The Hinterland of Emotions: Facing the Open-Microphone Challenge

Stefan Steidl
Lehrstuhl für Mustererkennung,
Friedrich-Alexander-Universität,
Erlangen-Nürnberg, Germany
steidl@informatik.uni-erlangen.de

Björn Schuller
Institute for Human-Machine Communication,
Technische Universität München,
München, Germany
schuller@tum.de

Anton Batliner
Lehrstuhl für Mustererkennung,
Friedrich-Alexander-Universität
Erlangen-Nürnberg, Germany
batliner@informatik.uni-erlangen.de

Dino Seppi
ESAT, Katholieke Universiteit
Leuven, Belgium
dino.seppi@esat.kuleuven.be

Abstract

We first depict the challenge to address all non-prototypical varieties of emotional states signalled in speech in an open microphone setting, i. e. using all data recorded. In the remainder of the article, we illustrate promising strategies, using the FAU Aibo Emotion Corpus, by showing different degrees of classification performance for different degrees of prototypicality, and by elaborating on the use of ROC curves, classification confidences, and the use of correlation-based analyses.

1. Introduction

Amongst the various conceptualizations of 'emotion', the concept of *prototypes* seems not to be the prevailing one; however, especially when faced with the necessity to model 'everything that comes in' it seems to be attractive and will therefore be introduced in the following. A *prototype* is a salient, central member of a category and typically most often associated with this category [9]. The well-known, big *n* emotions can be conceived as prototypes, cf. [7, 17]. If no external criterion is available, real-life data have to be annotated manually for obtaining a 'ground truth' as reference for automatic processing. Thus a straightforward operationalisation is to speak of 'prototypical' cases if the labellers agree. Non-prototypical – weak and/or mixed – emotions can be found when labellers annotate more than one emotion per item, or when we preserve the disagreement of several labellers in some sort of graded/mixed annotation. Irrespective of the type of annotation, we always can generate either categorical labels representing either pure or mixed cases, or we can generate a continuous representation by placing each case on some dimension, e. g. valence. Basically, it is always possible to convert a continuous representation into a categorical one, and vice versa.

Early mapping onto main classes (e. g. via majority voting) is a sort of *early quantisation*; late mapping is a sort of *late quantisation*, cf. early or late fusion in classification. If no early mapping has been conducted, in many applications, late mapping has to be done if the system should choose on-line between possible reactions. Late mapping is not necessary if we, for instance, are only interested in an 'emotion protocol' of conversations.

The *hinterland* is a less known, non-central, remote region beyond the coast or surrounding a town. In contrast to the big *n* emotions – anger, joy, sadness, despair, etc. – we can describe emotion-related, affective states such as interest, tiredness, etc. as constituting the hinterland of prototypical emotions ('proper' emotions), i. e. they are non-prototypical fringe instances. The same holds for *mixed* cases which can be combinations of – sometimes even antagonistic – different emotions, or *weak* instances of one emotion – which normally is as well a mixture, namely of the neutral, emotionally idle state with something else. Irrespective of the constituting nature of these events, they can be clearly recognizable or not, if speech is slurred, overlaid with technical noise, etc.

Emotions are *pervasive*, i. e. they are "[...] whatever is present in most life, but absent when people are emotionless [...]" [4], and *evasive* at the same time: pervasive, because in the non-prototypical case, there are many and frequent varieties, and evasive, because these varieties can neither be found, nor delimited from each other, easily. Thus life for the scientists is getting more troublesome if they do not preselect

nice, i. e. prototypical and clear, cases but record everything, i. e. in the case of the speech modality, if they use an *open microphone* setting; this challenge, however, is inevitable if we really want to employ emotion processing in real-life applications. Matters are similar in the other modalities that have been investigated such as facial gesture, gesture, body posture. When using such an 'open recording' setting, in all modalities, appropriate units have to be segmented. However, in emotion processing so far, normally a pre-segmented subset of a full recording is taken consisting of somehow clear, i. e. more or less prototypical cases with respect to inter-labeller agreement. This is not only a clever move to push classification performance, it simply has grown out from the problem of class assignment in emotion processing: there is no simple and unequivocal ground truth. Using 'realistic data', however, not only means using spontaneous data, it means as well using all these data as in a dialog system, media retrieval or surveillance tasks. This second, sort of 'quantitative aspect', has still been neglected by and large.

