Combining Efforts for Improving Automatic Classification of Emotional User States

Anton Batliner[1], Stefan Steidl[1], Björn Schuller[2], Dino Seppi[3], Kornel Laskowski[4], Thurid Vogt[5], Laurence Devillers[6], Laurence Vidrascu[6], Noam Amir[7], Loic Kessous[7], Vered Aharonson[8]

[1]**FAU:** Lehrstuhl für Mustererkennung, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
batliner@informatik.uni-erlangen.de, steidl@informatik.uni-erlangen.de

[2]**TUM:** Institute for Human-Machine Communication, Technische Universität München, Germany, *schuller@tum.de*

[3]**ITC:** ITC-irst, Trento, Italy, *seppi@itc.it*

[4]**UKA:** interACT, University of Karlsruhe, Germany, *kornel@cs.cmu.edu*

[5]**UA:** Multimedia Concepts and their Applications, University of Augsburg, Germany, *vogt@informatik.uni-augsburg.de*

[6]**LIMSI:** Spoken Language Processing Group, LIMSI-CNRS, Orsay Cedex, France, *devil@limsi.fr, vidrascu@limsi.fr*

[7]**TAU:** Dep. of Communication Disorders, Sackler Faculty of Medicine, Tel Aviv University, Israel,
noama@post.tau.ac.il, kessous@post.tau.ac.il
[8]**AFEKA:** Tel Aviv academic college of engineering,Tel Aviv, Israel, *vered@nexsig.com*

Abstract

Classification performance of emotional user states found in realistic, spontaneous speech is not very high, compared to the performance reported for acted speech in the literature. This might be partly due to the difficulty of providing reliable annotations, partly due to suboptimal feature vectors used for classification, and partly due to the difficulty of the task. In this paper, we present a co-operation between several sites, using a thoroughly processed emotional database. For the four-class problem *motherese*/*neutral*/*emphatic*/*angry*, we first report classification performance computed independently at each site. Then we show that by using all the best features from each site in a combined classification, and by combining classifier outputs within the ROVER framework, classification results can be improved; all feature types and features from all sites contributed.

Združevanje sil za boljše samodejno razvrščanje čustvenih stanj uporabnika:

Uspešnost samodejnega razvrščanja čustvenih stanj uporabnika, ki jih najdemo v realističnem, spontanem govoru, je v primerjavi s kakovostjo, ki jo v literaturi navajajo za igrani govor, precej nižja. To je lahko delno posledica težav pri zagotavljanju zanesljive anotacije, delno posledica uporabe podoptimalnih vektorjev značilk pri razvrščanju, delno pa posledica težavnosti te naloge. V prispevku predstavljamo sodelovanje med različnimi ustanovami na temeljito obdelani bazi podatkov. Za štiristopenjski problem *govor otroku*/*nevtralno*/*poudarjeno*/*jezno* najprej navedemo kakovost razvrščanja, kot so jo izračunali neodvisno na vsaki od sodelujočih ustanov. Nato poka-žemo, da lahko izboljšamo rezultate razvrščanja z uporabo najboljših značilk vsake izmed ustanov in z združevanjem rezultatov razvrščevalnikov znotraj ogrodja ROVER.

1. Introduction

In this paper, we present a co-operation between several sites dealing with classification of emotional user states conveyed via speech; this initiative was taken within the European Network of Excellence HUMAINE under the name CEICES (**C**ombining **E**fforts for **I**mproving automatic **C**lassification of **E**motional user **S**tates); as for an overview of emotion recognition in human-computer-interaction, cf. (Cowie et al., 2001). The database used is a German corpus with recordings of 51 ten- to thirteen-year old children communicating with Sony's AIBO pet robot. Conceptualization, design and recordings were done at the 'originator' site FAU[1]; results have been reported for example in (Steidl et al., 2005; Batliner et al., 2005b; Batliner et al., 2005a). The approach to be followed within CEICES looked like this: the originator site provided speech files, phonetic lexicon, manually corrected word segmentation (and, in the future, manually corrected F0 values), emotional labels, definition of train and test samples, etc. The data was annotated at the word level. We aimed at two different classification tasks: word-based and turn-based classification; for the latter, we mapped the word-based labels onto turn-based ones. All partners committed themselves to share with all the other partners their extracted feature values together with the necessary information (which feature models which acoustic or linguistic phenomenon, format of feature values, classifier used, etc.). Thus each site could assess the features provided by all other sites, together with their own features, aiming at a repertoire of optimal fea-

