INTERSPEECH 2009 BRIGHTON

The INTERSPEECH 2009 Emotion Challenge*

Björn Schuller[1], *Stefan Steidl*[2], *Anton Batliner*[2]

1 - Institute for Human-Machine Communication, Technische Universität München, Germany
2 - Chair of Pattern Recognition, Friedrich-Alexander University Erlangen-Nuremberg, Germany

Abstract

The last decade has seen a substantial body of literature on the recognition of emotion from speech. However, in comparison to related speech processing tasks such as Automatic Speech and Speaker Recognition, practically no standardised corpora and test-conditions exist to compare performances under exactly the same conditions. Instead a multiplicity of evaluation strategies employed – such as cross-validation or percentage splits without proper instance definition – prevents exact reproducibility. Further, in order to face more realistic scenarios, the community is in desperate need of more spontaneous and less prototypical data. This INTERSPEECH 2009 Emotion Challenge aims at bridging such gaps between excellent research on human emotion recognition from speech and low compatibility of results. The FAU Aibo Emotion Corpus [1] serves as basis with clearly defined test and training partitions incorporating speaker independence and different room acoustics as needed in most real-life settings. This paper introduces the challenge, the corpus, the features, and benchmark results of two popular approaches towards emotion recognition from speech.

Index Terms: emotion, challenge, feature types, classification

1. Motivation

The young field of emotion recognition from voice has recently gained considerable interest in the fields of Human-Machine Communication, Human-Robot Communication, and Multimedia Retrieval. Numerous studies have been seen in the last decade trying to improve on features and classifiers [2]. However, the research in these fields mostly lacks in two respects: small, preselected, prototypical, and often non-natural emotional data sets and low comparability of results. Early studies started with speaker dependent recognition of emotion, just as in the recognition of speech. But even today the lion's share of research relies on either subject dependent or percentage split and cross-validated test-runs. The latter however still may contain annotated data of the target speakers, as usually j-fold cross-validation with stratification (e. g. the VAM corpus) – if considered at all – yet random selection of instances is employed. However, only Leave-One-Subject-Out (found e. g. on the AVIC [3] and the Berlin Emotional Speech (EMO-DB) [4] databases) or Leave-One-Subject-Group-Out (as with the FAU AIBO corpus) cross-validation would ensure true subject independence. Or, even more realistic, cross-corpora tests.

*This work was partly funded by the European Union in the projects SEMAINE under grant agreement No. 211486 (FP7/2007-2013), PF-STAR under grant IST-2001-37599, and HUMAINE under grant IST-2002-50742. The authors would further like to thank the sponsors of the challenge, the HUMAINE Association and Deutsche Telekom Laboratories. The responsibility lies with the authors.

The arguably even bigger problem is the simple lack of exact reproducibility: reading results on randomly partitioned data, one does not know the exact configuration. As a consequence, figures are not comparable, even if 10-fold cross-validation is used by two different sites: the splits may be completely different. Likewise, clearly defined experiments by leaving subjects or groups out or provision of partitioning in some form of documentation as instance lists on the web will have to be employed in future studies. Ideally, such definition should be provided with database releases to avoid high diversity from the beginning – again, following the example set by speech recognition (cf. e. g. the WSJ or AURORA corpora).

In addition, practically all databases which have been used by different sites such as the freely available and highly popular EMO-DB do not contain realistic, non-prompted speech but prompted, acted speech. Occurrence and distribution of classes to be modelled for acted data and, by that, their acoustics cannot simply be transferred to realistic data.

As mentioned above, comparability between research results in the field is considerably low. Apart from the different evaluation strategies, the diversity of corpora is high, as many early studies report results on their individual and proprietary corpora. Additionally, there is practically no same feature set found twice: high diversity is not only found in the selection of low-level descriptors (LLD), but also in the perceptual adaptation, speaker adaptation, and – most of all – selection and implementation of functionals. This again opposes the more or less settled and clearly defined feature types MFCC, RASTA or PLP that allow for higher comparability in speech recognition. Further, different types of feature reduction are used with astonishingly low documentation on parameter configuration.