In spite of attempts towards defining the term 'emotion' in a more strict way in basic research, cf. [10], especially in application-oriented research on human-human and human-machine interaction, the term is used in a way that comprises all inhabitants of all hinterlands, i. e. all types of non-prototypical forms. Many studies, even – and maybe especially – those that want to define 'emotion' more strictly, used acted data although it is questionable that these acted productions really model realistic, non-prompted ones. Note that we do not plead in favour of abandoning all acted data: even in realistic scenarios, people can act in the sense of pretending being angry, to achieve their goal. Imagine a user calling a company and complaining about wrong delivery; it is at their liberty to be angry and show overt anger, to be angry but suppress this anger, not to be angry but pretend being it, or not to be angry and behave in a neutral way. We should not fall for a superhuman fallacy: a machine as dialogue partner will only be able to analyse the overt signs, be this signalled in the tone of voice (acoustics) or in the linguistic message, or in both. What we do plead for, however, is not to use prompted emotions but non-prompted data and recordings in realistic settings. Here, 'non-prompted' refers to recordings where we do not tell the subjects to act emotions, to behave emotionally, or to put themselves in some emotional mood.

Thus, after switching from acted to naturally occurring emotions and emotion-related states, from limited textual variation to spontaneous speech and reaching acceptable subject-independency, it is time to face one of the last barriers prior to integration of emotion recognition from speech into real-life technology: non-prototypicality in an open microphone setting. Crossing this barrier means facing a considerable performance loss, and means finding new paradigms to cope with this challenge as detection or spotting of emotion with potential garbage classes or decoding stages - potentially also including different quality measurement as ROC-curves with Equal Error Rates (EER) or the Area Under Curve (AUC), or correlation-based analyses.

In this paper, we want to deal with a highly realistic setting. We deal with non-prompted, spontaneous human-robot interaction. We do not aim at optimizing classification by using leave-one-speaker-out classification but partition the data into two independent sets for train and test with different room acoustics and (somehow) educational background – this can be seen if looking at the differences in vocabulary in the two sets, cf. below. We use all data recorded, by that simulating the open microphone setting in real applications. We do not use the spoken word chain but the results from automatic speech recognition (ASR). However, in order to be able to evaluate consistently our results, we do not use automatic but rule-based segmentation into processing units. Speech data, annotation, segmentation into meaningful units, and mapping onto two main classes is described in Section 2. Sections 3 and 4 describe our acoustic and linguistic feature vectors. Classifiers used and classification results are presented in Section 5. Correlation-based analyses are discussed in Section 6. In Section 7, we elaborate on different types of applications that require different types of decisions to be made in the classification or correlation space.

2. Database and annotation

The FAU Aibo Emotion Corpus comprises recordings of German children's interactions with Sony's pet robot Aibo; the speech data are spontaneous and emotionally coloured. The children were led to believe that the Aibo was responding to their commands, whereas the robot was actually controlled by a human operator. The wizard caused the Aibo to perform a fixed, predetermined sequence of actions; sometimes the Aibo behaved disobediently, thereby provoking emotional reactions. The data was collected at two different schools, MONT and OHM, from 51 children (age 10 - 13, 21 male, 30 female; about 8.9 hours of speech without pauses). Speech was transmitted with a high quality wireless head set and recorded with a DAT-recorder (16 bit, 48 kHz down-sampled to 16 kHz). The recordings were segmented automatically into 'turns' using a pause threshold of 1 s. Five labellers (advanced students of linguistics) listened to the turns in sequential order and annotated each word independently from each other as neutral (default) or as belonging to one of ten other classes. This procedure was iterative and supervised by an expert. Since many utterances are only short commands and rather long pauses can occur between words due to Aibo's reaction time, the emotional/emotion-related state of the child can change also within turns. Hence, the data is labelled on the word level. We resort to majority voting (MV): if three or more labellers agreed, the label was attributed to the word. In the following, the number of cases with MV is given in parentheses: *joyful* (101), *surprised* (0),

#	NEG	IDL	$\sum$
train	3 358	6 601	9 959
test	2 465	5 792	8 257
$\sum$	5 823	12 393	**18 216**

Table 1. Number of instances for the two classes

emphatic (2 528), *helpless* (3), *touchy*, i. e. irritated (225), *angry* (84), *motherese* (1 260), *bored* (11), *reprimanding* (310), *rest*, i. e. non-neutral, but not belonging to the other categories (3), *neutral* (39 169); 4 707 words had no MV; all in all, there were 48 401 words.

Classification experiments on a subset of the corpus [18, Table 7.22, p. 178] showed that the best unit of analysis is neither the word nor the turn, but some intermediate chunk being the best compromise between the length of the unit of analysis and the homogeneity of the different emotional/emotion-related states within one unit. Hence, manually defined chunks based on syntactic-prosodic criteria [18, Chap. 5.3.5] are used here. In contrast to other publications published recently, the whole corpus consisting of 18 216 chunks is used under the very same conditions as for the INTERSPEECH Emotion Challenge [15].

In this paper, we concentrate on the two-class problem consisting of the cover classes **NEG**ative (subsuming *angry*, *touchy*, *reprimanding*, and *emphatic*) and **IDL**e (consisting of all non-negative states); note that *emphatic* has to be conceived as a pre-stage of anger because on the valence dimension, it lies between neutral and anger, cf. [2]. A heuristic approach similar to the one applied in [18, Chap. 5.3.8] is used to map the raw labels of the five labelers on the word level onto one label for the whole chunk: If 50 % of these raw labels are **NEG**, then the whole chunk is labelled as **NEG**. Furthermore, the whole chunk is considered to be **NEG** as well if the following two conditions are fulfilled: 1) at least one third of all raw labels is **NEG**, and 2) the remaining raw labels are mostly pure *neutral*, i. e. at least 90 % of all raw labels are either negative (*angry*, *touchy*, *reprimanding*, *emphatic*) or *neutral*. By that a chunk is also considered to be **NEG** if only a few words are marked clearly as **NEG**. From a theoretical point of view, this actually makes sense; from a practical point, it helps to alleviate the problem of unbalanced classes. Nevertheless, the classes are still quite unbalanced since the whole corpus is used. Frequencies are given in Table 1. Speaker independence is guaranteed by using the data of one school (OHM, 13 male, 13 female) for training and the data of the other school (MONT, 8 male, 17 female) for testing.

As the label for the whole chunk is obtained by mapping the raw labels, i. e. the decisions of the five labelers on the word level, information of the prototypicality of the chunk is available. The prototypicality is defined as the proportion of raw labels matching the label for the whole chunk. There are two reasons for chunks with low emotional prototypicality:

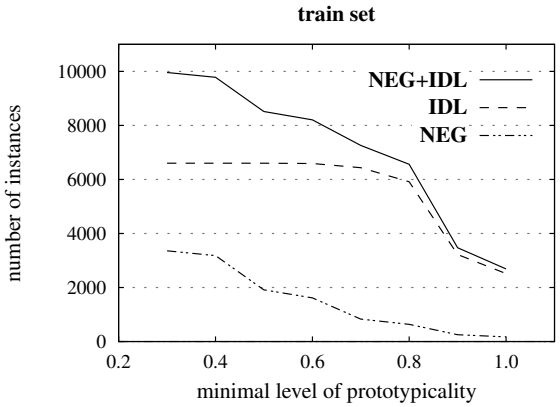

Figure 1. Number of instances in the train partition as function of minimal level of prototypicality.

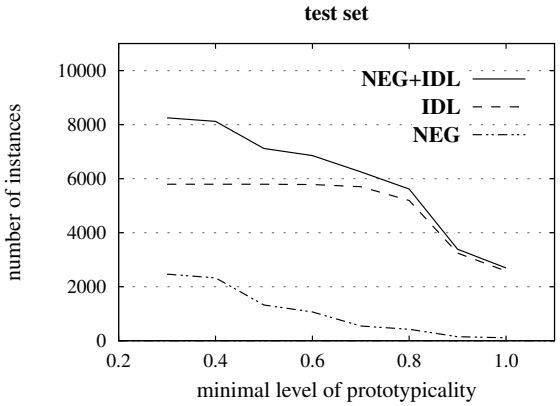

Figure 2. Number of instances in the test partition as function of minimal level of prototypicality.

1) not all words in a **NEG** chunk have to be **NEG** themselves; some words may also be produced in the state **IDL**. 2) Even if all words in a chunk are labeled as **NEG**, the agreement of the five labels for single words may be low, e. g. 3 out of 5. Of course, combinations of both phenomena can occur as well. Figures 1 and 2 depict the numbers of train and test instances as function of minimal prototypicality; for a given threshold, all chunks with a prototypicality lower than this threshold are discarded leading to less and less instances for higher levels of minimal prototypicality. In general, the prototypicality for **NEG** chunks is lower than for **IDL** chunks due to the chosen thresholds of the heuristic mapping algorithm. The distribution across levels of prototypicality is very similar for the train and the test set.

3. Acoustic features

A feature set is considered based on the findings in [11] by choosing the most common and at the same time promising feature types and functionals covering prosodic, spectral, and voice quality features. Further, we limit to a systematic generation of features using our open source feature

LLD (26 · 3)	**Functionals** (21)
(Δ/ΔΔ) ZCR	mean, abs. mean, centroid
(Δ/ΔΔ) DC, Min, Max	std. deviation, variance
(Δ/ΔΔ) RMS Energy	kurtosis, skewness
(Δ/ΔΔ) LOG Energy	*extremes:*
(Δ/ΔΔ) F0 frequency	value, rel. pos., range
(Δ/ΔΔ) F0 strength	*linear regression:*
(Δ/ΔΔ) F0 quality	offset, slope, MAE, MSE
(Δ/ΔΔ) HNR	*quadratic regression:*
(Δ/ΔΔ) MFCC 0-15	coeff. 1-3, MAE, MSE

Table 2. Acoustic features: low-level descriptors (LLD) and functionals.

