Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies (Extended Abstract)

Björn Schuller[1,2,3], Bogdan Vlasenko[4], Florian Eyben[3], Martin Wöllmer[3],
André Stuhlsatz[5], Andreas Wendemuth[6], and Gerhard Rigoll[7]
[1]Chair of Complex & Intelligent Systems, University of Passau, Germany
[2]Department of Computing, Imperial College London, UK
[3]audEERING UG, Gilching, Germany
[4]Idiap Research Institute, Martigny, Switzerland
[5]University of Applied Sciences Düsseldorf, Düsseldorf, Germany
[6]Cognitive Systems group, IESK, Otto-von-Guericke Universität (OVGU), Magdeburg, Germany
[7]Institute for Human-Machine Communication, Technische Universität München (TUM), Germany
Email: schuller@ieee.org

***Abstract*—As the recognition of emotion from speech has matured to a degree where it becomes applicable in real-life settings, it is time for a realistic view on obtainable performances. Most studies tend to overestimation in this respect: acted data is often used rather than spontaneous data, results are reported on pre-selected prototypical data, and true speaker disjunctive partitioning is still less common than simple cross-validation. A considerably more realistic impression can be gathered by inter-set evaluation: we therefore show results employing six standard databases in a cross-corpora evaluation experiment. To better cope with the observed high variances, different types of normalization are investigated. 1.8 k individual evaluations in total indicate the crucial performance inferiority of inter- to intra-corpus testing.**

I. Introduction

Since the dawn of emotion and speech research [1], [2], [3], [4], [5], [6], the usefulness of automatic recognition of emotion in speech seems increasingly agreed given hundreds of (commercially interesting) use-cases. Most of these, however, require sufficient reliability, which may not be given yet [7], [8], [9], [10], [11], [12], [13], [14]. A simplification that characterizes almost all emotion recognition performance evaluations is that systems are usually trained and tested using the same database. Even though speaker-independent evaluations have become quite common, other kinds of potential mismatches between training and test data, such as different recording conditions (including different room acoustics, microphone types and positions, signal-to-noise ratios, etc.), languages, or types of observed emotions, are usually not considered. Addressing such typical sources of mismatch all at once is hardly possible, however, we believe that a first impression of the generalization ability of today's emotion recognition engines can be obtained by simple cross-corpora evaluations. For emotion recognition, several studies already provide accuracies on multiple corpora – however, only very few consider training on one and testing on a completely different one (e. g., [15], and [16], where two, and four corpora are employed, respectively). In this article, we provide cross-corpus results employing six of the best known corpora in the field of emotion recognition. This allows us to discover similarities among databases which in turn can indicate what kind of corpora can be combined – e. g., in order to obtain more training material for emotion recognition systems as a means to reduce the problem of data sparseness. A specific problem of cross-corpus emotion recognition is that mismatches between training and test data not only comprise the aforementioned different acoustic conditions but also differences in annotation. Each corpus for emotion recognition is usually recorded for a specific task – and as a result of this, they have specific emotion labels assigned to the spoken utterances. For cross-corpus recognition this poses a problem, since the training and test sets in any classification experiment must use the same class labels. Thus, mapping or clustering schemes have to be developed whenever different emotion corpora are jointly used.

As classification technique, we follow the approach of supra-segmental feature analysis via Support Vector Machines by projection of the multi-variate time series consisting of Low-Level-Descriptors as pitch, Harmonics-to-Noise ratio (HNR), jitter, and shimmer onto a single vector of fixed dimension by statistical functionals such as moments, extremes, and percentiles [17]. To better cope with the described variation between corpora, we investigate four different normalization approaches: normalization to the speaker, the corpus, to both, and no normalization. As mentioned before, every considered database bases on a different model or subset of emotions. We therefore limit our analyses to employing only those emotions at a time that are present in the other data set, respectively. As recognition rates are comparably low for the full sets, we consider all available permutations of two up to six emotions by exclusion of remaining ones. In addition to exclusion, we also have a look at clustering to the two predominant types of general emotion categories, namely positive/negative valence, and high/low arousal. Four data sets are used for testing with an additional two that are used for training only. In total, we examine 23 different combinations of training and test data, leading to 409 different emotion class permutations. Together with 2 × 23 experiments on the discrimination of emotion categories (valence and arousal), we perform 455 different evaluations for four different normalization strategies, leading to 1 820 individual results. To best summarize the findings of this high amount of results, we show box-plots per test-database and the two most important measures: accuracy (i. e., recognition rate) and – important in the case of heavily unbalanced class distributions – unweighted average recall. For the evaluation of the best normalization strategy we calculate Euclidean distances to the optimum for each type of normalization over the complete results.

