Cross-Corpus Classification of Realistic Emotions – Some Pilot Experiments

Florian Eyben[1], Anton Batliner[2], Björn Schuller[1], Dino Seppi[3], Stefan Steidl[2]

[1]Institute for Human-Machine Communication, Technische Universität München,
[2]Pattern Recognition Lab, FAU Erlangen, [3]ESAT, Katholieke Universiteit Leuven,
[1]München, Germany, [2]Erlangen, Germany, [3]Leuven, Belgium
eyben@tum.de, batliner@informatik.uni-erlangen.de

Abstract

We use four speech databases with realistic, non-prompted emotions, and a large state-of-the-art acoustic feature vector, for cross-corpus classifications, in turn employing three databases for training and the fourth for testing. Categorical and continuous (dimensional) annotation is mapped onto a representation of valence with the three classes positive, neutral, and negative. This cross-corpus classification is compared with within corpus classifications. We interpret performance and most important features.

1. Introduction

The normal approach towards classifying emotion in speech is to subdivide one corpus into specific train, validation and test subsets, in the case of cross-validation with or without using specific validation sets. By that, many intervening variables such as microphone, room acoustics, speaker group, etc., are kept constant. However, we always have to keep in mind that we rather cannot generalize onto other corpora and settings when using this approach. A first step towards overcoming such restrictions and thus evaluating recognition of realistic emotions in a scenario which itself is more realistic, is doing cross-corpus classification. This will be pursued in this paper. First, in section 2., we introduce the four naturalistic emotion corpora used in this study. In section 3., we describe our acoustic feature set, and in section 4., we present results and describe the evaluation methods. Concluding remarks are given in section 5.

2. Corpora

Table 1 shows the basic statistics of the four naturalistic emotion corpora used in this study, namely the SmartKom Corpus (SmK), the FAU Aibo Emotion Corpus (Aibo), the Sensitive Artificial Listener Corpus (SAL), and the Vera-am-Mittag Corpus (VAM). One of the main difficulties of cross-corpus experiments in this field, besides the different content and acoustics, is the mismatch of annotations with respect to the labels considered. Each corpus was recorded more or less for a specific task – and as a result of this, they have specific emotion labels assigned to them. For cross-corpus recognition this poses a problem, since the training and test sets in any classification experiment must use the same class labels. This is especially problematic for corpora where annotations are made in terms of discrete class labels, such as SmartKom and Aibo. Corpora annotated in terms of affect dimensions such as valence and arousal are easier to match, although per corpus biases and different ranges can pose a problem.

In order to be able to perform cross-corpus valence recognition in this study, a standard set of classes has been defined for all corpora; we decided for three levels of valence: negative, idle (i. e. neutral), and positive. The labels used in each corpus can be mapped onto these three classes. This mapping can be found in the following subsections for each corpus. Moreover, a description of the notion of 'turn', which is used as unit of analysis, can be found in the corpus documentation in the following subsections. Table 1 reveals that there is indeed a considerable variation in turn duration and by that, most likely in consistency of valence.

2.1. SmartKom

SmartKom (SmK) is a multi-modal German dialogue system which combines speech with gesture and facial expression. The so called SmartKom-Public version of the system is a 'next generation' multi-modal communication telephone booth. The users can get information on specific points of interest, as, e.g., hotels, restaurants, cinemas. They delegate a task, for instance, finding a film, a cinema, and reserving the tickets, to a virtual agent which is visible on the graphical display. Users get the necessary information via synthesised speech produced by the agent, and on the graphical display, via presentations of lists of points of interest (e. g. hotels, restaurants, and cinemas), and maps of the inner city. For this system, data are collected in a large-scaled Wizard-of-Oz (WoZ) experiment. The dialogue between the (pretended) SmK system and the user is recorded with several microphones and digital cameras. Subsequently, several annotations are carried out. The recorded speech represents thus a special variety of non-prompted, spontaneous speech typical for man-machine-communication in general and for such a multi-modal setting in particular. More details on the recordings and annotations can be found in (Steininger et al., 2002; Batliner et al., 2003). The labellers could look at the persons' facial expressions, body gestures, and listen to his/her speech; they annotated the user states *joy/gratification, anger/irritation, helplessness, pondering/reflecting, surprise, neutral*, and *unidentifiable* episodes. *Joy* and *anger* were subdivided into the subclasses *weak* and *strong joy/anger*. The labelling was frame-based, i. e. beginning and end of an emotional episode was marked on the time axis. Turns are defined as dialogue moves, i. e. as everything produced by the user until the system takes over.

