Chapter 20
Emotion Recognition from Speech

Andreas Wendemuth, Bogdan Vlasenko, Ingo Siegert, Ronald Böck, Friedhelm Schwenker, and Günther Palm

Abstract Spoken language is one of the main interaction patterns in human-human as well as in natural, companion-like human-machine interactions. Speech conveys content, but also emotions and interaction patterns determining the nature and quality of the user's relationship to his counterpart. Hence, we consider emotion recognition from speech in the wider sense of application in Companion-systems. This requires a dedicated annotation process to label emotions and to describe their temporal evolution in view of a proper regulation and control of a system's reaction. This problem is peculiar for naturalistic interactions, where the emotional labels are no longer a priori given. This calls for generating and measuring of a reliable ground truth, where the measurement is closely related to the usage of appropriate emotional features and classification techniques. Further, acted and naturalistic spoken data has to be available in operational form (corpora) for the development of emotion classification; we address the difficulties arising from the variety of these data sources. Speaker clustering and speaker adaptation will as well improve the emotional modeling. Additionally, a combination of the acoustical affective evaluation and the interpretation of non-verbal interaction patterns will lead to a better understanding of and reaction to user-specific emotional behavior.

A. Wendemuth (✉)
Cognitive Systems Group, Otto von Guericke University, PF-4120, 39016 Magdeburg, Germany

Center for Behavioral Brain Sciences, 39118 Magdeburg, Germany
e-mail: andreas.wendemuth@ovgu.de

B. Vlasenko • I. Siegert • R. Böck
Cognitive Systems Group, Otto von Guericke University, PF-4120, 39016 Magdeburg, Germany
e-mail: bodgan.vlasenko@ovgu.de; ingo.siegert@ovgu.de; ronald.boeck@ovgu.de

F. Schwenker • G. Palm
Institute for Neural Information Processing, University of Ulm, 89069 Ulm, Germany
e-mail: friedhelm.schwenker@uni-ulm.de; guenther.palm@uni-ulm.de

S. Biundo, A. Wendemuth (eds.), *Companion Technology*, Cognitive Technologies,
DOI 10.1007/978-3-319-43665-4_20

20.1 Introduction

Human-machine interaction (HCI) has recently received increased attention. Besides making the operation of technical systems as simple as possible, a main goal is to enable a natural interaction with the user. However, today's speech-based operation still seems artificial, as only the content of the speech is evaluated. The way in which something is said remains unconsidered, although it is well known that also emotions are important to communicate successfully. "*Companion*-Systems" aim to fill this gap by adapting to the user's individual skills, preferences and emotions to be able to recognize, interpret and respond to emotional utterances appropriately (cf. Chap. 1).

A first prerequisite for emotion recognition is the availability of data for training and testing classifiers. In [43] it is pointed out that for a speech-based emotion recognition a rather straightforward engineering approach is usually used: "we take what we get, and so far, performance has been the decisive criterion". This very practical approach allows the evaluation of various feature extraction and classification methods. But to regulate and control a system's reaction towards a user adequately, the interpretation of the data labels can no longer be neglected. This problem has been addressed by many researchers in the community (cf. [4, 54, 63]), but a proper solution has not appeared yet.

Automatic emotion recognition is treated here as a branch of pattern recognition. It is data-driven—insights are gathered from sampled data, as it is difficult to rely on empirical evidence from emotion psychology: there is no universal emotion representation. This problem is arising for naturalistic[1] interactions. One has to rely on the emotional annotation of data, as the emotional labels are not given a priori. Furthermore, there are barely two datasets using the same emotional terms. Thus, for cross-corpora analyses researchers have to, for instance, combine different emotional classes to arrive at common labels.

Another issue that has only been rarely investigated for emotion recognition is speaker(-group) adaptation, to improve the emotional modeling. Although a model adaptation towards a specific group of speakers or an individual speaker has been used to improve automatic speech recognition [25], it has only rarely been used for emotion recognition. Additionally, these studies are conducted on databases of simulated affects only. Thus, there is no proof that these methods are suitable for natural interactions as well.

Furthermore, speech contains more than just content and emotions. It includes also "interaction patterns" determining the nature and quality of the user's relationship to his counterpart. These cues are sensed and interpreted by the humans, often without conscious awareness, but greatly influence the perception and interpretation of messages and situations. Thus a combination of the acoustical affective evaluation

[1] The term "naturalistic" is used to clarify the fact that a computer system always is a conversational partner less powerful than a human and thus HCI cannot be a natural interaction.

and the interpretation of these interaction patterns could enhance the prediction power for the process of naturalistic interactions.

From these considerations we derive the following research questions, which will be addressed in the following:

- How do we generate a reliable ground truth for emotion recognition from speech?
- How do we adapt models for user-specific emotional behavior?
- How do we combine acoustic and linguistic features for an improved emotion recognition?

In the next section we sketch common methods for speech-based emotion recognition and the utilized datasets. The novel and specialized methods which we used for the analysis of the research questions will be introduced in the later sections.

20.2 Methods for Acoustic Emotion Analysis

Speech-based emotion classifiers used in recent research publications include a broad variety [55] of different classification methods. There are two predominant emotion classification paradigms: frame-level dynamic modeling by means of Hidden Markov Models (HMMs) and turn-level static modeling [42]. In frame-level analysis, the speech data is divided into short frames of about 15–25 ms, where the human vocal apparatus generates a stable short-term spectrum; the extracted features are called segmental features or low-level features. For turn-level analyses, the whole turn, mostly the speaker's utterance, is investigated, and supra-segmental features are used to describe long-term acoustic characteristics (cf. [3]). It has to be taken into account that turn-level modeling in comparison with frame-level modeling does not provide good flexibility for modeling emotional intensity variability within a turn. But most emotional datasets provide an emotion annotation on the turn level, and generating a reliable annotation on the frame level is not as feasible.

Among dynamic modeling techniques, hidden Markov models are dominant. But so is "bag-of-frames" technique for multi-instance learning [45]. Further, dynamic Bayesian network architectures [29] could help to combine acoustic features on different time levels, such as the supra-segmental prosodic level, and spectral features on a frame-level basis. Regarding static modeling, the list of possible classification techniques seems endless: multi-layer perceptrons or other types of neural networks, Bayes classifiers, Bayesian decision networks, random forests, Gaussian mixture models (GMMs), decision trees, k-nearest neighbor distance classifiers, and support vector machines (SVM) are applied most often (cf. [24]).

Also, a selection of ensemble methods has been used, such as bagging, boosting, multi-boosting, and stacking with and without confidence scores. Newly developing approaches are long-short-term-memory recurrent neural networks, hidden conditional random fields, and tandem GMMs with support vector machines. They could further become more popular in the near future.