The rest of this article is structured as follows: we first deal with the basic necessities to get started: the six databases chosen (sec. II) with a general commentary on the present situation. We next get on track with features and classification (sec. III). Then we consider normalization to improve performance in sec. IV. Some comments will follow on evaluation (sec. V) before concluding this article (sec. VI).

II. Selected Databases

For the following cross-corpora investigations, we chose six among the most frequently used and well known. Only such available to the community were considered. These should cover a broad variety reaching from acted speech (the Danish and the Berlin Emotional Speech databases, as well as the eNTERFACE corpus) with acted fixed spoken content to natural with fixed spoken content represented by the SUSAS database, and to more modern corpora with respect to the number of subjects involved, naturalness, spontaneity, and free language as covered by the AVIC and SmartKom [18] databases. However, we decided to compute results only on those that cover a broader variety of more 'basic' emotions, which is why AVIC and SUSAS are exclusively used for training purposes. Naturally we have by that to leave out several emotional or broader affective states as frustration or irritation – once more databases cover such, one can of course investigate cross-corpus effects for such states as well. Note also that we did not exclusively focus on corpora that include non-prototypical emotions, since those corpora partly do not contain categorical labels (e. g., the VAM corpus). The corpus of the first comparative Emotion Challenge [17] – the FAU Aibo Emotion Corpus of children's speech – could regrettably also not be included in our evaluations, as it would be the only one containing exclusively children speech. We thus decided that this would introduce an additional severe source of difficulty for the cross-corpus tests.

An overview on properties of the chosen sets is found in Table II. Since all six databases are annotated in terms of emotion categories, a mapping was defined to generate labels for binary arousal/valence from the emotion categories. This mapping is given in Table I. In order to be able to also map emotions for which a binary arousal/valence assignment is not clear, we considered the scenario in which the respective corpus was recorded and partly re-evaluated the annotations (e. g., *neutrality* in the AVIC corpus tends to correspond to a higher level of arousal than it does in the DES corpus; *helpless* people in the SmartKom corpus tend to be highly aroused, etc.). The chosen sets provide a good variety reaching from acted (DES, EMO-DB) over induced (eNTERFACE) to natural emotion (AVIC, SmartKom, SUSAS) with strictly limited textual content (DES, EMO-DB, SUSAS) over more textual variation (eNTERFACE) to full textual freedom (AVIC, SmartKom). Further Human-Human (AVIC) as well as Human-Computer (SmartKom) interaction are contained. Three languages – English, German, and Danish – are comprised. However, these three all belong to the same family of Germanic languages. The speaker ages and backgrounds vary strongly, and so do of course microphones used, room acoustics, and coding (e. g., sampling rate reaching from 8 kHz to 44.1 kHz) as well as the annotators. Summed up, cross-corpus investigation will reveal performance as for example in a typical real-life media retrieval usage where a very broad understanding of emotions is needed.

TABLE I
Mapping of emotions for the clustering to a binary arousal/valence discrimination task.

AROUSAL	**Low**	**High**
AVIC	boredom	neutral, joyful
DES	neutral, sadness	anger, happiness, surprise
EMO-DB	boredom, disgust, neutral, sadness	anger, fear, joy
eNTERFACE	disgust, sadness	anger, fear, joy, surprise
SmartKom	neutral, pondering,	anger, helplessness, joy, surprise
SUSAS	neutral	high stress, medium stress, screaming, fear
VALENCE	**Negative**	**Positive**
AVIC	boredom	neutral, joyful
DES	angry, sadness	happiness, neutral, surprise
EMO-DB	anger, boredom, disgust, fear, sadness	joy, neutral
eNTERFACE	anger, disgust, fear, sadness	joy, surprise
SmartKom	anger, helplessness	joy, neutral, pondering, surprise, unidentifiable
SUSAS	high stress, screaming, fear	medium stress, neutral

III. Features and Classification

We decided for a typical state-of-the-art emotion recognition engine operating on supra-segmental level, and use a set of 1 406 systematically generated acoustic features based on 37 Low-Level-Descriptors as seen in Table III and their first order delta coefficients. These 37×2 descriptors are next smoothed by low-pass filtering with a simple moving average filter.

We derive statistics per speaker turn by a projection of each uni-variate time series – the Low-Level-Descriptors - onto a scalar feature independent of the length of the turn. This is done by use of functionals. 19 functionals are applied to each contour on the word level covering extremes, ranges, positions, first four moments, and quartiles as also shown in Table III. Note that three functionals are related to time (position in time) with the physical unit milliseconds.