One first cooperative experiment is found in the CEICES initiative [5], where seven sites compared their classification results under exactly the same conditions and pooled their features together for one combined unified selection process. This comparison was not fully open to the public, which motivates the INTERSPEECH 2009 Emotion Challenge to be conducted with strict comparability, using the same database. Three sub-challenges are addressed using non-prototypical five or two emotion classes (including a garbage model):

- The *Open Performance Sub-Challenge* allows contributors to find their own features with their own classification algorithms. However, they will have to stick to the definition of test and training sets.
- In the *Classifier Sub-Challenge*, participants may use a large set of standard acoustic features provided by the organisers (cf. Sec. 3) in the well known ARFF file format for classifier tuning. Features may be sub-sampled, altered and combined (e. g. by standardisation or analytical functions), the training bootstrapped, and several classifiers combined by e. g. ROVER or Ensemble Learning

or side tasks learned as gender, etc. However, the audio files may not be used in this task.

- In the *Feature Sub-Challenge*, participants are encouraged to upload their individual best features per unit of analysis with a maximum of 100 per contribution following the example of provided feature files. These features will then be tested by the organisers with equivalent settings in one classification task, and pooled together in a feature selection process. In particular, novel, high-level, or perceptually adequate features are sought-after.

The labels of the test set will be unknown, and all learning and optimisations need to be based only on the training material. However, each participant can upload instance predictions to receive the confusion matrix and results up to 25 times. The format will be instance and prediction and optionally additional probabilities per class. This allows a final fusion by e. g. ROVER or meta-classification of all participants' results to demonstrate the potential maximum by combined efforts (cf. [5]). As classes are unbalanced, the primary measure to optimise will be unweighted average (UA) recall, and secondly the weighted average (WA) recall (i. e. accuracy). The organisers will not take part in the sub-challenges but provide baselines.

This paper describes the data set (Sec. 2), and the features for the challenge (Sec. 3), the conditions for testing (Sec. 4), baseline results (Sec. 5), and concluding remarks (Sec. 6).

2. Emotional Speech Data

One of the major needs of the community ever since – maybe even more than in many related pattern recognition tasks – is the constant need for data sets. In the early days of the late 1990s, these have not only been few, but also small ($\sim$500 turns) with few subjects ($\sim$10), uni-modal, recorded under studio noise conditions, and acted. Further, the spoken content was mostly predefined (e. g. Danish Emotional Speech (DES), EMO-DB, Speech Under Simulated and Actual Stress (SUSAS) databases) [4]. These were seldom made public and few annotators – if any at all – labelled usually exclusively the perceived emotion. Additionally, these were partly not intended for analysis, but for quality measurement of synthesis (e. g. DES, EMO-DB). Today, there are more diverse emotions covered, more elicited or even spontaneous sets of many speakers, and larger amounts of instances (up to 10k and more) of more subjects (up to 50), that are annotated by more labellers (4 (AVIC) - 17 (VAM, [6])) and partly made publicly available. For acted data, equal distribution among classes is of course easily obtainable. Also transcription is becoming more and more rich: additional annotation of spoken content and non-linguistic interjections (e. g. AVIC, Belfast Naturalistic, FAU AIBO, SmartKom databases [4]), multiple annotator tracks (e. g. VAM), manually corrected pitch contours (FAU AIBO), additional audio tracks under different noise and reverberation conditions (FAU AIBO), phoneme boundaries and manual phoneme labelling (e. g. EMO-DB), different units of analysis, and different levels of prototypicality (e. g. FAU AIBO). At the same time these are partly also recorded under more realistic conditions (or taken from the media). Sadly, concrete test and training partitions have been defined still for only few databases. This will hopefully be overcome in future sets, following all named positive trends, yet adding multilinguality and subjects of diverse cultural backgrounds. Another positive recent trend is to use more than one database and even to report first cross-database results.