[1]The abbreviations for all sites can be found in the affiliations given in the title of this paper; AFEKA is subsumed under TAU.

tures. In this work, we look not only at acoustic but also at linguistic features.[2]

2. Material and annotation

The general framework for the database reported on in this paper is child-robot communication, and the elicitation and subsequent recognition of emotion-related user states. The robot is Sony's (dog-like) AIBO robot. The basic idea has been to combine a new type of corpus (children's speech) with 'natural' emotional speech within a Wizard-of-Oz task. The speech is intended to be 'natural' since children do not disguise their emotions to the same extent as adults. However, it is of course not fully 'natural' as it might be in an unsupervised setting. Furthermore the speech is spontaneous; the children were not told to use specific instructions but to talk to the AIBO as they would to a friend. In this experimental design, the child is led to believe that the AIBO is responding to his or her commands, but the robot is actually being controlled by a human operator, using the 'AIBO Navigator' software over a wireless LAN (the existing AIBO speech recognition module is not used). The wizard causes the AIBO to perform a fixed, pre-determined sequence of actions, which takes no account of what the child says. For the sequence of AIBO's actions, we tried to find a good compromise between obedient and disobedient behaviour: we wanted to provoke the children in order to elicit emotional behaviour but of course we did not want to run the risk that they break off the experiment. The children believed that the AIBO was reacting to their orders — albeit often not immediately. In fact, it was the other way round: the AIBO always strictly followed the same plot, and the children had to align their orders to its actions.

The data was collected from 51 children (age 10 - 13, 21 male, 30 female). The children are from two different schools (25 children from 'MONT' and 26 from 'OHM'); the recordings took place in the respective classrooms. The only persons in the room were the child, a supervisor who initially instructed the children, the wizard (behind the children, pretending to be doing the recordings) and a third assistant.[3] Each recording session took some 30 minutes. Because of the experimental setup, these recordings contain a huge amount of silence (the reaction time of the AIBO), which caused a noticeable reduction of recorded speech after raw segmentation; ultimately we obtained about 9.2 hours of speech. More details are given in (Steidl et al., 2005; Batliner et al., 2005b; Batliner et al., 2005a).

Five labellers (advanced students of linguistics) listened to the recordings and annotated independently of each other each word as *neutral* (default) or as belonging to one of ten other classes which were designed during earlier inspection of the data; we do not claim that these classes represent children's emotions in general, only that they are adequate for the modelling of these children's behaviour in this specific scenario. We resorted to majority voting (henceforth MV): if three or more labellers agreed, the label was attributed to the word; if four or five labellers agreed, we assumed a sort of prototype. The following raw labels were used — in parentheses, the number of cases with MV is given: *joyful* (101), *surprised* (0), *emphatic* (2528), *helpless* (3), *touchy*, i.e., irritated (225), *angry* (84), *motherese* (1260), *bored* (11), *reprimanding* (310), *other*, i.e. non-neutral, but not belonging to the other categories (3), and *neutral* (39169). 4707 words had no MV; all in all, there were 48401 words. *joyful* and *angry* belong to the 'big' emotions, the other ones rather to 'emotion-related/emotion-prone' user states and by that, to 'emotion' in its broader meaning.

The state *emphatic* has to be commented on especially: based on our experience with other emotion databases (Batliner et al., 2003), any marked deviation from a neutral speaking style can (but need not) be taken as a possible indication of some (starting) trouble in communication. If a user gets the impression that the machine does not understand her, she tries different strategies – repetitions, reformulations, other wordings, or simply the use of a pronounced, marked speaking style. Such a style does 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. There are three further — practical — arguments for the annotation of *emphatic*: firstly, it is to a large extent a prosodic phenomenon, and can thus be modelled and classified with prosodic features. Secondly, if the labellers are allowed to label *emphatic*, it may be less likely that they confuse it with other user states. Thirdly, we can try and model emphasis as an indication of (arising) problems in communication (Batliner et al., 2003).

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., 2005a). However, the distribution of classes is very unequal. Therefore, we downsampled *neutral* and *emphatic* to **N***eutral* and **E***mphatic*, respectively, and mapped *touchy*, *reprimanding*, and *angry* onto **A***ngry*[4], 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

[2] We expect improved recognition rates from this co-operation. However, it is an educated guess that, for instance, manual segmentation yields more reliable results for emotion recognition than automatic segmentation — we simply do not know yet whether and to what extent this will turn out to be a fact.

[3] Speech was transmitted with a wireless headset (UT 14/20 TP SHURE UHF-series with microphone WH20TQG) and recorded with a DAT-recorder. The sampling rate of the signals was 48 kHz, quantized at 16 bits. The data was downsampled to 16 kHz prior to processing.

[4] The initial letter is given boldfaced and recte; 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.

for **A***ngry* (**A**), 1224 words for **M***otherese* (**M**), 1645 words for **E***mphatic* (**E**), and 1645 for **N***eutral* (**N**) (Steidl et al., 2005). Cases where less than three labellers agreed were omitted as well as those cases where other than these four main classes were labelled. Interlabeller agreement is dealt with in (Steidl et al., 2005).

A last note on label names and terminology in general: names for the non-cognitive phenomena that we are dealing with are known not to be unequivocal or agreed upon. There is wide disagreement as to whether affect encompasses emotion or the other way around. In this paper, we follow a terminology widely adopted by HUMAINE. 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., 2005a). Instead of *motherese*, some people use 'child-directed speech'; this is, however, only feasible if there is in the respective database no negative counterpart such as *reprimanding* which is 'child-directed' as well. **A***ngry* was not named **N***egative* because we reserved **N** for **N***eutral*.

3. Pre-processing of the data

The word is a simple and rather unequivocal concept in speech processing; the basic unit of emotional speech might not be the word — nor the sentence — but something in between (clauses, noun phrases, etc.). By annotating words, we are able to map sequences of words onto larger emotion units later on. An automatic reverse top-down splitting — from turns to clauses — would not be possible. The processing of emotional speech, however, at almost any other site, resorts to turns as units which have been labelled as such. We therefore decided to start with turns as units of investigation; these 'turns' are physically stored and distributed as speech files which were extracted out of the recordings of the sessions using longer pauses as the automatic segmentation criterion.[5] This leaves us with the task of mapping word-based labels onto turn-based labels: a simple 50% threshold — for instance, if an **A** turn has 10 words, then 5 or more words have to be labelled as **A** — would be suboptimal because some words, especially function words, are likely not to be produced in an emotional manner; moreover, a longer turn can consist of one neutral clause, and one emotional clause — then chances are that the whole turn will be wrongly mapped onto neutral.

For the mapping onto turn-based labels, we employed the following strategy: fragments and auxiliaries were used as stop words.[6] For each turn, we pool together the labels given by our 5 labellers (for a turn of n words, we obtain 5 x n labels). For the turn to be mapped onto neutral, 70% of the labels have to be neutral; *joyful* and the other spurious labels are not taken into account for this computation. If 30% or more are non-neutral, then the turn is **A**, **M**, or **E**. If at least 50% of the non-neutral labels are **M**, the turn is mapped onto **M**. If **A** and **E** are equally distributed, the turn is mapped onto **A**. The remaining turns, which are neither **A** or **M**, are declared to be **E**. This means that we employ a sort of 'markedness' condition: **M** is more marked than **A**, and **A** is more marked than **E**, and all are more marked than **N**. This strategy yields the following turn-based label counts: 868 **A** (21.7 %), 1347 **E** (33.7 %), 495 **M** (12.4 %), and 1280 **N** (32.0 %), summing up to 3990 (100 %) turn labels.

Especially for the word-based classification to be reported on in a future work, to avoid automatic segmentation errors which certainly will be different at different sites, the automatic segmentation of all words belonging to these 3990 AMEN turns conducted at FAU was manually corrected by the first author. We hope that this will eliminate performance differences that might be traced back to different automatic segmentations.

4. Classification

For classification, we used 2-fold cross-validation: MONT vs. OHM and vice versa, and then average the two results. This way, we can guarantee strict speaker independence and, at the same time, easily compare results across sites by visual inspection — which would not be possible if we resorted to leave-one-speaker-out (i.e. 51-fold cross-validation). This 2-fold cross-validation is a more conservative strategy yielding lower recognition performance than leave-one-speaker-out. It might be argued that, in addition, we should define a validation sample, and that we should deal with the multiplicity effect, i.e. the repeated use of the same data, through significance testing using the Bonferroni adjustment. A practical argument against a validation sample is that it would reduce the number of cases — which is already low. There are some theoretical/methodological arguments against the Bonferroni adjustment (Pernegger, 1998); however, in our situation, when we are pursuing rather *I-wonder-what-will-happen* instead of *I-bet-this-will-happen* hypotheses, the Bonferroni adjustment might be appropriate — but only if we were to **claim** significance for our results. We prefer to conceive our experiments as what they indeed are: collecting cumulative evidence for trends that have to be corroborated anyway with other (types of) data. In Tables 1 to 3, we report the overall recognition rate RR (number of correctly classified cases divided by total number of cases or weighted average) and CL (a 'class-wise' computed recognition rate, i.e. the mean along the diagonal of the confusion matrix in percent, or unweighted average).

4.1. Separate Classification

In this section, we report on those initial experiments that were conducted at each site with different features and different classifiers, thereby providing a baseline for different automatic classification strategies. Essentially, one and

[5] Such a criterion is of course not based on syntactic considerations. Full-fledged sentences are, however, rather sparse in the register *'giving commands to a robot'*.

[6] For the turns containing our 6070 AMEN words, this means 17618 words in 3996 turns; stop words consisted of 596 fragments and 196 auxiliaries (some words both); this results in 16856 remaining words. Note that we could have identified more stop words, but this would be rather data-dependent and we chose to avoid that. For six turns containing only stop words, no turn-based labels were generated.

Site	# of features		# per type of features						domain		classification				
	original (4024)	selected (381)	prosodic	spectral	MFCC	POS	lexical	genetic search	turn	word	classifier	RR	CL	ROVER	features combined
FAU	303	87	19	-	-	6	62	-	√	√	Neural Networks	55.8	55.3	√	√
TUM	980	103	9	17	22	2	50	3	√	-	SVM	59.3	56.4	√	√
ITC	32	32	26	-	-	6	-	-	√	√	Random Forest (RF)	57.6	55.8	√	√
UKA	1320	25	6	-	5	-	14	-	√	-	Linear Regressor	59.1	54.8	-	√
UA	1289	84	10	1	73	-	-	-	√	-	Naive Bayes	50.9	52.3	√	√
LIMSI	76	26	9	9	-	5	3	-	√	-	SVM	54.9	56.6	√	√
TAU	24	24	24	-	-	-	-	-	√	-	Rule-based	48.9	46.6	-	√

Table 1: *Features and classifiers: per site, # of features before/after feature selection; # per type of features, and their domain; classifier used, weighted average recognition rate RR and non-weighted class-wise averaged recognition rate CL; used or not used (-) in ROVER and in classifiaction with all features; SVM = Support Vector Machines, POS = part-of-speech.*

the same database is independently used by each authoring site reporting different results. This effectively defines a range of performance for this task.

For the results given in Table 1, the 3990 cases, the labels, and training and test sets were identical across all sites; only the features and classifiers differed. The types of features included[7]:

- **prosodic**: F0, energy, duration, and other types of supra-segmental information such as jitter and shimmer;
- **spectral**: modelling Harmonics-to-Noise ratio, formants with band-width etc.;
- **MFCC**: the usual MFCC features plus derivatives;
- **part-of-speech (POS)**: based on coarse word classes such as nouns, particles, etc. provided by FAU;
- **lexical**: single words, or bag-of-word classes (Joachims, 1997);
- **genetic search**: features generated automatically, based on evolutionary alteration and combination.

Irrespective of the types of features and classifiers used, the results are roughly of the same order of magnitude; these figures are, for a 4-class problem and for realistic, spontaneous speech which does not only contain prototypical, very clear cases, in the expected range.[8] Our heuristic threshold of 70% for the definition of MV cases, cf. above, may have resulted in lower classification performance than a threshold of 50%. However, we were not interested in manipulating the data to obtain the highest possible recognition rates, but rather in a realistic setting which takes into account possible applications. For the same reasons, we avoided focusing on only those turns in which the labellers fully agreed, which could have led to a classification performance of up to 80% for our 4-class problem.

The results in Table 1 illustrate an initial range of performance for this task; they should not be conceived of as competing with each other. We found it hard to control all aspects of processing at the different sites which used, e.g. different feature normalization and selection procedures.[9] '√' in the last two columns means that these classifier outputs (cf. columns 13–14) and the features from columns 4–9 were put into ROVER and into a classification which combines features from all sites respectively. Our intention was that with this step, each site can reduce its own large feature set (sometimes $>$ 1000 features) to a smaller set with most of the relevant features.

4.2. Combining Classifiers

When multiple classifiers are available, it is possible to combine their independent results to obtain a composite output whose classification performance is higher than that of the individual systems. In automatic speech recognition (ASR), this is normally achieved using the ROVER framework described in (Fiscus, 1997). Basically, ROVER per-

[7]Note that at times, assignments of a feature to one of these feature cover classes is not unequivocal.

[8]There are some studies available describing realistic speech with two or three classes. As for the very few with four classes and classification performance (CL) well above 60%, it can be shown that the results were 'fine-tuned' somehow; such strategies are dealt with in (Batliner et al., 2005b).

[9]The results reported by TAU are obtained with one specific type of prosodic feature (intonation model pitch features) whereas the other sites used multiple prosodic feature types. FAU and ITC followed a two-stage strategy: they first computed word-based features using the manually corrected segmentation; in a second step, turn-based features were computed based on these word-based features, cf. column 11 in Table 1. Some of the LIMSI features were speaker-normalized. FAU independently selected acoustic/part-of-speech and lexical features each with sequential feature selection (SFS), LIMSI used several different methods, TAU none, all others used SFS based on all feature types. Feature selection has been done independently for the two computations in the 2-fold cross-validation, then the set union of the features was used again; these results are reported in Table 1. This procedure yields sub-optimal performance but guarantees that all possibly relevant features will be kept for the combined classification. As for features modelled and/or classifiers used, cf., in addition to the other references, (Schuller et al., 2005; Vogt and André, 2005; Devillers et al., 2005; Kießling, 1997).

forms a word alignment among independent ASR outputs, and later combines the best hypotheses and their confidence measures to find the most probable word. For our purposes, the alignment step can be skipped, while the scoring step is almost identical to that described in (Fiscus, 1997): the final label e^* is chosen using the following:

$$e^* = \arg\max_e \left[\alpha \cdot \left(\frac{N(e,i)}{\sum_i N(e,i)} \right) + (1-\alpha) \cdot C_k(e,i) \right]$$

where $N(e,i)$ is the frequency of label e in the i outputs, $C_k(e,i)$ is their combined confidence measure, and α is a weighting factor. $C_k(e,i)$ has been evaluated in $k = \{1,2,3\}$ different ways: the straightforward method assumes no weighting ($\alpha = 1$); C_2 is the mean of the confidence scores while C_3 is their maximum. For both these last two systems, α is usually chosen using a cross-validation data set. Due to data scarceness, we chose values of α that maximize RR on the training set, obtaining values for α between 0.7 and 0.9. In other words, when testing on the OHM subset of the data, we selected α by maximizing RR on the MONT subset, and vice-versa. As the original confidences for UKA and TAU were not available, we used altogether the output of five classifiers, cf. Table 1, column 15; results are given in Table 2.

confidence	k	α	RR	CL
C_1, none	1	1.0	62.8	61.9
C_2, mean	2	0.7 - 0.8	63.1	62.2
C_3, max	3	0.8 - 0.9	63.5	62.4

Table 2: *ROVER results obtained by combining the outputs and the confidences of 5 classifiers, cf. Table 1.*

4.3. Combining Features

We now report on classifications with all 381 'most relevant' features from all sites, cf. Table 1, columns 3–9. In Table 3, RR and CL are given for three different classifiers. Feature selection was performed independently for the two training sets MONT and OHM in the 2-fold cross-validation; the number (#) of 'surviving' features is given in columns 2–3. SVM and RF classifiers, using the surviving features, outperform all results obtained independently at each site. These two more sophisticated classifiers perform some percent points — but not considerably — better than the out-of-the-box LDA classifier which used considerably fewer features. The difference in performance may become more pronounced if we were to use a leave-one-speaker-out strategy.

A lack of space makes it prohibitive to fully explore the possible gain in knowledge from combining features and classifier outputs, but we attempt a cursory analysis in Table 4. We first give the number of features per type used by the three classifiers in the two 2-fold cross-validations MONT and OHM; the last line shows the number of features per type summing up to 381. Each feature type has been used throughout, and for each run, features from all sites were used. Note that the 'original' 4024 features were obtained with quite different methods – some by 'brute force' and automatic selection, some using prior knowledge. The SVM and LDA classifiers appear to use more lexical features in relation to RF which uses more acoustic features. Even if each additional feature contributes only negligibly in terms of performance — the size of the feature vectors in Table 3 grows much faster than classification accuracy — they may be valuable for subsequent interpretation.

classifier	# selected features		RR	CL
	MONT	OHM		
LDA	53	67	58.8	56.3
SVM	159	150	61.8	57.9
RF	299	284	60.8	58.7

Table 3: *Classification performance, combining 381 features from all sites, feature selection for 2-fold cross-validation on the training set, with 3 different classifiers; LDA = Linear Discriminant Analysis.*

classifier	training set	prosodic	spectral	MFCC	POS	lexical	gen. search
LDA	MONT	16	5	6	4	21	1
	OHM	19	2	11	3	31	1
SVM	MONT	47	14	37	7	53	1
	OHM	34	14	33	8	59	2
RF	MONT	102	27	100	18	49	3
	OHM	101	27	100	15	38	3
# original features		103	27	100	19	129	3

Table 4: *# of features used per type/per classifier/per training set.*

5. Discussion and Future Work

It is not very difficult to fine-tune classifier performance and obtain considerably higher recognition rates than those reported in this paper, by concentrating on prototypical cases for example — in (Batliner et al., 2005b), up to 75.5 % CL for the same 4-class problem with an LDA classifier — and/or by using leave-one-speaker-out. For prototypical exemplars, we could focus on only those cases where a majority of 4 or 5 out of 5 labellers agreed. In our opinion, to start with, it is more important to establish solid baselines such as those shown in Tables 2 and 3. With ROVER, we have shown an absolute improvement of up to 5.8 % with respect to the best independent site result for CL, cf. Table 2 versus Table 1. By combining features from all sites, we achieved up to 2.1 % absolute improvement for CL, cf. Table 3 versus Table 1. It appears that the combination of different classifiers with different (types of) features which is used by ROVER can model the distribution better than just the use of all 'surviving' features in one and the same classifier.[10]