Again, we choose the most frequently encountered solution (e. g., in [24], [25], [26], [27], [28]) for representative results

TABLE II

Details of the six emotion corpora. Content fixed/variable (spoken text). Number of turns per emotion category (# Emotion), binary arousal/valence, and overall number of turns (All). Emotions in corpus other than the common set (Else). Total audio time. Number of subjects (Sub), number of female (f) and male (m) subjects. Type of material (acted/natural/mixed) and recording conditions (studio/normal/noisy) (Type). Sampling rate (F_s). Emotion categories: anger (A), boredom (B), disgust (D), fear/screaming (F), joy(ful)/happy/happiness (J), neutral (N), sad(ness) (SA), surprise (SU); non-common further contained states: helplessness (he), high stress (hs), medium stress (ms), pondering (p), unidentifiable (u).

Corpus	**Content**	**# Emotion**								**# Arousal**		**# Valence**		**# All**	**Else**	**Time**	**# Sub**	**Type**	**Rate**
		N	**J**	**A**	**F**	**SU**	**SA**	**B**	**D**	–	+	–	+			h:mm			kHz
EMO-DB [19]	German fixed	78	64	127	55	–	53	79	38	248	246	352	142	494	–	0:22	5 f 5 m	acted studio	16
DES [20]	Danish fixed	85	86	85	–	79	84	–	–	169	250	169	250	419	–	0:28	2 f 2 m	acted normal	20
eNTERFACE [21]	English fixed	–	205	200	189	192	195	–	189	397	773	786	384	1 170	–	0:58	8 f 34 m	acted normal	16
SUSAS [22]	English fixed	701	–	–	484	–	–	–	–	701	484	701	484	1 185	hs, ms	0:20	3 f 4 m	mixed noisy	8
AVIC [23]	English variable	510	170	–	–	–	–	316	–	170	826	170	826	996	–	0:35	10 f 11 m	natural normal	44.1
SmartKom [18]	German variable	2 196	284	224	–	71	–	–	–	2 196	579	224	2 551	2 775	he, p, u	5:11	47 f 32 m	natural noisy	16
Total	–	**3 570**	**809**	**636**	**728**	**342**	**332**	**295**	**227**	**3 881**	**3 158**	**2 402**	**4 637**	**7 039**	–	**7:54**	**163**	–	–

TABLE III

Overview of Low-Level-Descriptors (2×37) *and functionals* (19) *for static supra-segmental modeling.*

Low-Level-Descriptors	**Functionals**
(Δ) Pitch	mean, centroid, stdandard deviation
(Δ) Energy	Skewness, Kurtosis
(Δ) Envelope	Zero-Crossing-Rate
(Δ) Formant 1–5 amplitude	quartile 1/2/3
(Δ) Formant 1–5 bandwidth	quartile 1 – min., quart. 2 – quart. 1
(Δ) Formant 1–5 position	quartile 3 – quart. 2, max. – quart. 3
(Δ) MFCC 1–16	max./min. value,
(Δ) HNR	max./min. relative position
(Δ) Shimmer	range max. – min.
(Δ) Jitter	position 95 % roll-off-point

in sections IV and V: Support Vector Machine (SVM) classification. Thereby we use a linear kernel and pairwise multi-class discrimination [29].

IV. Normalization

Speaker normalization is widely agreed to improve recognition performance of speech related recognition tasks. Normalization can be carried out on differently elaborated levels reaching from normalization of all functionals to, e. g., Vocal Tract Length Normalization of MFCC or similar Low-Level-Descriptors. However, to provide results with a simply implemented strategy, we decided for the first – speaker normalization on the functional level – which will be abbreviated SN. Thus, SN means a normalization of each calculated functional feature to a mean of zero and standard deviation of one. This is done using the whole context of each speaker, i. e., having collected some amount of speech of each speaker without knowing the emotion contained. As we are dealing with cross-corpora evaluation in this article, we further introduce another type of normalization, namely 'corpus normalization' (CN). Here, each database is normalized in the described way before its usage in combination with other corpora. This seems important to eliminate different recording conditions as varying room acoustics, different type of and distance to the microphones, and – to a certain extent – the different understanding of emotions by either the (partly contained) actors, or the annotators. These two normalization methods (SN and CN) can also be combined: after having each speaker normalized individually, one can additionally normalize the whole corpus, that is 'speaker-corpus normalization' (SCN). To get an impression upon improvement over no normalization, we consider a fourth condition, which is simply 'no normalization' (NN).