An important question which is investigated in the research community is about the choice and length of the best emotional unit. Unfortunately, this question has not been answered to date. Depending on the material and experimental setup, the emotion classification performance of sub-turn entries—automatically extracted quasi-stationary segments as well as manually marked syllables—falls behind models trained on turn-levels or sub-turn units that are related to the phonetic content clearly outperformed turn-level models; [3]. Lee et al. presented an acceptable acoustic-based emotion recognition performance using phoneme-class-dependent HMM classifiers with short-term spectral acoustic features [28]. The authors reached a classification performance of 76.1% for their four-class recognition problem, but used very expressive acted emotional data.

Still, most of the aforementioned phonetic pattern-dependent emotion classification techniques used forced alignment or manual annotation for the extraction of the phoneme borders. Just a few techniques faced real-life conditions by using automatic speech recognition (ASR) engines for the generation of phoneme alignments. Current ASR techniques, however, are not able to provide phoneme alignment on affective speech samples in quality comparable to that of manual phonetic transcription or alignment obtained with forced alignment. In order to exploit the advantages of ASR techniques and to meet real-life conditions, a phoneme-level emotion processing technique should use modified ASR methods for the phoneme time alignment. To be able to obtain the best possible phoneme alignment within real-world development tasks, we used an ASR system with acoustic models adapted on emotional speech samples.

For our experiments we applied a low-level feature modeling on a frame level for acoustic emotion recognition. The HMMs with Gaussian mixture models (GMMs) have been used for this purpose. Three different segments can be used for dynamic analysis: *utterance*, *chunk,* and *phoneme* (cf. [3]). We applied utterance- and phoneme-level analysis for our experiments. It is also possible to classify emotions with an average formant's value extracted from vowel segments. The phoneme boundaries' estimation is based on a *forced alignment*, provided by the Hidden Markov Toolkit (HTK) [62]. Within our experiments we use a simplified version of a BAS SAMPA with a set of 39 phonemes (18 vowels and 21 consonants). A list of emotion-indicative vowels with their corresponding instance number is given in [57].

The time evolution of emotions is another central question. We apply a state transition model for this purpose. Instead of the standard ASR task to deduce the most likely word sequence hypothesis Ω_k from a given acoustic vector sequence $\mathbf{O}$ of M observations $\mathbf{o}$, we recognize the speaker's emotional state. This is solved with standard Bayes' ASR recognition criteria, with a different argument interpretation: $P(\mathbf{O}|\Omega)$ is called the emotion acoustic model, $P(\Omega)$ is the prior user-behavior information and Ω is one of all system-known emotions.

20.3 Utilized Datasets of Emotional Speech

In this section, we shortly describe all datasets used in our experiments to be reported later. A broader overview of emotional speech databases can be found in the following survey articles [36] as well as [55]. Some important details are given in Table 20.1.

The Berlin Database of Emotional Speech (emoDB) [11] is one of the most common emotional acoustic databases. This corpus contains studio-recorded emotionally neutral German sentences for seven affective states and contains 494 phrases with a total length up to 20 min. The content is pre-defined and spoken by ten (five male, five female) actors. The age of the actors is in the range of 21–35.

The Vera am Mittag audio-visual emotional speech database (VAM) [22] contains spontaneous and unscripted discussions between two to five persons from a German talk show. The labeling uses Self-Assessment Manikins (cf. [34]). The recordings cover low and high expressive emotional speech, due to the nature of the origin as TV talk-show. This database contains 947 sentences derived with a total length of 47 min. The age of participants ranges from 16 to 58 years.

The LAST MINUTE CORPUS (LMC) (cf. [37, 38]) contains synchronous audio and video recordings in a so-called Wizard-of-Oz (WoZ) experiment including 130 participants with nearly 56 h. For our experiments, we selected those 79 participants with best signal-to-noise ratio, having 31 min of audio material. During the dialogue critical events are induced that could lead to a dialogue-break-off [20]. We focus on two key events: *baseline* (BSL) and *weight limit barrier* (WLB). A detailed description can be found in Chaps. 13 and 14.

The EmoRec corpus (EmoRec) [60] simulates a natural verbal human-computer interaction, also implemented as a WoZ experiment. The design of the trainer followed the principle of the popular game "Concentration". The procedure of emotion induction included differentiated experimental sequences during which the user passed through specific valence/arousal/dominance (PAD) [10] octants in a

Table 20.1 Overview of selected emotional speech corpora

Name	Emotions	HH:MM	# Speaker
Acted emotions			
emoDB [11]	Anger boredom disgust fear happiness neutral sadness	00:22	10
Excerpts of human-human interaction			
VAM [22]	Values of arousal and valence	00:48	47
UAH [12]	Anger boredom doubt neutral	02:30	60
Naturalistic interaction			
SAL [32]	Continuous traces	10:00	20
EmoRec [60]	Four quadrants of valence-arousal space	33:00	100
LMC [38]	Four dialogue barriers	56:00	130
EmoGest [5]	Happy, neutral, sad	12:00	32

controlled fashion. In our experiment, we investigate only ES-2, which is assumed to be positive, and ES-5, which is negative.

The UAH emotional speech corpus (UAH) [12] contains 85 dialogues from a telephone-based information system spoken in Andalusian dialect from 60 different users. They used four emotional terms to discern emotions. The annotation process was conducted by nine labelers assessing complete utterances.

The Belfast Sensitive Artificial Listener corpus (SAL) is built from emotionally colored multimodal conversations. With four different operator behaviors, the scenario is designed to evoke emotional reactions. To obtain annotations, trace-style continuous ratings were made on five core dimensions (valence, activation, power, expectation, overall emotional intensity) utilizing FEELTRACE [15]. The number of labelers varied between two and six.

The EmoGest corpus (EmoGest) [5] consists of audio and video recordings as well as Kinect data based on a linguistic experiment which consisted of two experimental phases: a musical emotion induction procedure and a gesture-eliciting task in dialogical interaction with a confederate partner. In total, the corpus provides roughly 12 h of multimodal material from 32 participants.

20.4 Ground Truth, Adaptation and Non-verbals in Emotion Recognition

20.4.1 Generating and Measuring a Reliable Ground Truth for Emotion Recognition from Speech

As stated in Sect. 20.1, finding appropriate emotional labels for spoken expressions in naturalistic interactions is a challenging issue. Thus, besides the support of the labelers with suitable emotional annotation tools like *ikannotate* [7] or ATLAS [33] (cf. Chap. 19), valid and well-founded emotion-labeling methods have also to be utilized (cf. [14, 21]).

The research community started with rather small datasets containing acted, studio-recorded, non-interactional, high-quality emotions like in emoDB [11] which are related to the early days of speech corpora generation. The next step in data collection was the emotional inducement. Emotional stimuli were presented to or induced in a subject, whose reactions were recorded [55]. For this, in (HCI) a Wizard-of-Oz setup is often used, providing optimal conditions to influence the recorded user (cf. e.g. Chap. 13). Therefore, the data can be directly used to analyze human reactions while interacting with a technical system [59, 60]. Further, an overview on emotional classes in selected corpora is given in Table 20.1.