We mapped the class *anger* to negative valence (**N**), the classes *helplessness, pondering*, and *neutral* to neutral va-

From: Proc. LREC 2010, ELRA, Valletta, Malta

Corpus	# of instances				Turn duration (s)			
	P	**I**	**N**	Overall	Mean	Stddev.	Min	Max
SmK	353	2963	219	3535	6.8	7.1	0.1	64.2
Aibo	495	11021	2215	13731	2.3	1.5	0.9	38.0
SAL	466	588	638	1692	3.5	3.0	0.9	26.8
VAM	16	511	420	947	3.0	2.2	0.4	17.7

Table 1: Number of instances in each of the four corpora; distribution of instances among the 3 valence classes (**N**: negative valence, **I**: idle, i. e. neutral valence, **P**: positive valence), and mean, minimum, and maximum turn duration per corpus.

lence (**I**), and *joy* and *surprise* to positive valence (**P**). The unidentifiable episodes were ignored, since they might contain episodes with positive or negative valence which could not be mapped onto the pre-defined classes.

2.2. Aibo

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. This WoZ 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 > 1 s). 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 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. 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), *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. *reprimanding*, *touchy*, and *angry* were mapped onto a main class *angry*. The mapping of word- onto turn-labels is described in (Steidl, 2009).

We map (based on the turn labels) the classes *angry* and emphatic to negative valence (**N**), the classes *neutral and rest* to neutral valence (**I**), and *motherese and joyful* to positive valence (**P**). *Helpless*, *surprised*, and *bored* did not occur amongst the turn based labels.

2.3. SAL

The Belfast Sensitive Artificial Listener (SAL) data is part of the final HUMAINE database (Douglas-Cowie et al., 2007). We consider the subset used e. g. in (Wöllmer et al., 2008) which contains 25 recordings in total from four speakers (2 male, 2 female) with an average length of 20 minutes per speaker. The data contains audio-visual recordings from human-computer conversations (WoZ scenario) that were recorded through a SAL interface designed to let users work through a range of emotional states. The data has been labelled continuously in real time by four annotators with respect to valence and activation using a system based on FEELtrace (Cowie et al., 2000): the annotators used a sliding controller to annotate both emotional dimensions separately whereas the adjusted values for valence and activation were sampled every 10 ms to obtain a temporal quasi-continuum. To compensate linear offsets that are present among the annotators, the annotations were normalised to zero mean globally. Further, to ensure common scaling among all annotators, each annotator's labels were scaled so that 98 % of all values are in the range from –1 to +1. The 25 recordings have been split into turns using an energy based Voice Activity Detection. Accordingly, a total of 1 692 turns is contained in the database. Labels for each turn are computed by averaging the frame level valence and activation labels over the complete turn.

We define the classes negative valence (**N**) for turns with an annotated valence below -0.25, neutral valence (**I**) from -0.25 to 0.25, and positive valence (**P**) for turns with an annotated valence above 0.25.

2.4. VAM

The Vera-Am-Mittag (VAM) corpus (Grimm et al., 2008) consists of audiovisual recordings taken from a German TV talk show. The set used contains 947 spontaneous and emotionally coloured turns from 47 guests of the talk show which were recorded from unscripted discussions. The topics were mainly personal issues such as friendship crises, fatherhood questions, or romantic affairs. To obtain non-acted data, a talk show in which the guests were not being paid to perform as actors was chosen. The speech extracted from the dialogues contains a large amount of colloquial expressions as well as non-linguistic vocalisations and partly covers different German dialects. For annotation of the speech data, the audio recordings were manually segmented into turns, each utterance containing at least one phrase. A large number of human labellers was used for annotation (17 labellers for one half of the data, six for the other). The labelling bases on a discrete five point scale for three dimensions mapped onto the interval of [–1,1]: the average results for the standard deviation are 0.29, 0.34, and 0.31 for valence, activation, and dominance. The averages for the correlation between the evaluators are 0.49, 0.72, and 0.61, respectively. The correlation coefficients for activation and dominance show suitable values, whereas the moderate value for valence indicates that this emotion primitive was more difficult to evaluate; it may partly also