V. Evaluation

Early studies started with speaker dependent recognition of emotion, just as in the recognition of speech [30], [31], [32]. But even today the lion's share of research presented relies on either subject dependent or percentage split and cross-validated test-runs, e. g., [33]. The latter, however, still may contain annotated data of the target speakers, as usually j-fold cross-validation with stratification, or random selection of instances is employed. Thus, only Leave-One-Subject-Out

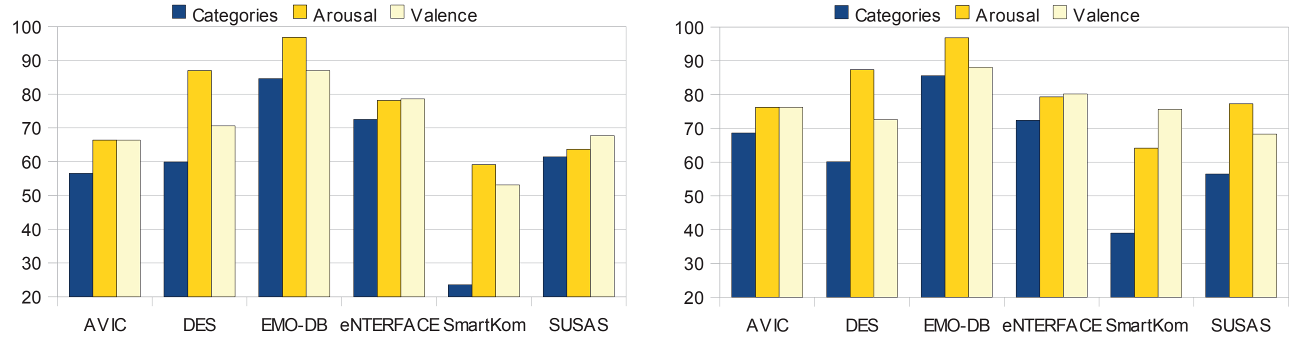
(LOSO) or Leave-One-Subject-Group-Out (LOSGO) cross-validation is next considered for 'within' corpus results to ensure true speaker independence (cf. [34]). Still, only cross-corpora evaluation encompasses realistic testing conditions which a commercial emotion recognition product used in every-day life would frequently have to face.

The within corpus evaluations' results – intended for a first reference – are sketched in Figures 1(a) and 1(b). As classes are often unbalanced in the oncoming cross-corpus evaluations, where classes are reduced or clustered, the primary measure is unweighted average recall (UAR, i. e., the accuracy per class divided by the number of classes without considerations of instances per class), which has also been the competition measure of the first official challenge on emotion recognition from speech [17]. Only where appropriate the weighted average recall (WAR, i. e., accuracy) will be provided in addition. For the inter-corpus results only minor differences exist between these two measures owed to the mostly acted and elicited nature of the corpora, where instances can easily be collected balanced among classes. The results shown in Figures 1(a) and 1(b) were obtained using LOSO (DES, EMO-DB, SUSAS) and LOSGO (AVIC, eNTERFACE, SmartKom) evaluations (due to frequent partitioning for these corpora). For each corpus classification of all emotions contained in that particular corpus is performed. A great advantage of cross-corpora experiments is the well definedness of test and training sets and thus the easy reproducibility of the results. Since most emotion corpora, in contrast to speech corpora for automatic speech recognition or speaker identification, do not provide defined training, development, and test partitions, individual splitting and cross validation are mostly found, which makes it hard to reproduce the results under equal conditions. In contrast to this, cross-corpus evaluation is well defined and thus easy to reproduce and compare. Table IV lists all 23 different training and test set combinations we evaluated in our cross-corpus experiments. As mentioned before, SUSAS and AVIC are only used for training, since they do not cover sufficient overlapping 'basic' emotions for the testing. Furthermore, we omitted combinations for which the number of emotion classes occurring in both, the training and the test set was lower than three (e. g., we did not evaluate training on AVIC and testing on DES, since only *neutral* and *joyful* occur in both corpora – see also Table II). In order to obtain combinations for which up to six emotion classes occur in the training and test set, we included experiments in which more than one corpus was used for training (e. g., we combined eNTERFACE and SUSAS for training in order to be able to model six classes when testing on EMO-DB). Dependent on the maximum number of different emotion classes that can be modeled in a certain experiment, and dependent on the number of classes we actually use (two to six), we get a certain number of possible emotion class permutations according to Table IV. For example, if we aim to model two emotion classes when testing on EMO-DB and training on DES, we obtain six possible permutations. Evaluating all permutations for all of the 23 different training-test combinations leads to 409 different experiments (sum of the last line in Table IV). Additionally, we evaluated the discrimination between positive and negative valence as well as the discrimination between high and low arousal for all 23 combinations, leading to 46 additional experiments.

TABLE IV
Number of emotion class permutations dependent on the used training and test set combination and the total number of classes used in the respective experiment.

