Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies

Björn Schuller, *Member*, *IEEE*, Bogdan Vlasenko, Florian Eyben, *Member*, *IEEE*, Martin Wöllmer, *Member*, *IEEE*, André Stuhlsatz, Andreas Wendemuth, *Member*, *IEEE*, and Gerhard Rigoll, *Senior Member*, *IEEE*

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 preselected prototypical data, and true speaker disjunctive partitioning is still less common than simple cross-validation. Even speaker disjunctive evaluation can give only a little insight into the generalization ability of today's emotion recognition engines since training and test data used for system development usually tend to be similar as far as recording conditions, noise overlay, language, and types of emotions are concerned. A considerably more realistic impression can be gathered by interset evaluation: We therefore show results employing six standard databases in a cross-corpora evaluation experiment which could also be helpful for learning about chances to add resources for training and overcoming the typical sparseness in the field. 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 intracorpus testing.

Index Terms—Affective computing, speech emotion recognition, cross-corpus evaluation, normalization

1 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]. When evaluating the accuracy of emotion recognition engines, obtainable performances are often overestimated since, usually, acted or elicited emotions are considered instead of spontaneous, "true" emotions, which in turn are harder to recognize. However, lately, language resources that respect such requirements have emerged and have been investigated repeatedly as the Audiovisual Interest Corpus (AVIC) [15], the FAU Aibo Emotion Corpus [16], the HUMAINE database [17], the Sensitive Artifical Listener (SAL) corpus [18], the SmartKom corpus [19], or the Vera am Mittag (VAM) database [20].

Besides such overestimation of obtainable accuracies due to acting, one usually observes a limitation to "prototypical" cases, that is consideration of only such phrases where n of N labelers agree, whereas $n > \frac{N}{2}$. However, an emotion recognition system in practical use has to process "all that comes in" and cannot be restricted to prototypical cases [16], [21], [22], [23], [24]. First light is shed on the difference in some recent studies, including the first comparative challenge on emotion recognition from speech [68].

Finally, another 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.

Cross-corpus evaluations are increasingly used in various machine learning disciplines: In [25] and [26], the usage of heterogeneous data sources for acoustic training of an ASR system is investigated. The authors thereby propose a cross-corpus acoustic normalization method that can be applied in systems using Hidden Markov Models. A selective pruning technique for statistical parsing using cross-corpus data is proposed in [27]. Further areas of research for which cross-corpus experiments are relevant include text classification [28] and sentence paraphrasing via multiple-sequence alignment [29]. In [30], cross-corpus data (elicited and spontaneous speech) is used for signal-adaptive ASR through variable-length time-localized features. For emotion recognition, several studies already

- *B. Schuller, F. Eyben, M. Wöllmer, and G. Rigoll are with the Institute for Human-Machine Communication, Technische Universität München, D-80333 München, Germany.*
E-mail: {schuller, eyben, woellmer, rigoll}@tym.de.
- *B. Vlasenko and A. Wendemuth are with the Cognitive Systems Group, IESK, Otto-von-Guericker Universität (OVGU), D-39106 Magdeburg, Germany. E-mail: {bogdan.vlasenko, andreas.wendemuth}@ovgu.de.*
- *A. Stuhlsatz is with the Laboratory for Pattern Recognition/Department of Electrical Engineering, University of Applied Sciences Düsseldorf, Germany. E-mail: andreas.stuhlsatz@fh-dusseldorf.de.*

Manuscript received 10 Dec. 2009; revised 25 May 2010; accepted 4 Aug. 2010; published online 17 Aug. 2010.
Recommended for acceptance by A. Batliner.
For information on obtaining reprints of this article, please send e-mail to: tac@computer.org, and reference IEEECS Log Number TAFFC-2009-12-0006.
Digital Object Identifier no. 10.1109/T-AFFC.2010.8.

1949-3045/10/$26.00 © 2010 IEEE Published by the IEEE Computer Society

provide accuracies on multiple corpora—however, only a very few consider training on one and testing on a completely different one (e.g., [31] and [32], 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 a classification technique, we follow the approach of supra-segmental feature analysis via Support Vector Machines by projection of the multivariate 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 [68].

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 the 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 (Section 2) with a general commentary on the present situation. We next get on track with features and classification (Section 3). Then, we consider normalization to improve performance in Section 4. Some comments will follow on evaluation (Section 5) before concluding this article (Section 6).

2 Selected Databases

One of the major needs of the community ever since—maybe even more than in many related pattern recognition tasks—is the constant wish for data sets [33], [34]. In the early days of the late 1990s, these were not only few, but also small (≈ 500 turns) with few subjects (≈ 10), unimodal, recorded in studio noise conditions, and acted. Further, the spoken content was mostly predefined (e.g., the Danish Emotional Speech Corpus (DES) [35], the Berlin Emotional Speech-Database (EMO-DB) [36], and the Speech Under Simulated and Actual Stress (SUSAS) database [37]). These were seldom made public and few annotators—if any at all—labeled, usually exclusively, the perceived emotion. Additionally, these were partly not intended for analysis, but for quality measurement of synthesis (e.g., the DES, EMO-DB databases). However, any data is better than none. Today we are happy to see more diverse emotions covered, more elicited or even spontaneous sets of many speakers, larger amounts of instances (5k-10k) of more subjects (up to more than 100), and multimodal data that is annotated by multiple labelers (4 (AVIC)-17 (VAM)). It thereby lies in the nature of collecting acted data that equal distribution among classes is easily obtainable. In more spontaneous sets this is not given, which forces one to either balance in the training or shift from reporting of simple recognition rates to F-measures or unweighted recall values, best per class (e.g., FAU Aibo Emotion Corpus and AVIC database). However, some acted and elicited data sets with predefined content are still seen (e.g., the eNTERACE corpus [38]), yet these also follow the trend of more instances and speakers. Positively, transcription is also becoming more and more rich: additional annotation of spoken content and nonlinguistic interjections (e.g., FAU Aibo Emotion Corpus, AVIC database), multiple annotator tracks (e.g., VAM corpus), or even manually corrected pitch contours (FAU Aibo Emotion Corpus) and additional audio tracks in different recordings (e.g., close-talk and room-microphone), phoneme boundaries and manual phoneme labeling (e.g., EMO-DB), different chunkings (e.g., FAU Aibo Emotion Corpus), as well as indications of the degree of inter-labeler-agreement for each speech turn. At the same time, these are also partly recorded under more realistic conditions (or taken from the media). However, in future sets, multilinguality and subjects of diverse cultural backgrounds will be needed in addition to all named positive trends.

For the following cross-corpora investigations, we chose six among the most frequently used and well known. Only those 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

TABLE 1
Mapping of Emotions for the Clustering to a Binary Arousal 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

TABLE 2
Mapping of Emotions for the Clustering to a Binary Valence Discrimination Task

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

respect to the number of subjects involved, naturalness, spontaneity, and free language, as covered by the AVIC and SmartKom [19] 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 therefore had to leave out several emotional or broader affective states such 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 [68]—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's 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 3. 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 Tables 1 and 2. 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.).

Next, we will briefly introduce the sets.