extraction[1] [5]. In detail, the slightly extended set in comparison to [15] comprises of 26 low-level descriptors: dc offset (DC), extremes (Min, Max), and zero-crossing-rate (ZCR) from the time signal, root mean square (RMS) and logarithmic (LOG) frame energy, pitch (F0, normalised to 500 Hz), strength, and quality as well as harmonics-to-noise ratio (HNR) by autocorrelation function, and Mel-frequency cepstral coefficients (MFCC) 0-15 in full accordance to HTK-based computation. To each of these, the delta and double delta coefficients are additionally computed. Next the 21 functionals mean, absolute mean, standard deviation, variance, kurtosis, skewness, minimum and maximum value, relative position, and range as well as two linear and three quadratic regression coefficients with their mean absolute (MAE) and square (MSE) errors are applied on a chunk basis as depicted in Table 2. Thus, the total feature vector per chunk contains $26 \cdot 3 \cdot 21 = 1\,638$ attributes. More details on feature implementation are found in [5].

4. Linguistic features

Linguistic analysis is based on the bag of words approach [8]: the idea behind this approach is the representation of text in a numeric feature space. Each feature thereby represents the occurrence of a specific word in a sentence. In past works, we had successfully ported it to the field of emotion [14] and interest [13] recognition from text and speech.

In analogy to bag of words, the bag of n-grams approach also represents text in a numeric feature space. The main difference is the observation of a series of consecutive words as semantic units of interest [12]. Whereas bag of words observes only single words for the mapping to a numeric feature space, the idea behind bag of n-grams provides a simple extension by observing n-grams of words. For n=1, only single words are observed which directly corresponds to the bag of words. For n=2, however, word bi-grams are observed, for n=3 tri-grams, etc. The approach allows to observe several n-grams together, determined by a minimum and a maximum n-gram length, similar to 'backing-off'. For each of these text units, the generation of a numeric feature is computed exactly as for the bag of words.

The search for good bag of n-grams parameters actually is not trivial because of the many different influence factors. However, we found that term frequency, inverse document frequency, chunk length, binary, and case lowering transformations had no influence for the corpus at hand: the average length of a chunk in terms of the number of words is as low as 2.66; **IDL** chunks are 2.82 words long on average, **NEG** chunks only 2.30 words. Only little influence of stemming was observed, which is why it is not used in the experiments. An optimum was further found for the n-gram length of one to three words, while one to two words produced only slightly lower results.

To obtain the spoken word chain from the speech signal, we use the ASR engine that has been developed within our speech group at the University Erlangen-Nuremberg. A recent overview is given in [19]. The acoustic features are the first 12 standard MFCC features (the first MFCC coefficient is replaced by the sum of the energies of the 22 Mel filterbanks) and their first derivatives. The features are computed every 10 ms over a Hamming window of 16 ms. Our ASR system is based on semi-continuous hidden Markov models (SC-HMM) modelling polyphones, i. e. an extension of the well-known triphones to model large context sizes. A polyphone is modelled by its own HMM if it can be observed at least 50 times in the training set. All HMM states share the same set of Gaussian densities; the size of the codebook is 500. By that, a smaller number of densities can be used, which is beneficial if – as in our case – only limited training data is available. Yet, full covariance matrices are used in contrast to most systems based on continuous HMMs. We use Baum-Welch re-estimation for training and Viterbi decoding. As language model we use back-off bi-grams. It is interesting to note that the higher level education school's children comprised in the training partition have a higher vocabulary of 703/253 words/fragments as opposed to the test set's vocabulary size at 383/158. The vocabulary of the ASR system consists of all words (but no word fragments) of both the training and the test set; all in all 813 words. Hence, 158 vocabulary words (types) of the test set are out of vocabulary (OOV), which amounts to a total of 2.1 % OOV events (tokens). The ASR engine, trained on the train set, yields a word accuracy of 77.48 % for the test set. To ensure that the linguistic emotion model is trained on the same type of phenomena, i. e. ASR errors, it has to face when dealing with the ASR output of the test set, we trained and subsequently tested the ASR engine on the train set, obtaining a word accuracy of 76.42 %. This ASR output of the train set is used for the training of the linguistic emotion model.

[1] http://sourceforge.net/projects/openSMILE

5. Classification analysis

The classifier of choice in this article is a discriminatively learned simple Bayesian Network, namely Discriminative Multinomial Naive Bayes (DMNB) [20], for the large feature space tasks together with Support-Vector Machines (SVM) and Random Forests for late fusion as also applied in our previous investigations [11,15]. The reason is two-fold: first, the mean recall values resulted in a slight absolute improvement over SVM in our experiments on the FAU Aibo Emotion Corpus: an improvement of 1.90 % / 2.01 % and 0.066 (unweighted/weighted average recall and AUC, cf. below) for acoustic features, and one of 2.05 % / -0.02 % and 0.059 for linguistic features. At the same time, DMNB requires lower memory and only a fraction of the computation time of SVM – Sequential Minimal Optimisation training of SVM with linear kernel demanded 200 times higher computation time than DMNB in parameterisation as below using [21] on an 8 GB RAM, 2.4 GHz, 64 Bit industry PC. Second, the parameter learning is carried out by discriminative frequency estimation, whereby the likelihood information and the prediction error are considered. Thus, a combination of generative and discriminative learning is employed. This method is known to work well in highly correlated spaces (as in our case), to converge quickly, and not to suffer from over-fitting.