Test set	**Training set**	**# classes**				
		2	3	4	5	6
EMO-DB	AVIC	3	1	0	0	0
	DES	6	4	1	0	0
	eNTERFACE	10	10	5	1	0
	SmartKom	3	1	0	0	0
	eNTERF.+SUSAS	15	20	15	6	1
	eNTERF.+SUSAS+DES	15	20	15	6	1
DES	EMO-DB	6	4	1	0	0
	eNTERFACE	6	4	1	0	0
	SmartKom	6	4	1	0	0
	EMO-DB+SUSAS	6	4	1	0	0
	EMO-DB+eNTERFACE	10	10	5	1	0
eNTERFACE	DES	6	4	1	0	0
	EMO-DB	10	10	5	1	0
	SmartKom	3	1	0	0	0
	EMO-DB+SUSAS	10	10	5	1	0
	EMO-DB+SUSAS+DES	15	20	15	6	1
SmartKom	DES	6	4	1	0	0
	EMO-DB	3	1	0	0	0
	eNTERF.	3	1	0	0	0
	EMO-DB+SUSAS	3	1	0	0	0
	EMO-DB+SUSAS+DES	6	4	1	0	0
	eNTERF.+SUSAS	6	4	1	0	0
	eNTERF.+SUSAS+DES	6	4	1	0	0
	SUM	163	146	75	22	3

We next strive to reveal the optimal normalization strategy from those introduced in section IV (refer to Table V for the results). The following evaluation is carried out: the optimal result obtained per run by any of the four test sets is stored as the maximum obtained performance as corresponding element in a maximum result vector v_{max}. This result vector contains the result for all tests and any permutation arising from exclusion and clustering of classes (see also Table IV). Next, we construct the vectors for each normalization strategy on its own, that is v_i with $i \in \{NN, SN, CN, SCN\}$. Subsequently each of these vectors v_i is element-wise normalized to the maximum vector v_{max} by $v_{i,norm} = v_i \cdot v_{max}^{-1}$. Finally, we calculate the Euclidean distance to the unit vector of the according dimension. Thus, overall we compute the normalized Euclidean distance of each normalization method to the maximum obtained performance by choosing the optimal strategy at a time. That is the *distance to maximum* (DTM) with $DTM \in [0, \infty[$ whereas $DTM = 0$ resembles the optimum ("this method has always produced the best result"). Note that the DTM as shown in Table V is a rather abstract performance measure, indicating the *relative* performance difference between the normalization strategies, rather than the *absolute* recognition accuracy. Here, we consider mean weighted average recall (=accuracy, Table V) and – as before – mean unweighted



(a) UAR

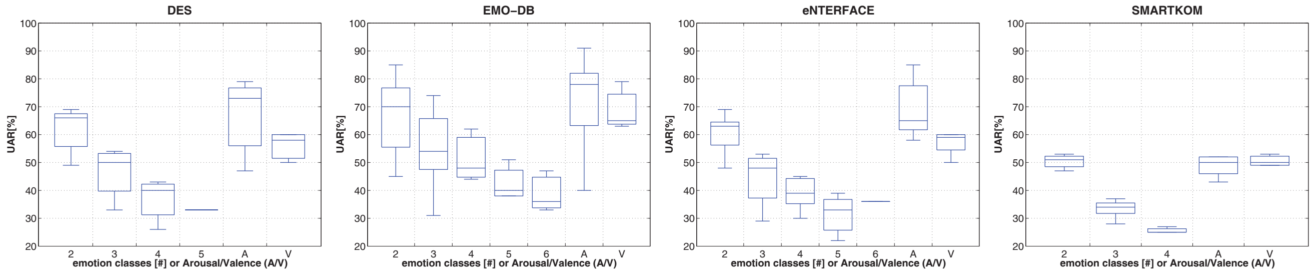
(b) WAR

Fig. 1. Unweighted and weighted average recall (UAR/WAR) in % of within corpus evaluations on all six corpora using corpus normalization (CN). Results for all emotion categories present within the particular corpus, binary arousal, and binary valence.

TABLE V

(Un-)Weighted average recall (UAR/WAR). Revealing the optimal normalization method: none (NN), speaker (SN), corpus (CN) or combined speaker, then corpus (SCN) normalization. Shown is the Euclidean distance to the maximum vector (DTM) of mean accuracy over the maximum obtained throughout all class permutations and for all tests. Detailed explanation in the text.

	DTM	**# classes**							
	[%]	2	3	4	5	6	V	A	**mean**
WAR	NN	1.24	1.82	1.96	0.69	0.71	0.98	1.43	1.26
	CN	0.67	0.87	0.94	0.87	0.90	0.63	0.86	0.82
	SN	0.61	0.82	**0.63**	**0.58**	**0.64**	0.57	0.72	**0.65**
	SCN	**0.47**	**0.78**	0.70	0.76	0.84	**0.32**	**0.71**	0.65
UAR	NN	0.78	1.32	1.51	0.99	0.81	0.50	0.94	0.98
	CN	0.83	0.82	1.09	1.07	0.90	0.44	0.62	0.82
	SN	**0.27**	**0.38**	**0.42**	**0.39**	**0.41**	0.43	**0.23**	**0.36**
	SCN	0.30	0.39	0.47	0.46	0.52	**0.42**	0.26	0.40



(a) DES, UAR

(b) EMO-DB, UAR

(c) eNTERFACE, UAR

(d) SMARTKOM, UAR

Fig. 2. Box-plots for unweighted average recall (UAR) in % for cross-corpora testing on four test corpora. Results obtained for varying number of classes (2–6) and for classes mapped to high/low arousal (A) and positive/negative valence (V).

recall (UAR) (Table V) for the comparison, as some data sets are not in balance with respect to classes (cf. Table II). In the case of accuracy, no significant difference [35] between speaker and combined speaker and corpus normalization is found. As the latter comprises increased efforts not only in terms of calculation but also in terms of needed data, the favorite seems clear, already. A secondary glance at UAR strengthens this choice: here solemnly normalizing the speaker outperforms the combination with the corpus normalization. Thus, no extra boost seems to be gained from additional corpus normalization. However, there is also some variance visible from the tables: the distance to the maximum (DTM in the tables) never resembles zero, which means that no method is always performing best. Further it can be seen that depending on the number of classes the combined version of speaker and corpus normalization partly outperforms speaker only. As a result of this finding, the further provided box-plots are based on speaker normalized results: to summarize the results of permutations over cross-training sets and emotion groupings, box-plots indicating the unweighted average recall are shown (see Figures 2(a) to 2(d)). All values are averaged over all constellations of cross-corpus training to provide a raw general impression of performances to be expected. The plots show the average, the first and third quartile, and the extremes for a varying number (two to six) of classes (emotion categories) and the binary arousal and valence tasks. First, the DES set is chosen for testing, as depicted in Figure 2(a). For training five different combinations of the remaining sets are used (see Table IV). As expected the weighted (i. e., accuracy – not shown) and unweighted recall monotonously drop on average with an increased number of classes. For the DES experience holds: arousal discrimination tasks are 'easier' on average. No big differences are further found between the weighted and unweighted recall plots. This stems from the fact that DES consists of acted data, which is usually found in more or less balanced distribution among classes. While the average results are constantly found considerably above chance level, it also becomes clear that only selected groups are ready for real-life application – of course allowing for some error tolerance. These are two-class tasks with an approximate error of 20 %. A very similar overall behavior is observed for the EMO-DB in Figure 2(b). This seems no surprise, as the two sets have very similar characteristics. For EMO-DB a more or less additive offset in terms of recall is obtained, which is owed to the known lower 'difficulty' of this set. Switching from acted to mood-induced, we provide results on eNTERFACE in Figure 2(c). However, the picture remains the same, apart from lower overall results: again a known fact from experience, as eNTERFACE is no 'gentle' set, partially for being more natural than the DES corpus or the EMO-DB.

Finally considering testing on spontaneous speech with non-restricted varying spoken content and natural emotion we note the challenge arising from the SmartKom set in Figure 2(d): as this set is – due to its nature of being recorded in a user-study – highly unbalanced, the mean unweighted recall is again mostly of interest. Here, rates are found only slightly above chance level. Even the optimal groups of emotions are not recognized in a sufficiently satisfying manner for a real-life usage. Though one has to bear in mind that SmartKom was annotated multimodally, i. e., the emotion is not necessarily reflected in the speech signal, and overlaid noise is often present due to the setting of the recording, this shows in general that the reach of our results is so far restricted to acted data or data in well defined scenarios: the SmartKom results clearly demonstrate that there is a long way ahead for emotion recognition in user studies (cf. also [17]) and real-life scenarios. At the same time, this raises the ever-present and in comparison to other speech analysis tasks unique question on ground truth reliability: while the labels provided for acted data can be assumed to be double-verified, as the actors usually wanted to portray the target emotion which is often additionally verified in perception studies, the level of emotionally valid material found in real-life data is mostly unclear relying on few labelers with often high disagreement among these.

VI. Concluding Remarks

Summing up, we have shown results for intra- and inter-corpus recognition of emotion from speech. By that we have learnt that the accuracy and mean recall rates highly depend on the specific sub-group of emotions considered. In any case, performance is decreased dramatically when operating cross-corpora-wise. As long as conditions remain similar, cross-corpus training and testing seems to work to a certain degree: the DES, EMO-DB, and eNTERFACE sets led to partly useful results. These are all rather prototypical, acted or mood-induced with restricted pre-defined spoken content. The fact that three different languages – Danish, English, and German – are contained, seems not to generally disallow inter-corpus testing: these are all Germanic languages, and a highly similar cultural background may be assumed. However, the cross-corpus testing on a spontaneous set (SmartKom) clearly indicated limitations of current systems. Here only few groups of emotions stood out in comparison to chance level. To better cope with the differences among corpora, we evaluated different normalization approaches, whereas speaker normalization led to the best results. For all experiments we had used supra-segmental feature analysis basing on a broad variety of prosodic, voice quality, and articulatory features and SVM classification. While an important step was taken in this study on inter-corpus emotion recognition a substantial body of future research will be needed to highlight issues like different languages. Future research will also have to address the topic of cultural differences in expressing and perceiving emotion. Cultural aspects are among the most significant variances that can occur when jointly using different corpora for the design of emotion recognition systems. Thus, it is important to systematically examine potential differences and develop strategies to cope with cultural manifoldness in emotional expression. To better cope with differences across corpora, adaptation of the feature sets [36], sub-sampling of the instances of the corpora rather than taking all data [37], adding unlabelled data to self-train the system [38], synthesizing of additional data [39], or employing transfer learning methods to make the data more 'similar' [40].