2.1 Danish Emotional Speech

The Danish Emotional Speech [35] database has been chosen as the first set as one of the "traditional representatives" for our study because it is easily accessible. Also, several results were already reported on it [39], [40], [41]. The data used in the experiments are nine Danish sentences, two words and chunks that are located between two silent segments of two passages of fluent text. For example: "*Nej*" (*No*), "*Ja*" (*Yes*), "*Hvor skal du hen?*" (*Where are you going?*). The total amount of data sums up to more than 400 speech utterances (i.e., speech segments between two silence pauses) which are expressed by four professional actors, two males and two females. All utterances are balanced for each gender, i.e., every utterance is spoken by a male and a female speaker. Speech is expressed in five emotional states: *anger, happiness, neutral, sadness, and surprise.* The actors were asked to express each sentence in all five emotional states. The sentences were labeled according to the state

TABLE 3
Details of the Six Emotion Corpora

Corpus	Content	# Emotion								# Arousal		# Valence		# All	Else	Time	# Sub	Type	Rate
		N	J	A	F	SU	SA	B	D	−	+	−	+			h:mm			kHz
EMO-DB	German fixed	78	64	127	55	–	53	79	38	248	246	352	142	494	–	0:22	5 f 5 m	acted studio	16
DES	Danish fixed	85	86	85	–	79	84	–	–	169	250	169	250	419	–	0:28	2 f 2 m	acted normal	20
eNTERFACE	English fixed	–	205	200	189	192	195	–	189	397	773	786	384	1 170	–	0:58	8 f 34 m	acted normal	16
SUSAS	English fixed	701	–	–	484	–	–	–	–	701	484	701	484	1 185	hs, ms	0:20	3 f 4 m	mixed noisy	8
AVIC	English variable	510	170	–	–	–	–	316	–	170	826	170	826	996	–	0:35	10 f 11 m	natural normal	44.1
SmartKom	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**	–	–

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 (Rate). Emotion categories: anger (A), boredom (B), disgust (D), fear/screaming (F), joy(ful)/happy/happiness (J), neutral (N), sad(ness) (SA), surprise (SU); noncommon further contained states: helplessness (he), medium stress (ms), pondering (p), unidentifiable (u).

they should be expressed in, i.e., one emotion label was assigned to each sentence. In a listening experiment, 20 participants (native speakers from 18 to 59 years old) verified the emotions with an average score rate of 67 percent in [35].

2.2 Berlin Emotional Speech Database

A further well-known set chosen with which to test the effectiveness of cross-corpora emotion classification is the popular studio recorded Berlin Emotional Speech Database (EMO-DB) [36], which covers *anger, boredom, disgust, fear, joy, neutral, and sadness* as speaker emotions. The spoken content is again predefined by 10 German emotionally neutral sentences like "*Der Lappen liegt auf dem Eisschrank*" (*The cloth is lying on the fridge.*). The actors were asked to express each sentence in all seven emotional states. The sentences were labeled according to the state they should be expressed in, i.e., one emotion label was assigned to each sentence. As DES, it thus provides a high number of repeated words in diverse emotions. Ten (five female) professional actors speak 10 sentences. While the whole set is comprised of around 900 utterances, only 494 phrases are marked as minimum 60 percent natural and minimum 80 percent agreement by 20 subjects in a listening experiment. This selection is usually used in the literature reporting results on the corpus (e.g., [42], [43], [44], and in this article). Mean accuracy of 84.3 percent is the result of the perception study for this limited "more prototypical" subset.

2.3 eNTERFACE

The eNTERFACE [38] corpus is a further public, yet audiovisual emotion database. It contains the induced emotions *anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise*. Forty-two subjects (eight female) from 14 nations are included. Contained are office environment recordings of predefined spoken content in English. Each subject was instructed to listen to six successive short stories, each of them intended to elicit a particular emotion. They then had to react to each of the situations by uttering previously read phrases that fit the short story. Five phrases are available per emotion, as "*I have nothing to give you! Please don't hurt me!*" in the case of fear. Two experts judged whether the reaction expressed the intended emotion in an unambiguous way. Only if this was the case was a sample (= sentence) added to the database. Therefore, each sentence in the database has one assigned emotion label, which indicates the emotion expressed by the speaker in this sentence. Overall, eNTERFACE consists of 1,170 instances. Research results are reported, e.g., in [45], [46], [47].

2.4 Speech under Simulated and Actual Stress

The SUSAS [37] database serves as a first reference for spontaneous recordings. As an additional challenge, speech is partly masked by field noise. We decided on the 3,663 "actual stress" speech samples recorded in "subject motion fear and stress tasks." Seven speakers, three of them female, in roller coaster and free fall situations are contained in this set. Next to *neutral* speech and *fear*, two different stress conditions have been collected: *medium stress* and *high stress*, which are not used in this article as they are specific to this set. SUSAS is also restricted to a predefined spoken text of 35 English air-commands, such as "*brake*," "*help*," or "*no*." Likewise, only single words are contained, similarly to DES, where this is also mostly the case. SUSAS is also popular with respect to the number of reported results (e.g., [39], [48], [49], [50], [51], [52], [53]).

2.5 Audiovisual Interest Corpus

In order to add spontaneous emotion samples of nonrestricted spoken content, we decided to include the Audiovisual Interest Corpus (AVIC) [15] in our experiments. It is a further audiovisual emotion corpus containing recordings during which a product presenter leads one of 21 subjects (10 female) through an English commercial presentation. The level of interest is annotated for every turn and reaches from *boredom* (subject is bored with listening and talking about the topic, very passive, does not follow the discourse), over *neutral* (subject follows and participates in the discourse, it cannot be recognized if she/he is interested in the topic) to *joyful* interaction (strong wish of the subject to talk and learn more about the topic). Four annotators listened to the turns and rated them in terms of these three categories. The overall rating of the turn was computed from the majority label of the four annotators. If no majority label exists, the turn is discarded and not included in the database, leaving 996 turns in the database. The AVIC corpus also includes annotations of the spoken content and non-linguistic vocalizations. For our evaluation we use the 996 phrases as, e.g., employed in [15], [24], [54], [55].

2.6 SmartKom

Finally, we include a second corpus of spontaneous speech and natural emotion in our tests: The SmartKom [19] multimodal corpus consists of Wizard-Of-Oz dialogues in German and English. For our evaluations we use German dialogues recorded during a public environment technical scenario. Street noise is present in all of the original recordings, in contrast to the SUSAS database, where noise is partly overlaid. The database contains multiple audio channels and two video channels (face, body from side). The primary aim of the corpus was the empirical study of human-computer interaction in a number of different tasks and technical setups. It is structured into sessions which contain one recording of approximately 4.5 min length with one person. The labelers could look at the persons' facial expressions, body gestures, and listen to his/her speech. The labeling was frame-based, i.e., the beginning and the end of an emotional episode was marked on the time axis and a majority voting was conducted to translate the frame-based labeling to a per-turn labeling, as it is used in this study. Utterances are labeled in seven broader emotional states: *neutral*, *joy*, *anger*, *helplessness*, *pondering*, and *surprise* are contained together with unidentifiable episodes.

The SmartKom data collection is used in over 250 studies, as reported in [56]. Some interesting examples include, e.g., [57], [58], [59].

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 used. 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.

3 Features and Classification

In the past, focus was placed on prosodic features, in particular pitch, durations, and intensity, where comparably small feature sets (10-100) were utilized [48], [60], [61], [62], [63], [64], [65]. Thereby, only sparse studies saw low-level feature modeling on a frame level as alternative: usually by Hidden Markov Models (HMM) or Gaussian Mixture Models (GMM) [63], [64], [66]. The higher success of static feature vectors derived by projection of the low-level contours as pitch or energy by descriptive statistical functional application such as lower order moments (mean, standard deviation) or extrema [67] is probably justified by the supra-segmental nature of the phenomena occurring with respect to emotional content in speech [24], [68]. In more recent research, however, voice quality features such as HNR, jitter, or shimmer, and spectral and cepstral as formants and MFCC have also become more or less the "new standard" [69], [70], [71], [72]. At the same time brute-forcing of features (1,000 up to 50,000) by analytical feature generation, partly also in combination with evolutionary generation, is seen increasingly often [73]. It seems as if this was, at the time, able to outperform hand-crafted features at a high number of such [68]. However, the individual worth of automatically generated features seems to be lower in return.

Further, linguistic features are often added these days, and will certainly also be in the future [74], [75], [76]. However, as our databases stem from the same language group, but different languages, these are of limited utility in this article. Further problems would certainly arise with respect to the recognition of cross-corpus recognition of *affective* speech, which in itself is still a mostly untouched topic [77].

Following these considerations, we decided on 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 4, and their first order delta coefficients. These 37×2 descriptors are next smoothed by low-pass filtering with a simple moving average filter.

These features already stood the test in manifold studies (e.g., [15], [41], [52],, [55] [78], [79], [80], [81], [82] [83], [84], [85], [86]).

We derive statistics per speaker turn by a projection of each univariate time series—the Low-Level-Descriptors—onto a scalar feature independent of the length of the turn. This is done by the use of functionals.

Nineteen functionals are applied to each contour on the word level covering extremes, ranges, positions, first four moments, and quartiles, as also shown in Table 4. Note that three functionals are related to time (position in time) with the physical unit milliseconds.

TABLE 4
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

Classifiers used in the literature comprise a broad variety [87]. Depending on the feature type considered for classification (cf. Section 3), either dynamic algorithms [88] for processing on a frame-level or static for higher level statistical functionals [67] are found. Among dynamic algorithms Hidden Markov Models are predominant (cf., e.g., [63], [64], [66], [88]). Also, Multi Instance Learning is found as a "bag-of-frames" approach on this level (e.g., [31]). A seldom applied alternative is Dynamic Time Warping, favoring easy adaptation. In the future, generally popular Dynamic Bayesian Network architectures [89] could help to combine features on different time levels as spectral on a per frame basis and prosodic being rather supra-segmental. With respect to static classification, the list of classifiers seems endless: neural networks (mostly Multi-Layer Perceptrons) [75], Bayes classifier [67], Baysian Networks [88], [90], Gaussian Mixture Models [71], [91], Decision Trees [92], Random Forests [93], k-Nearest Neighbor distance classifiers [94], and Support Vector Machines [88], [95], [96] are found most often. Also, a selection of ensemble techniques [97], [98] has been applied as Boosting, Bagging, Multiboosting, and Stacking with and without confidences. New emerging techniques such as Long-Short-Term-Memory Recurrent Neural Networks [18], Hidden Conditional Random Fields [18], Tandem Gaussian Mixture Models with Support Vector Machines [99], or GentleBoosting could further be seen more frequently soon. A promising side-trend is also the fusion of dynamic and static classification as inspired by [68], where more research on how to best model which types will reveal the true potential.

Again, following these considerations, we choose the most frequently encountered solution (e.g., in [24], [95], [96], [100], [101]) for representative results in Sections 4 and 5: Support Vector Machine (SVM) classification. Thereby, we use a linear kernel and pairwise multiclass discrimination [102].

4 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

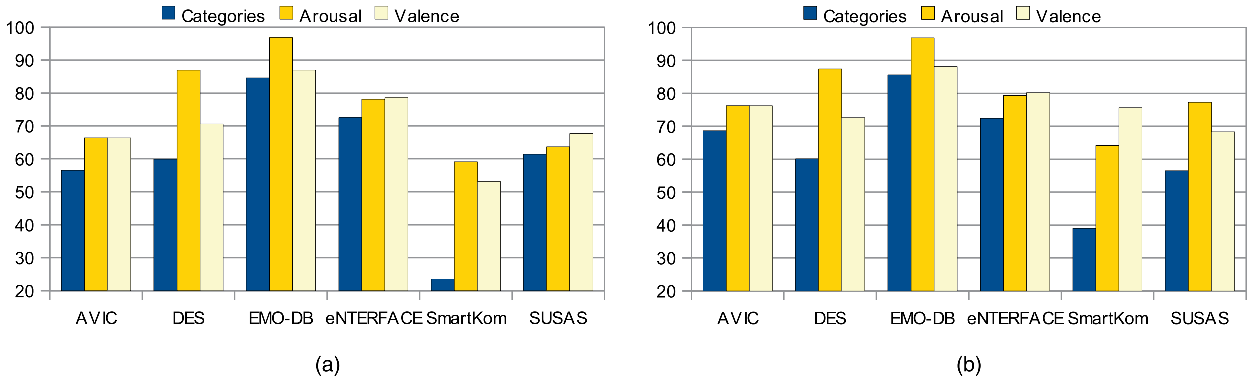

Fig. 1. (a) Unweighted and (b) weighted average recall (UAR/WAR) in percentage of within corpus evaluations on all six corpora using corpus normalization (CN). Results for all emotion categores present with the particular corpus, binary arousal, and binary valence.

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).

5 Evaluation

Early studies started with speaker dependent recognition of emotion, just as in the recognition of speech [91], [64], [69]. 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., [103]. 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. [104]). Still, only cross-corpora evaluation encompasses realistic testing conditions which a commercial emotion recognition product used in everyday life would frequently have to face.

The within corpus evaluations' results—intended for a first reference—are sketched in Figs. 1a and 1b. 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 [68]. Only where appropriate will the weighted average recall (WAR, i.e., accuracy) be provided in addition. For the inter-corpus results only minor differences exist between these two measures due to the mostly acted and elicited nature of the corpora, where instances can easily be collected balanced among classes. The results shown in Figs. 1a and 1b 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 5 lists all 23 different training and test set combinations we evaluated in our cross-corpus experiments. As mentioned before, AVIC and SUSAS 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 3). In order to obtain combinations for which up to six emotion classes occur in the training and test set, we included experiments in which

TABLE 5
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

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). Depending on the maximum number of different emotion classes that can be modeled in a certain experiment and depending on the number of classes we actually use (two to six), we get a certain number of possible emotion class permutations according to Table 5. 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 5). 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.

We next strive to reveal the optimal normalization strategy from those introduced in Section 4 (refer to Tables 6 and 7 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 the corresponding element in a maximum result vector v_{max}. This result vector contains the result for all tests and any permutation arising from the exclusion and clustering of classes (see also Table 5). 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 Tables 6 and 7 is a rather abstract performance measure, indicating the *relative* performance difference between the normalization strategies, rather than the *absolute* recognition accuracy.