be a result of the smaller variance of valence.
As for the SAL corpus, we define the classes negative valence (**N**) for turns with an annotated valence below -0.25, neutral valence (**I**) from -0.25 to 0.25, and positive valence (**P**) for turns with an annotated valence above 0.25.

2.5. Inconsistencies across corpora

When doing cross-corpus classification using these four corpora, we are facing several problems and inconsistencies that most certainly will not always be favourable for our classification performance:

- Three of the corpora are German, one, i. e. SAL, is English.
- In three corpora, speakers are adult, whereas in Aibo, speakers are children.
- The scenarios differ in several respect: SmK is about information queries in a human-WoZ setting; Aibo is about giving commands to a pet robot, again in a WoZ setting – however, the WoZ (Aibo) never talks; SAL is about the interaction with an emotional agent, again in a WoZ setting; VAM is about human-human interaction in an 'emotion-prone' talk show.
- Subjects are 'naive' in SmK and in Aibo, experts in SAL, and most likely belonging to some specific type of personality in VAM.
- Number of subjects, of labellers, and of items per class can differ considerably.
- The original units of annotation differ: frames in SmK, words in Aibo, turns in SAL and in VAM; different types of mappings onto the turn level had to be performed. Certainly, this goes along with less clear, 'smeared' classes, although this might have different impact in each of our four corpora.
- The emotional taxonomies differ: categories in SmK and in Aibo, dimensions in SAL and in VAM; again, we had to perform different types of mapping onto our three mutually exclusive valence classes. As a consequence, a few 'garbage' turns had to be mapped or skipped on a somehow arbitrary basis.
- Last but not least, valence – both as dimension or as categories – is notoriously more difficult to process and classify than, e. g. arousal, when only acoustic information is used, because a straightforward equation such as 'higher/longer/stronger means higher arousal, and vice versa' cannot be used.

3. Acoustic features

We use a set of 2 832 acoustic features extracted with the openEAR toolkit (Eyben et al., 2009). Thereby 59 acoustic low level descriptor (LLD) contours (see table 2) are computed at a rate of one every 10 ms. A Gaussian window ($\sigma = 0.25$) of size 50 ms is used for all LLD except for pitch and formants, where a window size of 75 ms is preferred. A pre-emphasis with a factor of $k = 0.97$ is applied to the 50 ms frames, and a de-emphasis with factor $k = 0.92$ is applied to the 75 ms frames.
First order delta regression coefficients are computed from all 59 LLD contours resulting in 118 LLD features in total. After applying the 24 functionals described in table 3 to each of the LLD, a 2 832 dimensional vector is obtained for each input instance (turn).

Feature Group	**Features in Group**
Pitch	F_0 in Hz via sub-harmonic sampling (F_0), smoothed F_0 contour (Hz) ($F_0 env$)
Energy	Intensity ($Intens$)
Formant	Formant frequency ($freq$) and bandwidth (bw) of F_1 to F_4 via LPC analysis, LPC gain
Voice Quality	Probability of voicing (p_{voice}), local Jitter (Jit_{loc}), differential Jitter (Jit_D), local Shimmer (Shi_{loc})
Spectral	Centroid, Entropy, Flux 90 % roll-off point (rop) and position of highest peak in spectrum ($specMaxPos$).
Mel-bands	Mel-frequency-bands (MFB) 0-25 (20-8000 Hz)
Cepstral	MFCC 1–12

Table 2: Set of 59 Low-Level Descriptors (LLD).