For optimal results we found it best to ignore the frequency information in the data and select a number of ten iterations for acoustic processing and one iteration for linguistic processing. Numeric variables are discretised using unsupervised ten-bin discretisation [21].

As mentioned above and carried out within [15], we split the FAU Aibo Emotion Corpus into train and test partitions by schools of recording. Thus, utmost independence of the speaker, room acoustics, general prosody and articulation patterns, and wording of the children is ensured. To better cope with this variety, all acoustic features are standardised per partition ('speaker group normalisation'). Due to the high imbalance among classes (cf. Table 1), balancing of the training instances is further mandatory to achieve reasonable values of unweighted recall and thus avoid overfitting of the strong **IDL** class [13]. The chosen straight-forward strategy is random up-sampling of the sparse **NEG** instances enforcing unit distribution while slightly increasing the total number of instances to 134 %. Note that the order of operations has an influence on (un)weighted recall figures [15]: we first balance and then standardise the training. Interestingly, only acoustic features benefited from balancing in terms of unweighted average recall (cf. below). Next we classify with DMNB as described.

To carry out late fusion of acoustic and linguistic predictions, we employ late fusion by a meta-classifier that learns 'which stream' to trust 'when'. Here, StackingC with linear regression on meta-level [21] and Support Vector Machines, Random Forests (30 trees), and the Naive Bayes on base-level serve as optimal choice over the base classifiers and further variants. To train, we needed to split the train partition into two: fold 1 comprises the first 13 speakers of the train set, i. e. 4 921 instances. Fold 2 consists of the remaining 13 speakers, and 5 038 instances. Acoustic and linguistic predictions were produced separately with the same settings as before – DMNB with 10 iterations for acoustics with up-sampling to uniform distribution (132 % for fold 1, 137 % for fold 2) and subsequent standardisation, DMNB with one iteration for linguistics with no processing – in cross-manner to obtain training predictions for the fusion. The measured performance for fold 1 is 70.40 % / 70.96 % and 0.764 for acoustic features, and 64.60 % / 69.61 % and 0.738 for linguistic features for unweighted/weighted average recall and AUC (cf. below). For fold 2, 64.60 % / 69.61 %, and 0.738 and 64.30 % / 69.38 % and 0.738 were observed, accordingly. Next, the likewise produced predictions per class (**IDL**, **NEG**) and the predicted label as binary feature by using exclusively the train partition were employed to fuse on basis of the formerly obtained predictions on the test partition; these results are shown in Table 3.

recall [%]	**UA**	**WA**
acoustic features	68.30	65.97
linguistic features	66.05	67.87
late fusion	**69.30**	71.47

Table 3. Recognition results for the test set by unweighted (UA) and weighted (WA) average recall for acoustic features, linguistic features, and late fusion.

Figures 3 and 4 display unweighted average recall (UA) – or 'class-wise' computed recognition rate (CL), i. e. the mean along the diagonal of the confusion matrix in percent, and weighted average recall (UA), i. e. the overall recognition rate (RR) or recall (number of correctly classified cases divided by total number of cases), for three sets of features: only acoustic features, only linguistic features, and both acoustic and linguistic predictions (late fusion).

Another method of assessing a classifier is the receiver operating characteristic (ROC) – a graphical method derived from signal detection theory [6]. The ROC plots the true positive rate (*TPR*) over the false positive rate (*FPR*) achieved by a binary classifier. The diagonal line *TPR* = *FPR* corresponds to a classifier that randomly guesses the positive class r percent of the time, resulting in a point (r, r) in the ROC space. Informally, the goal is to optimize a classifier towards producing a point in the upper triangle, close to the upper left corner of the graph. Such a classifier has a high *TPR* and a low *FPR* at the same time, the point $(0, 1)$ denoting a 'perfect' classifier. Classifiers that appear close to the x-axis can be thought of as 'conservative': they make few false positive errors, but they often have low true positive rates at the same time. Classifiers that produce a point on the upper right-hand side can be thought of as 'liberal':

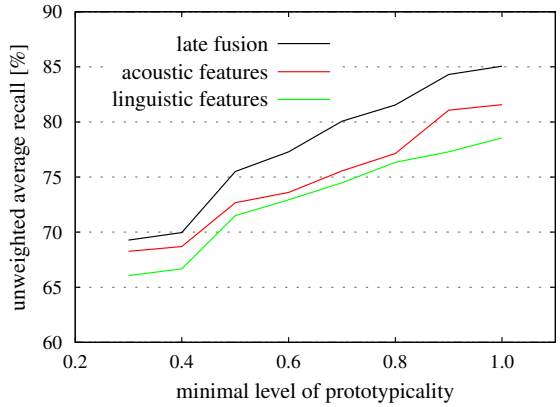

Figure 3. Unweighted average recall as function of minimal level of prototypicality in the test set. Depicted are acoustic and linguistic analyses and late fusion of these.