TABLE 6
Weighted Average Recall (WAR) = Accuracy

Accuracy DTM	# classes 2	3	4	5	6	V	A	mean
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

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 obtained throughout all class permutations and for all tests. Detailed explanation in the text.

Here, we consider mean weighted average recall (= accuracy, Table 6) and—as before—mean unweighted recall (UAR) (Table 7) for the comparison, as some data sets are not in balance with respect to classes (cf. Table 3). In the case of accuracy, no significant difference [105] between speaker and combined speaker and corpus normalization is found. As the latter is comprised of 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 Figs. 2a, 2b, 2c, and 2d). 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)

TABLE 7
Unweighted Average Recall (UAR)

Recall DTM	# classes 2	3	4	5	6	V	A	mean
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

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 recall rate over the maximum obtained throughout all class permutations and for all tests. Detailed explanation in the text.

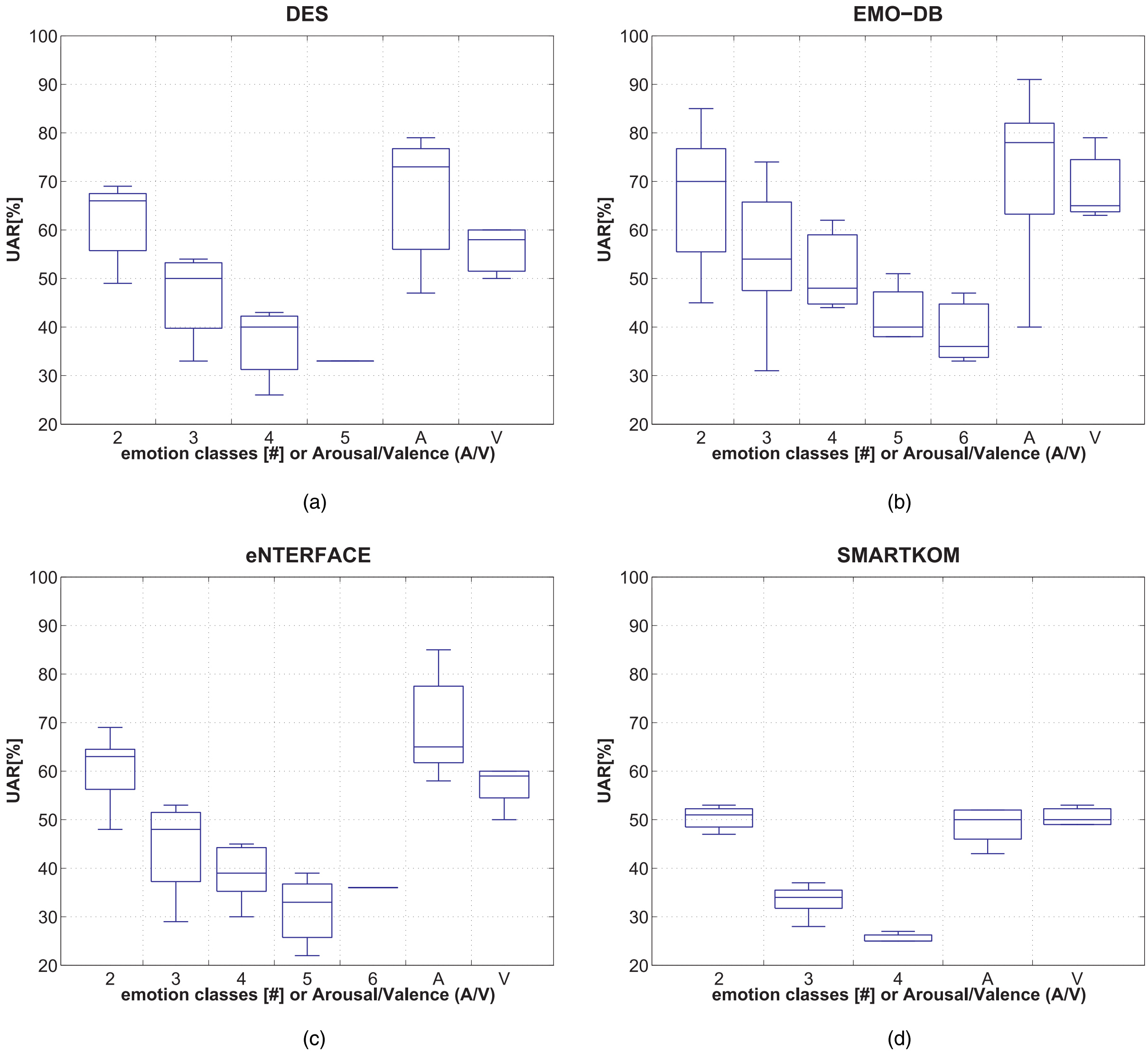

Fig. 2. Box-plots for unweighted average recall (UA) in percentage for cross-corpora testing on four test corpora. Results obtained for varying number of classes (2-6) and for classes mapped to high/low arousal (A) and positive/negative valence (V). (a) DES, UAR. (b) EMO-DB, UAR. (c) eNTERFACE, UAR. (d) SMARTKOM, UAR.

of classes (emotion categories) and the binary arousal and valence tasks.

First, the DES set is chosen for testing, as depicted in Fig. 2a. For training, five different combinations of the remaining sets are used (see Table 5). 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 percent.

A very similar overall behavior is observed for the EMO-DB in Fig. 2b. 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 due to the known lower "difficulty" of this set.

Switching from acted to mood-induced, we provide results on eNTERFACE in Fig. 2c. 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 nonrestricted varying spoken content and natural emotion we note the challenge arising from the SmartKom set in Fig. 2d: 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 [68]) 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 due to relying on a few labelers, often with high disagreement among these.

6 Concluding Remarks

Summing up, we have shown results for intra and inter-corpus recognition of emotion from speech. By that we have learned that the accuracy and mean recall rates highly depend on the specific subgroup 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 predefined spoken content. The fact that three different languages—Danish, English, and German—are contained in the tested corpora 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 the limitations of current systems. Here, only a few groups of emotions stood out in comparison to chance level.

To better cope with the differences among corpora, we evaluated different normalization approaches, wherein speaker normalization led to the best results. For all experiments, we 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.

Cross-corpus experiments and applications will also profit from techniques that automatically determine similarity between multiple databases (e.g., as in [106]). This in turn requires the definition of similarity measures in order to find out in what respect and to what degree it is necessary to adapt emotional speech data before it is used for training or evaluation. Furthermore, measuring similarity is useful to determine which corpora can be combined to overcome the ever-present sparseness of training data and which characteristics have to be modeled separately. Also, measures can be thought of to evaluate which corpora resemble each other most and by which emotions. By that, adaptation of a model with additional data from diverse further corpora can be improved by selecting only suited instances. An important criteria for corpus similarity that is specific to the area of emotion recognition is the issue of annotation: The ground truth labels assigned to different corpora are not only a result of subjective ratings but also depend on the task for which the respective corpus had been recorded. Thus, the "vocabulary" of annotated emotions varies from database to database and makes it difficult to combine multiple corpora. In order to provide a general basis of mapping annotation schemes onto each other, an interface definition will be needed (as, e.g., suggested by the Emotion Markup Language[1] or similar endeavors). Such definitions enable a unified relabeling of existing databases as a basis for future cross-corpora experiments. In addition to overcoming different "vocabularies," strategies will be needed to cope with the different units of annotation as frames, words, or turns.

Next, inter-corpus feature selection and verification of their merit will be needed in addition to the manifold studies evaluating feature values on single corpora.

Since cross-corpus experiments have already been conducted in many machine learning disciplines (e.g., [25], [26], [27], [28], [29], [30]), future research on increasing the generalization ability of systems for automatic emotion recognition should also focus on transferring adaptation strategies developed for other speech-related tasks to the area of emotion recognition. Examples for successful techniques can be found in the domain of signal-adaptive ASR [30] or cross-corpus acoustic normalization for HMMs [25]. GMM-based approaches toward emotion recognition might profit from adaptation techniques that are well known in the field of automatic speech recognition such as Maximum Likelihood Linear Regression (MLLR). However, the applicability of methods tailored for speech recognition will heavily depend on the classifier type that is used for emotion recognition.

Finally, acoustic training from multiple sources or corpora can be advantageous not only for emotion recognition: Using a broad variety of different corpora, e.g., for training detectors of nonlinguistic vocalizations, might result in better accuracies.

No linguistic feature information was used herein, opposing our very good experience with acoustic and linguistic feature integration [24]. However, inter-corpus ASR of emotional speech will have to be investigated first. Also, most of the corpora considered herein would not have allowed for reasonable linguistic information exploitation as they utilize predefined and highly limited spoken content. In this respect, more sets with natural speech will thus be needed.

Considering the fact that little experience with emotion recognition products in everyday life has so far been

1. http://www.w3.org/2005/Incubator/emotion/XGR-emotionml/.

gathered, we see that cross-corpus evaluation is a helpful method to thoroughly research the performance of an emotion recognition engine in real-life usage and the challenges which it faces. Using many different corpora allows benchmarking of factors from varying acoustic environment, recording conditions, interaction type (acted, spontaneous), textual content, to cultural and social background, and type of application.

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

Acknowledgments

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 with 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," *J. Acoustical Soc. Am.,* vol. 52, pp. 1238-1250, 1972.

[5] K.R. Scherer, "Vocal Affect Expression: A Review and a Model for Future Research," *Psychological Bull.,* vol. 99, pp. 143-165, 1986.

[6] C. Whissell, "The Dictionary of Affect in Language," *Emotion: Theory, Research and Experience,* vol. 4, *The Measurement of Emotions,* R. Plutchik and H. Kellerman, eds., pp. 113-131, Academic Press, 1989.

[7] R. Picard, *Affective Computing.* 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 People Really Talk and Why Engineers Should Care," *Proc. EUROSPEECH,* pp. 1781-1784, 2005.

[10] C.M. Lee and S.S. Narayanan, "Toward Detecting Emotions in Spoken Dialogs," *IEEE Trans. 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?" *Proc. Second Int'l Conf. Affective Computing and Intelligent Interaction,* pp. 440-451, 2007.

[12] A. Wendemuth, J. Braun, B. Michaelis, F. Ohl, D. Rösner, H. Scheich, and R. Warnem, "Neurobiologically Inspired, Multimodal Intention Recognition for Technical Communication Systems (NIMITEK)," *Proc. Fourth IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-based Systems,* pp. 141-144, 2008.

[13] M. Schrö, R. Cowie, D. Heylen, M. Pantic, C. Pelachaud, and B. Schuller, "Towards Responsive Sensitive Artificial Listeners," *Proc. Fourth Int'l Workshop on Human-Computer Conversation,* 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 Trans. Pattern Analysis and Machine Intelligence,* vol. 31, no. 1, pp. 39-58, Jan. 2009.

[15] B. Schuller, R. Müller, B. Hörnler, A. Höthker, H. Konosu, and G. Rigoll, "Audiovisual Recognition of Spontaneous Interest within Conversations," *Proc. Int'l Conf. Multimodal Interfaces,* pp. 30-37, 2007.

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

[17] 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, "The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data," *Proc. Int'l Conf. Affective Computing and Intelligent Interaction,* A. Paiva, R. Prada, and R.W. Picard, eds., pp. 488-500, 2007.

[18] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning Emotion Classes—Towards Continuous Emotion Recognition with Modelling of Long-Range Dependencies," *Proc. INTERSPEECH,* pp. 597-600, 2008.

[19] S. Steininger, F. Schiel, O. Dioubina, and S. Raubold, "Development of User-State Conventions for the Multimodal Corpus in Smartkom," *Proc. Workshop Multimodal Resources and Multimodal Systems Evaluation,* pp. 33-37, 2002.

[20] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German Audio-Visual Emotional Speech Database," *Proc. Int'l Conf. Multimedia & Expo,* pp. 865-868, 2008.

[21] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in Real-Life Emotion Annotation and Machine Learning Based Detection," *Neural Networks,* vol. 18, no. 4, pp. 407-422, 2005.

[22] L. Devillers and L. Vidrascu, "Real-Life Emotion Recognition in Speech," *Speaker Classification II,* pp. 34-42, Sept. 2007.

[23] A. Batliner, D. Seppi, B. Schuller, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, and V. Aharonson, "Patterns, Prototypes, Performance," *Proc. HSS-Cooperation Seminar Pattern Recognition in Medical and Health Eng.,* J. Hornegger, K. Höller, P. Ritt, A. Borsdorf, and H.P. Niedermeier, eds., pp. 85-86, 2008.

[24] 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 J.,* vol. 27, no. 12, pp. 1760-1774, 2009.

[25] S. Tsakalidis and W. Byrne, "Acoustic Training from Heterogeneous Data Sources: Experiments in Mandarin Conversational Telephone Speech Transcription," *Proc. IEEE Int'l Conf. Aacoustics, Speech, and Signal Processing,* 2005.

[26] S. Tsakalidis, "Linear Transforms in Automatic Speech Recognition: Estimation Procedures and Integration of Diverse Acoustic Data," PhD dissertation, 2005.

[27] D. Gildea, "Corpus Variation and Parser Performance," *Proc. Conf. Empirical Methods in Natural Language Processing,* pp. 167-202, 2001.

[28] Y. Yang, T. Ault, and T. Pierce, "Combining Multiple Learning Strategies for Effective Cross Validation," *Proc. 17th Int'l Conf. Machine Learning,* pp. 1167-1174, 2000.

[29] R. Barzilay and L. Lee, "Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment," *Proc. Human Language Technology Conf.-North Am. Chapter Assoc. Computational Linguistics Conf.,* pp. 16-23, 2003.

[30] K. Soenmez, M. Plauche, E. Shriberg, and H. Franco, "Consonant Discrimination in Elicited and Spontaneous Speech: A Case for Signal-Adaptive Front Ends in ASR," *Proc. Int'l Conf. Spoken Language Processing,* pp. 548-551, 2000.