Concluding, this article has shown ways and need of future research on the recognition of emotion in speech as it reveals fallbacks of current-date analysis and corpora.

Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 211486 (SEMAINE). The work has been conducted in the framework of the project "Neurobiologically Inspired, Multimodal Intention Recognition for Technical Communication Systems (UC4)" funded by the European Community through the Center for Behavioral Brain Science, Magdeburg. Finally, this research is associated and supported by the Transregional Collaborative Research Centre SFB/TRR 62 "Companion- Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG).

References

[1] E. Scripture, “A study of emotions by speech transcription,” *Vox*, vol. 31, pp. 179–183, 1921.

[2] E. Skinner, “A calibrated recording and analysis of the pitch, force, and quality of vocal tones expressing happiness and sadness,” *Speech Monographs*, vol. 2, pp. 81–137, 1935.

[3] G. Fairbanks and W. Pronovost, “An experimental study of the pitch characteristics of the voice during the expression of emotion,” *Speech Monographs*, vol. 6, pp. 87–104, 1939.

[4] C. Williams and K. Stevens, “Emotions and speech: some acoustic correlates,” *Journal of the Acoustical Society of America*, vol. 52, pp. 1238–1250, 1972.

[5] K. R. Scherer, “Vocal affect expression: a review and a model for future research,” *Psychological Bulletin*, vol. 99, pp. 143–165, 1986.

[6] C. Whissell, “The dictionary of affect in language,” in *Emotion: Theory, Research and Experience. Vol. 4, The Measurement of Emotions*, R. Plutchik and H. Kellerman, Eds. New York: Academic Press, 1989, pp. 113–131.

[7] R. Picard, *Affective Computing*. Cambridge, MA: MIT Press, 1997.

[8] 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.

[9] E. Shriberg, “Spontaneous speech: How peoply really talk and why engineers should care,” in *Proc. of EUROSPEECH 2005*, 2005, pp. 1781–1784.

[10] C. M. Lee and S. S. Narayanan, “Toward detecting emotions in spoken dialogs,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

[11] M. Schröder, L. Devillers, K. Karpouzis, J.-C. Martin, C. Pelachaud, C. Peter, H. Pirker, B. Schuller, J. Tao, and I. Wilson, “What should a generic emotion markup language be able to represent?” in *Proc. 2nd Int. Conf. on Affective Computing and Intelligent Interaction ACII 2007, Lisbon, Portugal*, vol. LNCS 4738. Springer Berlin, Heidelberg, 2007, pp. 440–451.

[12] A. Wendemuth, J. Braun, B. Michaelis, F. Ohl, D. Rösner, H. Scheich, and R. Warnemünde, “Neurobiologically inspired, multimodal intention recognition for technical communication systems (NIMITEK),” in *Proc. of the 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-based Systems (PIT 2008)*. Berlin, Heidelberg: Springer, 2008, vol. LNCS 5078, pp. 141–144.

[13] M. Schröder, R. Cowie, D. Heylen, M. Pantic, C. Pelachaud, and B. Schuller, “Towards responsive sensitive artificial listeners,” in *Proc. 4th Intern. Workshop on Human-Computer Conversation, Bellagio, Italy*, 2008.

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

[15] M. Shami and W. Verhelst, “Automatic classification of emotions in speech using multi-corpora approaches,” in *Proc. of the second annual IEEE BENELUX/DSP Valley Signal Processing Symposium (SPS-DARTS 2006)*, Antwerp, Belgium, 2006, pp. 3–6.

[16] ——, “Automatic classification of expressiveness in speech: A multi-corpus study,” in *Speaker Classification II*, ser. Lecture Notes in Computer Science / Artificial Intelligence, C. Müller, Ed. Heidelberg - Berlin - New York: Springer, 2007, vol. 4441, pp. 43–56.

[17] B. Schuller, S. Steidl, and A. Batliner, “The INTERSPEECH 2009 emotion challenge,” in *Proc. of INTERSPEECH 2009*, 2009.