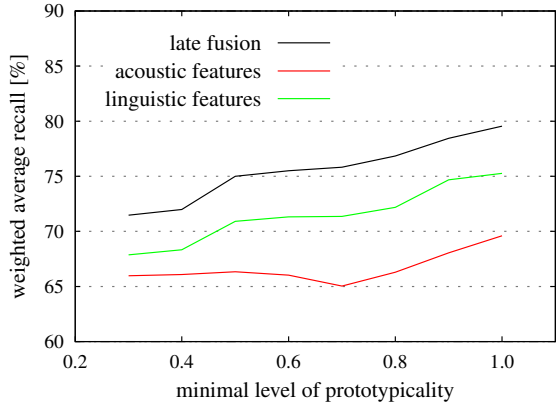

Figure 4. Weighted average recall as function of minimal level of prototypicality in the test set. Depicted are acoustic and linguistic analyses and late fusion of these.

they classify almost all positives correctly, but often at the cost of a high false positive rate [6]. In contrast to a binary classifier, the classifiers used in this work yield a score value for each instance, similar to a probability. The score can be used together with a discrimination threshold to produce a binary classifier. Here, we use the a-posteriori score of 0.5 for the results in Table 3, cf. also below Figure 7. Varying the threshold results in a number of points that form a curve in the upper triangle of the graph area.

Figure 5 depicts our ROC obtained for the detection of **NEG** for acoustic and linguistic analyses as well as their fusion. When comparing classifiers, we may want to reduce the two-dimensional ROC curve to a single scalar value. A common method is to calculate the area under the ROC curve, called AUC. The highest possible AUC is 1.0, equal to the whole graph area, and achievable only by a 'perfect' classifier. Random guessing has an AUC of 0.5 since it corresponds to the diagonal line in the ROC space. A reasonable classifier should therefore have an AUC that is significantly

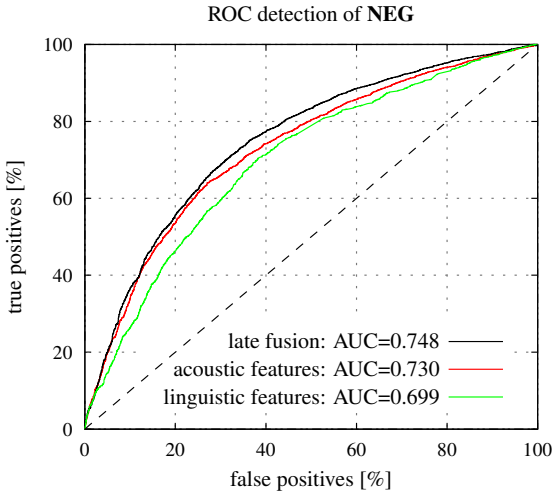

Figure 5. Receiver operating characteristic (ROC) for the detection of **NEG** instances for acoustic and linguistic analyses and late fusion of these.

greater than 0.5, with better classifiers yielding higher values. The values obtained are also shown in Figure 5.

6. Correlation analysis

So far, we have shown in a post festum analysis that instances with higher prototypicality are really classified better than those with a lower one, cf. Figures 3 and 4. In real-life, on-line processing, such an evaluation is of course not possible. However, for this classifier, the a-posteriori scores are available, a sort of confidence value between 0.0 and 1.0: the more extreme the value, i. e. closer to 0.0 or to 1.0, the more certain should the classifier be to have found the correct class assignment. Figure 6 displays the histogram of the a-posteriori scores obtained by late fusion of the acoustic and linguistic features; the black bars show the distribution of all instances of the test set (both **NEG** and **IDL**), the red bars (dotted lines, lower bars) show the distribution of only those instances that are labelled as **NEG**. As can be seen, the classifier is never 'absolutely sure': there are no values lower than 0.1 and higher than 0.75. The distribution for 'total' is more or less U-shaped, and the one for **NEG** predictions is left-skewed. Instances with a-posteriori scores higher than 0.5 are classified as **NEG**. Hence, the proportion of instances that actually belong to class **NEG** (ratio of the size of the red bar and the one of the black bar) should increase for higher a-posteriori scores. Based on the distribution given in Figure 6, Figure 7 displays the precision for **IDL** (a-posteriori scores below 0.5) and **NEG** (a-posteriori scores above 0.5). Note that the distribution at the 'turning point' at 0.5 is the basis for the classification results reported in Table 3. We thus can safely conclude that only relying on cases with more extreme a-posteriori scores (i. e. higher ones in the case of **NEG**) really yields better precision. Note that this means

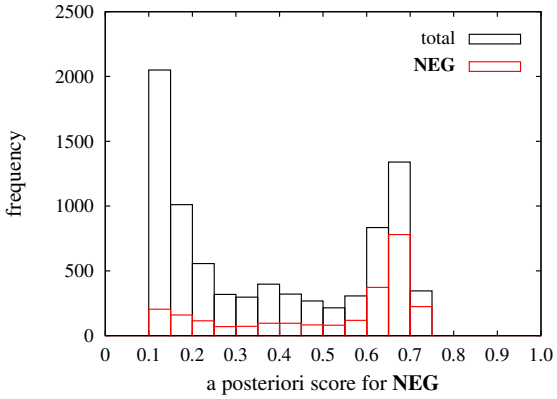

Figure 6. Distribution of late fusion a-posteriori scores for **NEG** instances; black lines (higher bars) all instances classified as **NEG**, i. e. scores above 0.5; red instances (lower bar) all instances labelled as **NEG**, i. e. reference.

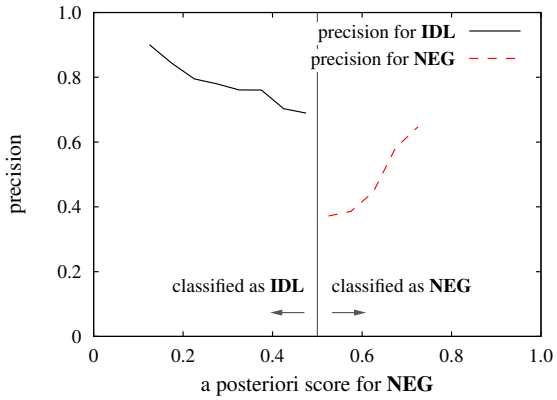

Figure 7. Precision for **IDL** (a-posteriori scores below 0.5) and **NEG** (a-posteriori scores above 0.5), based on the distribution displayed in Figure 6.

a lower number of **NEG** cases to be recognized as **NEG**, but a higher reliability of these decisions. These results corroborate earlier findings on prototypes found for the same database, cf. [3, 16] using related but not identical emotion classes and the spoken word chain.

We measured the correlation of the degree of prototypicality (DOP) as defined above with the a-posteriori scores of the test partition instances obtained by late fusion for **NEG** items, observing a positive and significant, albeit not very high correlation coefficient (CC) of 0.44. We now want to find out whether a regression can automatically reach this CC between the predicted scores and the continuous DOP, i. e. we investigate whether prototypicality can be determined as additional measure over the sheer binary classification result investigated so far. We thus carry out late fusion as described in Section 5 but with the DOP as target instead of the class label. SVM regression with linear kernel proved a good choice and led to a CC of 0.46 and mean linear error (MLE) of 0.18 (cf. [13]). While this number could probably

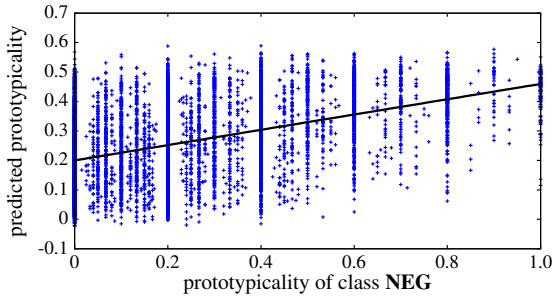

Figure 8. SVM regression: late fusion on class predictions.

be improved by producing regression predictions of acoustic and linguistic features for the fusion, it seems noteworthy that the calculated correlation between scores and degree of prototypicality can be reached by learning. Figure 8 displays the predicted over the actual DOP by plotting each instance of the test partition as individual '+'. The strong vertical lines at 0, 0.2, ... , 1.0 stem from the fact that five labellers are contained, and thus significant maxima are present in the histogram of the train partition instances (other values can exist as the labellers annotated each word of a chunk. However, apparently higher inter-labeller disagreement is present than intra-labeller and intra-chunk changes of labels). These are thus also assigned more likely. The upward linear trend line clearly indicates that DOP indeed can be predicted to a certain degree. Interestingly the MLE of 0.18 lies close to the above-named steps of 0.2, i. e. on average the DOP is predicted wrongly by 'one labeler'.