[31] M. Shami and W. Verhelst, "Automatic Classification of Emotions in Speech Using Multi-Corpora Approaches," *Proc. Second Ann. IEEE BENELUX/DSP Valley Signal Processing Symp.,* pp. 3-6, 2006.

[32] M. Shami and W. Verhelst, "Automatic Classification of Expressiveness in Speech: A Multi-Corpus Study," *Speaker Classification II,* C. Müller, ed., pp. 43-56, 2007.

[33] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional Speech: Towards a New Generation of Databases," *Speech Comm.,* vol. 40, nos. 1-2, pp. 33-60, 2003.

[34] D. Ververidis and C. Kotropoulos, "A Review of Emotional Speech Databases," *Proc. Panhellenic Conf. Informatics*, pp. 560-574, 2003.

[35] I.S. Engbert and A.V. Hansen, "Documentation of the Danish Emotional Speech Database DES," technical report, Center for PersonKommunikation, Aalborg Univ., Denmark, 2007.

[36] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," *Proc. INTERSPEECH*, pp. 1517-1520, 2005.

[37] J. Hansen and S. Bou-Ghazale, "Getting Started with SUSAS: A Speech under Simulated and Actual Stress Database," *Proc. EUROSPEECH*, vol. 4, pp. 1743-1746, 1997.

[38] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 Audio-Visual Emotion Database," *Proc. IEEE Workshop Multimedia Database Management*, 2006.

[39] D. Ververidis and C. Kotropoulos, "Fast Sequential Floating Forward Selection Applied to Emotional Speech Features Estimated on DES and SUSAS Data Collection," *Proc. European Signal Processing Conf.*, 2006.

[40] D. Datcu and L.J. Rothkrantz, "The Recognition of Emotions from Speech Using Gentleboost Classifier. A Comparison Approach," *Proc. Int'l Conf. Computer Systems and Technologies*, vol. 1, pp. 1-6, 2006.

[41] B. Schuller, D. Seppi, A. Batliner, A. Meier, and S. Steidl, "Towards More Reality in the Recognition of Emotional Speech," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, pp. 941-944, 2007.

[42] H. Meng, J. Pittermann, A. Pittermann, and W. Minker, "Combined Speech-Emotion Recognition for Spoken Human-Computer Interfaces," *Proc. IEEE Int'l Conf. Signal Processing and Comm.*, 2007.

[43] V. Slavova, W. Verhelst, and H. Sahli, "A Cognitive Science Reasoning in Recognition of Emotions in Audio-Visual Speech," *Int'l J. Information Technologies and Knowledge*, vol. 2, pp. 324-334, 2008.

[44] B. Schuller, M. Wimmer, L. Mösenlechner, C. Kern, and G. Rigoll, "Brute-Forcing Hierarchical Functionals for Paralinguistics: A Waste of Feature Space?" *Proc. IEEE Int'l Conf. Acoustics, Speech, and Sigal Processing*, pp. 4501-4504, 2008.

[45] D. Datcu and L.J. M. Rothkrantz, "Semantic Audio-Visual Data Fusion for Automatic Emotion Recognition," *Proc. Euromedia*, 2008.

[46] M. Mansoorizadeh and N.M. Charkari, "Bimodal Person-Dependent Emotion Recognition Comparison of Feature Level and Decision Level Information Fusion," *Proc. First Int'l Conf. Pervasive Technologies Related to Assistive Environments*, pp. 1-4, 2008.

[47] M. Paleari, R. Benmokhtar, and B. Huet, "Evidence Theory-Based Multimodal Emotion Recognition," *Proc. 15th Int'l Multimedia Modeling Conf. on Advances in Multimedia Modeling*, pp. 435-446, 2008.

[48] D. Cairns and J.H. L. Hansen, "Nonlinear Analysis and Detection of Speech under Stressed Conditions," *J. Acoustical Soc. Am.*, vol. 96, no. 6, pp. 3392-3400, Dec. 1994.

[49] L. Bosch, "Emotions: What Is Possible in the ASR Framework?" *Proc. ISCA Workshop Speech and Emotion*, pp. 189-194, 2000.

[50] G. Zhou, J.H.L. Hansen, and J.F. Kaiser, "Nonlinear Feature Based Classification of Speech under Stress," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 3, pp. 201-216, Mar. 2001.

[51] R.S. Bolia and R.E. Slyh, "Perception of Stress and Speaking Style for Selected Elements of the SUSAS Database," *Speech Comm.*, vol. 40, no. 4, pp. 493-501, 2003.

[52] B. Schuller, M. Wimmer, D. Arsic, T. Moosmayr, and G. Rigoll, "Detection of Security Related Affect and Behaviour in Passenger Transport," *Proc. INTERSPEECH*, pp. 265-268, 2008.

[53] L. He, M. Lech, N. Maddage, and N. Allen, "Stress and Emotion Recognition Based on Log-Gabor Filter Analysis of Speech Spectrograms," *Proc. Int'l Conf. Affective Computing and Intelligent Interaction*, 2009.

[54] B. Schuller, N. Köhler, R. Müller, and G. Rigoll, "Recognition of Interest in Human Conversational Speech," *Proc. INTERSPEECH*, pp. 793-796, 2006.

[55] B. Vlasenko, B. Schuller, K. Tadesse Mengistu, and G. Rigoll, "Balancing Spoken Content Adaptation and Unit Length in the Recognition of Emotion and Interest," *Proc. INTERSPEECH*, pp. 805-808, 2008.

[56] W. Wahlster, "Smartkom: Symmetric Multimodality in an Adaptive and Reusable Dialogue Shell," *Proc. Human Computer Interaction Status Conf.*, pp. 47-62, 2003.

[57] D. Oppermann, F. Schiel, S. Steininger, and N. Beringer, "Off-Talk—A Problem for Human-Machine-Interaction?" *Proc. EUROSPEECH*, pp. 2197-2200, 2001.

[58] A. Schweitzer, N. Braunschweiler, T. Klankert, B. Säiber;lich, and B. Möbius, "Restricted Unlimited Domain Synthesis," *Proc. EUROSPEECH*, pp. 1321-1324, 2003.

[59] T. Vogt and E. André, "Improving Automatic Emotion Recognition from Speech via Gender Differentiation," *Proc. Int'l Conf. Language Resources and Evaluation*, 2006.

[60] R. Banse and K.R. Scherer, "Acoustic Profiles in Vocal Emotion Expression," *J. Personality and Social Psychology*, vol. 70, no. 3, pp. 614-636, 1996.

[61] Y. Li and Y. Zhao, "Recognizing Emotions in Speech Using Short-Term and Long-Term Features," *Proc. Int'l Conf. Spoken Language Processing*, p. 379, 1998.

[62] G. Zhou, J.H.L. Hansen, and J.F. Kaiser, "Linear and Nonlinear Speech Feature Analysis for Stress Classification," *Proc. Int'l Conf. Spoken Language Processing*, 1998.

[63] T.L. Nwe, S.W. Foo, and L.C. De Silva, "Classification of Stress in Speech Using Linear and Nonlinear Features," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, vol. 2, pp. II-9-12, 2003.

