Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 1062–1087
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge

Björn Schuller [a,*], Anton Batliner [b], Stefan Steidl [b], Dino Seppi [c]

[a] Institute for Human–Machine Communication, Technische Universität München, Germany
[b] Pattern Recognition Lab, University of Erlangen-Nuremberg, Germany
[c] ESAT, Katholieke Universiteit Leuven, Belgium

Available online 5 February 2011

Abstract

More than a decade has passed since research on automatic recognition of emotion from speech has become a new field of research in line with its 'big brothers' speech and speaker recognition. This article attempts to provide a short overview on where we are today, how we got there and what this can reveal us on where to go next and how we could arrive there. In a first part, we address the basic phenomenon reflecting the last fifteen years, commenting on databases, modelling and annotation, the unit of analysis and prototypicality. We then shift to automatic processing including discussions on features, classification, robustness, evaluation, and implementation and system integration. From there we go to the first comparative challenge on emotion recognition from speech – the INTERSPEECH 2009 Emotion Challenge, organised by (part of) the authors, including the description of the Challenge's database, Sub-Challenges, participants and their approaches, the winners, and the fusion of results to the actual learnt lessons before we finally address the ever-lasting problems and future promising attempts.

Keywords: Emotion; Affect; Automatic classification; Feature types; Feature selection; Noise robustness; Adaptation; Standardisation; Usability; Evaluation

1. Setting the scene

This special issue will address new approaches towards dealing with the processing of realistic emotions in speech, and this overview article will give an account of the state-of-the-art, of the lacunas in this field, and of promising approaches towards overcoming shortcomings in modelling and recognising realistic emotions. We will also report on the first emotion challenge at INTERSPEECH 2009, constituting the initial impetus of this special issue; to end with, we want to sketch future strategies and applications, trying to answer the question 'Where to go from here?'

The article is structured as follows: we first deal with the basic phenomenon briefly reflecting the last 15 years, commenting on databases, modelling and annotation, the unit of analysis and prototypicality. We then proceed to automatic processing (Section 2) including discussions on features, classification, robustness, evaluation, and implementation and system integration. From there we go to the first Emotion Challenge (Section 3) including the description of the Challenge's database, Sub-Challenges, participants and their approaches, the winners, and the fusion of results to the lessons learnt, before concluding this article (Section 4).

1.1. The phenomenon

The title of this survey article reflects the problems in defining 'emotion': The phrase 'emotion and affect' does

* Corresponding author. Tel.: +49 89 289 28548; fax: +49 89 289 28535.
E-mail address: schuller@tum.de (B. Schuller).

0167-6393/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2011.01.011

not mean that these are two different phenomena that easily can be told apart. It rather should be read as 'some people call it emotion, some affect; in fact, we do not really care'. We do not care because there is an abundance of competing definitions; thus, to find the common ground, the only meaningful way of defining the phenomenon is *'ex negativo'*: to describe what it is *not*. Therefore, we follow the definition of (Cowie et al., 2010) that 'emotion' is what is "present in most of life but absent when people are emotionless"; this is the concept of *pervasive emotion*. This definition encompasses on the one hand the prevailing narrow concept of (basic) emotions such as anger, fear, sadness, etc., on the other hand rather vague concepts such as emotion-related/affective states, and everything else that might be associated with a broad concept of *'emotional intelligence'* – which does not only mean that you are able to show your anger or fear but your interest and your tiredness as well, and that you are able to interact with other humans in a social manner, beyond pure cognitive reasoning and actions. Thus, there is no clear-cut borderline towards the new field of 'social signals' (Vinciarelli et al., 2008) which might have been (re-)established to circumvent problems of definition and 'garden fencing'.

Seen from an application-oriented point of view, the focus is on enabling future non-human interaction partners – be this computers, robots, or something else – at least to come closer to the normal human abilities to recognise 'emotions' and to respond in an 'emotionally intelligent way'. The concentration on a few (acted) basic emotions, especially in the early phase of this field, was rather not due to some theoretical considerations but simply to the fact that acted data were way easier to obtain than more realistic ones. This brings us to another salient concept in the title of this article, i.e., *realistic* emotions and affect: the missing links between some big, acted emotions that (too) often constituted databases so far, and the requirements of real-life situations are too many and too spurious. This holds for the modelling of both realistic human–human and human–machine scenarios.

The term 'realistic' – sometimes, 'naturalistic' is used instead – is rather vague, and again, it might best be defined ex negativo, as 'non-prompted' (Schiel, 1999), i.e., the speakers are not prompted explicitly to produce specific emotions. Often, 'acted' is used in the same sense as 'prompted' but we can imagine someone acting without intentionally producing specific emotions. The ideal is of course to record speakers who are not aware that they are being recorded, in a natural setting; due to privacy reasons, this is seldom possible – and if it is, such data cannot be distributed freely. The task is therefore to design a scenario that is as close as possible to a real natural setting. Note that anyway, there will be no *generic* natural setting because every setting is grounded in and coloured by varying factors such as selection of subjects, specificities of tasks, etc. This is a lesson to be learnt by the community, the same way as the Automatic Speech Recognition (ASR) community had to learn that there is a multiplicity of registers and varieties, if it comes to non-read speech.

1.2. The last 15 years

We can sub-categorise the time-line of this field during the last 15 years into three phases: some spurious papers on recognition of emotion in speech during the second half of the 90s (less than 10 per year), a growing interest until 2004 (maybe some 30 per year), and then, a steep rise until today (>100 per year).[1] The last special issue on emotion processing in speech in Speech Communication appeared in 2003, i.e., before automatic emotion processing really turned from 'yet another exotic speech topic' into mainstream. Some of the articles in that issue were dealing with basic topics such as preliminaries in this field and modelling of emotion, some with the synthesis of emotion, and three were dealing with different aspects of automatically recognising emotions in speech: acted vs. non-acted emotions (Batliner et al., 2003a), recognition of stress (Fernandez and Picard, 2003), and emotion recognition seen from the perspective of automatic speech recognition (tenBosch, 2003). At that time, it might have been difficult to compile a whole issue with high-quality articles dealing only with the automatic recognition of realistic emotions in speech. These times have changed. However, what still can be observed nowadays is on the one hand, a more and more sophisticated employment of statistical procedures; on the other hand, these procedures do not keep up with the requirements of processing application-oriented, realistic speech data, and too often, the data used are still un-realistic, i.e., prompted and acted, and thus not representative for real-life. Moreover, in comparison to related speech processing tasks such as Automatic Speech and Speaker Recognition, practically no standardised corpora and test-conditions exist to compare performances under exactly the same conditions. Instead a multiplicity of evaluation strategies employed – such as cross-validation or percentage splits without proper instance definition – prevents exact reproducibility. Most importantly, in order to face more realistic use cases, the community is in desperate need of more spontaneous and less prototypical data.

1.3. Databases

A coarse taxonomy of emotional databases can subdivide into *type of speech* and *type of scenario*. Type of speech can be *prompted* (mostly acted) emotions, or *non-prompted* emotions, elicited/obtained in specific scenarios. Prompted emotions are often out-of-context, and the speakers are explicitly told to produce specific emotions while often

[1] These rough figures are based on a search for 'emotion AND speech AND recognition' in Scopus; the tendency does not change even if we take into account that the coverage of Scopus might have increased during these years.

producing segmentally identical utterances. It is hoped that by that, generic (maybe universal) expressions of emotions can be obtained. The decisive disadvantages are first that deliberately acting emotions is not the same as producing 'spontaneous' emotions (Erickson et al., 2004). Second, the types of emotions that normally are prompted are definitely not the same as one would encounter in realistic scenarios. Sometimes, some context is given: The subjects have to watch film clips, or to imagine specific emotion-prone situations. Such settings might enable (induce) more realistic productions, but we cannot be sure as for the extent of naturalness; moreover, the resulting speech is monologic and not dialogic/interactive. Non-prompted emotions are elicited in more or less application-prone scenarios. Apart from several problems pertaining the design of these experiments, or the selection of recordings out of real-life databases, these data are most likely *non-generic*, i.e., not representing some universal emotions but confined to specific scenarios: In scenarios that are interesting for potential applications, we will seldom encounter emotions such as disgust or sadness but rather something like disappointment/frustration, or interest vs. boredom, i.e., not pure, full-blown emotions but rather emotion-related states. (Note that with a broader coverage of representative scenarios and emotion-related states that we will hopefully encounter in some future, the distinction between a few universal emotions and scenario-specific states will give room for a better conceptualisation of 'representativity' and 'genericity'.)

Here, we will only give a short typology of emotional databases and refer to some representative examples; catalogues can be found, for instance, in (Ververidis and Kotropoulos, 2003; Cowie et al., 2005; Schuller et al., 2009e). Still widely used – because freely available – acted databases are The Danish Emotional Speech Database (DES) (Engberg et al., 1997), and The Berlin Emotional Speech Database (BES) (Burkhardt et al., 2005), as well as The Speech Under Simulated and Actual Stress (SUSAS) Database (Hansen and Bou-Ghazale, 1997). DES and BES are representative for the 'early' databases in the nineties but still serve as exemplars for acted emotional databases, for instance in other languages. For English, there is the 2002 Emotional Prosody Speech and Transcripts acted database (Hirschberg et al., 2003).

Scenarios for recording non-prompted emotions can consist of *human–human* or *human–machine* interaction; the machine can be a 'real' machine/robot, or can be played by a human operator (so called Wizard-of-Oz scenario). Human–machine interaction normally takes place in the laboratory or in carefully controlled settings 'in the wild'. Human–human interaction can take place in the laboratory as well, or in some specific real-life settings where recording is possible or already given. Representatives for such real-life settings are TV-recordings (The Vera-Am-Mittag (VAM) Corpus (Grimm et al., 2008)), call-centre interactions (Devillers et al., 2005b; Liscombe et al., 1845), multi-party interaction (the ICSI meeting corpus (Laskowski, 2009)), or The Speech In Minimal Invasive Surgery Database (SIMIS) (Schuller et al., 2010a). Human–machine call-centre interaction (Ang et al., 2002; Batliner et al., 2004) is representative for telephone conversations in real-life settings. Typical human–machine interactions in the laboratory are stress detection in a driving simulation (Fernandez and Picard, 2003), tutoring dialogues (Litman and Forbes, 2003; Liscombe, 2005), information systems (Batliner et al., 2003b), or human–robot communication (Batliner et al., 2008b); a still emerging field is interaction with virtual agents (Schröder et al., 2008).

Today, more databases are used for exploiting textual analysis, either on its own for document processing, or – more frequently – in combination with acoustic information. Moreover, video information such as facial expression or, in a few cases, also body postures and gestures is exploited, followed by fewer studies on audiovisual processing (e.g., Chen et al., 1998; Pantic and Rothkrantz, 2003; Busso et al., 2004; Wang and Guan, 2005; Schuller et al., 2007b, 2009b; Zeng et al., 2007a,b, 2009). At the same time a smaller group of studies dealing with analysis of biosignals, such as heart rate, skin conductivity, or brain-waves are found (e.g., de Gelder et al., 1999; de Gelder and Vroomen, 2000; Picard et al., 2001; Kim et al., 2004, 2005; Nasoz et al., 2004; Takahashi, 2004; Wagner et al., 2005), also in combination with speech. Due to the general trend to multimodal processing and availability of more complex algorithms to this aim such as asynchronous Hidden Markov Models (HMM), Dynamic Bayesian Networks (DBN), general Graphical Models or Multidimensional Dynamic Time Warp (DTW) and Meta-Classification (Bengio and Frasconi, 1995; Tomlinson et al., 1996; Fiscus, 1997; Lizhong et al., 1999; Nefian et al., 2002; Bengio, 2003; Boda, 2004; Gunes and Piccardi, 2005; