average) and CL ('class-wise' computed recognition rate, i.e. mean of diagonal of
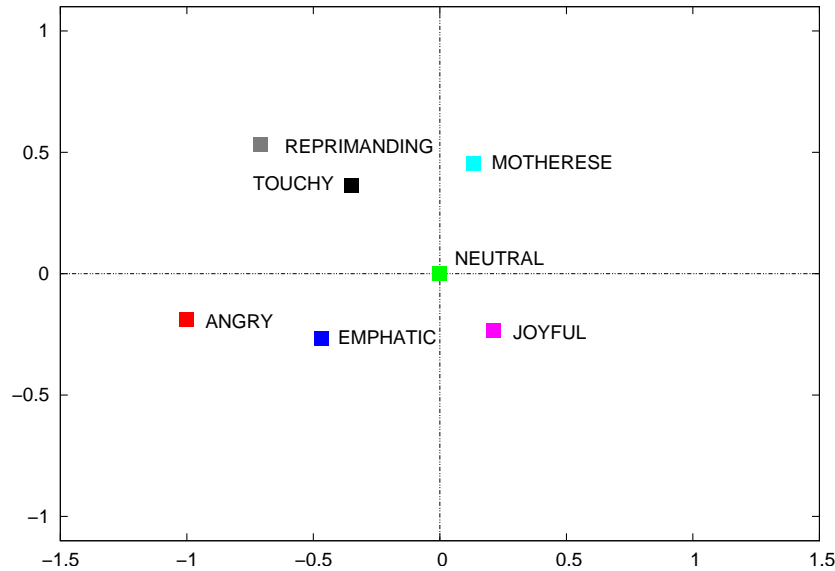
Manuscript

Figure 10. NMDS solution for MFCC features, RR = 36.7, CL = 45.2; 2 dimensions; stress: .16, RSQ = .91

confusion matrix in percent, also known as unweighted average).[15] As the default class *neutral* is by far most frequent, its percentage cannot be used as chance level; instead, we assume a chance level of 14.3%, i.e. equal random assignment to all seven classes.

In Figures 9 to 11, the edge items *angry* and *joyful* as well as *reprimanding* and *motherese* denote the well known two dimensions VALENCE and INTERACTION. In all three figures, *emphatic* is not that close to *neutral* as it is in Fig. 6, esp. not in Fig. 9. Obviously, the acoustic, esp. the prosodic manifestations of *angry* and *emphatic* are similar. In Fig. 9 and 10 (prosodic-spectral and MFCC features), *touchy* is closer to *reprimanding*, in Fig. 11, it is closer to *angry*. This might indicate that different information is encoded in the different feature vectors: the semantics, i.e. the wording, of *touchy* might be similar to the one of *angry* whereas its acoustic realisation is not; throughout, *touchy* seems to be the least stable label - this might mirror the fact that it is a stage in between slight irritation and full anger.

---

[15] Note that classification rates for leave-one-out are a bit too optimistic w.r.t. leave-one-speaker-out. In comparison, in (Batliner et al., 2005b) we report, for a feature vector which is very similar to our prosodic-spectral feature vector, a CL of 44.5% for a strict separation of speakers into training and test sample. As for classification performance, the difference between seen and unseen speakers thus amounts to some four percent points.
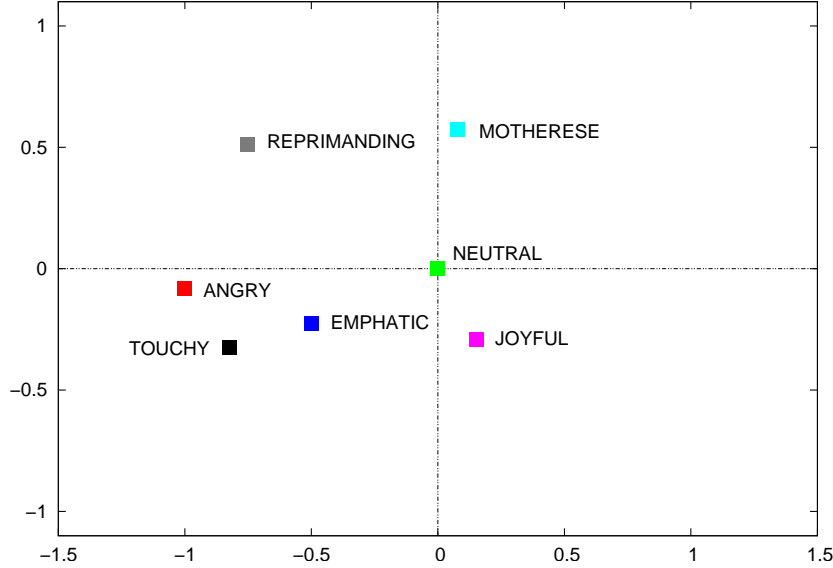
Manuscript