Various emotional labels are used in different corpora, situations, tasks, and setups [12, 35]. The emotion recognition community is aware of these difficulties and in [47, 50] a comparative study is conducted. The authors stated that for naturalistic interactions the emotion labels are generated by a quite complex and

Table 20.2 Qualitative assessment of labeling qualities for different labeling methods on a 5-item scale in the range of −− to ++

Method	Usability	Emotion coverage	Label reproducibility
Basic emotions	++	−−	+
GEW	−	++	++
SAM	+	+	−−

++: valid to high degree, −−: not valid at all

time-consuming annotation process. Such an annotation should cover the full range of observed emotions and, further, be proper for the labeling process.

Which labels are needed for labeling emotions in speech? To answer this question, three mainly used emotion assessment methods were compared (cf. [46]), namely Basic Emotions [17], Geneva Emotion Wheel (GEW) [39], and self-assessment manikins (SAMs) [22], and additionally questionnaires assessing the methods were applied. Comparative results are shown in Table 20.2. The main results can be summarized as follows (cf. [46]): Basic Emotions are not sufficient for emotional labeling of spontaneous speech since more variations are observable than covered by Basic Emotions. SAMs are able to cover these variations. On the other hand, labelers have to identify the three values of valence, arousal and dominance and, further, non-trained labelers have difficulties identifying valence or dominance only from speech. Siegert et al. suggest using GEW, which provides a mapping of Basic Emotions into a subset of the (valence-)arousal-dominance space and thus a possible clustering (cf. [46]). Labelers could cover nearly all variations with GEW, which can become quite complex, as the annotator has to chose 1 of 17 emotion families, each with five graduations of intensity. Though evaluated labeling tools exist, the selection of a proper labeling method highly depends on the established system and on the scenario. Although this comparison is not complete, as several other emotion assessment methods and tools exist, it gives a first advice which methods to prefer.

To assess the annotation and the appropriateness of the methods themselves, measures which allow a statement concerning the correctness of the found phenomena and, thus, the reliability, should be investigated. Since reliability values are usually low regarding naturalistic speech, a new interpretation is needed (cf. [50]). For this purpose, the inter-rater-reliability (IRR) is a good measure where the general ideas are presented in [2]. In particular, inter-rater reliability determines the extent to which two or more coders obtain the same result when measuring a certain object, called agreement.

A good reliability measure must fulfill the demands of stability, reproducibility, and accuracy, where reproducibility is the strongest demand. To calculate the IRR, mostly a kappa-like statistic is used. Siegert et al. [50] decided on Krippendorff's alpha since this is generally more applicable than the κ statistics, but the same scheme to interpret the values can be used. Further, it has the advantage of incorporating several distance metrics (ordinal, nominal, ...) for the utilized labels.

Table 20.3 Comparison of different agreement interpretations of kappa-like coefficients utilized in content analysis

Agreement	Landis and Koch [27]	Altmann [1]	Fleiss [19]	Krippendorff [26]
Poor	>0	>0.2	>0.4	>0.66
Slight	0.0–0.2	–	–	–
Fair	0.2–0.4	0.2–0.4	–	–
Moderate/Good	0.4–0.6	0.4–0.6	0.4–0.75	0.67–0.8
Substantial	0.6–0.8	0.6–0.8	–	–
Excellent/Very good	>0.8	>0.8	>0.75	>0.8

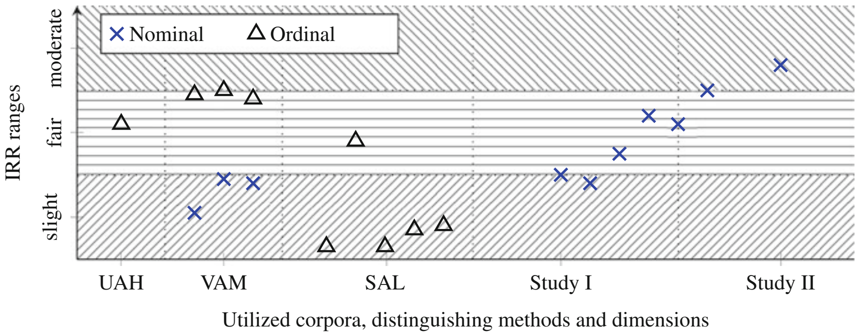

Fig. 20.1 Compilation of IIRs reported in [50], plotted against the agreement interpretation by Landis and Koch [27]

Table 20.3 presents the interpretation schema for IRR, which is widely accepted. Comparing the IRRs for the presented corpora, we notice that for all annotation methods and types of material, the reported reliabilities are far away from the values regarded as reliable. Even well-known and widely used corpora like VAM and SAL reveal a low inter-rater agreement. In particular, nominal alpha is between 0.01 and 0.34. Also, an ordinal metric increased alpha only up to 0.48 at best. Both cases are interpreted as a slight to fair reliability (cf. Table 20.3). Furthermore, comparing the three different annotation methods shows that the methods themselves only have a small impact on the reliability value. Even the use of additional information did not increase the reliability (cf. [50]). The quite low IRR values are due to the subjective nature of emotion perception and emphasize the need for well-trained labelers.

In [50] four corpora, namely UAH, VAM, SAL, and LMC (cf. Table 20.1), are investigated in terms of IRR, considering the previously mentioned annotation paradigms. Furthermore, two approaches are presented to increase the reliability values on LMC. At first, both audio and video recordings of the interaction (denoted as Study I in Fig. 20.1) as well as the natural time order of the interaction (denoted as Study II in Fig. 20.1) were used to increase the reliability. Also, training the annotators with preselected material avoided the Kappa paradoxes [13, 18] and, further, improved the IRR.