In future work, we hope to address the following topics:

[10] Note that RFs are an exception, as they are actually a multi-classifier system (Breiman, 2001) composed of a large set of classification trees, each one working on a randomly sampled sub-

- pre-processing: various strategies such as automatic versus manual segmentation and F0 extraction, and forced alignment versus processing based on word-hypothesis graphs[11];

- units and context: turn- versus word-based processing; mapping chunks of words onto 'emotionally significant' units; taking into account of session context in turn-based classification;

- phonetic and linguistic 'substance': which features and types of features are most relevant, and which are not, and why is this the case?

- pattern classification: optimization of classifiers, comparison of performance with and without a loss matrix; possibly automatic feature generation, genetic programming and boosting, and decorrelation of features with PCA; using other knowledge sources such as language models.

6. Concluding Remarks

The idea behind this CEICES endeavour has been to co-operate closely by assembling and evaluating together all kinds of features, both acoustic and linguistic, rather than to compete between sites as in the more common assessment and evaluation procedures (Gibbon et al., 1997). The small performance differences between the authoring sites (Table 1) have to be traced back to differences in either features, classifiers, or feature space optimization. We have shown that co-operation leads to improvements if we simply accumulated and evaluated all features from all sites together at the input level of classification, cf. Table 3. However, results were 'only' up to some two percent points better than the best results obtained at any single site. Further impovements are possible by combining different sets of features with different types of classifiers, cf. the results obtained with ROVER in Table 2. Markedly better classification peformance might not be possible with a further fine-tuning of features and classifiers; in addition we should take into account some of the aspects mentioned in section 5.

7. Acknowledgements

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 responsibilty lies with the authors.[12]

set of features. It is surprising that they are outperformed by the ROVER framework. A possible interpretation is that the diversity of classifier types, as used by ROVER, is crucial. Furthermore, we must stress that the two approaches — in general all results reported in Tables 2 and 3 — cannot be directly compared as they do not rely on the same features, cf. columns 15–16 in Table 1.

[11]For fully automatic processing, some features such as bag-of-words or part-of-speech have to be extracted from a word-hypothesis graph and will not always be correct.

[12]The ROVER computation of section 4.2. was done at ITC, the combined classification of section 4.3. at FAU, TUM, and ITC.

8. References

A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. 2003. How to Find Trouble in Communication. *Speech Communication*, 40:117–143.

A. Batliner, S. Steidl, C. Hacker, E. Nöth, and H. Niemann. 2005a. Private Emotions vs. Social Interaction - towards New Dimensions in Research on Emotion. In *Proceedings of a Workshop on Adapting the Interaction Style to Affective Factors, 10th International Conference on user Modelling*, pages 8 pages, no pagination, Edinburgh.

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

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

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

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

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

Dafydd Gibbon, Roger Moore, and Richard Winski, editors. 1997. *Handbook of Standards and Resources for Spoken Language Systems*. Mouton de Gruyter, Berlin.

T. Joachims. 1997. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Technical report, LS-8 Report 23, Dortmund, Germany.

A. Kießling. 1997. *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung*. Shaker, Aachen.

Thomas V. Pernegger. 1998. What's wrong with Bonferroni adjustment. *British Medical Journal*, 316:1236–1238.

B. Schuller, R. Müller, M. Lang, and G. Rigoll. 2005. Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In *Proceedings of Interspeech 2005*, pages 805–808, Lisbon, Portugal.

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

Thurid Vogt and Elisabeth André. 2005. Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition. In *Proceedings of IEEE International Conference on Multimedia & Expo (ICME 2005)*, Amsterdam, The Netherlands.