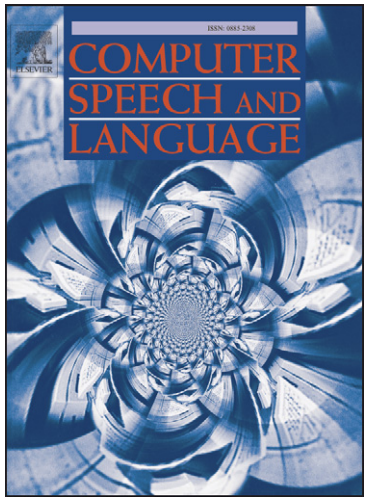
Whodunnit - Searching for the Most Important Feature Types Signalling Emotion-Related User States in Speech

Anton Batliner, Stefan Steidl, Björn Schuller, Dino Seppi, Thurid Vogt, Johannes Wagner, Laurence Devillers, Laurence Vidrascu, Vered Aharonson, Loic Kessous, et al.



Whodunnit – Searching for the Most Important Feature Types Signalling Emotion-Related User States in Speech

Anton Batliner, Stefan Steidl, Björn Schuller, Dino Seppi, Thurid Vogt, Johannes Wagner, Laurence Devillers, Laurence Vidrascu, Vered Aharonson, Loic Kessous, Noam Amir

PII: S0885-2308(09)00074-6
DOI: 10.1016/j.csl.2009.12.003
Reference: YCSLA 438

To appear in: *Computer Speech and Language*

Received Date: 31 July 2008
Revised Date: 26 April 2009
Accepted Date: 24 December 2009

Please cite this article as: Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Aharonson, V., Kessous, L., Amir, N., Whodunnit – Searching for the Most Important Feature Types Signalling Emotion-Related User States in Speech, *Computer Speech and Language* (2010), doi: 10.1016/j.csl.2009.12.003

Whodunnit – Searching for the Most Important Feature Types Signalling Emotion-Related User States in Speech

Anton Batliner [a] Stefan Steidl [a] Björn Schuller [b] Dino Seppi [c] Thurid Vogt [d] Johannes Wagner [d] Laurence Devillers [e] Laurence Vidrascu [e] Vered Aharonson [f] Loic Kessous [g] Noam Amir [g]

[a] *FAU: Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany*

[b] *TUM: Institute for Human-Machine Communication, Technische Universität München, Germany*

[c] *FBK: Fondazione Bruno Kessler – irst, Trento, Italy*

[d] *UA: Multimedia Concepts and their Applications, University of Augsburg, Germany*

[e] *LIMSI-CNRS, Spoken Language Processing Group, Orsay Cedex, France*

[f] *AFEKA: Tel Aviv Academic College of Engineering, Tel Aviv, Israel*

[g] *TAU: Dep. of Communication Disorders, Sackler Faculty of Medicine, Tel Aviv University, Israel*

Abstract

In this article, we describe and interpret a set of acoustic and linguistic features that characterise emotional/emotion-related user states – confined to the one database processed: four classes in a German corpus of children interacting with a pet robot. To this end, we collected a very large feature vector consisting of more than 4000 features extracted at different sites. We performed extensive feature selection (Sequential Forward Floating Search) for seven acoustic and four linguistic types of features, ending up in a small number of 'most important' features which we try to interpret by discussing the impact of different feature and extraction types. We establish different measures of impact and discuss the mutual influence of acoustics and linguistics.

Key words: feature types, feature selection, automatic classification, emotion

1 Introduction

The manifestations of affective/emotional states in speech have become the subject of great interest in recent years. In this article, we refrain from attempting to define terms such as 'affect' vs. 'emotion', and to attribute classes in a general way to the one term or the other. For those definitions we refer to the literature, e. g. to Cowie and Cornelius (2003); Ortony *et al.* (1988); Picard (1997). Furthermore, the phenomena we are interested in are partly cognitive. We therefore follow the convention of the HUMAINE project and employ the term 'pervasive emotion' in a broader sense encompassing "... whatever is present in most of life, but absent when people are emotionless ...", cf. Cowie *et al.* (2010); this term includes pure emotions and 'emotion-related states' such as 'interpersonal stances' which are specified as "affective stance taken towards another person in a specific interaction, colouring the interpersonal exchange in that situation" in Scherer (2003). Human-machine interaction will certainly profit from including these aspects, becoming more satisfactory and efficient.

Amongst the different and basically independent modalities of emotional expression, such as gesture, posture, facial expression and speech, this article will focus on speech alone. Speech plays a major role in human communication and expression, and distinguishes humans from other creatures. Moreover, in certain conditions such as communication via the phone, speech is the only channel available.

To prevent fruitless debates, we use the rather vague term 'emotion-related user states' in the title of this paper to point out that we are interested in empirically observable states of users within a human-machine communication, and that we are employing the concept of pervasive emotion in a broad sense. In the text, we will often use 'emotion' as the generic term, for better readability. This resembles the use of generic 'he' instead of 'he/she'; note, however, that in our context, it is not a matter of political correctness that might make a more cumbersome phrasing mandatory, it is only a matter of competing theoretical approaches, which are not the topic of the present article. The focus of this article is on methodology: we establish taxonomies of acoustic and linguistic features and describe new evaluation procedures for using very large feature sets in automatic classification, and for interpreting the impact of different feature types.

Email address: batliner@informatik.uni-erlangen.de (**Anton Batliner).**

[1] The initiative to co-operate was taken within the European Network of Excellence (NoE) HUMAINE under the name CEICES (Combining Efforts for Improving automatic Classification of Emotional user States). This work was partly funded by the EU in the projects PF-STAR under grant IST-2001-37599 and HUMAINE under grant IST-2002-50742. The responsibility lies with the authors.

1.1 Background

The study of 'Speech and Affect/Emotion' during the recent years can be characterised by three trends: (1) striving for more natural(istic), real-life data, (2) taking into account not only some 'prototypical', big n emotions but also emotion-related, affective states in a broader sense, and (3) the trend towards a thorough exploitation of the feature space, resulting in hundreds or even thousands of features used for classification. Note that (2) is conditioned by (1) – researchers simply realised that most of the full-blown, prototypical emotions that could easily be addressed and modelled for acted speech, were absent in realistic databases. Thus the set of emotion classes found in realistic databases normally consists of pervasive emotions in the broad sense, e. g. interest, boredom, etc., and of no or only a few prototypical emotions such as anger.

Relatively few studies have been conducted using more than one database, cf. Devillers and Vidrascu (2004); Shami and Verhelst (2007); Schuller *et al.* (2007b); Batliner *et al.* (2008a); Vidrascu and Devillers (2008), discussing similar or different characteristics of different databases; however, similar trends are sometimes pointed out across different studies. No study, however, has been able, or will be able in the foreseeable future, to exploit fully the huge feature space that models all possibly relevant factors, or to come up with a choice of 'real-life', realistic databases displaying representative samples of all emotional states. In this study, we concentrate on one specific database; this means that we cannot generalise our findings. On the other hand, we can safely compare across features and types because everything else can be kept constant. Results reported in Batliner *et al.* (2006) showed that pooling together features extracted at different sites indeed improved classification performance; Schuller *et al.* (2007a) was a first attempt at comparing feature types and their relevance for emotion classification. The present article will give a systematic account of the different steps – such as feature taxonomy and selection – that had to be taken in order to obtain a set of most relevant features and types of features.

The 'holy grail' of automatic classification is to find 'the' optimal set of the most important independent features. The task is difficult, due to factors such as the huge number of possible features that can be extracted from speech signals, and due to the computationally demanding methods needed for classifying such high-dimensional features spaces. The latter difficulty could be dealt with feature space de-correlation and reduction, e. g. through transformations like Principal Component Analysis (PCA). However, in this article we do not follow this approach because it would not provide the answer to the question which *types* of features contribute to classification performance, and to what extent; this information is crucial for understanding and mod-

elling the phenomenon we are interested in. Neither did we opt for comparing selection and classification results obtained at each site separately; instead, feature selection and classification were performed on a pooled set of features to enable a more reliable comparison between feature types. The various sites are rooted in different traditions; some focus on acoustics only, and other on a combination of acoustics and linguistics; some sites follow a 'brute-force' method of exploiting the feature space, while other sites compute features in a 'knowledge-based' way. Sometimes, hybrid strategies are used as well. In this article, we concentrate on feature types (Low Level Descriptors (LLDs) and functionals), and study their respective impact on classification performance.

1.2 State of the Art

In the 'pre-automatic' phase of emotion modelling, cf. Frick (1985), the inventory of features was more or less pre-defined or at least inspired by basic (phonetic) research. Hence, until the nineties of the last century, features were rather 'hand-picked', expert-driven, and based on phonetic knowledge and models; this was especially true for pitch (contour) features which were often based on intonation models, cf. Mozziconacci (1998). To give some examples of developments during the last years: at the beginning of 'real' automatic processing of emotion, Dellaert *et al.* (1996) for instance used 17 pitch features. McGilloway *et al.* (2000) reduced 375 measures to 32 variables as robust markers of emotion. Batliner *et al.* (2000a) used 27 prosodic features on the utterance level, Oudeyer (2003) 200 features and Information gain for feature reduction, Schuller *et al.* (2005) 276 features and SVM-SFFS (cf. below) for reduction, and Vogt and André (2005) 1280 features and correlation based feature subset selection (CFS).

More recently, expert-driven feature selection has often been replaced by the automatic generation and combination of features within the so called 'brute-force' approach. It is easy to create a feature vector which encompasses thousands of features, cf. Schuller *et al.* (2006). However, just using such large feature vectors is very time consuming; moreover, finding interesting and relevant features has simply been post-poned: while in the previous approaches, the selection of features was based on general considerations and took place before classification, in the newer ones, it is either an integral step of classification or has to be done after feature extraction and before classification. Dealing with such large feature vectors, one has to circumvent the curse of dimensionality: even if some statistical procedures are rather robust if there are too many features in relation to the number of items to be classified, it is definitely advisable to use some feature selection procedure.

1.3 CEICES: the Approach

Sites dealing with the automatic processing of emotion are rooted in specific traditions – such as a general engineering background, automatic speech recognition, or basic research (phonetics, psychology, etc.); thus their tools as well as the types of features they use, differ. For instance, linguistic information is normally only used by sites having some expertise in word recognition; on the other hand, features modelling aspects of intonation theories are normally only used by sites coming from basic (phonetic) research. The idea behind CEICES ('Combining Efforts for Improving automatic Classification of Emotional user States') was to overcome the 'fossilisation' at each site and to combine heterogeneous expertise in a sort of – metaphorically speaking – genetic approach: different features were separately extracted at different sites and subsequently combined in late or early fusion. [2] After agreeing on the training and the test set, the CEICES co-operation started with classification runs, independently at each site. The results are documented in Batliner *et al.* (2006). Basically, the classification performance was comparable across sites: the class-wise computed recognition rate in percent (this measure is described in Sec. 5.1 below) for the sites was: FAU 55.3, TUM 56.4, FBK 55.8, UA 52.3, LIMSI 56.6, and TAU/AFEKA 46.6. [3] We realized, however, that a strict comparison of the impact of different features and feature types was not possible with such 'benchmark-like' procedures, as too many factors were not constant across sites. To start with, a necessary prerequisite was an agreed-upon, machine readable representation of extracted features. Note that the idea of combining heterogeneous knowledge sources or representations is not new and has been pursued in approaches such as ROVER, Stacking, Ensemble Learning, etc. As the European Network of Excellence HUMAINE (2004–2007) was conceived as a network bringing together different branches of science dealing with emotion, a certain diversity was already given; moreover, sites from outside HUMAINE were invited to take part in the endeavour.

We want to point out that the number of features used in the present study (or in any other study) is of course not a virtue in itself, automatically paying off in classification performance; cf. Batliner *et al.* (2006) where we have seen that one site, using 'only' 32 features, produced a classification performance in the same range as other sites, using more than 1000 features. It is simply more convenient to automatise feature selection, and more importantly, this method ensures that we do not overlook relevant features.

[2] Late fusion was done in Batliner *et al.* (2006) by combining independent classifier output in the so-called ROVER approach; the early fusion will be reported on in this article.

[3] TAU/AFEKA used only rather specific pitch features and not multiple acoustic features as all other sites.

1.4 Overview

The present article deals with the following topics: in Sec. 2, we describe experimental design, recording, and emotion annotation. The segmentation into meaningful chunks as units of analysis, based on syntactic, semantic, and prosodic criteria, is presented in Sec. 3. In Sec. 4, we depict the features extracted at the different sites and the mapping of these features onto feature *types*; for that purpose an exhaustive *feature coding scheme* has been developed. In Sec. 5, we address the classifier and the feature selection procedure chosen, discuss classification performance (overall and separately for each acoustic and linguistic feature type), and introduce some specific performance measures.

In Sec. 6, we summarise the findings and discuss some important, general topics. In order to focus the presentation, we decided not to give a detailed account of all stages of processing if a stage is not pivotal for the topic of this article; as for details we refer to Steidl (2009). [4]

2 The Database

2.1 Design and Recording

The database used is a German corpus of children communicating with Sony's pet robot AIBO, the *FAU Aibo Emotion Corpus* [5]. This database can be considered as a corpus of spontaneous speech, because the children were not given specific instructions. They were just told to talk to AIBO as they would talk to a friend. Emotional, affective states conveyed in this speech are not elicited explicitly (prompted) but produced by the children in the course of their interaction with the AIBO; thus they are fully natural(istic). The children were led to believe that AIBO was responding to their commands, whereas the robot was actually controlled by a human operator (Wizard-of-Oz, WoZ) using the 'AIBO Navigator' software over a wireless LAN; the existing AIBO speech recognition module was not used, and the AIBO did not produce speech. The WoZ caused the AIBO to perform a fixed, pre-determined sequence of actions; sometimes the AIBO behaved disobediently, thus provoking emotional reactions. The data was collected at two different schools from 51 children

[4] The book is available online at:
http://www5.cs.fau.de/en/our-team/steidl-stefan/dissertation

[5] As there are other 'Aibo' corpora with emotional speech, cf. Tato *et al.* (2002); Küstner *et al.* (2004), the specification 'FAU' is used.

(age 10 - 13, 21 male, 30 female). Speech was transmitted via a wireless head set (UT 14/20 TP SHURE UHF-series with microphone WH20TQG) and recorded using a DAT-recorder (sampling rate 48 kHz, quantisation 16 bit, down-sampled to 16 kHz). Each recording session took some 30 minutes. Due to this experimental setup, these recordings contained a huge amount of silence (reaction time of the AIBO), which caused a noticeable reduction of recorded speech after raw segmentation; ultimately we obtained about 8.9 hours of speech.

In planning 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, while being careful not to risk their breaking off the experiment. The children believed that the AIBO was reacting to their orders – albeit often not immediately. In reality, the scenario was the opposite: the AIBO always followed strictly the same plot, and the children had to modify their orders to its actions. By this means, it was possible to examine different children's reactions to the very same sequence of AIBO's actions. Examples for the tasks to be fulfilled and for the experimental design can be found in Steidl (2009), p. 73ff.

In each of the other five tasks of the experiment, the children were instructed to direct the AIBO towards one of several cups standing on the carpet. One of these cups was allegedly 'poisoned' and had to be avoided. The children applied different strategies to direct the AIBO. Again, all actions of the AIBO were pre-determined. In the first task, the AIBO was 'obedient' in order to make the children believe that the AIBO would understand their commands. In the other tasks, the AIBO was 'disobedient'. In some tasks the AIBO went directly towards the 'poisoned' cup in order to evoke emotional speech from the children. No child broke off the experiment, although it could be clearly seen towards the end that many of them were bored and wanted to put an end to the experiment – a reaction that we wanted to provoke. Interestingly, in a post-experimental questionnaire, all the children reported that they had much fun and liked it very much. At least two different conceptualisations could be observed: in the first, the AIBO was treated as a sort of remote-control toy (commands like *turn left, straight on, to the right*); in the second, the AIBO was addressed as a pet dog (commands like *Little Aibo doggy, now please turn left - well done, great!*) or *Get up, you stupid tin box!*), cf. Batliner *et al.* (2008b).

2.2 Manual Processing

The recordings were segmented automatically into 'utterances' or 'turns'[6] using a pause threshold of 1 s., Steidl (2009), p. 76ff. Each turn was transliterated, i. e. orthographically transcribed, by one annotator and cross-checked by another. In addition to the words, other 'non-linguistic' events such as breathing, laughing, and (technical) noise were annotated. For the experiments reported on in this article, we aimed at an optimal representation of the acoustic data. After a forced alignment using the spoken word chain, the automatic word segmentation of the subset used in this study was therefore corrected manually by the first author. Automatic pitch extraction was corrected manually by the first author as well; this procedure is described in more detail in Batliner *et al.* (2007b) and in Steidl (2009), p. 83ff.

2.3 Annotation

In the past, typical studies on emotion in speech used segmentally identical – and mostly, semantically neutral – utterances, produced in different emotions by actors. These utterances were processed as a whole; no word segmentation and/or eventual automatic word recognition were carried out. Recently, some researchers claimed that a combination of utterance level features along with segment-level features yields better performance, cf. Shami and Verhelst (2007). For establishing such 'segments', units smaller than the whole utterance must be defined: syllables, voiced/unvoiced parts, segments of fixed length or fixed proportion of the whole utterance. Although we believe that such strategies normally do pay off, a more promising approach is to incorporate word processing from the very beginning. After all, in a fully developed 'emotional system', not only acoustic information should be used for recognition but all linguistic information should be used for interaction, i. e. for understanding and generation/synthesis. In such a full end-to-end system, word recognition is an integral part, cf. Batliner *et al.* (2000b, 2003a).

In realistic speech databases with long stretches of speech, the word itself is normally not the optimal emotion unit to be processed. It is more reasonable to use larger units (termed here 'chunks') comprising one or up to several words, establishing syntactically/semantically meaningful units, and/or units repre-

[6] Note that 'turn' and 'utterance' are vague concepts: a turn is defined by 'turn-taking', i.e. change of speakers; an utterance can be defined by pauses before and after. As the AIBO does not speak, we rather have to do with 'action turns'. The length of such speech units can thus vary between one word and hundreds of words. We therefore aim at a more objective criterion using syntactic-prosodic information, cf. Sec. 3 below.

senting dialogue acts/moves. It has been shown that there is a high correlation between all these units, cf. Batliner *et al.* (1998, 2003a). Thus a reasonable strategy could be devised to segment the data in a pre-processing step into such units to be presented to the annotators for labelling emotions. However, this would require an a-priori knowledge on how to define the optimal unit – which we do not have yet. In order not to decide beforehand on the units to be processed, we decided in favour of a word-based labelling: each word had to be annotated with one emotion label. Later on, this makes it possible to explore different chunk sizes and different degrees of prototypicality.

The labels to be used for annotating emotional user states were data-driven. We started with a set that has been used for another realistic emotional database, cf. Batliner *et al.* (2004); the adaptation to FAU Aibo was done iteratively, in several steps, and supervised by an expert. Our five labellers (advanced students of linguistics) first listened to the whole interaction in order to become 'fine-tuned' to the children's baseline: some children sounded bored throughout, others 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 the AIBO's actions. Independently from each other, the annotators labelled each word as neutral (default) or as belonging to one of ten other classes. In the following list, we summarize the annotation strategy for each label:

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

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

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

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

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 the AIBO.

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

helpless: the child is hesitant, seems not to know what to tell the 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 modelling these children's behaviour in this specific scenario. We do claim, however, that it is an adequate strategy to use such a data-driven approach instead of one based on abstract theoretical models. Note that a more 'in-depth' approach followed by a few other studies would be first to establish an exhaustive list of classes (up to > 100, i. e. labels, or lists of both classes and dimensions, cf. Devillers *et al.* (2005). However, for automatic processing, this large list has to be reduced necessarily to fewer cover classes – we know of studies reporting recognition using up to seven discrete categories, for instance in Batliner *et al.* (2003b, 2008b) and eventually, if it comes to 'real' classification, three or two, e. g., *neutral* and *negative*. Some studies relying on the dimensional approach may obtain more classes by discretising the axes of the emotional space, cf. Grimm *et al.* (2007). Moreover, our database demonstrates that confining oneself to the 'classic' dimensions AROUSAL/INTENSITY and VALENCE might not be the best thing to do because the first one is not that important, and another one, namely (social) INTERACTION, comes to the fore instead, cf. Batliner *et al.* (2008b). Instead of putting too much effort into the earlier phases of establishing 'emotional dictionaries', we decided to concentrate on later stages of annotation, e. g., on manual correction of segmentation and pitch, and on the annotation of the interaction between the child and the AIBO.

If three or more labellers agreed, the label was attributed to the word (Majority Voting, MV); in parentheses, the number of cases with MV is given: *joyful* (101), *surprised* (0), *emphatic* (2528), *helpless* (3), *touchy*, i. e. irritated (225), *angry* (84), *motherese* (1260), *bored* (11), *reprimanding* (310), *rest*, i. e. non-neutral, but not belonging to the other categories (3), *neutral* (39169). 4707 words had no MV; all in all, there were 48401 words.

Some of the labels are very sparse; if we only take labels with more than 50 MVs, the resulting 7-class problem is most interesting from a methodological point of view, cf. the new dimensional representation of these seven categorical labels in Batliner *et al.* (2008b). However, the distribution of classes is highly non-homogeneous. Therefore, we randomly down-sampled *neutral* and *emphatic* to ***N****eutral* and ***E****mphatic*, respectively, and mapped *touchy*, *reprimanding*, and *angry* onto ***A****ngry*[7], as representing different but closely related kinds of negative attitude. This more balanced 4-class problem, which we refer to as AMEN, consists of 1557 words for ***A****ngry* (**A**), 1224 words for

[7] The initial letter is given boldfaced; this letter will be used in the following for referring to these four cover classes. 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.

***M**otherese* (**M**), 1645 words for ***E**mphatic* (**E**), and 1645 for ***N**eutral* (**N**), cf. Steidl *et al.* (2005). Cases where less than three labellers agreed were omitted, as well as cases labelled with other than these four main classes. This mapping onto cover classes is corroborated by the two- and one-dimensional Nonmetric Multidimensional Scaling solutions presented in Batliner *et al.* (2008b).

Angry belongs to the 'big', 'basic' emotions, cf. Ekman (1999), whereas the other ones are rather 'emotion-related/emotion-prone' user states and therefore represent 'pervasive emotions' in a broader meaning; most of them are addressed in Ortony *et al.* (1988) such as boredom, surprise, and reproach (i. e. *reprimanding*). *Touchy* is nothing else than *weak anger*. [8] The state *emphatic* has been introduced because it can be seen as a possible indication of some (starting) trouble in communication and by that, as a sort of 'pre-emotional' state, cf. Batliner *et al.* (2003a, 2005), or even as *weak anger*: any marked deviation from a neutral speaking style can (but does not need to) be taken as a possible indication of some (starting) trouble in communication. If a user gets the impression that the machine does not understand him/her, he/she tries different strategies – repetitions, re-formulations, other wordings, or simply the use of a pronounced, marked speaking style. Such a style does not necessarily indicate any deviation from a neutral user state, but it suggests a higher probability that the (neutral) user state will be changing soon. Of course, it can be something else as well: a user idiosyncrasy, or a special style – 'computer talk' – that some people use while speaking to a computer, like speaking to a non-native listener, to a child, or to an elderly person who is hard of hearing. Thus the fact that *emphatic* is observed can only be interpreted meaningfully if other factors are considered; note that we only annotated *emphatic* if this was not the default way of speaking. There are three further – practical – arguments for the annotation of *emphatic*: first, it is to a large extent a prosodic phenomenon, and can thus be modelled and classified with prosodic features. Second, if the labellers are allowed to label *emphatic*, it may be less likely that they confuse it with other user states. Third, as mentioned above, we can try and model emphasis as an indication of (arising) problems in communication, cf. Batliner *et al.* (2003a).

For assessing inter-rater reliability, weighted kappa for multi-raters, cf. Fleiss *et al.* (1969); Davies and Fleiss (1982), was computed for the four-class AMEN problem and for six classes splitting the ***A**ngry* cover class into the original classes *touchy*, *reprimanding*, and *angry*. The weighted version of kappa allows to penalise confusions of dissimilar emotion categories more than confusions of

[8] It is interesting that *motherese* has, to our knowledge, not really mentioned often in such listings of emotion terms, although child-directed speech has been addressed in several studies. We can speculate that researchers have been more interested in negative states such as reproach (*reprimanding*), i. e. in the negative pendant to *motherese*.

similar ones. Therefore, nominal categories have to be aligned on a linear scale such that the distances between categories can be meaningfully interpreted as dissimilarities. In order to employ an objective measure for the weighting, we used the co-ordinates derived from a one-dimensional Non-Metrical Dimensional Scaling (NMDS) solution based on the confusion matrix of the five labellers; cf. for details Batliner *et al.* (2008b). The distance measure used is based on squared differences. Weighted kappa is 0.59 for four classes, and 0.61 for six classes. (This is a rather small difference, presumably because in the one-dimensional NMDS solution, *touchy* and *reprimanding* have been given almost identical values, cf. Batliner *et al.* (2008b) p. 188.) Overall, kappa values are satisfactory albeit not very high – this could be expected, given the difficulty and subjectivity of the task. Another, entropy-based measure of inter-labeller agreement and agreement between labellers and automatic classification is dealt with in Steidl *et al.* (2005). [9]

2.4 Children's Speech

Our database might seem to be atypical since it deals with children's speech; however, children represent just one of the usual partitions of the world's population into sub-groups such as women/men, upper/lower class, or different dialects. Of course, automatic procedures have to adapt to this specific group – children's speech is a challenge for an Automatic Speech Recognition (ASR) system, cf. Blomberg and Elenius (2003), as both acoustic and linguistic characteristics differ from those of adults, cf. Giuliani and Gerosa (2003). However, this necessity to adapt to a specific sub-group is a frequent issue in speech processing. Pitch, formant positions, and not yet fully developed co-articulation vary strongly, especially for younger children due to anatomical and physiological development, cf. Lee *et al.* (1999). Moreover, until the age of five/six, expression and emotion are strongly linked: children express their emotions even if no one else is present; the expression of emotion can be rather intense. Later on, expressions and emotions are decoupled, cf. Holo-

[9] A note on label names and terminology in general: 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.* (2004). Instead of *motherese*, some people use 'child-directed speech'; this is, however, only feasible if the respective database does not contain any negative counterpart such as *reprimanding* which is 'child-directed' as well. ('Parentese' or 'fatherese' might be more politically correct but are descriptively and historically less adequate.) Our nomenclature is sometimes arbitrary – for example, we could exchange ***A****ngry* with ***N****egative* – which we had to avoid because we reserved **N** for ***N****eutral*. A methodological decision has been taken in favour of a categorical and not a dimensional representation. However, in Batliner *et al.* (2008b) we show how the one can be mapped onto the other.

dynski and Friedlmeier (2006), when children start to control their feelings. So far, we found no indication that our children (age 10-13) behave differently from adults in a *principled* way, as far as speech/linguistics in general or emotional states conveyed via speech are concerned. It is known, for example, that children in this age do not yet have full laryngeal control. Thus, they might produce more irregular phonation, but we could not find any evidence that they employ these traits differently from adults, cf. Batliner *et al.* (2007a). Moreover, our database is similar to other realistic, spontaneous (neutral and) emotional speech: although it is rather large, we are faced with the well-known sparse data problem which makes a mapping of sub-classes onto cover classes necessary – neutral is by far the most frequent class. The linguistic structure of the children's utterances is not too uniform, as it might have been if only pure, short commands were used; on the other hand, it displays specific traits, for instance, many *Aibo* vocatives because these are often used adjacent to commands. All this can, however, be traced back to this specific scenario and not to the fact that our subjects are children.

3 Segmentation

Finding the appropriate unit of analysis for emotion recognition has not posed a problem in studies involving acted speech with different emotions, using segmentally identical utterances, cf. Burkhardt *et al.* (2005); Engberg *et al.* (1997). In realistic data, a large variety of utterances can be found, from short commands in a well-defined dialogue setting, where the unit of analysis is obvious and identical to a dialogue move, to much longer utterances. In Batliner *et al.* (2003a) it has been shown that in a WoZ-scenario (appointment scheduling dialogues), it is beneficial not to model whole turns but to divide them into smaller, syntactically and semantically meaningful chunks. Our scenario differs in one pivotal aspect from most of the other scenarios investigated so far: there is no real dialogue between the two partners; only the child is speaking, and the AIBO is only acting. Thus it is not a 'tidy' stimulus-response sequence that can be followed by tracking the very same channel; we are using only the recordings of the children's speech. When annotating, we therefore do not know what the AIBO is doing at the corresponding time, or has been doing shortly before or after the child's utterance. Moreover, the speaking style is rather special: there are not many 'well-formed' utterances but a mixture of some long and many short sentences and one- or two-word utterances which are often commands.[10] We observe neither 'integrating' prosody as in the

[10] The statistics of the observable turn lengths (in terms of the number of words) for the whole database is as follows: 1 word (2538 times), 2 words (2800 times), 3 words (2959 times), 4 words (2134 times), 5 words (1190 times), 6-9 words (1560 times), ≥ 10 words (461 times). We see that on the one hand, the threshold for segmentation

case of reading, nor 'isolating' prosody as in the case of TV reporters. Many pauses of varying length are found which can be hesitation pauses – the child produces slowly while observing the AIBO's actions – or pauses segmenting into different dialogue acts – the child waits until he/she reacts to the AIBO's actions.

Note that in earlier studies, we found out that there is a rather strong correlation of up to $> 90\%$ between prosodic boundaries, syntactic boundaries, and dialogue act boundaries, cf. Batliner *et al.* (1998). Using only prosodic boundaries as chunk triggers might not result in (much) worse classification performance (in Batliner *et al.* (1998), some 5 percent points lower). However, from a practical point of view, it would be more cumbersome to time-align the different units – prosodic, i. e. acoustic units, and linguistic, i. e. syntactic or dialogue units, based on automatic speech recognition and higher level segmentation – at a later stage in an end-to-end processing system, and to interpret the combination of these two different types of units accordingly. [11]

A detailed account of our segmentation principles can be found in Steidl (2009), p. 89ff; in Batliner *et al.* (2009), different types of emotion units, based on different segmentation principles, are compared. In our segmentation, we basically annotated a chunk boundary after higher syntactic units such as main clauses and free phrases; after lower syntactic units such as coordinate clause and dislocations, we only introduced such a boundary when the pause is longer than 500 ms. By that, we could chunk longer turns – we obtained turns containing up to 50 words – into meaningful smaller units. The following example illustrates such a long turn, divided into meaningful syntactic units; the boundary is indicated by a pipe symbol. The German original of this example and further details can be found in Steidl (2009), p.90, and in Batliner *et al.* (2009).

English translation with chunk boundaries: *and stop Aibo stand still | go this way | to the left towards the street | well done Aibo and now go on | well done Aibo | and further on | and now turn into the street to the left | to the blue cup | no Aibo no | stop Aibo no | no Aibo stop | stand still | Aibo stand still |*

of 1 s is meaningful; on the other hand, there are still many turns having more than 5 words per turn. This means that they tend to be longer than one intonation unit, one clause, or one elementary dialogue act unit, which are common in this restricted setting 'giving commands'.

[11] Preliminary experiments with chunks of different granularity, i. e. length, showed that using our longer turns actually results in sub-optimal classification performance, while the chunking procedure presented below which was used for the experiments dealt with in this article, results in better performance. This might partly result from the fact that more training instances are available, but partly as well from the fact that shorter units are more 'consistent'.

Now we had to map our word-based labels onto chunk-based labels. A simple majority vote on the raw labels (the decisions of the single labellers for each word in the turn or chunk) does not necessarily yield a meaningful label for the whole unit. A whole turn which, for example, consists of two main clauses – one clause which is labelled as ***N****eutral* and one slightly shorter clause which is labelled as ***A****ngry* by the majority – would be labelled as ***N****eutral*. A chunk consisting of five words, two of them clearly labelled as ***M****otherese*, three of them being ***N****eutral*, can be reasonably labelled as ***M****otherese* although the majority of raw labels yields a different result – after all, we are interested in deviations from the default ***N****eutral*. Thus, the mapping of labels from word level onto higher units is not as obvious as one might expect. A more practical problem of a simple majority vote is that the sparse data problem, which already exists on the word level, becomes aggravated on higher levels since the dominating choice of the label *neutral* on the word level yields an even higher proportion of *neutral* chunks and turns.

We developed a simple heuristic algorithm. It uses the raw labels on the word level mapped onto the cover classes ***N****eutral*, ***E****mphatic*, ***A****ngry*, and ***M****otherese*. Due to their low frequency, labels of the remaining two cover classes ***J****oyful* and ***R****est* (*other*) are ignored. If the proportion of the raw labels for ***N****eutral* is above a threshold $\theta_{\mathbf{N}}$, the whole unit is considered to be ***N****eutral*. This threshold depends on the length of the unit; the longer the unit is, the higher the threshold need to be set. For our chunks, it is set to 60 %. If this threshold is not reached, the frequency of the label ***M****otherese* is compared to the sum of the frequencies of ***E****mphatic* and ***A****ngry* which are pooled since *emphatic* is considered as a possible pre-stage of *anger*. If ***M****otherese* prevails, the chunk is labelled as ***M****otherese*, provided that the relative frequency of ***M****otherese* w. r. t. the other three cover classes is not too low, i. e. it is above a certain threshold $\theta_{\mathbf{M}} = 40\,\%$. If not, the whole unit is considered to be ***N****eutral*. If ***M****otherese* does not prevail, the frequency of ***E****mphatic* is compared to the one of ***A****ngry*. The label of the whole unit is the one of the prevailing class, again provided that the relative frequency of this class w. r. t. the other three cover classes is above a threshold $\theta_{\mathbf{EA}} = 50\,\%$. The thresholds are set heuristically by checking the results of the algorithm for a random subset of chunks and have to be adapted to the average length of the chosen units. A structogram describing the exact algorithm can be found in Steidl (2009), p. 101.

If all 13642 turns are split into chunks, the chunk triggering procedure results in a total of 18216 chunks. 4543 chunks contain at least one word of the original AMEN set. Compared to the original AMEN set, where the four emotion labels on the word level are roughly balanced, the frequencies of the chunk labels for this subset differ to a larger extent: 914 ***A****ngry*, 586 ***M****otherese*, 1045 ***E****mphatic*, and 1998 ***N****eutral*. Nevertheless, in the training phase of a machine classifier, these differences can be easily equalised by up-sampling of the less

frequent classes. On average, the resulting 4543 chunks are 2.9 words long; in comparison, there are 3.5 words per turn on average in the whole *FAU Aibo corpus*.

The basic criteria of our chunking rules have been formulated in Batliner *et al.* (1998); of course, other thresholds could be imagined if backed by empirical results. The rules for these procedures can be automated fully; in Batliner *et al.* (1998), multi-layer perceptrons and language models have successfully been employed for an automatic recognition of similar syntactic-prosodic boundaries, yielding a class-wise average recognition rate of 90% for two classes (boundary vs. no boundary). Our criteria are 'external' and objective, and are not based on intuitive notions of an 'emotional' unit of analysis as in the studies by Devillers *et al.* (2005); Inanoglu and Caneel (2005); de Rosis *et al.* (2007). Moreover, using syntactically motivated units makes processing in an end-to-end system more straightforward and adequate.

4 Features

In Batliner *et al.* (2006), we combined for the first time features extracted at different sites. These were used both for *late* fusion using the ROVER approach, cf. Fiscus (1997), and for *early* fusion, by combining only the most relevant features from each site within the same classifier. We were only interested in classification performance, thus an unambiguous assignment to feature types and functionals was not yet necessary. It turned out, however, that we had to establish a uniform taxonomy of features, which could also be processed fully automatically. To give one example of a possible point of disagreement: the temporal aspects of pitch configurations are often subsumed under 'pitch' as well. However, the positions of pitch extrema on the time axis clearly represent duration information, cf. Batliner *et al.* (2007b). Thus on the one hand, we decided to treat these positions as belonging to the duration domain across all sites; on the other hand, we of course wanted to encode this hybrid status as well. For our encoding scheme, we decided in favour of a straightforward ASCII representation: one line for each extracted feature; each column is attributed a unique semantics. This encoding can be easily converted into a Markup Language such as the one envisaged by Schröder *et al.* (2007), cf. as well Schröder *et al.* (2006). [12]

[12] A documentation of the scheme can be downloaded from:
http://www5.cs.fau.de/en/our-team/steidl-stefan/materials

4.1 Types of Feature Extraction

Before characterising the feature types, we give a broad description of the extraction strategy employed by each site. More specifically we can identify three different approaches generating three different sets of features: the *'selective'* approach is based on phonetic and linguistic knowledge, cf. Kießling (1997); Devillers *et al.* (2005); this could be called 'knowledge-based' in its literal meaning. The number of features per set is rather low, compared to the number of features in sets based on *'brute-force'* approaches. There, a strict systematic strategy for generating the features is chosen; a fixed set of functions is applied to time series of different base functions. This approach normally results in more than 1 k features per set, cf. the figures given at the end of this section. From a 'technical' point of view, the differences between the two approaches can be seen in the feature selection step: in the selective approach, the main selection takes place before putting the features into the classification process; in the brute-force approach, an automatic feature selection is mandatory.[13] Moreover, for the computation of some of our selective features, FAU/FBK use manually corrected word segmentation, by that employing additional knowledge. (This is, of course, not a necessary step; as for a fully automatic processing of this database, cf. Schuller *et al.* (2007b).) The approach of FAU/FBK will be called *'two-layered'*: in a first step, word-based features are computed; in a second step, functionals such as mean values of all word-based features are computed for the chunks. In contrast, a *'single-layered'* approach is used by all other sites, i.e. features are computed for the whole chunk. The following arrangement into types of feature extraction has to be taken with a grain of salt; it rather describes the starting point and the basic approach. FAU for instance uses a selective approach for the computation of word-based features, and then a systematic approach for the subsequent computation of chunk-based features; some of UA's feature computations could be called two-layered because functionals are applied twice. To sum up, there are three different *types of feature extraction*:[14]

set I: FAU/FBK; selective, two-layered; 118 acoustic and 30 linguistic features.

[13] It is an empirical question which type of extraction yields better performing features, and there are at least the following aspects to be taken into account: (1) given the same number of features, which set performs better? (2) Which features can be better interpreted? (3) Which features are more generic, i. e. can be used for different types of data without loosing predictive power? The last aspect might be most important but has not been addressed yet: typically, feature evaluation is done within one study, for one database.

[14] Note that w. r. t. to Schuller *et al.* (2007a), we have changed the terminology presented in this section to avoid ambiguities.

set II: TAU/LIMSI; selective, single-layered; 312 acoustic and 12 linguistic features.

set III: UA/TUM; brute-force, single-layered; 3304 acoustic and 489 linguistic features.

In the following, we shortly describe the features extracted at each site.

4.1.1 Two-layered, 'selective' computation: chunk features, based on word statistics

FAU: 92 *acoustic features:* word-based computation (using manually corrected word segmentation) of pauses, energy, duration, and F0; for energy: maximum (max), minimum (min), mean, absolute value, normalised value, and regression curve coefficients with mean square error; for duration: absolute and normalised; for F0: min, max, mean, and regression curve coefficients with mean square error, position on the time axis for F0 onset, F0 offset, and F0 max; for jitter and shimmer: mean and variance; normalisation for energy and duration based on speaker-independent mean phone values; for all these word-based features, min, max, and mean chunk values computed based on all words in the chunk. 24 *linguistic features:* part-of-speech (POS) features: AUX (auxiliaries), PAJ (particles, articles, and interjections), VERB (verbs), APN (adjectives and participles, not inflected), API (adjectives and participles, inflected), and NOUN (nouns, proper nouns), annotated for the spoken word chain (# of classes per chunk and normalised as for # of words in chunk); higher semantic features (SEM): vocative, positive valence, negative valence, commands and directions, interjections, and rest (# of classes per chunk and normalised as for # of words in chunk).

FBK: 26 *acoustic features:* similar to the ones of FAU, with the following difference: no F0 onset and offset values, no jitter/shimmer; normalisation of duration and energy done on the training set without backing off to phones but using information on the number of syllables in addition, cf. Kießling (1997); 6 *linguistic features:* POS features.

4.1.2 Single-layered, 'selective' computation of chunk features

LIMSI: 90 *acoustic features:* min, max, median, mean, quartiles, range, standard deviation for F0; the regression curve coefficients in the voiced segments, its slope and its mean square error; calculations of energy and of the first 3 formants and their bandwidth; duration features (speaking rate, ratio of the voiced and unvoiced parts); voice quality (jitter, shimmer, Noise-to-Harmonics Ratio (NHR), Harmonics-to-Noise Ratio (HNR), etc.), cf. Devillers *et al.*

(2005); 12 *linguistic features:* POS, nonverbals and disfluencies.

TAU/AFEKA: 222 *acoustic features:* five families of features: pitch based, duration based, intensity based, spectral, and voice quality based; different levels of functionals applied to the raw contours: from basic statistics to curve fitting methods to methods based on perceptual criteria. Several duration features computed on the lengths of voiced segments and pauses, and spectral features based on Mel Frequency Cepstral Coefficients (MFCC) and Long Term Average Spectrum (LTAS).

4.1.3 Single-layered, 'brute force' computations of chunk features

UA: 1586 *acoustic features:* pitch, energy, 12 MFCCs, 10 cepstral coefficients based on wavelet transformation, HNR and short-term spectra, as well as different views on the time series such as considering only local maxima or minima, or distances, magnitudes and steepness between adjacent extrema. From each of these series of values, mean, max, min, range, median, first quartile, third quartile, interquartile range, and variance. Chunk length added to the vector as a durational feature. The proportion of voiced to unvoiced frames, several normalised and positional features of pitch and energy.

TUM: 1718 *acoustic features:* a systematic generation by acoustic Low Level Descriptors (LLD) extraction, filtering, derivation, and application of functionals on the chunk level. As LLDs pitch, HNR, jitter, shimmer, energy, MFCCs 1-16, formants 1-7 with amplitude, position, and bandwidth, and a selection of spectral features; derived LLDs comprising derivatives and crossed LLDs; functionals covering the first four moments, extremes, quartiles, ranges, zero-crossings, roll-off, and higher level analysis. 489 *linguistic features:* frequencies of bag of words (BOW), cf. Joachims (1998), using the manual transliteration of the spoken word chain, POS, non-verbals, and disfluencies.

4.2 Types of Features

In the following clustering of the 4244 features used in this study [15], we shortly describe the breakdown into types of LLDs. We concentrate on a characterisation in phonetic and linguistic terms (*what* has been extracted); in parentheses, we indicate by which set, i. e. by which extraction type (which sites) this feature type is used:

[15] Note that 21 acoustic features could not be attributed un-equivocally to one of the types; the final figures is thus 112 for set I, 297 for set II, and 3304 for set III.

duration: these features model temporal aspects; the basic unit is milliseconds for the 'raw' values. Different types of normalisation are applied. Positions of prominent energy or F0 values on the time axis are attributed to this type as well (set I, II, III).

energy: these features model intensity, based on the amplitude in different intervals; different types of normalisation are applied. Energy features can model intervals or characterising points (set I, II, III).

pitch: the acoustic equivalent to the perceptual unit pitch is measured in Hz and often made perceptually more adequate by logarithmic transformation and by different normalizations. Intervals, characterising points, or contours are being modelled (set I, II, III).

spectrum: formants (spectral maxima) model spoken content, esp. lower ones. Higher ones also represent speaker characteristics. Each one is fully represented by position, amplitude and bandwidth. As further spectral features band-energies, roll-off, centroid or flux are used. Long term average spectrum over a chunk averages out formant information, giving general spectral trends (set II, III).

cepstrum: **MFCC** features – as homomorphic transform with equidistant band-pass-filters on the Mel-scale – tend to strongly depend on the spoken content. Yet, they have been proven beneficial in practically any speech processing task. They emphasise changes or periodicity in the spectrum, while being relatively robust against noise (set II, III).

voice quality: jitter/shimmer and other measures of microprosody, NHR/HNR and autocorrelation. They are based in part on pitch and intensity but reflect voice quality such as breathiness or harshness (set I, II, III).

wavelets: A wavelet packet decomposition is applied to divide the speech signal into 24 frequency bands. For every sub-band the log of the average Teager energy is found for a frame length of 8 ms and inverse DCT transformation is applied to obtain 10 cepstrum coefficients. For the most part the procedure is similar to the extraction of MFCCs. However, it is based on multi-resolution analysis to provide the application of a Teager energy operator in order to reflect nonlinear vortex-flow interactions, which has been found useful to classify stressed vs. neutral speech (set III).

bag of words (BOW): well known from document retrieval tasks, showing good results for emotion recognition as well, cf. Batliner *et al.* (2006). Each term within a vocabulary is represented by an individual feature modelling the term's (logarithmic and normalised) frequency within the current phrase. Terms are clustered with Iterated Lovins Stemming, cf. Lovins (1968) (set III).

part of speech (POS): this is a coarse taxonomy of six lexical and morphological main word classes based on the spoken word chain; given are frequencies in the chunk, raw or normalised (set I, II, III).

higher semantics (SEM): this is a coarse taxonomy into six classes, (partly

scenario-specific) most relevant words, word classes, and emotional valence (negative vs. positive), based on the spoken word chain; given are frequencies in the chunk, raw or normalised (set I).

varia: disfluencies/non-verbals such as breathing or laughter (set II, III).

The following types of functionals have been employed:

sequential and combinatorial: functionals of any type under the premise that a minimum of two functionals has been applied in either a sequential way (e. g. mean of max) or combinatorial way (e. g. ratio of mean of two different LLDs).

extremes: min/max by value, min/max position, range, and slope min/max, as well as on-/off-position.

means: first moment by arithmetic mean and centroid.

percentiles: quartiles 1/2/3, quartile ranges lower/upper/total and other percentiles.

higher statistical moments: standard deviance, variance, skewness, kurtosis, length, and zero-crossing-rate.

specific functions (distributional, spectral, regressional): a blend of several more 'unusual' functionals: several complex statistical functionals, micro variation, number of segments/intervals/reversal points, ratio, error, linear/quadratic regression coefficients, and DCT coefficients 1-5.

not attributable: features cannot be attributed un-equivocally to one of the other types.

It is well known that acoustic parameters like energy and pitch excursions can be relevant for emotion modelling; the same holds for single words, which are modelled in the BOW approach. Interestingly, even the very coarse syntactic/morphological POS taxonomy already displays marked differences: Table 1 displays a cross-tabulation of our four emotion categories with POS classes illustrating the high impact of POS due to the un-balanced distribution: more adjectives (API, APN) for ***M****otherese* (example: *good boy*), more verbs (VERB) for ***E****mphatic* (example: *stop*), and more nouns (NOUN) and less particles (PAJ) for ***A****ngry* (example: vocative *Aibo*).

POS	# of tokens	NOUN	API	APN	VERB	AUX	PAJ
M*otherese*	1240	10.6	5.0	25.2	16.0	1.5	41.8
N*eutral*	7750	15.5	0.7	2.2	24.4	1.3	55.9
E*mphatic*	1889	10.9	0.3	0.4	40.6	0.1	47.8
A*ngry*	1588	45.7	1.8	0.9	23.6	0.6	27.4

Table 1
Cross-tabulation of emotion labels (**MNEA**) and POS labels; displayed is frequency in percent per emotion label

Our feature vector combines different state-of-the-art feature types. There is no explicit dynamic modelling, though, for instance with Hidden Markov Models (HMM), Dynamic Bayesian Networks (DBN) or Dynamic Time Warp (DTW), which has been used as well for emotion classification, cf. Nose *et al.* (2007); Wagner *et al.* (2007); Inanoglu and Caneel (2005); Schuller *et al.* (2003); Nwe *et al.* (2003); Nogueiras *et al.* (2001). Note, however, that dynamic information over time is represented by 'static' features as well: minimum, maximum, onset, offset, position of extreme values, regression with mean square error, etc. can represent a full pitch or energy contour, albeit somehow 'quantified'. Thus it is rather the manner of modelling and not the representation of pertinent information per se that is different between dynamic and static modelling. So far, however, dynamic modelling has only be carried out for frame-level features. As far as we can see, there is no evidence that it yields better performance than static features representing dynamic information, cf. Wagner *et al.* (2007); Schuller *et al.* (2003); Vlasenko *et al.* (2007): this most likely stems from the fact that in frame-level features, the phonetic content is over-modelled. To ascertain this assumption, more in-depth studies are needed. Apart from dynamic modelling, some other feature types are not included which could be used, such as TRAPs, cf. Hermansky and Sharma (1998).

An increase in accuracy has been reported for the combination of different time levels, combining frame- and turn-level features, cf. Vlasenko *et al.* (2007). However, this effect was especially true for data with pre-defined spoken content, again demonstrating the dependency on phonetic structure for small time-units. Other work deals with different time levels that would allow for dynamic modelling, cf. Schuller *et al.* (2008). However, no results have been reported in this respect, yet. In our **set I**, we use features combining word- and chunk-level information.