[64] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov Model-Based Speech Emotion Recognition," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, pp. 1-4, 2003.

[65] C.M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "Emotion Recognition Based on Phoneme Classes," *Proc. Int'l Conf. Spoken Language Processing*, 2004.

[66] B. Vlasenko and A. Wendemuth, "Tuning Hidden Markov Model for Speech Emotion Recognition," *Proc. DAGA*, Mar. 2007.

[67] D. Ververidis and C. Kotropoulos, "Automatic Speech Classification to Five Emotional States Based on Gender Information," *Proc. EUSIPCO*, pp. 341-344, 2004.

[68] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," *Proc. INTERSPEECH*, 2009.

[69] 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," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing*, vol. 1, 2006.

[70] 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," *Proc. INTERSPEECH*, pp. 2253-2256, 2007.

[71] M. Lugger and B. Yang, "An Incremental Analysis of Different Feature Groups in Speaker Independent Emotion Recognition," *Proc. Int'l Congress Phonetic Sciences*, pp. 2149-2152, Aug. 2007.

[72] B. Schuller, M. Wöllmer, F. Eyben, and G. Rigoll, *The Role of Prosody in Affective Speech*, pp. 285-307. Peter Lan Publishing Group, 2009.

[73] B. Schuller, M. Wimmer, L. Mösenlechner, C. Kern, D. Arsic, and G. Rigoll, "Brute-Forcing Hierarchical Functionals for Paralinguistics: A Waste of Feature Space?" *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, Apr. 2008.

[74] L. Devillers, L. Lamel, and I. Vasilescu, "Emotion Detection in Task-Oriented Spoken Dialogs," *Proc. Int'l Conf. Multimedia & Expo*, July 2003.

[75] B. Schuller, G. Rigoll, and M. Lang, "Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine-Belief Network Architecture," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, vol. 1, 2004.

[76] B. Schuller, R. Jiménez Villar, G. Rigoll, and M. Lang, "Meta-Classifiers in Acoustic and Linguistic Feature Fusion-Based Affect Recognition," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, pp. 325-328, 2005.

[77] T. Athanaselis, S. Bakamidis, I. Dologlou, R. Cowie, E. Douglas-Cowie, and C. Cox, "ASR for Emotional Speech: Clarifying the Issues and Enhancing Performance," *Neural Networks*, no. 18, pp. 437-444, 2005.

[78] A. Batliner, B. Schuller, S. Schaeffler, and S. Steidl, "Mothers, Adults, Children, Pets—Towards the Acoustics of Intimacy," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, pp. 4497-4500, 2008.

[79] B. Schuller, "Speaker, Noise, and Acoustic Space Adaptation for Emotion Recognition in the Automotive Environment," *Tagungsband 8.—ITG-Fachtagung Sprachkommunikation 2008*, vol. ITG 211, VDE, 2008.

[80] B. Schuller, G. Rigoll, S. Can, and H. Feussner, "Emotion Sensitive Speech Control for Human-Robot Interaction in Minimal Invasive Surgery," *Proc. 17th Int'l Symp. Robot and Human Interactive Comm.*, pp. 453-458, 2008.

[81] B. Schuller, B. Vlasenko, D. Arsic, G. Rigoll, and A. Wendemuth, "Combining Speech Recognition and Acoustic Word Emotion Models for Robust Text-Independent Emotion Eecognition," *Proc. Int'l Conf. Multimedia & Expo*, 2008.

[82] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "On the Influence of Phonetic Content Variation for Acoustic Emotion Recognition," *Proc. Fourth IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems*, 2008.

[83] B. Schuller, B. Vlasenko, R. Minguez, G. Rigoll, and A. Wendemuth, "Comparing One and Two-Stage Acoustic Modeling in the Recognition of Emotion in Speech," *Proc. IEEE Workshop Automatic Speech Recognition and Understanding*, pp. 596-600, 2007.

[84] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Combining Frame and Turn-Level Information for Robust Recognition of Emotions within Wpeech," *Proc. INTERSPEECH*, pp. 2249-2252, 2007.

[85] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Frame vs. Turn-Level: Emotion recognition from Speech Considering Static and Dynamic Processing," *Proc. Int'l Conf. Affective Computing and Intelligent Interaction*, A. Paiva, ed., pp. 139-147, 2007.

[86] 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," *Proc. Fifth Slovenian and First Int'l Language Technologies Conf.*, pp. 240-245, 2006.

[87] D. Ververidis and C. Kotropoulos, "Emotional Speech Recognition: Resources, Features, and Methods," *Speech Comm.*, vol. 48, no. 9, pp. 1162-1181, Sept. 2006.

[88] R. Fernandez and R.W. Picard, "Modeling Drivers' Speech under Stress," *Speech Comm.*, vol. 40, nos. 1-2, pp. 145-159, 2003.

[89] C. Lee, C. Busso, S. Lee, and S. Narayanan, "Modeling Mutual Influence of Interlocutor Emotion States in Dyadic Spoken Interactions," *Proc. INTERSPEECH*, pp. 1983-1986, 2009.

[90] I. Cohen, N. Sebe, F.G. Gozman, M.C. Cirelo, and T.S. Huang, "Learning Bayesian Network Classifiers for Facial Expression Recognition Both Labeled and Unlabeled Data," *Proc. IEEE CS Conf Computer Vision and Pattern Recognition*, vol. 1, pp. 595-601, June 2003.

[91] M. Slaney and G. McRoberts, "Baby Ears: A Recognition System for Affective Vocalizations," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, vol. 2, pp. 985-988, 1998.

[92] C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion Recognition Using a Hierarchical Binary Decision Tree Approach," *Proc. INTERSPEECH*, pp. 320-323, 2009.

[93] T. Iliou and C.-N. Anagnostopoulos, "Comparison of Different Classifiers for Emotion Recognition," *Proc. Panhellenic Conf. Informatics*, pp. 102-106, 2009.

[94] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing Emotions in Speech," *Proc. Int'l Conf. Spoken Language Processing*, vol. 3, pp. 1970-1973, 1996.

[95] 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," *Proc. Fifth Slovenian and First Int'l Language Technologies Conf.*, pp. 240-245, 2006.

[96] F. Eyben, M. Wöllmer, and B. Schuller, "openEAR—Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit," *Proc. Int'l Conf. Affective Computing and Intelligent Interaction*, pp. 576-581, 2009.

[97] B. Schuller, M. Lang, and G. Rigoll, "Robust Acoustic Speech Emotion Recognition by Ensembles of Classifiers," *Proc. DAGA*, vol. I, pp. 329-330, 2005.

[98] D. Morrison, R. Wang, and L.C.D. Silva, "Ensemble Methods for Spoken Emotion Recognition in Call-Centres," *Speech Comm.*, vol. 49, no. 2, pp. 98-112, 2007.

[99] M. Kockmann, L. Burget, and J. Cernocky, "Brno University of Technology System for Interspeech 2009 Emotion Challenge," *Proc. INTERSPEECH*, 2009.

[100] B. Schuller, D. Arsic, F. Wallhoff, and G. Rigoll, "Emotion Recognition in the Noise Applying Large Acoustic Feature Sets," *Proc. Speech Prosody*, 2006.