Functionals	**Abbrv.**
Maximum and Minimum value	max/min
Range (Max.–Min.)	range
Arithmetic Mean (of non-0 values)	(nz)amean
Relative pos. of global max. value	maxpos
Centroid	centroid
Linear regression coefficients and corresp. quad. approximation error	qregc1–3 qregerr
Quadratic regression coefficients and corresp. quad. approximation error	linregc1–2 linregerr
Number of non-zero values	nnz
Standard deviation	stddev
Skewness, kurtosis	skew/kurt
Number of peaks	numPeaks
Arithmetic mean of peaks	peakMean
Mean distance between peaks	meanPeakDist
Rel. time below 25% of range	downleveltime25
Rel. time above 75%/90% of range	upleveltime75/90

Table 3: Set of 24 functionals applied to LLD contours and delta coefficients of LLD contours. Abbreviations as used in the following tables.

4. Classification and Results

In total we perform three experiments: within-corpus classification, cross-corpus classification (leave one corpus

[UAR %]	PI-N	P-IN	PN-I	P-N	Avg.
SmK	50.7*	49.5	56.3*	58.9	53.9
Aibo	57.9*	57.6*	59.9*	76.0*	62.9*
SAL	61.9	53.1	46.5	69.5*	57.8
VAM	60.2*	(50.0)	58.1	(50.0)	(54.6)
Avg.	57.8*	52.5	55.2	63.6	57.3

Table 4: Within corpus UAR obtained on four corpora with SVM (SMO). * indicates significant improvement ($\alpha = 0.01$) over random guess. 10-fold SCV.

[UAR %]	PI-N	P-IN	PN-I	P-N	Avg.
SmK	51.1*	52.4	47.9	55.0*	51.6
Aibo	52.0	55.7	50.7	54.9*	53.3
SAL	52.2*	49.0	51.9	48.6	50.4
VAM	56.2*	63.4*	53.8*	59.1*	58.1*
Avg.	52.9	55.1	51.1	54.4*	53.4

Table 5: cross-corpus UAR obtained on four corpora as test sets (leave-one-corpus-out) with SVM (SMO). * indicates significant improvement ($\alpha = 0.01$) over random guess.

out), and cross-corpus feature ranking. For establishing a coarse, preliminary reference for within corpus classification, we perform 10-fold cross validation (note that this is not speaker independent). We use Support Vector Machines (SVM) trained with the Sequential Minimal Optimisation algorithm (SMO) as implemented in WEKA 3 (Witten and Frank, 2005) for all experiments. In order to obtain a somewhat generalised and classifier independent feature ranking, we use Discriminative Multinominal Bayes (DMNB) (Su et al., 2008) and Support Vector Machines as implemented by LibSVM (Chang and Lin, 2001) in addition to the SMO SVM.

To investigate the performance for classifying different aspects of valence independently, we map the three classes **P**, **I**, and **N** onto four binary class sets: **P** and **I** vs. **N** (**PI-N**), **P** vs. **N** (**P-N**), **P** vs. **I** and **N** (**P-IN**), and **P** and **N** vs. **I** (**PN-I**). Doing that, on the one hand we subdivide the valence axis at two different points: between positive and rest (**I** and **N**), and between negative and rest (**I** and **P**). On the other hand we know that neutral (**I**) is very often confused with either positive or negative, cf. (Batliner et al., 2008), thus we evaluate the performance of contrasting **P** vs. **N** leaving aside **I**, and telling apart **I** from emotional (**P** and **N**).

4.1. Within corpus evaluation

For each of these four sets we compute within corpus recognition results in terms of unweighted average class-wise recall rate (UAR) by 10-fold stratified cross validation (SCV); UAR is computed as the mean value of the numbers of correctly recognised instances per class divided by the total number of instances per class. By that the resulting numbers are not biased by the distribution of instances among classes. These results are given in table 4. For this preliminary within corpus classification, we decided not to balance the number of instances for this experiment, since for VAM only 16 **P** instances exist and thus balancing via sub-sampling is not feasible. This, however, yields non-informative results for the sets **P-IN** and **P-N** on the VAM corpus (last line of table 4). Leaving aside VAM, the quality of the performance is positively correlated with the turn length (cf. table 1): the shorter the turns are, the more likely it is that they are emotionally consistent, i. e. that the emotion is constant throughout the turn. Note that due to the small number of corpora, this is no hard proof yet but an indication worthwhile to be pursued further.

4.2. Cross-corpus evaluation

Next, we report cross-corpus results in table 5. One of four corpora was used for testing while the other three were combined as training set (the sets are speaker disjunctive, thus the results indicate speaker and corpus independent performance). For this experiment, the training set is balanced by randomly sub-sampling all classes to the number of instances in the smallest class. The distribution of instances among classes in the test set, however, is not balanced. Thus, we prefer the unweighted average class-wise recall rate (UAR) as an evaluation metric.

In contrast to within-corpus classification, there is no clear-cut correlation between performance and emotional consistency. This could be expected because training and test set differ with respect to several factors as detailed in section 2.5. Moreover, the average length of turns differ within the training set and across training and test set.

4.3. Cross-corpus feature ranking

The two experiments described so far were conducted with the full set of 2 832 features. We now want to find generic, corpus independent acoustic features relevant for revealing valence in general, and for each of the four binary class sets in particular. Since an exhaustive search on a set of 2 832 features is not feasible in a decent amount of time with today's hardware, we perform a quick estimation of the impact of each low-level descriptor and each functional separately. For this we evaluate the classification performance (UAR) of 142 individual feature sets. 118 sets are created by extracting single LLD with all 24 functionals applied to them. The remaining 24 sets are created by applying each of the 24 functionals separately to all 118 LLD. We then rank the 118 and 24 sets by UAR and thus obtain two rankings, one for LLD and one for functionals. For each of the three classifiers we obtain a separate ranking, as well as for each of the four corpora. Thus we obtain $3 \cdot 4 = 12$ rankings of LLD and functionals for each binary class set. We then compute the mean rank of each feature over all 12 rankings to obtain a unified ranking for each binary class set. The mean rank over all four class sets gives the overall rank of features for valence recognition. For the final sets of relevant features we select only those which by themselves achieve an UAR performance of ≥ 0.51. Table 6 shows the top 5 selected functionals; in table 7 we show the top 10 selected LLD for the four class sets. This roughly amounts to $\frac{1}{5}$ of the 24 functionals and the 59 LLD.

Since this feature ranking is only uni-variate and features with similar rank may still be correlated, this contribution should be considered as a pilot study, and a more thorough

Set	Functionals	# sel.
All	upleveltime75, downleveltime25, amean, kurtosis, min, ...	18
PI-N	nzamean, min, amean, downleveltime25, upleveltime75, ...	17
P-IN	upleveltime75, upleveltime90, qregerr, range, qregc1	5
PN-I	min, max, numPeaks, linregerr, amean, ...	8
P-N	upleveltime75, downleveltime25, skewness, kurtosis, range, ...	22

Table 6: Top 5 of selected functionals, and number of selected functionals in total (individual UAR ≥ 0.51).

Set	LLD	# sel.
All	MFCC 1,3–5,8,10,12 MFB 19,20, spec. Flux	56
PI-N	MFCC 1–5,8,10–12, MFB 20, spec. Flux	33
P-IN	MFB 6,8–12,18,23-25	32
PN-I	MFCC 1,4,5,6,7,8,12 MFB 20,21, spec. Flux	22
P-N	P_{voice}, MFCC 1,3,4,10, MFB 14,19, $F_{2,3,freq.}$, spec. Flux	79

Table 7: Top 10 of selected LLD, and number of selected LLD in total (individual UAR ≥ 0.51).

search for the best feature set must be conducted in future work. However, from the rankings of the LLD and the functionals, a slight tendency across all class configurations can be observed. For the functionals up-/downlevel-times (esp. upleveltime75) prevail in the top 5 for all configurations except for **PN-I**. This configuration is quite different with respect to selected functionals. This seems logical, since **PN-I** is about discriminating positive/negative valence from neutral, while the other three configurations are about detecting positive or neutral valence vs. the rest. For the top 10 LLD, the picture is quite different. The **PN-I** configuration is not exceptional. Instead, for the **P-IN** configuration a high relevance of only MFB is observed. For the three other configurations, spectral flux (i. e. the spectral difference between consecutive frames), higher order MFB (above 19, or 20) and lower order MFCC, esp. 1, and 4, occur frequently in the top 10 list. The higher order MFB correspond to frequencies in the 4–6 kHz region, where the upper formants are found.