Figure 11. NMDS solution for semantic (bag-of-words) features, RR = 33.1, CL = 38.6; 2 dimensions; stress: .22, RSQ = .80

A fourth classification was done using all three feature types together. The NMDS solution for the resulting confusion matrix is given in Fig. 12 which closely resembles Figures 9 and 10; this might indicate that our *SEMANTIC* classes on the one hand contribute to performance but are on the other hand too coarse-grained for a detailed modelling of the space.

If we do emotion recognition using acoustic and linguistic features, we understand emotion as information that can be transmitted via these different channels. All these channels contribute to decoding this information; these features and/or feature groups are obviously — sometimes highly – correlated with each other, although the difference is most pronounced between the semantic features on the one hand and the acoustic features on the other hand. This is corroborated by the example-based argumentation in section 9.3.

9. Discussion

In this section, we want to discuss some additional aspects and questions in more detail.
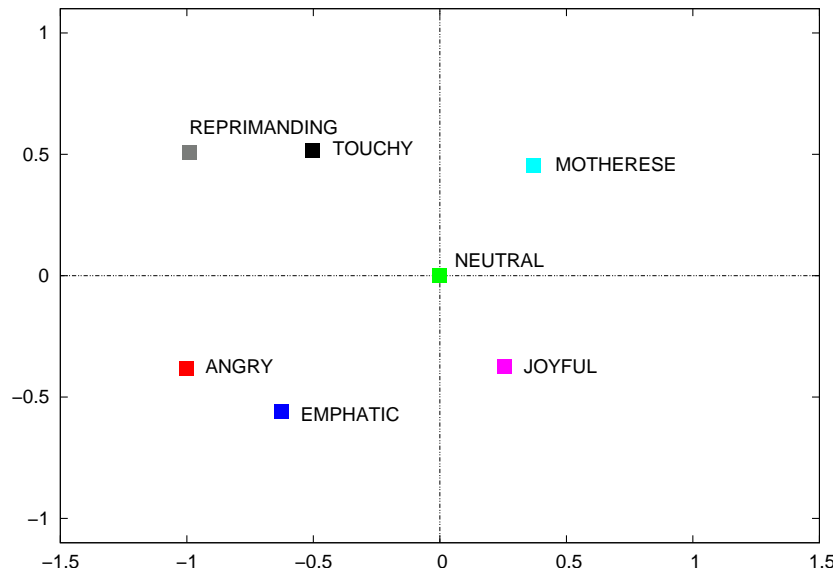
Manuscript

*Figure 12. NMDS solution for all three features types combined, RR = 70.3, CL =
53.4; 2 dimensions; stress: .17, RSQ = .88*

9.1. ASSESSMENT OF NMDS SOLUTIONS

The rule of thumb is that stress values below .2 and RSQ values above
.8 are OK. Note that this should be taken only as a rough guide: it
strongly depends on the type of data. Studies cannot be compared in
a strict sense; however, it is plausible that more artifical and by that,
more controlled data will, other things being equal, result in a better
quality. For instance, acted facial expressions in (Lyons et al., 1998)
yielded better stress and RSQ values, and the resp. values are very
good in (Jäger and Bortz, 2001) even in a 1-dimensional solution for
smilies which of course do have very unequivocal characteristic traits.
In contrast, we can expect much more 'white noise' in our realistic
data especially if the emotional states are not full-blown but mixed. In
(Batliner et al., 2005b) we show that for our AIBO data, there obviously
are more or less clear cases: the better performance of **prototypes** in
automatic classification indicates that the emotional user states labelled
are either a graded or a mixed phenomenon – or both.

There is some 'critical mass' w.r.t. number of items in an NMDS, and
number of different labellers: if the number of items is too small w.r.t.
the dimensionality, then the solution is not stable. If the number of
labellers is too small, then spurious and random factors might influence
computation. The one and/or the other factor might be responsible for
the constellations in Figures 6 and 8. However, it is reassuring that

Manuscript

different computations yield similar solutions in the case of Figures 3, 5 and 7.

9.2. How to Annotate, how to Process

There are indications that emotion-related user states (encompassing the states that we could find in our data) are more or less continuous. This does not tell us the best way how to annotate these phenomena, and it does not tell us either whether we will process them in an automatic system as dimensional entities or not. It has been our experience in fully developed end-to-end systems, cf. the SmartKom system (Batliner et al., 2003c; Portele, 2004; Streit et al., 2006), that the highly complex processing makes it necessary to map any fine-grained scale onto some very few states - two or three. Early/late mapping and/or fusion can be imagined. It might be a matter of practicability and not of theoretical considerations whether we want to use categorical or graded labels as input into such systems. Moreover, if we go over to large-scaled collections of realistic databases, it might not be feasible to employ several labellers using a very elaborated annotation system.

9.3. Mixed Cases

In Table VII we give two interesting examples of a relative majority for mixed cases; in the left row, the German words belonging to one utterance are given; non-standard forms such as *ne* instead of *nein*, are starred. In the right row, the English translation is given. In between, the labels given by labeller one (L1) to five (L5) are displayed. We can see that in the first example, *motherese* alternates with *reprimanding* (and *neutral*). Thus, INTERACTION is clearly positive, although VALENCE is not that clear. Obviously, if *motherese* is labelled, the 'tone of voice' was the discriminating feature, if *reprimanding* was labelled, the semantics of *'no'* played a greater role. In the second example, the negative VALENCE is clear, the detailed classes obviously not. A mapping onto a cover class *negative* or **A**ngry thus suggests itself, cf. as well the similarities of these negative labels in Table III. The cases are thus 'interesting', but – at least for our data – not necessarily representative. By using pre-emphasis, we do account for such mixed cases in our NMDS solutions as well.

9.4. Different Conceptualizations

Figure 13 shows for our 4-class AMEN problem a scatterplot with the distribution of **M**otherese vs. **A**ngry per speaker (leaving aside one

Manuscript

Table VII. Examples for Relative Majority = 2.

German	L1	L2	L3	L4	L5	English
mixed VALENCE, clear INTERACTION						
*ne	M	R	N	M	R	no
*ne	M	R	N	M	R	no
*ne	M	R	N	M	R	no
so	M	R	N	M	N	so
weit	M	R	N	M	N	far
*simma	M	R	N	M	N	we are
noch	M	R	N	M	N	yet
nicht	M	R	N	M	N	not
aufstehen	M	R	N	N	R	get up
clear VALENCE, unclear categories						
nach	A	T	E	E	N	to
links	A	T	E	E	R	the left
Aibo	A	T	T	R	R	Aibo
nach	A	T	T	E	N	to
links	A	T	T	E	R	the left
Aibolein	A	T	E	A	R	little Aibo
ganz	A	T	E	A	R	very
böser	A	T	T	A	N	bad
Hund	A	T	T	A	N	dog

outlier subject which displays very high frequencies for both). Spearman's rho (non-parametric correlation) 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.[16] Here, the **target prototypes** are thus at the origin (no

---

[16] A fourth type only displaying **M**otherese would constitute something like a resource-oriented, therapeutic interaction; naturally enough, our children do not display it.

Manuscript

interactive behaviour at all, only commands), high on the y-axis and low on the x-axis (showing only **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 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 t-test; as this difference was not formulated as alternative hypothesis, we had to use the two-tailed test.

It is clear that these different conceptualizations lead to different or missing dimensions: if subjects do not use **M***otherese* then the NMDS will not find our second dimension INTERACTION. And if subjects neither use **M***otherese* nor **A***ngry* (i.e., *touchy*, *reprimanding*, or *angry*), then we possibly will not find our first dimension VALENCE either.
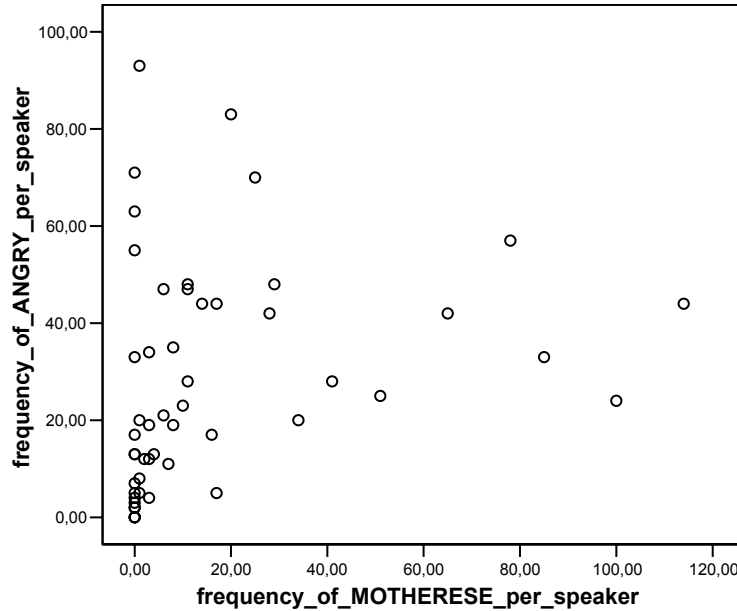

Figure 13. Scatterplot: Distribution of Motherese and Angry per Speaker; displayed is # of cases

Manuscript

9.5. USER ADAPTATION

In this paper, we have stressed passim the importance of frequency for the automatic processing of emotion. Only phenomena that can be observed frequently enough can constitute a representative training database which is necessary for optimal recognition performance. However, this performance nowadays is not much better than 80% for a two-class problem or 60% for a four-class problem; note that this seems to be close to the performance of single labellers (Steidl et al., 2005). Even if better modelling and larger databases will yield better performance in future, we cannot expect perfect recognition rates. Thus, a wrong system reaction to single instances — based on erroneous recognition — can yield rather unfavorable results in human-machine interaction. Instead, the system could monitor the user's emotional states using cumulative evidence and make decisions after an initial phase. We want to demonstrate this possibility by assuming that the system monitors the use of *motherese* by children interacting with the AIBO. If for a certain amount of time, the frequency of — correctly or incorrectly recognized — instances of *motherese* exceeds a threshold, different attenuating or reinforcing system reactions could be triggered: if the child does not use *motherese* at all, the AIBO could be triggered to display a more pronounced pet-like behaviour in order to elicit social interaction (wag its tail, dance around, etc.). If, however, the child uses too much *motherese*, by that forgetting the concrete task he/she has to complete, then AIBO could reduce it's pet-like behaviour.

In order to assess such a procedure, we computed for the four-class problem AMEN a two-fold cross-validation using mainly prosodic features. The non-parametric correlation coefficient Spearman's rho between the sum of *motherese* instances produced and the sum of correctly recognized *motherese* instances per speaker is 0.94; the rho value for the realistic correlation between the sum of *motherese* instances produced and the sum of recognized *motherese* instances — be these correctly recognized or false alarms — is still fairly high, namely 0.83. Even higher correlations can be expected by incorporating additional knowledge sources such as linguistic information. Thus it is possible not to use a maybe erroneous 'single instance decision' for processing in the application but the frequency of the recognized instances of a label (including false alarms) for modelling user behaviour over time and by that, users' conceptualizations, and for an appropriate adaption of the system's behaviour.

Manuscript

9.6. WHICH DIMENSIONS

The dimensions that best model specific types of scenarios depend crucially on at least: firstly, the subjects and their conceptualizations; secondly, the communication structure, e.g., whether it is symmetric or not; thirdly, in which setting the emotions are observed. Due to the observer's paradox (Labov, 1970; Batliner et al., 2003b), the threshold for displaying emotions might be higher, the more likely it is that the subjects are being observed by a third party, meaning that some type of general public is present.

It might as well be that for some data, no clear picture emerges. This can be due to insufficient size of the database, or simply to a constellation where no clear dimensional solution can emerge. The dimensions we can find will definitely be affected by the sparse data problem: for our SympaFly data we decided not to take into account labels with a frequency below 50 in order to ensure a half-decent robustness of our solution. By that, we excluded user states like *angry* and *panic* from our analysis; with these emotions, we probably could have obtained AROUSAL as first or second dimension. Thus what we get is an indication of those emotional user states we will encounter in applications if – and only if – the distribution of our phenomena and by that, labels, can be transferred to real applications. Of course, we cannot say anything about the emotions our subjects — or any other subject — will display in other situations or scenarios. For instance, in the scenario 'medical emergency call center' (Devillers and Vidrascu, 2006) with the classes *anger*, *fear*, *relief*, and *sadness*, AROUSAL might turn out to be amongst the most important dimensions, cf. the positions of *angry* and *sad* in Fig. 1.

It will certainly not be meaningful to create a new dimensional space each time we deal with a new scenario. As far as we can see, it might often be the case that only a certain **sub-space** can be modelled with those categories that can be found and labelled in specific databases. We therefore do not want to claim that we span the complete dimensional space of VALENCE and INTERACTION using our seven items — maybe we have to distinguish at least three types of anger that have to be located at different positions on a full INTERACTION axis: first a 'private', non-social, non-interactive: you are angry because it rains (Poggi et al., 2001), cf. the discussion in section 6; second a still private but more social one when you are in a genuinely social setting because here, showing your anger means at the same time to communicate your anger, cf. the 'impossibility of not communicating' (Watzlawick et al., 1967); third, a socially mediated one in the 'disguise' of *reprimanding* with a manifested intention to force your communication partner to

Manuscript

behave differently. This can explain why *reprimanding* is on the VA-
LENCE axis less negative than *angry*: the speaker tells AIBO that it
behaves disobediently but appeals to it's co-operation at the same time.

Even if it might be possible to map any new category onto the tradi-
tional dimensions VALENCE and AROUSAL etc., this will, however,
not be a very wise strategy because in many cases, this solution will
not turn out to be stable and adequate.

Almost all of the studies which contributed to the notion of emo-
tion dimensions so far have been conducted with elicited, somehow
acted emotions. Thus the social, interactive aspect (the so-called 'pull-
effects') has rather been neglected, the so-called 'push-effects' have been
primary object of investigation.[17] With more realistic data, it might
turn out that for different modalities – and of course, different situations
– different dimensions prevail: maybe the face is better at displaying the
'classic' dimension AROUSAL, whereas in real-life speech, our INTER-
ACTION dimension will be observed more frequently.[18] Based on the
results described in this paper, we can try to tentatively represent our
two dimensions and our seven categorical labels in Fig. 14, in analogy
to Fig. 1.

The outcome of visualisation techniques such as NMDS or the Sam-
mon transform (Sammon, 1969) can be conceptualized at different lev-
els: first, it can simply be taken as a help in interpreting the data; in our
case, this means a convenient way to interpret confusion matrices and
find an objective and optimal mapping onto few cover classes — which
often will be necessary because of sparse data and suboptimal classifi-
cation performance for too many different classes. Moreover, dialogue
systems will often not be able to model more than only a few emotional
user states. Second, it can be taken as guidelines for building meaningful
applications or decision steps within such applications. Third, it can be
taken as a representation of a cognitive and/or emotional space. This
last alternative could be called the 'strong dimensional hypothesis'. As

---

[17] cf. (Scherer, 1996) and *http://emotion-research.net/wiki/Glossar*: "push effect:
the biologically determined externalization of naturally occurring internal processes
of the organism, particularly information processing and behavioral preparation;
pull effects: socioculturally determined norms or moulds concerning the signal char-
acteristics required by the socially shared codes for the communication of internal
states and behavioral intentions."

[18] The fact that non-interactive emotional speech has been by far more investi-
gated than interactive speech is a scientific 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.
As for considerations along similar lines, cf. (Campbell, 2006). (Reidsma et al., 2006)
report that the 'classical' two-dimensional approach has not been well suited for
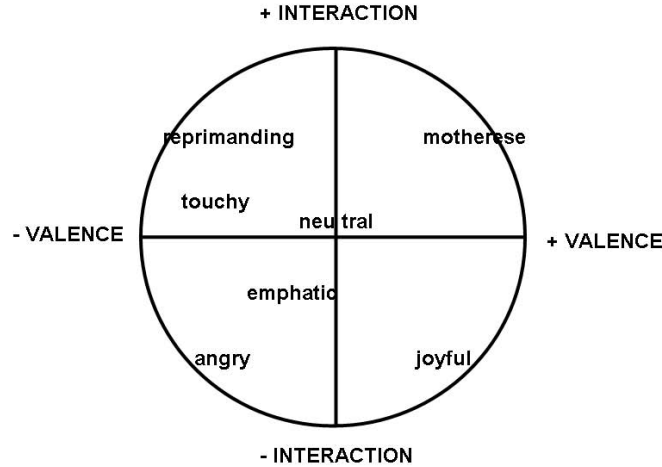meeting data with their interaction between participants.

Manuscript

Figure 14. Graphical representation of the two dimensions VALENCE and INTER-ACTION with our seven categorical labels

far as we can see, there is no convincing theoretical or methodological evidence in favour of or against this strong version yet.

10. Concluding remarks

We might not exactly be on the verge of a classic paradigm shift but we definitely are mid stream: turning from theoretical playgrounds towards demands put forth by applications. In this situation, we favour a rather data-driven, 'roving' approach such as the one described in this paper, i.e., realistic, non-acted data and non pre-defined sets of labels. Even if possibly, new models based on frequency distribution and combining emotion with the interaction aspect might be grounded in such studies, our more modest goal is for the moment simply to get at a clearer picture of the data we will have to deal with in possible applications: an additional characterisation in terms of some few dimensions might be more informative than just using a list of categorical labels.

   In conclusion and coming back to the title of this paper 'private emotions vs. social interaction': 'typical' emotions are to a large extent rather private and therefore, we might not be able to observe them as often, esp. in 'public' settings. Instead, it might be necessary to model social interaction in more detail.

Manuscript

Acknowledgements

References

Ai, H., D. J. Litman, K. Forbes-Riley, M. Rotaru, J. Tetreault, and A. Purandare: 2006, 'Using System and User Performance Features to Improve Emotion Detection in Spoken Tutoring Dialogs'. In: *Proceedings of the International Conference on Spoken Language Processing (Interspeech 2006 – ICSLP)*. Pittsburgh, pp. 797–800.

Ang, J., R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke: 2002, 'Prosody-based automatic detection of annoyance and frustration in human-computer dialog'. In: *Proceedings of the International Conference on Spoken Language Processing (Interspeech 2002 – ICSLP)*. Denver, pp. 2037–2040.

Batliner, A., S. Biersack, and S. Steidl: 2006a, 'The Prosody of Pet Robot Directed Speech: Evidence from Children'. In: *Proceedings of Speech Prosody 2006*. Dresden, pp. 1–4.

Batliner, A., K. Fischer, R. Huber, J. Spilker, and E. Nöth: 2000a, 'Desperately Seeking Emotions: Actors, Wizards, and Human Beings'. In: *Proceedings of the ISCA Workshop on Speech and Emotion*. Newcastle, Northern Ireland, pp. 195–200.

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, S. D'Arcy, M. Russell, and M. Wong: 2004a, '"You stupid tin box" - children interacting with the AIBO robot: A cross-linguistic emotional speech corpus'. In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2004)*. Lisbon, pp. 171–174.

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: *Proceedings of an ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*. Chateau d'Oex, pp. 5–10.

Batliner, A., C. Hacker, S. Steidl, E. Nöth, and J. Haas: 2004b, 'From Emotion to Interaction: Lessons from Real Human-Machine-Dialogues'. In: E. André, L. Dybkiaer, W. Minker, and P. Heisterkamp (eds.): *Affective Dialogue Systems, Proceedings of a Tutorial and Research Workshop*, Vol. 3068 of *Lecture Notes in Artificial Intelligence*. Berlin: Springer, pp. 1–12.

Manuscript

Batliner, A., R. Huber, H. Niemann, E. Nöth, J. Spilker, and K. Fischer: 2000b, 'The Recognition of Emotion'. In: W. Wahlster (ed.): *Verbmobil: Foundations of Speech-to-Speech Translations*. Berlin: Springer, pp. 122–130.

Batliner, A. and B. Möbius: 2005, 'Prosodic Models, Automatic Speech Understanding, and Speech Synthesis: Towards the Common Ground?'. In: W. Barry and W. Dommelen (eds.): *The Integration of Phonetic Knowledge in Speech Technology*. Dordrecht: Springer, pp. 21–44.

Batliner, A., S. Steidl, C. Hacker, E. Nöth, and H. Niemann: 2005a, 'Private Emotions vs. Social Interaction - towards New Dimensions in Research on Emotion'. In: *Proceedings of a Workshop on Adapting the Interaction Style to Affective Factors, 10th International Conference on User Modelling*. Edinburgh. no pagination.

Batliner, A., S. Steidl, C. Hacker, E. Nöth, and H. Niemann: 2005b, 'Tales of Tuning – Prototyping for Automatic Classification of Emotional User States'. In: *Proceedings of the European Conference on Speech Communication and Technology (Interspeech 2005 - Eurospeech)*. Lisbon, pp. 489–492.

Batliner, A., S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson: 2006b, 'Combining Efforts for Improving Automatic Classification of Emotional User States'. In: *Proceedings of Language Technologies (IS-LTC 2006)*. Ljubljana, Slovenia, pp. 240–245.

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: *Proceedings of the European Conference on Speech Communication and Technology (Interspeech 2003 – Eurospeech)*. Geneva, pp. 733–736.

Campbell, N.: 2006, 'A Language-Resources Approach to Emotion: the Analysis of Expressive Speech'. In: *Proceedings of a Satellite Workshop of the International Conference on Language Resources and Evaluation (LREC 2006) on Corpora for Research on Emotion and Affect*. Genoa, pp. 1–5.

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

Cowie, R., E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder: 2000, ''FEELTRACE': An instrument for recording perceived emotion in real time'. In: *Proceedings of the ISCA Workshop on Speech and Emotion*. Newcastle, Northern Ireland, pp. 19 – 24.

Cowie, R. and M. Schröder: 2004, 'Piecing Together the Emotion Jigsaw'. In: S. Bengio and H. Bourlard (eds.): *Machine Learning for Multimodal Interaction, First International Workshop, MLMI 2004, Martigny, Switzerland, June 21-23, 2004*, Vol. 3361 of *Lecture Notes in Computer Science*. Berlin: Springer, pp. 305 – 317.

Devillers, L. and L. Vidrascu: 2006, 'Real-life emotions detection with lexical and paralinguistic cues on Human-Human call center dialogs'. In: *Proceedings of the International Conference on Spoken Language Processing (Interspeech 2006 – ICSLP)*. Pittsburgh, pp. 801–804.

Devillers, L., L. Vidrascu, and L. Lamel: 2005, 'Challenges in real-life emotion annotation and machine learning based detection'. *Neural Networks* **18**, 407–422.

D'Mello, S.K., S.D. Craig, A. Witherspoon, B. McDaniel, and A. Graesser: 2008, 'Automatic Detection of Learner's Affect from Conversational Cues'. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research* **XX**, ??–??

Manuscript

Gratch, J., W. Mao, and S. Marsella: 2006, 'Modeling Social Emotions and Social Attributions'. In: R. Sun (ed.): *Cognitive Modeling and Multi-agent Interactions.* Cambridge: Cambridge University Press, pp. 219–251.

Jäger, R. and J. Bortz: 2001, 'Rating scales with smilies as symbolic labels – determined and checked by methods of Psychophysics'. In: *70. Annual Meeting of the International Society for Psychophysics.* Leipzig. no pagination.

Kehrein, R.: 2002, *Prosodie und Emotionen.* Tübingen: Niemeyer.

Kruskal, J. and M. Wish: 1978, *Multidimensional Scaling.* Beverly Hills and London: Sage University.

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

Laskowski, K. and S. Burger: 2006, 'Annotation and Analysis of Emotionally Relevant Behavior in the ISL Meeting Corpus'. In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2006).* Genoa, pp. 1111–1116.

Lee, C., S. Narayanan, and R. Pieraccini: 2001, 'Recognition of Negative Emotions from the Speech Signal'. In: *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU'01).* no pagination.

Lyons, M., S. Akamatsu, M. Kamachi, and J. Gyoba: 1998, 'Coding Facial Expressions with Gabor Wavelets'. In: *Proceedings of the 3rd International Conference on Face & Gesture Recognition (FG '98), Nara, Japan.* pp. 200–205.

Müller, S. and W. Kasper: 2000, 'HPSG Analysis of German'. In: W. Wahlster (ed.): *Verbmobil: Foundations of Speech-to-Speech Translations.* Berlin: Springer, pp. 238–253.

Neiberg, D., K. Elenius, and K. Laskowski: 2006, 'Emotion Recognition in Spontaneous Speech Using GMMs'. In: *Proceedings of The International Conference on Spoken Language Processing (Interspeech 2006 – ICSLP).* Pittsburgh, pp. 809–812.

Ortony, A., G. L. Clore, and A. Collins: 1988, *The Cognitive Structure of Emotion.* Cambridge: Cambridge University Press.

Osgood, C., G. Suci, and P. Tannenbaum: 1957, *The measurement of meaning.* Urbana: University of Illinois Press.

Picard, R.: 1997, *Affective Computing.* Cambridge, MA: MIT Press.

Poggi, I., C. Pelachaud, and B. D. Carolis: 2001, 'To Display or Not To Display? Towards the Architecture of a Reflexive Agent'. In: *Proceedings of the 2nd Workshop on Attitude, Personality and Emotions in User-adapted Interaction. User Modeling 2001.* Sonthofen, Germany, pp. 13–17.

Portele, T.: 2004, 'Interaction Modeling in the SmartKom system'. In: E. André, L. Dybkiaer, W. Minker, and P. Heisterkamp (eds.): *Affective Dialogue Systems, Proceedings of a Tutorial and Research Workshop*, Vol. 3068 of *Lecture Notes in Artificial Intelligence.* Berlin: Springer, pp. 89–94.

Reidsma, D., D. Heylen, and R. Ordelman: 2006, 'Annotating Emotions in Meetings'. In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2006).* Genoa, pp. 1117–1122.

Sammon, J.: 1969, 'A nonlinear mapping for data structure analysis'. *IEEE Transactions on Computers* **C-18**, 401–409.