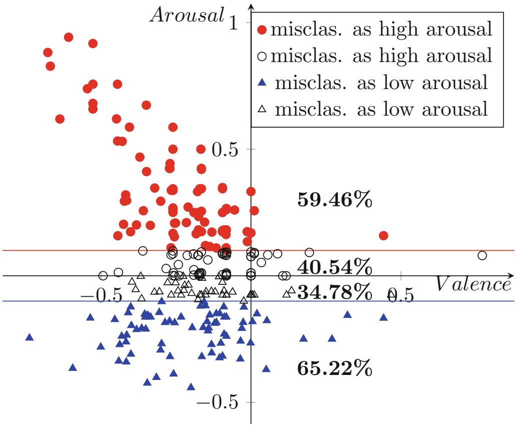

Fig. 20.2 Misclassified emotional instances for a two-class emotion classification task. *Filled markers* represent instances with arousal level in ranges [−1, −0.1) and (0.1, 1]. The plot is adapted from [56]

During the annotation, human labelers tend to a subjective view on the material. This results in a variance in the agreement of samples into classes. Utilizing learning methods, such variance can be identified and, further, a proper number of classes can be derived. In several studies, subspaces of the valence-arousal-dominance (VAD) space were considered (cf. e.g. [43]). In [56] a two-class emotion classification was applied to suggest a grouping of emotional samples in a valence-arousal space. The optimal classification performance on emoDB was obtained with 31 GMMs. To establish a cross-corpus grouping of emotions the classifier was tested on VAM, which results in three classes in terms of the valence-arousal space, namely high-arousal, neutral, and low-arousal (cf. [56]). The applied measure is the number of misclassified emotional instances in VAM utilizing the emoDB models. As shown in Fig. 20.2 about 40.54% of misrecognized low-arousal instances and about 34.78% of misrecognized high-arousal samples are located in the range (−0.1, 0.1) in the arousal dimension. Therefore, a third class which covers the emotionally neutral arousal instances should be introduced. This approach improves the recognition performance in a densely populated arousal subspace by about 2.7% absolute.

Up to this point, we have discussed whether and how the annotation method influences the reliability of the labeling. Further, we proposed a grouping of emotions based on cross-corpora evaluation of emoDB and VAM which provides an approach to handle observed data in the beginning of an interaction (cf. [56]). On the other hand, as stated in [9], the annotation process as such is of interest and, thus, should be considered to be objectified.

We know that emotions are usually expressed by multiple modalities like speech, facial expressions, and gestures. Thus, an annotation system, which will work in a semi-automatic fashion, can rely on a large amount of data that argues for an efficient way in the annotation. Therefore, Böck et al. [9] presented an approach towards semi-automatic annotation in a multimodal environment. The main idea is as follows: Automatic audio analyses are used to identify relevant affective sequences which are aligned with the corresponding video material. This indicates

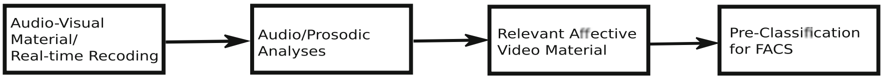
Audio-Visual Material/ Real-time Recoding
Audio/Prosodic Analyses
Relevant Affective Video Material
Pre-Classification for FACS

Fig. 20.3 Workflow to establish a semi-automatic annotation based on an audio-based pre-classification of video material as proposed in [9]

a pre-classification for the facial annotation (cf. Fig. 20.3). Since each person shows emotions in (slightly) different ways, the utilized audio features should be relatively general, so that a wide range of domains and audio conditions are covered. Suitable features are investigated in [6, 31], resulting in Mel-Frequency-Cepstral-Coefficients (MFCCs) and prosodic features. A GMM-based identification method is proposed and tested in [9]. For this, an objective way of annotation can be established since a classification system does not tend to interpret user reactions differently in several situations. Further, the approach reduces the manual effort as human annotators are just asked to label debatable sequences. This approach can also be used to preselect emotional material for the improvement of IRR.

20.4.2 User-Specific Adaptation for Speech-Based Emotion Recognition

An issue that has only been rarely investigated is the user-specific adaptation for speech-based emotion recognition to improve the corresponding emotional modeling [52]. For speech-based emotion recognition, an adaptation onto the speaker's biological gender has been used, for instance, in [16]. The authors utilized a gender-specific UBM-MAP approach to increase the recognition performance, but did not apply their methods on several corpora. In [58] a gender differentiation is used to improve the automatic emotion recognition from speech. The authors achieved an absolute difference of approx. 3% between the usage of correct and automatically recognized gender information.

All these publications only investigate the rather obvious gender-dependency. No other factors such as, for instance, age are considered, although it is known that age has an impact on both, the vocal characteristics (cf. [23]) and the emotional response (cf. [30]). It has not been investigated whether the previously mentioned improvements are dependent on the utilized material, as most of these studies are conducted on databases of acted emotions. Thus, a proof of whether these methods are suitable for natural interactions is still missing.

Therefore, in our experiments (cf. [48, 52]), we investigated whether a combination of age and gender can further improve the classification performance. As the analyses were conducted on different corpora, comparatively general statements can be derived. Being able to compare our results on emoDB and VAM, we used the two-class emotional set generated by Schuller et al. [42]. They defined combinations of emotional classes to cluster into `low arousal (A-)` and `high arousal`

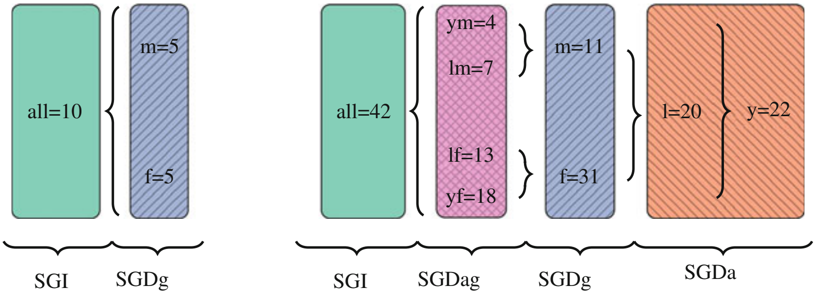

Fig. 20.4 Distribution of speaker groupings and their abbreviations on emoDB (*left*) and VAM (*right*). *SGI* speaker group independent, *SGD* speaker group dependent, *a* age, *g* gender

`(A+)` for several datasets (cf. Sect. 20.3). The different age-gender groupings together with the number of corresponding speakers are depicted in Fig. 20.4. The combination of both grouping factors led to sub-groups according to the speakers' gender: **m**ale vs. **f**emale; according to their age: mid-***l***ife vs. **y**oung speakers; and combinations of both: i.e. **y**oung **m**ale speakers.

The following acoustic characteristics are utilized as features: 12 mel-frequency cepstral coefficients, zeroth cepstral coefficient, fundamental frequency, and energy. The Δ and $\Delta\Delta$ regression coefficients of all features are used to include contextual information. As channel normalization technique, relative spectral (RASTA)-filtering is applied. GMMs with 120 mixture components utilizing four iteration steps are used as classifiers. For validation we use a leave-one-speaker-out (LOSO) strategy. As performance measure, the unweighted average recall (UAR) is applied. The UAR indicated the averaged recall taking into account the recognition results for each class independently. More details about the parameter optimization can be found in [52]. Afterwards, we performed the experiments on the SGI set as well as the SGDa, SGDg, and SGDag sets. For this, the subjects are grouped according to their age and gender in order to train the corresponding classifiers in a LOSO manner. To allow a comparison between all speaker groupings, we combined the different results. For instance, the results for each male and female speaker are merged to obtain the overall result for the SGDg set. This result can be directly compared with results gained on the SGI set. The outcome is shown in Fig. 20.5.

The SGD results on a two-class problem are outperforming the classification result of 96.8% from [42]. In comparison to the SGI results the SGD classifiers achieved an absolute improvement of approx. 4%. This improvement is significant ($F = 4.48238$, $p = 0.0281$). Utilizing VAM allows us to examine a grouping of the speakers on different age ranges, namely young adults (`y`) and mid-life adults (`l`). These groupings comprise the speakers' age (SGDa), the speakers' gender (SGDg), and the combination of both (SGDag); see Fig. 20.5. For all three combinations, a substantial improvement was achieved in comparison to the BSL classification (SGI). Unfortunately, SGDa and SGDag achieve lower results than the classification using only the gender characteristic. This is mostly caused by the

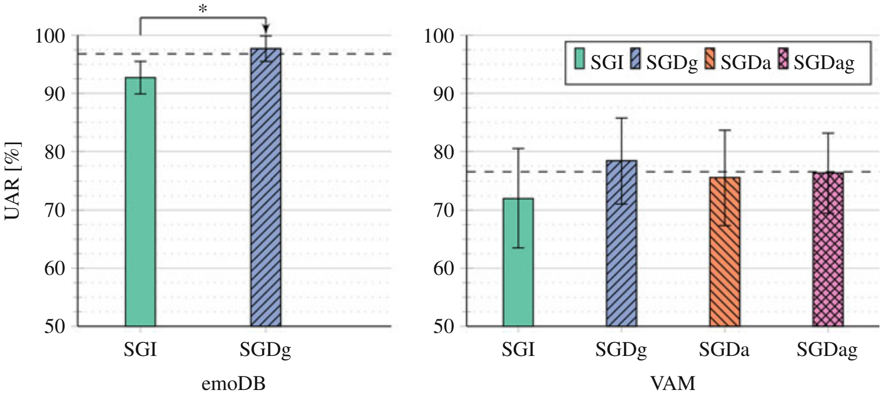

Fig. 20.5 UARs in percent for the two-class problem on emoDB and VAM comparing SGI and SGDg utilizing LOSO validation on different feature sets. For comparison, the best results from [42] are marked with a *dashed line*. The *star* denotes the significance level: * ($p < 0.05$)

declined performance for the m group. It has to be further investigated whether this can be attributed to the small amount of material available or to the fact that the present acoustical differences within the mid-life adults are larger than those in the young adults' group (cf. [52]).

If we take into account datasets with high differences in the age grouping, we can observe that the SGDag grouping outperforms the results of the other models. More details about this investigation can be found in Chap. 14. In general, it can be stated that for the investigated groups of young adults and mid-life adults the gender grouping is the dominant factor. Taking also corpora into account, where a high age difference can be observed, a combination of age and gender groups is needed.

In addition to the speaker-dependent characteristics like age and gender, we also see that a combination of several sources of information could provide a gain in the classification performance of emotions (cf. Chap. 19). Therefore, we conducted another study (cf. [8]). In this experiment, we prove the performance of chosen modalities in emotion recognition. The most prominent modalities are speech, facial expression, and, as they are speaker-dependent, biopsychological characteristics like skin conductance, heart rate, etc. In this case, a comparison of intraindividual against interindividual classification of emotions is highly recommended; cf. Table 20.4. Intraindividual analysis means that only the data of a particular speaker is analyzed and used for both training and test. In contrast, interindividual analysis directly compares the performance of a classifier training on all material except this of a certain person against the material of this particular speaker (test material). For the experiments, we need a corpus which provides material for all three modalities and, thus, allows a comparative study. Therefore, the EmoRec data set was used (cf. Sect. 20.3). Indeed, these evaluations are necessary to investigate the effect of personalization. Using biopsychological features for classification, a calibration of the features is necessary since the intraindividual characteristics of each speaker

Table 20.4 Recognition rates on the EmoRec Corpus for intraindividual and interindividual classification of ES-2 vs. ES-5 in percent with HTK, taken from [8]

Subject	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
ES-2 (**intra**individual)	50.0	50.0	38.5	50.0	50.0	65.0	68.4	71.4	100.0	81.3
ES-5 (**intra**individual)	66.7	75.0	75.0	70.8	84.6	70.0	84.0	100.0	100.0	63.2
ES-2 (**inter**individual)	84.0	66.7	88.5	74.2	82.1	86.2	73.1	72.4	90.5	50.0
ES-5 (**inter**individual)	33.3	36.5	3.8	34.0	22.6	24.3	26.4	86.7	13.3	63.2
Subject	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
ES-2 (**intra**individual)	100.0	100.0	80.0	43.9	68.8	58.5	40.0	40.6	70.0	61.1
ES-5 (**intra**individual)	93.3	72.5	55.8	59.1	68.8	83.8	65.0	81.6	68.0	59.1
ES-2 (**inter**individual)	21.4	28.1	40.6	14.3	55.2	31.2	97.0	79.3	10.7	54.5
ES-5 (**inter**individual)	83.8	85.0	75.0	97.7	66.7	71.4	9.3	16.7	73.9	80.0

influence the features as such. This means that for a particular person the heart rate corresponding to an emotion can be higher rather in a calm situation than for another person. Unfortunately, the calibration is difficult due to the unknown baseline of emotional behavior for each user; this means obtaining a value for each feature which represents a neutral emotion. Therefore, the idea is to use a classifier based on other modalities to provide an emotional rating which can be used to calibrate the biopsychological observations. In the experiment (cf. [8]), two emotional situations—for short positive and negative (for details, cf., e.g., [8])—were distinguished that are given by the design of the data recording scenario (cf. [60]). For this, no external annotation was conducted.

In the case of audio classification we found that intraindividual and interindividual systems provided good results. Since we compared our analyses to multimodal investigations (cf. [8]), we can state: Biopsychological classifiers, for instance neural networks (cf. [61]), in particular, if they are calibrated (details cf. [60]), showed, in general, recognition accuracies of more than 75%. Comparable results were gained by video analysis. Given these results, a personalization process for a technical system should be based on interindividual approaches, first using material which is clearly detected to adapt the system, and finally adapting classifiers which are adjusted to a certain user. This is guiding us to several issues: Which modality will influence the phases of the system adaptation? Which is the best combination of modalities in information fusion? Such aspects are discussed in Chap. 19.

20.4.3 Combination of Acoustic and Linguistic Features for Emotion Recognition

In this section, we compare emotion recognition of acoustic features containing both spectral and prosodic information combined with the linguistic bag-of-words features.

Several research groups have used two feature sets defined in the following (cf. Table 20.5): The first set consists of 12 MFCCs and the zeroth coefficient with corresponding Delta and Acceleration coefficients. The first three formants, the corresponding bandwidths, intensity, jitter, and pitch were accompanied into the acoustic feature vector. The second acoustic feature set is a subset of the first and is focused on spectral features only, namely MFCC (cf. MFCC_0_D_A in Table 20.5). Both feature sets have recently been heavily applied in the research community (cf. [4, 41]). Furthermore, they are also applied in classification of naturalistic and/or spontaneous emotion recognition from speech (cf. e.g. [9]).