Trying to meet the utmost of these requirements, the FAU AIBO database [1, Chap. 5] was chosen for the challenge: it is a corpus with recordings of children interacting with Sony's pet robot Aibo. The corpus consists of spontaneous, German speech that is 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 9.2 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. 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), *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.

#	A	E	N	P	R	$\sum$
train	881	2,093	5,590	674	721	9,959
test	611	1,508	5,377	215	546	8,257
$\sum$	1,492	3,601	10,967	889	1,267	**18,216**

Table 1: Number of instances for the 5-class problem

#	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 2: Number of instances for the 2-class problem

Classification experiments on a subset of the corpus [1, 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 [1, Chap. 5.3.5] are used here. In contrast to other publications published recently, the whole corpus consisting of 18,216 chunks is used for this challenge. For the five-class classification problem, the cover classes **A**nger (subsuming *angry*, *touchy*, and *reprimanding*) **E**mphatic, **N**eutral, **P**ositive (subsuming *motherese* and *joyful*), and **R**est are to be discriminated. The two-class problem consists of the cover classes **NEG**ative (subsuming *angry*, *touchy*, *reprimanding*, and *emphatic*) and **IDL**e (consisting of all non-negative states). A heuristic approach similar to the one applied in [1, Chap. 5.3.8] is used to map the labels of the five labelers on the word level onto one label for the whole chunk. Since the

whole corpus is used in this challenge, the classes are highly unbalanced. The frequencies for the five- and the two-class problem are given in Table 1 and Table 2, respectively. 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. In the training set, the chunks are given in sequential order and the chunk id contains the information which child the chunk belongs to. In the test set, the chunks are presented in random order without any information about the speaker. Additionally, the transliteration of the spoken word chain of the training set and the vocabulary of the whole corpus is provided allowing for ASR training and linguistic feature computation.

3. Acoustic Features

The main focus was on prosodic features in the past, in particular pitch, durations and intensity [7]. Comparably small feature sets (10-100) were first utilised. In only a few studies, low-level feature modelling on a frame level was pursued, usually by HMM or GMM. The higher success of static feature vectors derived by projection of the LLD such as pitch or energy by descriptive statistical functional application such as lower order moments (mean, standard deviation) or extrema is probably justified by the supra-segmental nature of the phenomena occurring with respect to emotional content in speech. In more recent research, also voice quality features such as HNR, jitter, or shimmer, and spectral and cepstral features such as formants and MFCC have become more or less the "new standard". At the same time, brute-forcing of features (1,000 up to 50,000), e. g. by analytical feature generation, partly also in combination with evolutionary generation, has become popular. It seems as if this (slightly) outperforms hand-crafted features while the individual worth of automatically generated features seems to be lower. Within expert-based hand-crafted features, perceptually more adequate features have been investigated, reaching from simple log-pitch to Teager energy or more complex features such as articulatory features (e. g. (de-)centralisation of vowels). Further, linguistic features are often added these days, and will certainly also be in the future. However, these demand for robust recognition of speech in the first place.

For the *Classifier Sub-Challenge*, a feature set is provided that shall best cover the described gained knowledge. We therefore stick to the findings in [8] 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. The creation of more elaborate features will be left as subject of the *Feature Sub-Challenge*. For highest transparency we utilise the open source openSMILE feature extraction[1]. In detail, the 16 low-level descriptors chosen are: zero-crossing-rate (ZCR) from the time signal, root mean square (RMS) frame energy, pitch frequency (normalised to 500 Hz), harmonics-to-noise ratio (HNR) by autocorrelation function, and mel-frequency cepstral coefficients (MFCC) 1-12 in full accordance to HTK-based computation. To each of these, the delta coefficients are additionally computed. Next the 12 functionals mean, standard deviation, kurtosis, skewness, minimum and maximum value, relative position, and range as well as two linear regression coefficients with their mean square error (MSE) are applied on a chunk basis as depicted in Table 3. Thus, the total feature vector per chunk contains $16 \cdot 2 \cdot 12 = 384$ attributes.

[1] F. Eyben, M. Wöllmer, B. Schuller (2009): Speech and Music Interpretation by Large-Space Extraction, `http://sourceforge.net/projects/openSMILE`.

LLD ($16 \cdot 2$)	**Functionals** (12)
(Δ) ZCR	mean
(Δ) RMS Energy	standard deviation
(Δ) F0	kurtosis, skewness
(Δ) HNR	extremes: value, rel. position, range
(Δ) MFCC 1-12	linear regression: offset, slope, MSE

Table 3: *Features for the* Classifier Sub-Challenge*: low-level descriptors (LLD) and functionals.*

4. Realism and Classification Performance

With this challenge, we aim at two related, but independent objectives: comparability and realism. Above, we have dealt with the first objective. As for the second one: a fully realistic approach towards automatic processing of emotions in speech means dealing with non-acted, non-prompted, realistic (i. e. spontaneous) data, many speakers, and *all* data obtained. During the last years, not only many authors agreed that we do need realistic data, a few – but of course, not enough – realistic databases have been recorded and processed as well. As far as we can see, however, almost never the full data set has been processed. This is different from the transition from read speech to spontaneous speech in automatic speech recognition (ASR) where normally, always all data have been employed apart from, for instance, non-linguistic vocalisations, etc., which are treated as garbage; but they are still treated and not removed from the signal before processing. In emotion processing, normally a subset of the full database is taken consisting of somehow clear, i. e. more or less prototypical cases. 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. In ASR, however, a lexicion can be established containing all the words or word forms found in the database. We have elaborated on these differences in [9]. So far, we have defined majority voting cases for our FAU AIBO database as well and concentrated on this sub-set. For a four-class problem, unweighted average recall was in the range of above 65 %; using only very prototypical cases, an unweighted average recall close to 80 % could be obtained. By mapping onto a two-class-problem, classification performance could be pushed above 90 %, [9] and unpublished experiments. Using 'realistic data', however, not only means using spontaneous data, it means as well using *all* these data. Whereas the first, qualitative aspect, has been taken into account by several studies yet, the second, sort of 'quantitative' aspect, has still been neglected by and large. With this challenge, we therefore want to initiate work dealing with this second aspect as well by employing the full database; in our opinion, this constitutes the last step before automatic emotion processing (including full ASR and automatic chunking) really can be used 'in the wild', i. e. in real applications. This means at the same time to scale down our expectations for classification performance. In ASR, a rough estimate for the difference between read and spontaneous data was that, at least to start with, one could expect an error rate for spontaneous data twice the size than the one for read data [10]. Of course, we cannot simply transfer this empirically obtained estimate onto our classification problem. Yet we definitely will have to deal with a plainly lower classification performance.

	#States	Recall [%]		Precision [%]	
		UA	WA	UA	WA
	1	**62.3**	71.7	65.2	69.8
2-class	3	**62.9**	57.5	61.1	70.0
	5	**66.1**	65.3	63.6	71.3
	1	**35.5**	50.8	29.6	57.1
5-class	3	**35.2**	34.7	27.5	57.2
	5	**35.9**	37.2	29.3	59.0

Table 4: *Baseline results employing the low-level descriptors of the Classifier Sub-Challenge tasks by dynamic modelling.*

	Process		Recall [%]		Precision [%]	
	1	2	**UA**	WA	UA	WA
	-	-	**62.7**	72.6	66.4	70.6
2-class	S	B	**67.6**	68.3	65.2	72.3
	B	S	**67.7**	65.5	64.8	72.7
	-	-	**28.9**	65.6	35.5	57.0
5-class	S	B	**38.2**	39.2	30.0	59.7
	B	S	**38.0**	32.2	29.4	59.8