5 Feature Relevance Analysis

5.1 Procedures

In the last decade, almost all standard classifiers have been employed for automatic emotion classification. Following this practice, in Batliner *et al.* (2006), for *FAU Aibo* we have used Neural Networks, Support Vector Machines SVMs, Random Forests RF, Linear Regression, Linear Discriminant Classification, Naïve Bayes, and Rule-based classifiers. A more rigorous comparison of SVMs and RFs, cf. in Schuller *et al.* (2007a); Batliner *et al.* (2008a), confirmed that both have approximately the same performance. Although not necessarily the best classifiers for every constellation, cf. Meyer *et al.* (2002), SVMs provide

very good generalisation properties, cf. McGilloway *et al.* (2000); Lee *et al.* (2002); Chuang and Wu (2004); Devillers *et al.* (2005); You *et al.* (2006); Morrison *et al.* (2007); Hu *et al.* (2007). Since the topic of this study is not a comparison of different classifiers, we decided to use only SVMs. Note that a thorough parallel use of several classifiers for selection and classification of features would have increased the computational effort by an order of magnitude (more than a 100 k additional runs per classifier to avoid bias between selection and classification, cf. below). Feature evaluation is usually done following two possible strategies: the closed-loop "wrapper" method, which trains and re-evaluates a given classifier at each search step using its accuracy as objective function, and the open-loop "filter" method, which maximises simpler objective functions. While a wrapper can consist of any classifier, filter objective functions are usually measures such as inter-feature and feature-class correlation, cf. Hall (1998). As an exhaustive search through all possible feature combinations is unfeasible considering typical database and feature space sizes, faster but sub-optimal search functions are usually chosen. Typical conservative hill-climbing procedures are sequential search methods as Sequential Forward Selection (SFS): at each step the feature reporting the best wrapper or filter accuracy is chosen. SFS has been commonly used for emotion recognition, cf. Lee *et al.* (2001); Lee and Narayanan (2005); Kwon *et al.* (2003). Sequential floating forward selection SFFS, cf. Pudil *et al.* (1994); Jain and Zongker (1997),is an improved SFS method in the sense that at each step, previously selected features are checked and can be discarded from the optimal group to overcome nesting effects. Experiments show SFFS to be superior to other methods, cf. Jain and Zongker (1997). Note that a good feature selection should de-correlate the feature space to optimize a set of features as opposed to sheer ranking of features. This is particularly the case for wrapper-search, which also usually demands considerably higher computational effort. Some studies combine feature selection with feature generation to find better representations and combinations of features by simple mathematical operations such as addition, multiplication or reciprocal value of features, cf. Batliner *et al.* (2006).

Determining the most relevant features may lead to overall higher accuracy, especially for our conglomerate of many and partly highly correlated features – classifiers are susceptible to dimensionality. In addition to accuracy, a low-dimensional set obviously saves computational load at several stages such as extraction, training, and classification. Furthermore, feature selection results in a reduced, interpretable set of significant features; their counts and weights in the selection set allow to draw conclusions on the relevance of the feature types they belong to. In this article, we focus on interpretation and not on optimising classification performance. We search the most important features in the original untransformed feature space directly by using SFFS for three groups: acoustics only, linguistics only, and both acoustics and linguistics together. This resembles the more usual approach to feature selection in the

field of emotion recognition, cf. Vogt and André (2005). Yet it omits a first de-correlation step, which often clusters features of the same LLD or functional. This may distort the outcome if the feature space reduction is very large. Nevertheless, this approach is necessary, because de-correlating preprocessing, such as PCA or LDA, would project the original feature space into a new one, thus making it even more difficult to interpret the impact of different feature types. To freeze the number throughout feature sets and cross validation splits, we selected the 50 best features per split; using a cut-off criterion like Receiver Operating Characteristic (ROC) curves would inject variance on several layers because they are often overlaid with statistical noise or show very flat slopes. The wrapper classifier for SFFS is Support Vector Machines, known as SVM-SFFS, with linear kernel, one-against-one multi-class discrimination, and Sequential Minimal Optimisation (SMO), cf. Witten and Frank (2005). The wrapper runs in turn on three speaker-independent cross-validation splits; two of them are used for training/validation (two-fold cross-validation), the third for latter performance analysis. The final space of selected features is the union (thus allowing for potential repetitions of features which are discarded subsequently) obtained by applying SFFS on the original (untransformed) space, on each of these three splits separately. This approach leads to a reduced set of 150 features (50 per split). The data partitioning meets the following requirements (in order of priority): no splitting of within-subject chunks, similar distribution of labels, balance between the two schools, and balance between genders. For the training splits, we up-sampled all classes but ***N****eutral*: 3x ***M****otherese*, 2x ***E****mphatic*, and 2x ***A****ngry* [16] . SFFS is thereby carried out in 2-fold, cross-fold selection on two of the three partitions having one split for training and one for validation. The third split is held out completely and is used for eventually testing each feature type separately.

For interpretation, we refrain from discussing single 'most important' features because of both technical reasons (i. e. parameterisation of the SFFS selection algorithm, local minima, etc.) and the very high dimensionality of the feature space; at this stage, it seems to be safer to discuss distribution of feature types, their classification performance, and their proportion in relation to all types and their frequencies. To this aim we report an F-MEASURE as introduced in Batliner *et al.* (2006) which is used for having a unique classification performance measure; here, the F-MEASURE is defined as the uniformly weighted harmonic mean of RR and CL: $2 \cdot CL \cdot RR/(CL + RR)$. RR is the overall recognition rate or recall (number of correctly classified cases divided by total number of cases or weighted average); CL is the 'class-wise' computed recognition rate, i. e. the mean along the diagonal of the confusion matrix in percent, or unweighted average. This F-MEASURE represents a trade-off between CL and RR. This is a slightly different definition from the standard one, given

[16] Unlike random up-sampling, such a rule-based strategy guarantees easy and unequivocal reproducibility in case of latter parallel experiments.

in Makhoul *et al.* (1999); however, it seems to be more adequate for a multi-classification problem. Note that F-MEASURE values displayed henceforth are calculated on the reduced feature sets separately for each feature type (i. e. using its surviving features only), and extraction type.

Another rough but useful indicator of the relevance of a feature type is the number (#) of the features selected by SFFS. As the reduced set is fixed to 150 we can further refine the count by normalising it: with SHARE we define the number of each feature type normalised by 150: #/150; with PORTION we also introduce the same number normalised by the cardinality of a feature type in the original feature set: (#/#*total*). SHARE displays for each feature type its percentage in modelling our 4-class problem, summing up to 100% modulo rounding errors, across the three splits. It shows the contribution of single types under the (strictly speaking, contrafactual) assumption that the types are modelled exhaustively and 'fair' across types;[17] PORTION shows the contribution of single types weighted by the (reciprocal) number of tokens per type: the higher it is, the better this type can be exploited. To give examples for both SHARE and PORTION: in Tab. 2, SHARE for duration is the number of duration features surviving the feature selection (28) divided by the total number of features in the reduced set (150): (28/150) = 18.7; PORTION for duration is the number of duration features surviving the feature selection (28) divided by the total number of duration features (391): (28/391) = 7.2.

[17] This caveat refers on the one hand to the points addressed in Sec. 1, esp. to the fact that even our large set with its more than 4 k features is of course not exhaustive; on the other hand, set I with its two-layered approach has some advantage over the single-layered sets II and III because the acoustic features of set I are computed based on manually corrected word boundaries. Thus the sets spectrum, cepstrum, and wavelets which are all not computed for set I, cf. Sec. 4.1.1, might be slightly 'handicapped'. We do believe, however, that this fact has only a small impact in the two-layered approach because computation on the first (word-) level is not used directly but serves only as input to the computation on the second (chunk-) level.

		duration	energy	pitch	spectrum	cepstral	voice quality	wavelets	all
	# total	391	265	333	656	1699	153	216	3713
SFFS	#	28	33	23	17	23	11	15	150
	F-MEASURE	54.9	56.9	46.7	49.9	50.4	41.5	44.9	63.4
	SHARE	18.7	22.0	15.3	11.3	15.3	7.3	10.0	100.0
	PORTION	7.2	12.5	6.9	2.6	1.4	7.2	6.9	4.0

BOW	POS	SEM	VAR	all
476	31	12	12	531
94	27	27	2	150
53.2	54.9	57.9	–	62.6
62.7	18.0	18.0	0.1	100.0
19.7	87.1	225.0	16.7	28.2

Table 2
Summary of feature selection results exploiting acoustics only **(a)**, left, and linguistics only LLDs **(b)**, right. Last columns 'all' report results using all feature types together. # lines are counts; F-MEASURE, SHARE, and PORTION are %. 'VAR' is 'disfluencies' plus 'non-verbal'. Percent values can be $> 100\%$ for SHARE and PORTION because the same features can be used in any of the three cross-validations. Further explanation is given in text.

		duration	energy	pitch	spectrum	cepstral	voice quality	wavelets	BOW	POS	SEM	VAR	all
# total		391	265	333	656	1699	153	216	476	31	12	12	4244
SFFS	#	10	32	16	15	16	7	5	25	7	17	0	150
	F-MEASURE	49.6	56.3	46.8	46.2	46.4	38.7	35.3	37.4	48.1	56.0	–	65.5
	SHARE	6.7	21.3	10.7	10.0	10.7	4.7	3.4	16.7	4.7	11.3	0.0	100.0
	PORTION	2.6	12.1	4.8	2.3	1.0	4.6	2.3	5.3	22.6	141.7	0.0	3.5

Table 3
Summary of feature selection results exploiting acoustic plus linguistic LLDs, in the same selection pass. Last columns 'all' reports the result using all feature types together. # lines are counts; F-MEASURE, SHARE, and PORTION are %. 'VAR' is 'disfluencies' plus 'non-verbal'. Percent values can be $> 100\%$ for SHARE and PORTION because the same features can be used in any of the three cross-validations. Further explanation is given in text.

		duration	energy	pitch	spectrum	cepstral	voice quality	wavelets	BOW	POS	SEM	VAR
	# total	391	265	333	656	1699	153	216	476	31	12	12
SFFS	Δ#	18	1	7	2	7	4	10	69	24	10	12
	ΔF-MEASURE	5.3	0.6	-0.1	3.7	4.0	2.8	9.9	15.8	6.8	1.9	–
	ΔSHARE	12.0	0.7	4.6	1.3	4.6	2.6	6.6	46.0	13.3	6.7	–
	ΔPORTION	4.6	0.4	2.1	0.3	0.4	2.6	4.6	4.4	64.5	83.3	16.7

Table 4
Absolute differences per feature type between 'acoustics only' and 'linguistics only', and 'acoustics plus linguistics'. # lines are counts; F-MEASURE, SHARE, and PORTION are %. 'VAR' is 'disfluencies' plus 'non-verbal'. Further explanation is given in text.

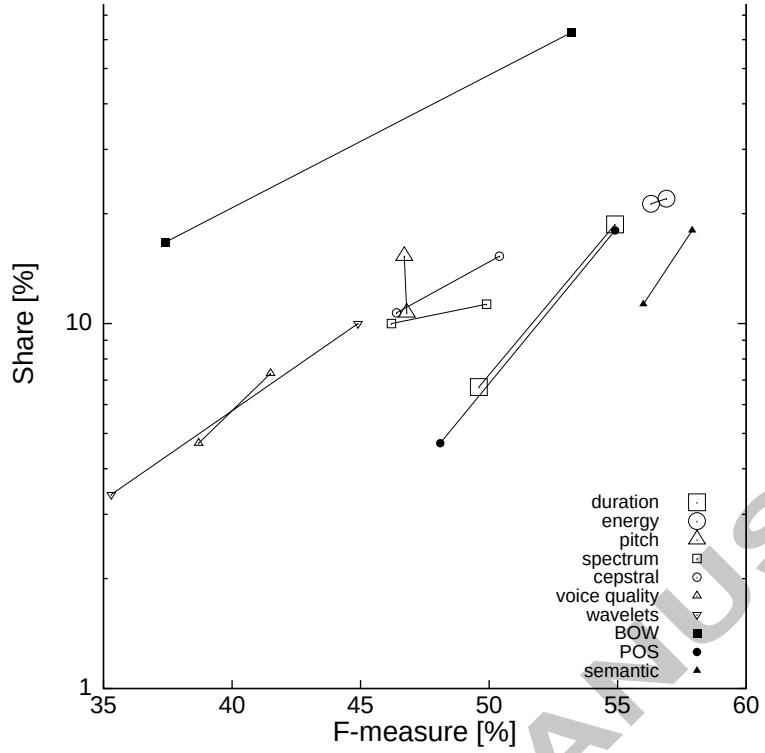


Fig. 1. Feature types: different impact for separate or combined modelling of acoustics and linguistics; for each type, F-MEASURE and SHARE are plotted for a separate modelling of acoustics and linguistics, and for a combined modelling; the two points are connected with a straight line to indicate the distance; the y-axis is scaled logarithmically

5.2 Types vs. Functionals

From Tables 2 and 3, we try to gain more insight into the effectiveness of each feature type and the differences between types in modelling our 4-class problem. As introduced in the last section, we compare types by looking at F-MEASURE, SHARE, and PORTION.