The GMMs provided by HTK [62] were trained for all feature sets given in Table 20.5. Notice that for each emotional class, in particular, happy and sad, a separate GMM was generated, representing the characteristics of the class, and a final decision was established by passing through the models using the Viterbi decoding algorithm. The classifiers with nine (cf. [9]), 81 (cf. [42]), and 120 (cf. [52]) Gaussian mixtures gained the best performance on the EmoGest corpus.

Table 20.6 shows that the achieved classification performance has two maxima. Employing a LOSO strategy, the ability to classify emotions on the speaker-independent level can be shown. In particular, remarkable results were obtained on the full feature set (i.e., MFCC_0_D_A_F_B_I_J_P) with 0.785 recall (cf. Table 20.6). For the classifier with 120 Gaussian mixtures the variance has also

Table 20.5 The two feature sets (with total number of features) used in the experiments with the corresponding applied features

Feature set	Number	Applied features
MFCC_0_D_A	39	MFCC, zeroth cepstral coefficient, delta, acceleration
MFCC_0_D_A_F_B_I_J_P	48	MFCC_0_D_A, formants 1–3, bandwidths, intensity, jitter, pitch

Table 20.6 Experimental results for the two acoustic features sets (cf. Table 20.5 on page 422) employing GMMs with different numbers of mixtures (#mix)

Feature set	#mix	Recall	Variance
MFCC_0_D_A	9	0.822	0.375
	81	0.826	0.380
	120	0.822	0.375
MFCC_0_D_A_F_B_I_J_P	9	0.755	0.324
	81	0.782	0.327
	120	0.785	0.322

Recall with corresponding variance is given its minimum. Since we oriented ourselves on the experiment presented in [52], we have not tested more than 120 mixtures, which was the maximum number utilized.

One determines that the spectral feature set outperforms the combined feature set. The recall rates are more than (absolute) 4% higher, staying with comparable variance rates. We conclude that spectral features can better distinguish or cover the two emotions sad and happy. This leads to the discussion about which acoustic features are the most meaningful ones for emotions [6, 10, 40]. In [44], the authors state that, usually, optimal feature sets are highly dependent on the evaluated dataset.

A further application of feature selection is to automatically find potential significant dialog turns. For the recognition of WLB events we analyze acoustic, linguistic and non-linguistic content with a static classification setup. We investigated several feature sets: a full set consisting of 54 acoustic and 124 linguistic features, two reduced sets focusing on the most relevant features, and four sets corresponding to a different combination of linguistic and acoustic features. The proposed classification techniques were evaluated on the LAST MINUTE corpus. The dataset provides emotional instances for WLB and BSL classes. An unweighted average recall of 0.83 was achieved by using the full feature set.

Using Bag-of-words (BoW) features usually leads to a high-dimensional sparse feature vector, as only a limited vocabulary is used in the utterances we are interested in. We pre-selected only those words which appear in at least three turns over the database. Finally, this resulted in 1242 BoW features. Therefore we employed a feature ranking method to select the most informative ones. Using WEKA, we employed an information gain attribute evaluator in conjunction with a feature ranker as the search method. The top 100 BoW features selected were added to the combined (acoustic, linguistic) feature set described above, giving us 278 features in total. The resulting feature vector is referenced hereinafter as *full*. In addition to this, we constructed two further feature vectors, denoted by *top 100* and *top 50*, corresponding to the top 100 and top 50 meaningful features from the full set.

In Table 20.7 our experimental results are presented. The evaluation has been conducted using tenfold speaker-independent cross-validation. Classification results are slightly better for the BSL class, which reflects the grouped material belonging to the barrier called BSL, denoting the experiment's part where the first excitement has been gone (cf. [20]). In terms of the feature selection, the top 100 features provide only 1–2% absolute improvement compared to the full feature set; however, the differences are not significant.

Among 50 (from *top 100* to *top 50*) reduced features, 45 belong to BoW. Such reduction is accompanied with a severe performance degradation, which suggests that BoW features contribute much to the final recognition result.

Therefore, these results do not answer the question of the development of an optimal feature set, which goes beyond the scope of the current research.

Table 20.7 Classification performance for speaker-independent evaluation on the LAST MINUTE corpus

Feature set	Full 178	Top 100	Top 50
General results			
Weighted avg recall	0.86	0.86	0.81
UAR	0.83	0.84	0.76
Weighted F-score	0.86	0.86	0.80
Class WLB			
Precision	0.77	0.79	0.72
Recall	0.77	0.77	0.62
F-score	0.77	0.78	0.67
Class BSL			
Precision	0.90	0.89	0.84
Recall	0.89	0.91	0.89
F-score	0.90	0.90	0.86

Full 178, Top 100, Top 50 feature set

20.5 Conclusion and Outlook

We have considered emotions and other interaction patterns in spoken language, determining the nature and quality of the users' interaction with a machine in the wider sense of application in *Companion*-Systems. Acted and naturalistic spoken data in operational form (corpora) have been used for the development of emotion classification; we have addressed the difficulties arising from the variety of these data sources. A dedicated annotation process to label emotions and to describe their temporal evolution has been considered, and its quality and reliability have been evaluated. The novelty here is the consideration of naturalistic interactions, where the emotional labels are no longer a priori given. We have further investigated appropriate emotional features and classification techniques. Optimized feature sets with combined spectral, prosodic and Bag-of-Words components have been derived. Speaker clustering and speaker adaptation have been shown to improve the emotional modeling, where detailed analyses on gender and age dependencies have been given.

In summary, it can be said that acoustical emotion recognition techniques have been considerably advanced and tailored for naturalistic environments, approaching a *Companion*-like scenario.

In the future, emotional personalization will be the next stage, following the presented speaker grouping. Also, the further incorporation of established methods in psychology will improve the labeling and annotation process towards higher reliability and ease-of-use for annotators. A combination of the acoustical affective evaluation and the interpretation of non-verbal interaction patterns will lead to a better understanding of and reaction to user-specific emotional behavior (cf. [49, 51, 53]). The meaning and usage of acoustic emotion, disposition and intention in the general human-machine context will be further investigated.

Acknowledgements This work was done within the Transregional Collaborative Research Centre SFB/TRR 62 "*Companion*-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG).