Scherer, K. and G. Ceschi: 2000, 'Criteria for Emotion Recognition From Verbal and Nonverbal Expression: Studying Baggage Loss in the Airport'. *Personality and Social Psychology Bulletin* **26**, 327–339.

Manuscript

Scherer, K. R.: 1996, 'Adding the Affective Dimension: A New Look in Speech Analysis and Synthesis'. In: *Proceedings of the International Conference on Spoken Language Processing (ICSLP 1996)*. Philadelphia, PA. no pagination.

Scherer, K. R.: 2001, 'The Nature and Study of Appraisal: A Review of the Issues'. In: K. R. Scherer, A. Schorr, and T. Johnstone (eds.): *Appraisal processes in emotion: Theory, Methods, Research*. Oxford University Press, pp. 369–391.

Scherer, K. R.: 2003, 'Vocal communication of emotion: A review of research paradigms'. *Speech Communication* **40**, 227–256.

Schlosberg, H.: 1941, 'A scale for judgment of facial expressions'. *Journal of Experimental Psychology* **29**, 497–510.

Schlosberg, H.: 1952, 'The description of facial expressions in terms of two dimensions'. *Journal of Experimental Psychology* **44**, 229–237.

Schlosberg, H.: 1954, 'Three dimensions of emotion'. *Psychological Review* **61**, 81–88.

Schröder, M.: 2004, *Speech and Emotion Research. An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis*, Vol. 7 of *Reports in Phonetics, University of the Saarland*. Institute for Phonetics, University of Saarbrücken.

Slaney, M. and G. McRoberts: 1998, 'Baby Ears: A Recognition System for Affective Vocalizations'. In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1998)*. Seattle, pp. 985–988.

Steidl, S., M. Levit, A. Batliner, E. Nöth, and H. Niemann: 2005, '"Of All Things the Measure is Man": Automatic Classification of Emotions and Inter-Labeler Consistency'. In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*. Philadelphia, pp. 317–320.

Streit, M., A. Batliner, and T. Portele: 2006, 'Emotions Analysis and Emotion-Handling Subdialogues'. In: W. Wahlster (ed.): *SmartKom: Foundations of Multimodal Dialogue Systems*. Berlin: Springer, pp. 317–332.

Watzlawick, P., J. Beavin, and D. D. Jackson: 1967, *Pragmatics of Human Communications*. New York: W.W. Norton & Company.

Wundt, W.: 1896, *Grundriss der Psychologie*. Leipzig: Engelmann.

Zeißler, V., J. Adelhardt, A. Batliner, C. Frank, E. Nöth, P. Shi, and H. Niemann: 2006, 'The Prosody Module'. In: W. Wahlster (ed.): *SmartKom: Foundations of Multimodal Dialogue Systems*. Berlin: Springer, pp. 139–152.

Manuscript

Author's vitae

Anton Batliner

Anton Batliner received his M.A. degree in Scandinavian languages in 1973 and his doctoral degree in phonetics in 1978, both from the University of Munich, Germany. Since 1997 he is senior researcher at the Institute of Pattern Recognition at Friedrich-Alexander University Erlangen-Nuremberg. His research interests are the modelling and automatic recognition of emotional user states, all aspects of prosody in speech processing, focus of attention, and spontaneous speech phenomena such as disfluencies, irregular phonation, etc.

Stefan Steidl

Stefan Steidl is a Ph. D. candidate in Computer Science at the Institute of Pattern Recognition at Friedrich-Alexander University Erlangen-Nuremberg, where he also received his Diploma degree in 2002. His primary interests lie in the area of automatic classification of naturalistic emotional user states from speech. Previous research has also included work in speech recognition and speaker adaptation.

Christian Hacker

Christian Hacker is member of the research staff at the Institute of Pattern Recognition at Friedrich-Alexander University Erlangen-Nuremberg. There, he received his Diploma degree in 2002 and is now Ph. D. candidate. His primary interests lie in the area of speech recognition and analysis, which comprises assessment of non-native children, classification of emotional user states, and multi-modal classification of the user's focus of attention.

Elmar Nöth

Elmar Nöth obtained his Diploma degree and his doctoral degree from the Institute of Pattern Recognition at the University of Erlangen-Nuremberg in 1985 and 1990, respectively. Since 1990 he is an Associate Professor and the head of the speech group at the same institute. His current research activities concern prosody, the detection of emotion and user state, multi-modal human-machine interaction, and the automatic analysis of pathologic speech.

Manuscript

Manuscript