With respect to the number of selected LLD/functionals, the **P-N** configuration is leading, which is in line with the finding that **P-N** performs best in overall classification (second for cross-corpus and best for within corpus), when we consider the uni-variate selection process.

As expected, the within corpus results are better than the cross-corpus results, yet the difference is only approx. 4% on average. Within corpus recognition for the **P-IN** constellation is below cross-corpus performance. The biggest difference can be observed again for the **P-N** configuration. This is another indicator that the separation of the classes **P** and **N** is the most doable.

5. Discussion and concluding Remarks

We have presented pilot experiments in a novel field: cross-corpus recognition of naturalistic emotions (here: valence) from acoustic features. Significant improvements over random guess were observed in at least a few cases, which indicates that cross-corpus recognition – even, for acoustic feature based approaches, of the most challenging dimension valence – is feasible in principle; however, it needs more effort to mature to a usable stage. Separation of positive vs. neutral valence gave best results, while a neutral (idle) vs. rest scenario showed lower recall rates. This indicates a fundamental problem with naturalistic emotion recognition: emotions are a continuum. Tagging emotions with discrete classes works for prototypical emotions (such as **P** and **N** valence), but yields inherent quantisation errors when dealing with naturalistic emotions, where there is no fixed class border and thus confusions between adjacent classes are common.

It is generally known that valence recognition from acoustic cues alone is challenging and perhaps not possible perfectly. The within corpus recognition results support this, as well as other studies on the SAL and VAM databases (Wöllmer et al., 2008; Grimm et al., 2007). Thus, future studies might need to investigate linguistic features as well as other modalities, such as vision. Moreover, these studies need to consider word-level chunking, which also has been proven to yield better results (Batliner et al., 2010).

A necessary step towards improving classification performance will be to take care of the inconsistencies listed in section 2.5. Some of these inconsistencies are given (different languages/scenarios) or can only be minimized with a high effort, such as differences in type of labels and number of annotators. However, we can try and find out which corpora are 'good', and which are 'bad' to be included in such cross-corpus evaluations; in other words, which are generic enough, and which are too specific. And we can define the same and optimal type of unit of analysis, for instance words or syntactically well-defined chunks, across all corpora.

6. Acknowledgment

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

7. References

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

A. Batliner, S. Steidl, C. Hacker, and E. Nöth. 2008. Private emotions vs. social interaction — a data-driven

approach towards analysing emotions in speech. *User Modeling and User-Adapted Interaction*, 18:175–206.

A. Batliner, D. Seppi, S. Steidl, and B. Schuller. 2010. Segmenting into adequate units for automatic recognition of emotion-related episodes: a speech-based approach. *Advances in Human-Computer Interaction*, 2010. doi:10.1155/2010/782802.

C.-C. Chang and C.-J. Lin, 2001. *LibSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

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

E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilan, A. Batliner, N. Amir, and K. Karpousis. 2007. The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data. In Ana Paiva, Rui Prada, and Rosalind W. Picard, editors, *Affective Computing and Intelligent Interaction*, pages 488–500, Berlin-Heidelberg. Springer.

F. Eyben, M. Wöllmer, and B. Schuller. 2009. openear - introducing the munich open-source emotion and affect recognition toolkit. In *Proc. ACII*, pages 576–581, Amsterdam.

M. Grimm, K. Kroschel, and S. Narayanan. 2007. Support vector regression for automatic recognition of spontaneous emotions in speech. In *Proc. ICASSP*, pages IV–1085–IV, Honolulu.

M. Grimm, Kristian Kroschel, and Shrikanth Narayanan. 2008. The Vera am Mittag German Audio-Visual Emotional Speech Database. In *Proc. of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 865–868, Hannover, Germany.

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

S. Steininger, F. Schiel, O. Dioubina, and S. Raubold. 2002. Development of user-state conventions for the multimodal corpus in smartkom. In *Proc. Workshop on Multimodal Resources and Multimodal Systems Evaluation*, pages 33–37, Las Palmas.

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

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

M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie. 2008. Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. In *Proc. Interspeech*, pages 597–600, Brisbane.