References

1. Altman, D.G.: Practical Statistics for Medical Research. Chapman & Hall, London (1991)
2. Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Comput. Linguist. **34**, 555–596 (2008)
3. Batliner, A., Seppi, D., Steidl, S., Schuller, B.: Segmenting into adequate units for automatic recognition of emotion-related episodes: a speech-based approach. Adv. Hum. Comput. Interact. **2010**, 15 (2010)
4. Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Aharonson, V., Kessous, L., Amir, N.: Whodunnit – searching for the most important feature types signalling emotion-related user states in speech. Comput. Speech Lang. **25**, 4–28 (2011)
5. Bergmann, K., Böck, R., Jaecks, P.: Emogest: investigating the impact of emotions on spontaneous co-speech gestures. In: Proceedings of the Workshop on Multimodal Corpora 2014, pp. 13–16. LREC, Reykjavik (2014)
6. Böck, R., Hübner, D., Wendemuth, A.: Determining optimal signal features and parameters for HMM-based emotion classification. In: Proceedings of the 15th IEEE MELECON, Valletta, Malta, pp. 1586–1590 (2010)
7. Böck, R., Siegert, I., Vlasenko, B., Wendemuth, A., Haase, M., Lange, J.: A processing tool for emotionally coloured speech. In: Proceedings of the 2011 IEEE ICME, p. s.p, Barcelona (2011)
8. Böck, R., Limbrecht, K., Walter, S., Hrabal, D., Traue, H., Glüge, S., Wendemuth, A.: Intraindividual and interindividual multimodal emotion analyses in human-machine-interaction. In: Proceedings of the IEEE CogSIMA, New Orleans, pp. 59–64 (2012)
9. Böck, R., Limbrecht-Ecklundt, K., Siegert, I., Walter, S., Wendemuth, A.: Audio-based pre-classification for semi-automatic facial expression coding. In: Kurosu, M. (ed.) Human-Computer Interaction. Towards Intelligent and Implicit Interaction. Lecture Notes in Computer Science, vol. 8008, pp. 301–309. Springer, Berlin/Heidelberg (2013)
10. Böck, R., Bergmann, K., Jaecks, P.: Disposition recognition from spontaneous speech towards a combination with co-speech gestures. In: Böck, R., Bonin, F., Campbell, N., Poppe, R. (eds.) Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction. Lecture Notes in Artificial Intelligence, vol. 8757, pp. 57–66. Springer, Cham (2015)
11. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A database of German emotional speech. In: Proceedings of the INTERSPEECH-2005, Lisbon, pp. 1517–1520 (2005)
12. Callejas, Z., López-Cózar, R.: Influence of contextual information in emotion annotation for spoken dialogue systems. Speech Comm. **50**, 416–433 (2008)
13. Cicchetti, D., Feinstein, A.: High agreement but low kappa: II. Resolving the paradoxes. J. Clin. Epidemiol. **43**, 551–558 (1990)
14. Cowie, R., Cornelius, R.R.: Describing the emotional states that are expressed in speech. Speech Comm. **40**, 5–32 (2003)
15. Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M.: FEELTRACE: an instrument for recording perceived emotion in real time. In: Proceedings of the SpeechEmotion-2000, Newcastle, pp. 19–24 (2000)
16. Dobrišek, S., Gajšek, R., Mihelič, F., Pavešić, N., Štruc, V.: Towards efficient multi-modal emotion recognition. Int. J. Adv. Robot. Syst. **10**, 1–10 (2013)
17. Ekman, P.: Are there basic emotions? Psychol. Rev. **99**, 550–553 (1992)

18. Feinstein, A., Cicchetti, D.: High agreement but low kappa: I. The problems of two paradoxes. J. Clin. Epidemiol. **43**, 543–549 (1990)
19. Fleiss, J.: Measuring nominal scale agreement among many raters. Psychol. Bull. **76**, 378–382 (1971)
20. Frommer, J., Rösner, D., Haase, M., Lange, J., Friesen, R., Otto, M.: Detection and Avoidance of Failures in Dialogues – Wizard of Oz Experiment Operator's Manual. Pabst Science Publishers, Lengerich (2012)
21. Grimm, M., Kroschel, K., Mower, E., Narayanan, S.: Primitives-based evaluation and estimation of emotions in speech. Speech Comm. **49**, 787–800 (2007)
22. Grimm, M., Kroschel, K., Narayanan, S.: The Vera am Mittag German audio-visual emotional speech database. In: Proceedings of the 2008 IEEE ICME, Hannover, pp. 865–868 (2008)
23. Harrington, J., Palethorpe, S., Watson, C.: Age-related changes in fundamental frequency and formants: a longitudinal study of four speakers. In: Proceedings of the INTERSPEECH-2007, Antwerp, vol. 2, pp. 1081–1084 (2007)
24. Iliou, T., Anagnostopoulos, C.N.: Comparison of different classifiers for emotion recognition. In: Proceedings of the Panhellenic Conference on Informatics, pp. 102–106 (2009)
25. Kelly, F., Harte, N.: Effects of long-term ageing on speaker verification. In: Vielhauer, C., Dittmann, J., Drygajlo, A., Juul, N., Fairhurst, M. (eds.) Biometrics and ID Management. Lecture Notes in Computer Science, vol. 6583, pp. 113–124. Springer, Berlin/Heidelberg (2011)
26. Krippendorff, K.: Content Analysis: An Introduction to Its Methodology, 3rd edn. SAGE, Thousand Oaks (2012)
27. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**, 159–174 (1977)
28. Lee, C.M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., Narayanan, S.: Emotion recognition based on phoneme classes. In: Proceedings of the INTERSPEECH 2004, Jeju Island, pp. 889–892 (2004)
29. Lee, C., Busso, C., Lee, S., Narayanan, S.: Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions. In: Proceedings of the INTERSPEECH 2009, pp. 1983–1986 (2009)
30. Lipovčan, L., Prizmić, Z., Franc, R.: Age and gender differences in affect regulation strategies. Drustvena istrazivanja: J. Gen. Soc. Issues **18**, 1075–1088 (2009)
31. Maganti, H.K., Scherer, S., Palm, G.: A novel feature for emotion recognition in voice based applications. In: Affective Computing and Intelligent Interaction, pp. 710–711. Springer, Berlin/Heidelberg (2007)
32. McKeown, G., Valstar, M., Cowie, R., Pantic, M., Schroder, M.: The SEMAINE database: annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE Trans. Affect. Comput. **3**, 5–17 (2012)
33. Meudt, S., Bigalke, L., Schwenker, F.: ATLAS – an annotation tool for HCI data utilizing machine learning methods. In: Proceedings of the 1st APD, San Francisco, pp. 5347–5352 (2012)
34. Morris, J.D.: SAM: the self-assessment manikin an efficient cross-cultural measurement of emotional response. J. Adv. Res. **35**, 63–68 (1995)
35. Palm, G., Glodek, M.: Towards emotion recognition in human computer interaction. In: Neural nets and surroundings, pp. 323–336. Springer, Berlin/Heidelberg (2013)
36. Pittermann, J., Pittermann, A., Minker, W.: Handling Emotions in Human-Computer Dialogues. Springer, Amsterdam (2010)
37. Prylipko, D., Rösner, D., Siegert, I., Günther, S., Friesen, R., Haase, M., Vlasenko, B., Wendemuth, A.: Analysis of significant dialog events in realistic human–computer interaction. J. Multimodal User Interfaces **8**, 75–86 (2014)
38. Rösner, D., Frommer, J., Friesen, R., Haase, M., Lange, J., Otto, M.: LAST MINUTE: a multimodal corpus of speech-based user-companion interactions. In: Proceedings of the 8th LREC, Istanbul, pp. 96–103 (2012)