Table 5: *Baseline results Classifier Sub-Challenge task by static modelling. Diverse pre-processing strategies: training balancing (B) by SMOTE and standardisation (S).*

5. Baseline Results

For provision of baseline results, we consider the two predominant architectures within the field: Firstly, dynamic modelling of LLD as pitch, energy, MFCC, etc. by hidden Markov models. Secondly, static modelling using supra-segmental information obtained by statistical functional application to the same LLD on the chunk level. We decided to entirely rely on two standard publicly available tools widely used in the community: the Hidden Markov Model Toolkit (HTK)[2] in the case of dynamic, and the WEKA 3 Data Mining Toolkit[3] in the case of static modelling. This ensures easy reproducibility of the results and reduces description of parameters to a minimum: unless specified, defaults are used. Constantly picking the majority class would result in an accuracy (WA recall) of 70.1 % for the two-class problem and 65.1 % for the five-class problem, while the chance level for UA recall is simply 50 % and 20 %, respectively. As instances are unequally distributed among classes, we consider balancing of the training material to avoid classifier over-fitting. We decided for applying the Synthetic Minority Oversampling TEchnique (SMOTE). Note that up-sampling does not have any influence in the case of dynamic modelling: for each class one HMM is trained individually and equal priors are assumed. Table 4 depicts these results for the two- and five-class tasks (classification by linear left-right HMM, one model per emotion, diverse number of states, 2 Gaussian mixtures, 6+4 Baum-Welch re-estimation iterations, Viterbi) by recall and precision. Further standardisation of the whole sets, individually, is considered to cope with biases occurring to different room acoustics, etc. In Table 5 results are shown employing the whole feature set of the *Classifier Sub-Challenge* with support vector machine classification (sequential minimal optimisation learning, linear kernel, pairwise multi-class discrimination). Thereby the influence of these two pre-processing steps is seen. Note that the order of operations is crucial, as the standardisation leads to different results if classes are balanced.

6. Conclusions

The need for a first official challenge on emotion recognition from speech was motivated and the conditions were laid out. Though not aiming at maximum performance, the baseline results clearly demonstrate the difficulty of the switch from performances in the lab to dealing with "everything that comes in".

Clearly, more such comparisons and challenges as NIST evaluations in the field of speech processing, or the MIREX challenges in the field of music retrieval will be needed. Thereby feature extraction or classification code will have to be made available for the organisers or – even better – to the community. In particular future challenges may address audiovisual emotion recognition, or lead to even more realism by considering cross-corpora, multi-cultural, and multi-lingual evaluations. The first step is done – may the next ones follow soon.

[2] `http://htk.eng.cam.ac.uk/docs/docs.shtml`
[3] `http://www.cs.waikato.ac.nz/ml/weka/`

7. References

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

[2] Z. Zeng, M. Pantic, G. I. Rosiman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[3] B. Schuller, R. Müller, B. Hörnler, A. Höthker, H. Konosu, and G. Rigoll, "Audiovisual recognition of spontaneous interest within conversations," in *Proc. ICMI*, Nagoya, Japan, 2007, ACM SIGCHI, pp. 30–37.

[4] D. Ververidis and C. Kotropoulos, "A state of the art review on emotional speech databases," in *1st Richmedia Conference*, Lausanne, Switzerland, 2003, pp. 109–119.

[5] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "Combining Efforts for Improving Automatic Classification of Emotional User States," in *Proc. IS-LTC 2006*, Ljubliana, 2006, pp. 240–245.

[6] M. Grimm, Kristian Kroschel, and Shrikanth Narayanan, "The Vera am Mittag German Audio-Visual Emotional Speech Database," in *Proc. ICME*, Hannover, Germany, 2008, pp. 865–868.

[7] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[8] 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*, Antwerp, 2007, pp. 2253–2256.

[9] 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*, Brisbane, Australia, 2008.

[10] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, no. 1, pp. 1–16, 1997.