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

[19] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, “A database of german emotional speech,” in *Proc. of INTERSPEECH 2005*, 2005, pp. 1517–1520.

[20] I. S. Engbert and A. V. Hansen, “Documentation of the danish emotional speech database des,” Center for PersonKommunikation, Aalborg University, Denmark, Tech. Rep., 2007.

[21] O. Martin, I. Kotsia, B. Macq, and I. Pitas, “The enterface’05 audio-visual emotion database,” in *Proc. of the IEEE Workshop on Multimedia Database Management*, Atlanta, 2006.

[22] J. Hansen and S. Bou-Ghazale, “Getting started with susas: A speech under simulated and actual stress database,” in *Proc. of EUROSPEECH 1997*, vol. 4, Rhodes, Greece, 1997, pp. 1743–1746.

[23] B. Schuller, R. Müller, B. Hörnler, A. Höthker, H. Konosu, and G. Rigoll, “Audiovisual recognition of spontaneous interest within conversations,” in *Proc. of ICMI 2007*, 2007, pp. 30–37.

[24] 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. of IS-LTC 2006*, Ljubliana, 2006, pp. 240–245.

[25] B. Schuller, D. Arsic, F. Wallhoff, and G. Rigoll, “Emotion recognition in the noise applying large acoustic feature sets.” in *Proc. of Speech Prosody 2006*. ISCA, May 2006.

[26] F. Eyben, B. Schuller, and G. Rigoll, “Wearable assistance for the ballroom-dance hobbyist - holistic rhythm analysis and dance-style classification,” in *Proc. ICME 2007*, 2007.

[27] 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, Elsevier*, vol. 27, no. 12, pp. 1760 – 1774, 2009.

[28] F. Eyben, M. Wöllmer, and B. Schuller, “openEAR - Introducing the Munich open-source Emotion and Affect Recognition toolkit,” in *Proc. of ACII 2009*. IEEE, 2009, pp. pp. 576 – 581.

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

[30] M. Slaney and G. McRoberts, “Baby ears: a recognition system for affective vocalizations,” in *Proc. of ICASSP 1998*, vol. 2, May 1998, pp. 985–988 vol.2.

[31] B. Schuller, G. Rigoll, and M. Lang, “Hidden markov model-based speech emotion recognition,” in *Proc. of ICASSP 2003*, vol. II. IEEE, 2003, pp. 1–4.

[32] R. Barra, J. M. Montero, J. Macias-Guarasa, L. F. D’Haro, R. San-Segundo, and R. Cordoba, “Prosodic and segmental rubrics in emotion identification,” in *Proc. of ICASSP 2006*, vol. 1, May 2006, pp. I–I.

[33] M. Grimm, K. Kroschel, and S. Narayanan, “Support vector regression for automatic recognition of spontaneous emotions in speech,” in *Proc. of ICASSP 2007*, vol. 4. IEEE, Apr. 2007, pp. IV–1085–IV.

[34] S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann, ““of all things the measure is man”: Automatic classification of emotions and inter-labeler consistency,” in *Proc. of ICASSP 2005*, Philadelphia, 2005, pp. 317–320.

[35] L. Gillick and S. J. Cox, “Some statistical issues in the comparison of speech recognition algorithms,” in *Proc. of ICASSP 1989*, vol. I, 1989, pp. 23–26.

[36] F. Eyben, A. Batliner, B. Schuller, D. Seppi, and S. Steidl, “Cross-Corpus Classification of Realistic Emotions - Some Pilot Experiments,” in *Proc. of 3rd International Workshop on EMOTION: Corpora for Research on Emotion and Affect, satellite of LREC 2010*. Valletta, Malta: ELRA, 2010, pp. 77–82.

[37] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, “Selecting Training Data for Cross-Corpus Speech Emotion Recognition: Prototypicality vs. Generalization,” in *Proc. of 2011 Speech Processing Conference*. Tel Aviv, Israel: AVIOS, 2011, 4 pages.

[38] Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller, “Unsupervised Learning in Cross-Corpus Acoustic Emotion Recognition,” in *Proc. of 12th Biannual IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2011*. Big Island, HY: IEEE, 2011, pp. 523–528.

[39] B. Schuller, Z. Zhang, F. Weninger, and F. Burkhardt, “Synthesized Speech for Model Training in Cross-Corpus Recognition of Human Emotion,” *International Journal of Speech Technology*, vol. 15, no. 3, pp. 313–323, 2012.

[40] J. Deng, Z. Zhang, and B. Schuller, “Linked Source and Target Domain Subspace Feature Transfer Learning – Exemplified by Speech Emotion Recognition,” in *Proc. of 22nd International Conference on Pattern Recognition (ICPR 2014)*. Stockholm, Sweden: IAPR, 2014, pp. 761–766.