Most of the time, all types contribute (SHARE), although some of them might be more important than others, e. g. for acoustics only (Table 2, left) energy, duration, pitch, and MFCCs. The picture is a bit different if we look at the F-MEASURE: highest ranks energy, then duration, MFCCs, and spectrum. Note that 'important' here does not imply a general statement on the LLDs concerning classification performance; it means that obviously, the surviving features belonging to this specific LLD are compact and good at exploiting relevant information. This can be illustrated with the linguistic types: for linguistics only (Table 2, right), BOW has the highest SHARE value; as for PORTION,

SEM and, to a lesser extent, POS features obviously model (positive or negative) valence and syntactic/semantic salience to a high extent. This might be the reason for the difference in F-MEASURE: SEM with highest F-MEASURE, then POS, and third position BOW. POS features might be most robust if it comes to dealing with ASR output, cf. Seppi *et al.* (2008): the difference between content words and function words is highly correlated with length of words, i. e. duration, cf. below. Thus word recognition can be suboptimal, and only word length has to be approximated. Moreover, in Schuller *et al.* (2007a) an SVM classification yielded for all BOW features an F-MEASURE of 62.6%, compared to the 53.2% shown in Table 2, right; thus, BOW is better at performing if all terms in the utterance can be considered while SEM is better at 'compactness': when you are interested in performance, take the brute force set BOW; when you are interested in a generic interpretation, have a look at the selective, compact SEM set.

Actually, the behaviour of the different types is not much different if used in acoustics or linguistics alone, or in the full set, if we look at relevance within the 'cover type' – acoustics or linguistics – the types belong to. If we look at the F-MEASURE values in Table 2 and 3, energy and SEM are most important within acoustics and linguistics respectively, followed by duration and part-of-speech features (POS). The SHARE of BOW features is high; this indicates that they are providing some useful, detailed information in addition without being as compact as the other two linguistic types. This could be expected: SEM and POS features are condensed cover classes of BOW features. Pitch, spectrum, and MFCCs come second in acoustics (Table 2, left), with somehow equal importance. Least important – but still providing some information – are voice quality and wavelets. We have stressed in the introduction that these results have to be seen in the context of the emotion classes that we can model, based on our data: the high impact of duration and energy might be partly due to ***E****mphatic* being one of our four classes. The necessity to map *reprimanding*, *touchy*, and *angry* onto a negative cover class could be responsible for the low impact of voice quality features which might be better at modelling particular differences. However, there is evidence that they are multi-functional and more susceptible to speaker-idiosyncrasies and therefore maybe of higher impact only in a personalised setting, cf. Batliner *et al.* (2007a). [18]

[18] Scherer (1986) stresses the high impact of voice quality. Evidence for such an impact comes very often from experiments using acted data or synthesised speech and perceptive evaluation, cf. Gobl and Ní Chasaide (2003). If subjects have to listen to stimuli where just one parameter is manipulated such as voice quality, they certainly attribute some function to such differences. This does not tell us how this parameter behaves in a natural setting and/or for speaker-independent evaluation. Moreover, the estimation of the vocal source is a major problem. Thus the low rank for voice quality can be due to the fact that it is a highly speaker-dependent feature, and/or to the fact that it cannot be extracted reliably.

In acoustics plus linguistics (Table 3), SHARE for the acoustic features should be roughly one third lower than in acoustics alone, because one third (49) of the 150 surviving features are now linguistic ones. Most pronounced is a deviation for duration (lower, i. e. only one third instead of two-thirds) and for energy (higher, i. e. three-thirds). To assess the mutual influence of acoustics and linguistics in a more objective way, we display in Table 4 the absolute differences between the values of #, F-MEASURE, SHARE, and PORTION in Table 2 and 3: we simply subtract the values from Table 3 from those of Table 2. All differences but one (-0.1 for pitch) are positive; this could be expected because adding another important knowledge source, namely linguistic information, lowers the impact of acoustics. It is interesting, however, that there are practically no exceptions. Fig. 1 display this difference in a plot of SHARE vs. F-MEASURE; for each of the types, two positions are given, one for the separate modelling of acoustics and linguistics, one for the combined modelling. Apart from pitch, all types display the same behaviour: they loose some impact – i.e. they are closer to the origin because F-MEASURE and SHARE values are lower – if in combination. The highest 'loss' can be observed for the two linguistic types BOW and POS, and for duration and wavelets; this is indicated by the longer lines connecting the respective two points. Between duration and linguistic information, there is a 'trading' (i. e., a complementary) relation because duration implicitly encodes linguistic information: longer words are content words, and shorter ones are function words; short utterances tend more towards denoting ***E****mphatic* or ***A****ngry*, cf. below. It might be that adding linguistic information makes energy more salient, especially for our classes which entail ***E****mphatic*. Obviously, wavelets and MFCCs model as well linguistics to a higher extent than spectrum and esp. pitch and energy. In Schuller *et al.* (2007a), the feature types were used alone and not in combination with all other feature types; classification performance using SVM and individual Information Gain Ratio for selecting the 150 most important features for each type was between 51.6% for voice quality and 60.6% for duration. The ranking from highest to lowest was: duration, energy, cepstral, wavelets, spectrum, pitch, and voice quality. Thus wavelets are 'in the middle of the field' if used alone, but lag behind if used together with other features.

As for the linguistic types, BOW is most affected, second come POS and SEM. The very low impact of VAR, i. e. disfluencies and non-verbals, might be due to the simple fact that they do not occur that often in this scenario: mostly waiting for the AIBO's actions, the children do have enough time to plan their rather short utterances. Thus planning does not interfere with production.

Acoustics only and Linguistics only yield roughly the same performance for 'all': 63.4% vs. 62.6% in Table 2. If taken together, cf. Table 3, more acoustic features survive. There might be at least two reasons for that: first, there

label	**M**	**N**	**E**	**A**
acoustics only				
Motherese	**64.0**	25.6	2.7	7.7
Neutral	11.3	**66.9**	12.3	9.6
Emphatic	2.6	15.4	**62.0**	20.0
Angry	4.5	16.5	19.9	**59.1**
linguistics only				
Motherese	**69.8**	19.3	4.1	6.8
Neutral	11.0	**62.5**	17.6	8.9
Emphatic	0.9	16.6	**70.1**	12.4
Angry	3.1	20.6	28.0	**48.4**
acoustics and linguistics				
Motherese	**69.3**	22.4	2.2	6.1
Neutral	10.5	**67.1**	15.1	7.4
Emphatic	0.6	15.9	**66.9**	16.7
Angry	2.3	17.6	22.0	**58.1**

Table 5
Confusion matrix for our 4 classes in percent correctly classified, for the three different feature type combinations acoustics only (above) linguistics only (middle), and acoustics and linguistics (below). Correct recall values in the diagonal are given in bold.

are simply many more acoustic features than linguistic features to choose amongst. The encoding of linguistic information in acoustic features such as duration plus some added value based on acoustics alone might be a second reason. The outcome that acoustics and linguistics alone yield roughly the same performance might of course depend on the type of data: trivially, in written data, there is no acoustics. In very restricted codes (e. g., commands only) acoustics is the channel that can be altered most. In acted data, if segmental structure is identical, the impact of linguistics is null. We believe that the equilibrium found in our data is characteristic for 'normal', realistic data, where both acoustic and linguistic means are available for the speakers; a similar equilibrium was obtained for realistic call-center data and for lexical vs. prosodic cues in Devillers and Vidrascu (2004). It is not a full equilibrium, though, because the scenario 'giving commands to pet robots' is certainly more restricted than, e. g. free human-human conversations. (Note that the impact of linguistics will be certainly lower if real ASR with noisy data is applied.)

With Table 4 and Fig. 1, we discussed the interrelation of acoustics with lin-

guistics by looking at the different feature types. With Table 5, we discuss it by looking at the differences of recalls (accuracy) per class. First we see that the values confirm the NMDS scaling found in Batliner *et al.* (2008b), cf. section 2.3, and by that, corroborate the order **M**, **N**, **E**, **A**: with a few and small exceptions, esp. for **M** vs. **E** and **A**, values are falling monotonously, starting from the correct recall. Second, **M** and **E** are classified better with acoustics only, **N** and **A** with linguistics only. Recall for **A** is lowest – the reason might be that it is composed of different categories with different arousal levels: high for *angry*, lower for *reprimanding* and *touchy*.

It should be born in mind at this point that an up-sampling was carried out to cope with the high imbalance of classes. Clearly, by different balancing different behaviours will be observed with respect to the highest recall. However, the chosen up-sampling leads as close as possible to a uniform distribution by a straightforward and easily re-doable rule-based processing. By up-sampling, recall rates are more balanced for all four classes.

	sequential & combinatorial	extremes	mean	percentiles	higher statistics	specific functions & regression	not attributable	all
# total	218	1132	427	1196	547	153	40	3713
SFFS #	50	30	18	15	21	4	12	150
SFFS F-MEASURE	58.5	49.9	56.1	43.1	54.7	35.5	41.1	63.4
SFFS SHARE	33.3	20.0	12.0	10.0	14.0	2.7	8.0	100
PORTION	22.9	2.7	4.2	1.3	3.8	2.6	30.0	4.0

Table 6
Summary of feature selection results, for acoustic features only, breakdown into functional types. # lines are counts; F-MEASURE, SHARE, and PORTION are %. Further explanation is given in the text.

Table 6 displays in a parallel way F-MEASURE, SHARE, and PORTION, broken down into functional types, for acoustic features only. We see that highest comes 'sequential and combinatorial', second 'extremes', 'means', and 'higher statistics'. This is no proof yet that 'sequential and combinatorial' features are best because they are based on many other – but not on all – types of functionals, cf. the description given above; they are therefore sort of unrivalled. 'Specific functions & regression' as well as 'percentiles' rank low, compared to 'extremes', 'means', and 'higher statistics'; it might be that the latter types are either more robust such as 'means', or more 'characteristic' such as 'extremes': extreme values are more prone to extraction errors than mean values but if they are correct they can be good predictors, such as higher F0 excursion indicating higher arousal. For assessing the real impact of 'sequential and combinatorial' features, we have to use another strategy which is described in the next paragraph.

To find out up to what extent the three different sets (*types of feature extraction*) described in Sec. 4.1 result in different impact, we computed for them the same figures as in Table 2 and 3; results are given in Table 7. This was done only for the acoustic features because the linguistic ones show considerably lower variability. Further they are all based on information on the spoken word chain. We can see that the F-MEASURE is highest for the two-layered set I, and lower for the brute-force set III; for set II (one-layered and selective) it is a bit lower than for set III (brute-force). As for SHARE, set I (selective, two-layered) and set III (brute-force, single layered) change places. PORTION tells us that set I is exploited much better than set II; as expected, PORTION for set III is lowest. Due to the differences in processing described in Sec. 4.1, this is no exact benchmark: we used the features from each site as is and did not try to optimize performance independently. In set II, spectral and MFCC features are used but not in set I. On the other hand, set I used manually corrected word information. However, in Steidl *et al.* (2008) it is shown for the same database that there is no difference between manually corrected and automatically extracted word boundaries, as far as classification performance is concerned. It seems rather to be information based on the spoken word chain (word length, how many words per chunk, etc.) and the mere fact that sub-chunk units are taken into account, which somehow is modelled in these two-layered acoustic features as well. Thus most likely, a two-layered processing has some added value, and within the brute force approaches, the single features might not model 'too much', i. e. they might be only relevant within a whole network of many features. A combination of a 'many feature' approach, be this selective, i.e. knowledge-based, or brute-force features, with a two-layered processing, cf. Schuller *et al.* (2008), might be promising.

We should stress that comparing values of our measures SHARE and PORTION is not fully 'fair', for different reasons: first, some features are not extracted (fully) automatically: this holds for all linguistic features, and for the

two-layered approach of FAU and FBK which uses manually corrected word boundaries; note, however, that in Schuller *et al.* (2007b) we have shown that automatic segmentation based on ASR output does not deteriorate sharply classification performance for our 4-class problem. Second, our measures are 'unfair' against very small or very large feature sets: if all 31 POS features were entailed in all three SFFS sets, this highest possible SHARE would amount to $31 \cdot 3/150 = 62\%$; more is not possible, even in theory. On the other hand, the MFCC features could never have a PORTION higher than some 8.8% because of their overall frequency. However, if all (types of) features were of exactly equal importance, so that choosing amongst them would simply be by chance, then the proportion of SHARE and PORTION for extraction and feature types would tend to be roughly equivalent to the frequency of tokens in the respective set. This is not the case. Thus we are allowed to formulate some interesting hypotheses: first, that a combination of lower and higher units (two-layered word and chunk processing) is promising – even for non-'cheating' ASR processing, cf. Schuller *et al.* (2007b). Second, that features computed within brute force approaches may really need larger 'feature networks' than selective features. This is a daring hypothesis, but an interesting one which can be tested. Third, that all types of features contribute, albeit up to a different extent, cf. the differences in performance of pure types described in Schuller *et al.* (2007a) – but we maybe only need all of them if we are interested in utmost performance, or in a generic feature set. Fourth, that acoustics on its own definitely does not tell the whole story; note that our labellers annotated sequentially, taking into account all context given – after all, this is a realistic modelling of the interaction: speech is not only acoustics but linguistics as well.