[101] F. Eyben, B. Schuller, and G. Rigoll, "Wearable Assistance for the Ballroom-Dance Hobbyist—Holistic Rhythm Analysis and Dance-Style Classification," *Proc. Int'l Conf. Multimedia & Expo*, 2007.

[102] I.H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, second ed. Morgan Kaufmann, 2005.

[103] M. Grimm, K. Kroschel, and S. Narayanan, "Support Vector Regression for Automatic Recognition of Spontaneous Emotions in Speech," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, vol. 4, 2007.

[104] 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," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, pp. 317-320, 2005.

[105] L. Gillick and S.J. Cox, "Some Statistical Issues in the Comparison of Speech Recognition Algorithms," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, vol. I, pp. 23-26, 1989.

[106] M. Brendel, R. Zaccarelli, B. Schuller, and L. Devillers, "Towards Measuring Similarity between Emotional Corpora," *Proc. Third Int'l Workshop EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, pp. 58-64, 2010.

Bjöorn Schuller received the diploma and doctoral degrees in electrical engineering and information technology from the Technische Universität Munich (TUM) in Munich, where he is tenured as a senior researcher and lecturer on speech processing and pattern recognition. At present he is a visiting researcher in the Imperial College London's Department of Computing in London. Previously, he lived in Paris and worked in the CNRS-LIMSI Spoken Language Processing Group in Orsay, France. He has (co)authored more than 170 publications in peer reviewed books, journals, and conference proceedings in this field. Best known are his works advancing audiovisual processing in the areas of Affective Computing. He serves as a member of the steering committee of the *IEEE Transactions on Affective Computing* and as a guest editor and reviewer for several other scientific journals and as an invited speaker, session organizer and chairman, and program committee member of numerous international workshops and conferences. He is an invited expert in the W3C Emotion and Emotion Markup Language Incubator Groups, and has been repeatedly elected a member of the HUMAINE Association Executive Committee, where he chairs the Special Interest Group on Speech that organized the INTERSPEECH 2009 Emotion Challenge and the INTERSPEECH 2010 Paralinguistic Challenge. He is a member of the IEEE.

Bogdan Vlasenko received the BSc (2005) and MSc (2006) degrees from the National Technical University of Ukraine Kyiv Polytechnic Institute, Kiev, Ukraine, all in electrical engineering and information technology. Since 2006 he has been pursuing the PhD degre in the Department of Cognitive Systems at the Institute for Electronics, Signal Processing and Communications Technology, Otto-von-Guericke-University Magdeburg, Germany. From 2002 to 2005, he worked as a researcher at the International Research/Training Centre for Information Technologies and Systems (IRTC ITS), Kiev, Ukraine.

Florian Eyben received the diploma in information technology from the Technische Universität Munich (TUM). He works on a research grant within the Institute for Human-Machine Communication at TUM. His teaching activities consist of pattern recognition and speech and language processing. His research interests include large scale hierarchical audio feature extraction and evaluation, automatic emotion recognition from the speech signal, and recognition of nonlinguistic vocalizations. He has several publications in various journals and conference proceedings covering many of his areas of research. He is a member of the IEEE.

Martin Wöllmer received the diploma in electrical engineering and information technology from the Technische Universität Munich (TUM), where his current research and teaching activity includes the subject areas of pattern recognition and speech processing. He works as a researcher funded by the European Community's Seventh Framework Programme project SEMAINE at TUM. His focus lies in multimodal data fusion, automatic recognition of emotionally colored and noisy speech, and speech feature enhancement. He is a reviewer for various publications, including the *IEEE Transactions on Audio, Speech, and Language Processing*. Publications of his in various conference proceedings cover novel and robust modeling architectures for speech and emotion recognition such as switching linear dynamic models or long short-term memory recurrent neural nets. He is a member of the IEEE.

André Stuhlsatz received a diploma degree in electrical engineering from the Duesseldorf University of Applied Sciences, Germany, in 2003. Since 2004, he was a postgraduate with the Chair for Cognitive Systems at the Otto-von-Guericke-University, Magdeburg, Germany, and received the doctoral degree for his work on "Machine Learning with Lipschitz Classifiers" (2010). From 2005 to 2008, he was a research scientist at the Fraunhofer Institute for Applied Information Technology, Germany, with focus on virtual and augmented environments. At the same time, he was also a research scientist at the Laboratory for Pattern Recognition, Department of Electrical Engineering at the Duesseldorf University of Applied Sciences, Germany. Currently, he is with the Institute of Informatics, Department of Mechanical and Process Engineering at the Duesseldorf University of Applied Sciences, Germany. His research interests include machine learning, statistical pattern recognition, face and speech recognition, feature extraction, classification algorithms and optimization.

Andreas Wendemuth received the Master of Science degree (1988) from the University of Miami, Florida, and the diploma in physics (1991) and electrical engineering (1994) from the University of Giessen, Germany, and Hagen, Germany, respectively. He received the Doctor of Philosophy degree (1994) from the University of Oxford, United Kingdom, for his works on "Optimisation in Neural Networks." In 1991, he worked at the IBM development centre in Sindelfingen, Germany, before his postdoctoral stays in Oxford (1994) and Copenhagen (1995). From 1995 to 2000, he worked as a researcher at the Philips Research Labs in Aachen, Germany, on algorithms and data structures in automatic speech recognition, as EC-Project Manager of the group "Content-Addressed Automatic Inquiry Systems in Telematics," and on the design and setup of dialogue systems and automatic telephone switchboards. Since 2001, he has been a professor of cognitive systems and speech recognition at the Otto-von-Guericke-University Magdeburg, Germany, at the Institute for Electronics, Signal Processing and Communications Technology. He published three books on signal and speech processing, as well as manifold peer-reviewed papers and articles in these fields. He is a member of the IEEE.

Gerhard Rigoll received the diploma in technical cybernetics (1982), the PhD degree in the field of automatic speech recognition (1986), and the Dr.-Ing. habil. degree in the field of speech synthesis (1991) from the University of Stuttgart, Germany. He worked for the Fraunhofer-Institute Stuttgart, Speech Plus in Mountain View, California, and Digital Equipment in Maynard, spent a postdoctoral fellowship at the IBM T.J. Watson Research Center, Yorktown Heights, New York (1986-1988), headed a research group at the Fraunhofer-Institute Stuttgart, and spent a two-year research stay at NTT Human Interface Laboratories in Tokyo, Japan (1991-1993) in the area of neurocomputing, speech and pattern recognition until he was appointed a full professor of computer science at Gerhard-Mercator-University Duisburg, Germany (1993) and of human-machine communication at the Technische Universität Munich (TUM) (2002). He is a senior member of the IEEE and has authored or coauthored more than 400 publications in the field of signal processing and pattern recognition. He served as an associate editor for the *IEEE Transactions on Audio, Speech, and Language Processing* (2005-2008), and is currently a member of the Overview Editorial Board of the IEEE Signal Processing Society. He serves as an associate editor and reviewer for many other scientific journals, has been a session chairman and a member of the program committee for numerous international conferences, and was the general chairman of the DAGM-Symposium on Pattern Recognition in 2008.