7. Discussion and concluding remarks

Classification performance obtained is above the baseline proposed in [15] but, due to our open microphone setting, of course lower than for prototypical and selected classes. Note that linguistic features, based on ASR and not on the spoken word chain, contributed to the higher performance in late fusion. Given the less pronounced characteristics and, consequently, the low inter-labeller consistency of our classes, it is not very likely that some fancy new classification procedure will yield much higher performance. We therefore presented results analysing different degrees of prototypicality, and procedures that have not been frequently used in emotion processing so far, such as ROC analysis. For both procedures, we can choose different thresholds, suited for different application scenarios. Let us assume that we are interested in **NEG**: taking into account a-posteriori scores (cf. Figure 7) we can choose a higher degree of prototypicality, or we can choose either a lower value for *TPR* and by that, a low value for *FPR*, or a higher value for both *TPR* and *FPR*, cf. Figure 5. In applications, we have to tell apart single instance detection in on-line interaction from post festum, summarizing estimations, e. g. of the proportion of **NEG** instances in an interaction. Let us come back to the call

center scenario addressed in Section 1: if on-line detection of **NEG** is critical because the user might get upset being told that she is angry but she is not, we should aim at high prototypicality and low false alarm rate. If it is critical not to detect an angry user, we should aim at a high *TPR*; in the case of a high *FPR*, the action taken by the system should be to pass the conversation on to a human operator – this procedure is costly but not detrimental. If we are interested in off-line analyses of the felicity of call center interactions, we can live with a high amount of false instances as long as we get a reliable estimate for the quality of each interaction in total [1].

We have shown both categorical analyses based on classification results and dimensional analyses based on correlation procedures. As discussed in Section 1, it will be a matter of different types of applications whether to choose a categorical or a dimensional analysis, and whether to employ early quantisation (i. e. categories), or late quantisations (i. e. dimensional values that eventually are transferred onto categories), or no quantisation at all. We do not know much on the performance of these alternatives yet; this is an interesting topic to be addressed in future research.

8. Acknowledgments

The research leading to these results has received funding from the European Community under grant No. IST-2001-37599 (PF-STAR), grant No. IST-2002-50742 (HUMAINE), and grant (FP7/2007-2013) No. 211486 (SEMAINE). The responsibility lies with the authors.

References

[1] A. Batliner, F. Burkhardt, M. van Ballegooy, and E. Nöth. A Taxonomy of Applications that Utilize Emotional Awareness. In *Proceedings of IS-LTC 2006*, pages 246–250, Ljubliana, 2006.

[2] A. Batliner, S. Steidl, C. Hacker, and E. Nöth. Private Emotions vs. Social Interaction – a Data-Driven Approach towards Analysing Emotions in Speech. *User Modeling and User-Adapted Interaction*, 18:175–206, 2008.

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

[4] R. Cowie, N. Sussman, and A. Ben-Ze'ev. Emotions: concepts and definitions. In P. Petta, editor, *HUMAINE handbook on emotion*. Springer, 2009. to appear.

[5] F. Eyben, M. Wöllmer, and B. Schuller. openEAR – Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. In *Proc. ACII*. IEEE, 2009. this volume.

[6] T. Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[7] B. Fehr and J. A. Russel. Concept of emotion viewed from a prototype perspective. *Journal of Experimental Psychology: General*, 113:464–486, 1984.

[8] T. Joachims. 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, 1998. Springer, Heidelberg.

[9] E. Rosch. Cognitive Representations of Semantic Categories. *Journal of Experimental Psychology: General*, 104(3):192–233, 1975.

[10] K. R. Scherer. Emotion. In M. Hewstone and W. Stroebe, editors, *Introduction to Social Psychology: A European perspective*, pages 151–191. Blackwell, Oxford, 2000.

[11] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson. 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, 2007.

[12] B. Schuller, A. Batliner, S. Steidl, and D. Seppi. Emotion Recognition from Speech: Putting ASR in the Loop. In *Proc. ICASSP*, pages 4585–4588, Taipei, Taiwan, 2009. IEEE.

[13] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu. Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-Life Application. *Image and Vision Computing Journal (IMAVIS), Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior*, 2009. 17 pages, in print.

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

[15] B. Schuller, S. Steidl, and A. Batliner. The INTERSPEECH 2009 Emotion Challenge. In *Proc. Interspeech*, Brighton, UK, 2009.

[16] D. Seppi, A. Batliner, B. Schuller, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, and V. Aharonson. Patterns, Prototypes, Performance: Classifying Emotional User States. In *Proc. Interspeech*, pages 601–604, Brisbane, 2008.

[17] P. Shaver, J. Schwartz, D. Kirson, and C. O'Connor. Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology*, 52:1061–1086, 1987.

[18] S. Steidl. *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. Logos Verlag, Berlin, 2009.

[19] G. Stemmer. *Modeling Variability in Speech Recognition*. Logos Verlag, Berlin, 2005.

[20] J. Su, H. Zhang, C. X. Ling, and S. Matwin. Discriminative Parameter Learning for Bayesian Networks. In *Proc. ICML*, pages 1016–1023, Helsinki, 2008.

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