In Schuller *et al.* (2007a), we pursued a similar, but not identical approach towards the relevance of feature types: using the same taxonomy as described in Sec. 4.1, we processed separately each type using the 150 features per each type with the highest individual Information Gain Ratio; F-MEASURE was always computed for these 150 'most relevant' features per set. Thus in Schuller *et al.* (2007a), we addressed the question 'how good is this feature type alone?'; in the present article, we addressed the (more realistic) question 'how much does this specific feature type contribute if used together with all other feature types?'. The order of relevance (classification performance) for the different types is very much alike in these two different approaches; this is a reassuring result because differences in feature *selection strategy* seem not to be that important as differences in *phonetic or linguistic content* of the feature types.

It could be argued that we should apply tests of significance on our results to find out which of them are significantly different and which are not. To correctly apply inferential statistics, we should, however, deal with the multiplicity effect, i. e. the repeated use of the same data, through significance testing using, e. g. the Bonferroni adjustment: the errors made by our computations are not independent; chunks can be from the same turn, and the data are the

same throughout, cf. Salzberg (1997) – note that we certainly have done more than 100 k classification tests with these data throughout the years. Thus the Bonferroni adjustment would simply invalidate any result. (There are some theoretical/methodological arguments against the Bonferroni adjustment, cf. Pernegger (1998).) A common use of tests of significance in speech processing – albeit, strictly speaking, an incorrect one – is to disregard the multiplicity effect and to apply tests of significance to get an indication whether the results are marked enough, given the difference in performance and the size of the two samples: the smaller the samples, the higher the difference in performance has to be to yield 'significant' results. (Already Eysenck (1960); Rozeboom (1960) suggested to use significance not in the inferential meaning but as a sort of descriptive device.) As described previously, the performance of a feature set was measured as a proportion of correctly classified instances over a population. Thus we test the significance of the differences of RRs. If we assume RR1 and RR2 as independent, then we can adopt the Z-test for a proportion, namely Δ RR = RR1 - RR2. This test allows to investigate the significance of Δ RR compared to the standard normal distribution, in specific using a two-tailed test. In Fig. 2 we draw the significance threshold for Δ RR given two typical values of α: 0.01 and 0.05.[19] To give two examples: F-MEASURE for duration exploiting acoustics only in Tab. 2 is 54.9; F-MEASURE for duration exploiting acoustics plus linguistics in Tab. 3 is 49.6. The difference amounts to 5.3 and is above the value needed for 'significance' at the 5% level which is 2.0, and above the value needed for 'significance' at the 1% level which is 2.7, cf. Fig. 2. For SEM in the analogous constellation, the difference of $(57.9 - 56.0) = 1.9$ is slightly below the values needed for significance at both levels (2.0 or 2.7).

A last note on 'perceptual and cognitive adequacy': can we infer from our hierarchy of relevance to human processing? This question is normally being addressed by perception experiments – a task impossible to accomplish with that many features. Thus the argumentation can only be indirect: if we can model the behaviour of our labellers with our selection of features, chances are that perception and classification do have something in common. Even for such a rather general statement, one caveat has to be made: our features are, up to a great extent, extracted automatically. This means in turn that they all are, again up to a certain extent, erroneous. We have seen in Batliner *et al.* (2007b) that for automatically extracted pitch features, mean values are most important; for manually corrected pitch features, however, it is features

[19] The assumption that RR1 and RR2 are independent is of course not completely true, because the figures were computed on the same test data: although the features from the two algorithms are different, both methods exploit, for instance, the same classifier and they can therefore share a number of errors. However, a more precise significance test (e. g. McNemar's test) would simply be less conservative, yielding more 'significant' differences. Therefore, the values reported in Fig. 2 can be seen as a lower bound of the significance values.

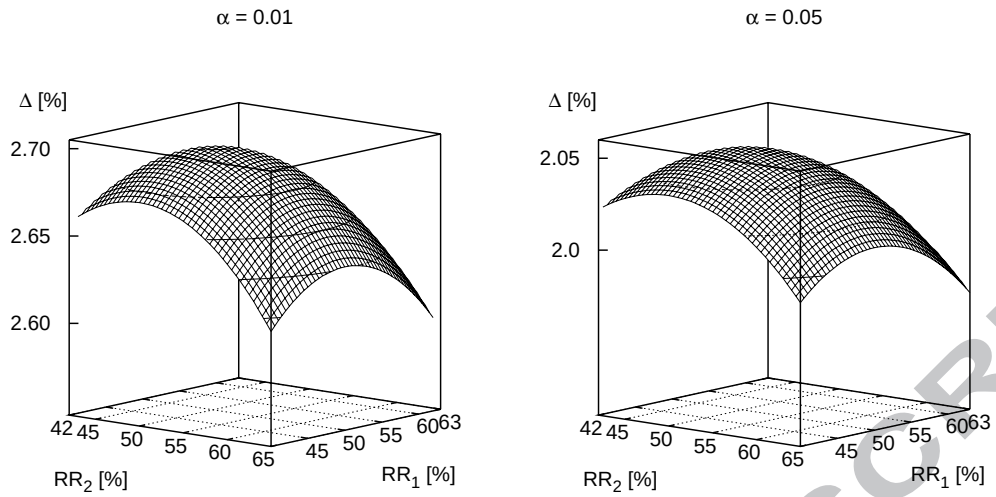


Fig. 2. Significance threshold for Δ RR given two typical values of α: 0.01 (left) and 0.05 (right); two-tailed test

		set I	set II	set III	all
	total #	112	297	3304	3713
SFFS	#	54	32	64	150
	F-measure	58.8	53.3	54.9	63.4
	share	36.0	21.3	42.7	100.0
	portion	48.2	10.8	1.9	4.0

Table 7
Summary per type of feature extraction (acoustic features only): set I (selective, two-layered), set II (selective, single layered), set III (brute force, single-layered), and all. Explanation of measures is given in the text.

modelling the shape of the pitch contour. This can be explained easily because mean values are less prone to octave errors than shape values. But we still do not know whether human perception (always) fine-tunes to shapes or uses less 'precise' but more robust features such as mean values as well.

6 Concluding Remarks

To our knowledge, the present article describes the most extensive approach so far towards examining the impact of different types of features on the performance of automatic recognition of emotions/emotion-related user states within the two sign systems acoustics and linguistics, both separately and in combination. We have shown that the performance of all acoustic and of all linguistic features is comparable, and that a combination of both sets still improves recognition performance. The ranking within the acoustic feature sets corresponds to most of the studies using realistic data and a fair coverage of feature types conducted so far; a strict comparison is not easy because of several trading relations within and across acoustic and linguistic feature types; however, a separate evaluation of the sets in Schuller *et al.* (2007a) yielded a comparable ranking.

The size of our feature vector was a necessary pre-requisite for obtaining a high coverage of different feature types and functionals but it makes it way more difficult to interpret single features. We therefore introduced the new and more global measures SHARE and PORTION, besides F-MEASURE. Another necessary pre-requisite for the co-operation between different sites was an un-equivocal, agreed upon feature encoding scheme.

It is clear that even if our scenario is realistic and as such, representative, the specific emotion/emotion-related classes we have found are not the only ones; thus the ranking of relevance for the different feature types might not be representative. Note that we always have to speak about a ranking, not about any irrelevance of single feature types: every type taken alone – either within the approach chosen for the present study, or within the one chosen in Schuller *et al.* (2007a) – yields a classification performance above chance level. The less feature values are a direct result of physiological conditions (for instance, high arousal producing large pitch ranges for *angry* or *joyful*), and the more they are conventionalised and individualised, the less plausible might a 'universal' ranking be.

The caveat has to be made that, strictly speaking, our statements on the relative importance of feature types have to be confined to our data. This caveat holds, of course, for every study. It seems to be stricter in our case due to the lesser amount of prototypical representation of emotional user states: we do not model the 'big n' emotions but a mixture of big emotions/emotion-related states (interpersonal stances in Scherer's terminology, cf. Scherer (2003)), and 'semantic-related' phenomena such as emphasis. Such a strict caveat holds, however, only against the background of an ideal world where emotions could be investigated 'in nuce'. The classes we have found might be, on the contrary, rather representative for realistic scenarios and applications – 'repre-

sentative' because they do not model some 'big n' emotions but a selection out of the much larger set of user states that can be found in realistic application scenarios. Note that most studies giving a survey of features most relevant for specific emotions such as, e. g. Frick (1985); Banse and Scherer (1996); Cowie *et al.* (2001), are based mostly on acted data – simply because at that time, most of the studies on acoustics and emotion were based on acted data. However, acted or synthesised data can be used as sort of 'heuristic inspiration' but must not simply be transferred generically onto realistic data. As Erickson *et al.* (2004), p. 20 put it: "In acted emotion, the speaker is volitionally changing the acoustic signal to impart to the listener a mental or emotional state (paralanguage) while in spontaneous emotion the speaker is working at maintaining the acoustic signal to convey the intended message even through emotional interruptions (nonlanguage)." It simply cannot be granted that conveying emotions by acting yields the very same acoustics and prosody as conveying a message, while being emotional.

In everyday, neutral speech, expressing salience with emphasis needs not co-occur with expression of emotions. But if we express emotions, we most certainly express salience besides because this in a genuine trait of speech: expressing emotions within the paralinguistic system is **modulated onto** expressing semantics within the linguistic system; we do not use only an 'emotion system', forgetting about linguistics. This is only the case if we act emotions using always the same, segmentally identical utterances, and maybe in specific situations with a very high arousal level. Thus the sparsity of emphasis as category in emotion studies so far might partly be due to the fact that it simply has not been addressed (*you don't get what you're not looking for*). However, in realistic applications, we have to model both phenomena in order not to confound the two sign systems linguistics, on the one hand, with its means of accentuation and emphasising, and para-linguistics, with its means of expressing emotional states. It will be an empirical question whether for a specific application, an integrated approach with an acoustic, an emotion and a linguistic module interacting with each other, or a sequential approach, strictly separating these modules from each other, is more applicable. In practice, often only parts of a sequential approach will be implemented, due to complexity considerations.

References

Banse, R. and Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, **70**(3), 614–636.

Batliner, A., Kompe, R., Kießling, A., Mast, M., Niemann, H., and Nöth, E. (1998). M = Syntax + Prosody: A syntactic–prosodic labelling scheme for

large spontaneous speech databases. *Speech Communication*, **25**(4), 193–222.

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

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

Batliner, A., Fischer, K., Huber, R., Spilker, J., and Nöth, E. (2003a). How to find trouble in communication. *Speech Communication*, **40**, 117–143.

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

Batliner, A., Hacker, C., Steidl, S., Nöth, E., and Haas, J. (2004). From Emotion to Interaction: Lessons from Real Human-Machine-Dialogues. In *Affective Dialogue Systems, Proceedings of a Tutorial and Research Workshop*, pages 1–12, Kloster Irsee.

Batliner, A., Steidl, S., Hacker, C., Nöth, E., and Niemann, H. (2005). Tales of Tuning – Prototyping for Automatic Classification of Emotional User States. In *Proc. Interspeech*, pages 489–492, Lisbon.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Laskowski, K., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., and Aharonson, V. (2006). Combining Efforts for Improving Automatic Classification of Emotional User States. In *Proceedings of IS-LTC 2006*, pages 240–245, Ljubliana.

Batliner, A., Steidl, S., and Nöth, E. (2007a). Laryngealizations and Emotions: How Many Babushkas? In *Proceedings of the International Workshop on Paralinguistic Speech – between Models and Data (ParaLing'07)*, pages 17–22, Saarbrücken.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., and Aharonson, V. (2007b). The Impact of F0 Extraction Errors on the Classification of Prominence and Emotion. In *Proc. ICPhS*, pages 2201–2204, Saarbrücken.

Batliner, A., Schuller, B., Schaeffler, S., and Steidl, S. (2008a). Mothers, Adults, Children, Pets — Towards the Acoustics of Intimacy. In *Proc. ICASSP 2008*, pages 4497–4500, Las Vegas.

Batliner, A., Steidl, S., Hacker, C., and Nöth, E. (2008b). Private emotions vs. social interaction — a data-driven approach towards analysing emotions in speech. *User Modeling and User-Adapted Interaction*, **18**, 175–206.

Batliner, A., Seppi, D., Steidl, S., and Schuller, B. (2009). Segmenting into adequate units for automatic recognition of emotion-related episodes: a speech-based approach. *Advances in Human-Computer Interaction*. in press.

Blomberg, M. and Elenius, D. (2003). Collection and recognition of children's speech in the PF-Star project. In *Proc. of Fonetik 2003*, pages 81–84, Umeå,

Sweden.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., and Weiss, B. (2005). A Database of German Emotional Speech. In *Proc. Interspeech*, pages 1517–1520, Lisbon.

Chuang, Z.-J. and Wu, C.-H. (2004). Emotion recognition using acoustic features and textual content. In *Proc. ICME*, pages 53–56, Taipei.

Cowie, R. and Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication*, **40**(1-2), 5–32.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, **18**(1), 32–80.

Cowie, R., Sussman, N., and Ben-Ze'ev, A. (2010). Emotions: concepts and definitions. In P. Petta, editor, *HUMAINE handbook on emotion*. Springer. to appear.

Davies, M. and Fleiss, J. L. (1982). Measuring agreement for multinomial data. *Biometrics*, **38**, 1047–1051.

de Rosis, F., Batliner, A., Novielli, N., and Steidl, S. (2007). 'You are Sooo Cool, Valentina!' Recognizing Social Attitude in Speech-Based Dialogues with an ECA. In A. Paiva, R. Prada, and R. W. Picard, editors, *Affective Computing and Intelligent Interaction*, pages 179–190, Berlin-Heidelberg. Springer.

Dellaert, F., Polzin, T., and Waibel, A. (1996). Recognizing emotion in speech. In *Proc. ICSLP*, pages 1970–1973, Philadelphia.

Devillers, L. and Vidrascu, L. (2004). Reliability of Lexical and Prosodic Cues in two Real-life Spoken Dialog Corpora. In *LREC*, Lisbon.

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

Ekman, P. (1999). Basic emotions. In T. Dalgleish and M. Power, editors, *Handbook of Cognition and Emotion*, pages 301–320. John Wiley, New York.

Engberg, I. S., Hansen, A. V., Andersen, O., and Dalsgaard, P. (1997). Design, recording and verification of a Danish emotional speech database. In *Proc. Eurospeech*, pages 1695–1698, Rhodes.

Erickson, D., Yoshida, K., Menezes, C., Fujino, A., Mochida, T., and Shibuya, Y. (2004). Exploratory Study of Some Acoustic and Articulatory Characteristics of *Sad* Speech. *Phonetica*, **63**, 1–25.

Eysenck, H. (1960). The Concept of Statistcal Significance and the Controversy about One-Tailed Tests. *Psychological Review*, **67**, 269–271.

Fiscus, J. (1997). A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Proc. ASRU*, pages 347–352, Santa Barbara.

Fleiss, J., Cohen, J., and Everitt, B. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, **72**(5), 323–327.

Frick, R. (1985). Communicating emotion: the role of prosodic features. *Psychological Bulletin*, **97**, 412–429.

Giuliani, D. and Gerosa, M. (2003). Investigating recognition of children's speech. In *Proc. of ICASSP 2003*, volume 2, pages 137–140, Hong Kong.

Gobl, C. and Ní Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, **40**, 189–212.

Grimm, M., Kroschel, K., Harris, H., Nass, C., Schuller, B., Rigoll, G., and Moosmayr, T. (2007). On the Necessity and Feasibility of Detecting a Driver's Emotional State While Driving. In A. Paiva, R. Prada, and R. W. Picard, editors, *Affective Computing and Intelligent Interaction*, pages 126–138, Berlin-Heidelberg. Springer.

Hall, M. A. (1998). *Correlation-based feature selection for machine learning*. Ph.D. thesis, Hamilton, NZ: Waikato University, Department of Computer Science.

Hermansky, H. and Sharma, S. (1998). Traps - classifiers of temporal patterns. In *Proc. ICSLP*, pages 1003–1006, Sydney.

Holodynski, M. and Friedlmeier, W. (2006). *Development of emotions and emotion regulation*. Springer, New York.

Hu, H., Xu, M.-X., and Wu, W. (2007). GMM Supervector Based SVM with Spectral Features for Speech Emotion Recognition. In *Proceedings of Interspeech*, pages 413–416, Antwerp.

Inanoglu, Z. and Caneel, R. (2005). Emotive alert: HMM-based emotion detection in voicemail messages. In *Proceedings of the 10th International Conference on Intelligent User Interfaces*, pages 251–253, San Diego, California.

Jain, A. and Zongker, D. (1997). Feature selection: evaluation, application and small sample performance. *PAMI*, **19**(2), 153–158.

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. In C. Nédellec and C. Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pages 137–142, Chemnitz. Springer, Heidelberg.

Kießling, A. (1997). *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung*. Berichte aus der Informatik. Shaker, Aachen.

Küstner, D., Tato, R., Kemp, T., and Meffert, B. (2004). Towards Real Life Applications in Emotion Recognition. In E. André, L. Dybkiaer, W. Minker, and P. Heisterkamp, editors, *Affective Dialogue Systems, Proceedings of a Tutorial and Research Workshop*, volume 3068 of *Lecture Notes in Artificial Intelligence*, pages 25–35, Berlin. Springer.

Kwon, O.-W., Chan, K., Hao, J., and Lee, T.-W. (2003). Emotion recognition by speech signals. In *Proc. Interspeech*, pages 125–128.

Lee, C., Narayanan, S., and Pieraccini, R. (2001). Recognition of Negative Emotions from the Speech Signal. In *Proc. ASRU*, Madonna di Campiglio. no pagination.

Lee, C. M. and Narayanan, S. S. (2005). Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, **13**(2), 293–303.

Lee, C. M., Narayanan, S. S., and Pieraccini, R. (2002). Combining acoustic

and language information for emotion recognition. In *Proc. Interspeech*, pages 873–376, Denver.

Lee, S., Potamianos, A., and Narayanan, S. (1999). Acoustic of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America (JASA)*, **105**(3), 1455–1468.

Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, **11**, 22–31.

Makhoul, J., Kubala, F., Schwartz, R., and Weischedel, R. (1999). Performance measures for information extraction. In *Proceedings of DARPA Broadcast News Workshop*, Herndon, VA, USA.

McGilloway, S., Cowie, R., Doulas-Cowie, E., Gielen, S., Westerdijk, M., and Stroeve, S. (2000). Approaching automatic recognition of emotion from voice: A rough benchmark. In *Proceedings of the ISCA workshop on Speech and Emotion*, pages 207–212, Newcastle.

Meyer, D., Leisch, F., and Hornik, K. (2002). Benchmarking Support Vector Machines. Report Series No. 78, Adaptive Informations Systems and Management in Economics and Management Science. 19 pages.

Morrison, D., Wang, R., Xu, W., and Silva, L. C. D. (2007). Incremental learning for spoken affect classification and its application in call-centres. *International Journal of Intelligent Systems Technologies and Applications*, **2**, 242–254.

Mozziconacci, S. (1998). *Speech variability and emotion: production and perception*. Ph.D. thesis, Technical University Eindhoven.

Nogueiras, A., Moreno, A., Bonafonte, A., and Mariño, J. B. (2001). Speech emotion recognition using hidden markov models. In *Proc. Eurospeech*, pages 2267–2270, Aalborg.

Nose, T., Kato, Y., and Kobayashi, T. (2007). Style estimation of speech based on multiple regression hidden semi-markov model. In *Proc. Interspeech*, pages 2285–2288, Antwerp.

Nwe, T., Foo, S., and Silva, L. D. (2003). Speech emotion recognition using hidden markov models. *Speech Communication*, **41**, 603–623.

Ortony, A., Clore, G. L., and Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge University Press, Cambridge, New York.

Oudeyer, P.-Y. (2003). The production and recognition of emotions in speech: features and algorithms. *International Journal of Human-Computer Studies*, **59**(1-2), 157–183.

Pernegger, T. V. (1998). What's wrong with Bonferroni adjustment. *British Medical Journal*, **316**, 1236–1238.

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

Pudil, P., Novovicova, J., and Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters*, **15**, 1119–1125.

Rozeboom, W. (1960). The Fallacy of the Null-Hypothesis Significance Test. *Psychological bulletin*, **57**, 416–428.

Salzberg, S. (1997). On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, **1**(3), 317–328.

Scherer, K. R. (1986). Vocal affect expression: a review and a model for future research. *Psychological Bulletin*, **99**, 143–165.

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

Schröder, M., Pirker, H., and Lamolle, M. (2006). First Suggestions for an Emotion Annotation and Representation Language. In L. Devillers, J.-C. Martin, R. Cowie, E. Douglas-Cowie, and A. Batliner, editors, *Proc. of a Satellite Workshop of LREC 2006 on Corpora for Research on Emotion and Affect*, pages 88–92, Genoa.

Schröder, M., Devillers, L., Karpouzis, K., Martin, J.-C., Pelachaud, C., Peter, C., Pirker, H., Schuller, B., Tao, J., and Wilson, I. (2007). What Should a Generic Emotion Markup Language Be Able to Represent? In A. Paiva, R. Prada, and R. W. Picard, editors, *Affective Computing and Intelligent Interaction*, pages 440–451, Berlin-Heidelberg. Springer.

Schuller, B., Rigoll, G., and Lang, M. (2003). Hidden markov model-based speech emotion recognition. In *Proc. ICASSP*, pages 1–4, Hong Kong.

Schuller, B., Müller, R., Lang, M., and Rigoll, G. (2005). Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensemble. In *Proc. Interspeech*, pages 805–808, Lisbon.

Schuller, B., Reiter, S., and Rigoll, G. (2006). Evolutionary feature generation in speech emotion recognition. In *Proc. Int. Conf. on Multimedia and Expo ICME 2006*, pages 5–8, Toronto, Canada.

Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., and Aharonson, V. (2007a). The Relevance of Feature Type for the Automatic Classification of Emotional User States: Low Level Descriptors and Functionals. In *Proc. Interspeech*, pages 2253–2256, Antwerp.

Schuller, B., Seppi, D., Batliner, A., Meier, A., and Steidl, S. (2007b). Towards more Reality in the Recognition of Emotional Speech. In *Proc. ICASSP*, pages 941–944, Honolulu.

Schuller, B., Wimmer, M., Mösenlechner, L., Kern, C., Arsic, D., and Rigoll, G. (2008). Brute-Forcing Hierarchical Functionals for Paralinguistics: a Waste of Feature Space? In *Proc. ICASSP*, pages 4501–4504, Las Vegas.

Seppi, D., Gerosa, M., Schuller, B., Batliner, A., and Steidl, S. (2008). Detecting Problems in Spoken Child-Computer Interaction. In *Proceedings of the 1st Workshop on Child, Computer and Interaction*, Chania, Greece.

Shami, M. and Verhelst, W. (2007). Automatic Classification of Expressiveness in Speech: A Multi-corpus Study. In C. Müller, editor, *Speaker Classification II*, volume 4441 of *Lecture Notes in Computer Science / Artificial Intelligence*, pages 43–56. Springer, Heidelberg - Berlin - New York.

Steidl, S. (2009). *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. Logos Verlag, Berlin. PhD thesis.

Steidl, S., Levit, M., Batliner, A., Nöth, E., and Niemann, H. (2005). "Of All Things the Measure is Man": Automatic Classification of Emotions and Inter-Labeler Consistency. In *Proc. ICASSP*, pages 317–320, Philadelphia.

Steidl, S., Batliner, A., Nöth, E., and Hornegger, A. (2008). Quantification of Segmentation and Fo Errors and Their Effect on Emotion Recognition. In *Text, Speech and Dialogue, 11th International Conference, TSD 2008*, pages 525–534.

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

Vidrascu, L. and Devillers, L. (2008). Anger detection performances based on prosodic and acoustic cues in several corpora. In L. Devillers, J.-C. Martin, R. Cowie, E. Douglas-Cowie, and A. Batliner, editors, *Proceedings of the Workshop on Corpora for Research on Emotion and Affect at LREC 2008*, pages 23–27, Marrakech.

Vlasenko, B., Schuller, B., Wendemuth, A., and Rigoll, G. (2007). Frame vs. Turn-Level: Emotion Recognition from Speech Considering Static and Dynamic Processing. In A. Paiva, R. Prada, and R. W. Picard, editors, *Affective Computing and Intelligent Interaction*, pages 139–147, Berlin-Heidelberg. Springer.

Vogt, T. and André, E. (2005). Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In *Proc. Multimedia and Expo (ICME05)*, pages 474–477, Amsterdam.

Wagner, J., Vogt, T., and André (2007). A Systematic Comparison of different HMM designs for Emotion Recognition from Acted and Spontaneous Speech. In A. Paiva, R. Prada, and R. W. Picard, editors, *Affective Computing and Intelligent Interaction*, pages 114–125, Berlin-Heidelberg. Springer.

Witten, I. H. and Frank, E. (2005). *Data mining: Practical machine learning tools and techniques, 2nd Edition*. Morgan Kaufmann, San Francisco.

You, M., Chen, C., Bu, J., Liu, J., and Tao, J. (2006). Emotion recognition from noisy speech. In *Proc. ICME*, pages 1653–1656, Toronto.