39. Scherer, K.R.: Unconscious Processes in Emotion: The Bulk of the Iceberg, pp. 312–334. Guilford Press, New York (2005)
40. Scherer, S., Kane, J., Gobl, C., Schwenker, F.: Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification. Comput. Speech Lang. **27**(1), 263–287 (2013)
41. Schuller, B., Steidl, S., Batliner, A.: The INTERSPEECH 2009 emotion challenge. In: Proceedings of the INTERSPEECH-2009, Brighton, pp. 312–315 (2009)
42. Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., Wendemuth, A.: Acoustic emotion recognition: a benchmark comparison of performances. In: Proceedings of the IEEE ASRU-2009, Merano, pp. 552–557 (2009)
43. Schuller, B., Batliner, A., Steidl, S., Seppi, D.: Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Comm. **53**, 1062–1087 (2011)
44. Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R., Pantic, M.: AVEC 2011–the first international audio/visual emotion challenge. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.C. (eds.) Affective Computing and Intelligent Interaction. Lecture Notes in Computer Science, vol. 6975, pp. 415–424. Springer, Berlin/Heidelberg (2011)
45. Shami, M., Verhelst, W.: Automatic classification of emotions in speech using multi-corpora approaches. In: Proceedings of the 2nd IEEE Signal Processing Symposium, Antwerp, pp. 3–6 (2006)
46. Siegert, I., Böck, R., Philippou-Hübner, D., Vlasenko, B., Wendemuth, A.: Appropriate emotional labeling of non-acted speech using basic emotions, Geneva emotion wheel and self assessment manikins. In: Proceedings of the 2011 IEEE ICME, p. s.p, Barcelona (2011)
47. Siegert, I., Böck, R., Wendemuth, A.: The influence of context knowledge for multi-modal affective annotation. In: Kurosu, M. (ed.) Human-Computer Interaction. Towards Intelligent and Implicit Interaction. Lecture Notes in Computer Science , vol. 8008, pp. 381–390. Springer, Berlin/Heidelberg (2013)
48. Siegert, I., Glodek, M., Panning, A., Krell, G., Schwenker, F., Al-Hamadi, A., Wendemuth, A.: Using speaker group dependent modelling to improve fusion of fragmentary classifier decisions. In: Proceedings of 2013 IEEE CYBCONF, Lausanne, pp. 132–137 (2013)
49. Siegert, I., Hartmann, K., Philippou-Hübner, D., Wendemuth, A.: Human behaviour in HCI: complex emotion detection through sparse speech features. In: Salah, A., Hung, H., Aran, O., Gunes, H. (eds.) Human Behavior Understanding. Lecture Notes in Computer Science, vol. 8212, pp. 246–257. Springer, Berlin/Heidelberg (2013)
50. Siegert, I., Böck, R., Wendemuth, A.: Inter-rater reliability for emotion annotation in human-computer interaction – comparison and methodological improvements. J. Multimodal User Interfaces **8**, 17–28 (2014)
51. Siegert, I., Haase, M., Prylipko, D., Wendemuth, A.: Discourse particles and user characteristics in naturalistic human-computer interaction. In: Kurosu, M. (ed.) Human-Computer Interaction. Advanced Interaction Modalities and Techniques. Lecture Notes in Computer Science , vol. 8511, pp. 492–501. Springer, Berlin/Heidelberg (2014)
52. Siegert, I., Philippou-Hübner, D., Hartmann, K., Böck, R., Wendemuth, A.: Investigation of speaker group-dependent modelling for recognition of affective states from speech. Cogn. Comput. **6**(4), 892–913 (2014)
53. Siegert, I., Prylipko, D., Hartmann, K., Böck, R., Wendemuth, A.: Investigating the form-function-relation of the discourse particle "hm" in a naturalistic human-computer interaction. In: Bassis, S., Esposito, A., Morabito, F. (eds.) Recent Advances of Neural Network Models and Applications. Smart Innovation, Systems and Technologies, vol. 26, pp. 387–394. Springer, Berlin/Heidelberg (2014)
54. Strauß, P.M., Hoffmann, H., Minker, W., Neumann, H., Palm, G., Scherer, S., Schwenker, F., Traue, H., Walter, W., Weidenbacher, U.: Wizard-of-oz data collection for perception and interaction in multi-user environments. In: International Conference on Language Resources and Evaluation (LREC) (2006)

55. Ververidis, D., Kotropoulos, C.: Emotional speech recognition: resources, features, and methods. Speech Comm. **48**, 1162–1181 (2006)
56. Vlasenko, B., Wendemuth, A.: Location of an emotionally neutral region in valence-arousal space. Two-class vs. three-class cross corpora emotion recognition evaluations. In: Proceedings of 2014 IEEE ICME (2014)
57. Vlasenko, B., Philippou-Hübner, D., Prylipko, D., Böck, R., Siegert, I., Wendemuth, A.: Vowels formants analysis allows straightforward detection of high arousal emotions. In: Proceedings of 2011 IEEE ICME, Barcelona (2011)
58. Vogt, T., André, E.: Improving automatic emotion recognition from speech via gender differentiation. In: Proceedings of the 5th LREC, p. s.p, Genoa (2006)
59. Wahlster, W. (ed.): SmartKom: Foundations of Multimodal Dialogue Systems. Springer, Heidelberg/Berlin (2006)
60. Walter, S., Scherer, S., Schels, M., Glodek, M., Hrabal, D., Schmidt, M., Böck, R., Limbrecht, K., Traue, H., Schwenker, F.: Multimodal emotion classification in naturalistic user behavior. In: Jacko, J. (ed.) Human-Computer Interaction. Towards Mobile and Intelligent Interaction Environments. Lecture Notes in Computer Science, vol. 6763, pp. 603–611. Springer, Berlin/Heidelberg (2011)
61. Walter, S., Kim, J., Hrabal, D., Crawcour, S., Kessler, H., Traue, H.: Transsituational individual-specific biopsychological classification of emotions. IEEE Trans. Syst. Man Cybern. Syst. Hum. **43**(4), 988–995 (2013)
62. Young, S., Evermann, G., Gales, M., Hasin, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK Book (for HTK Version 3.4). Engineering Department, Cambridge University, Cambridge (2009)
63. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. **31**, 39–58 (2009)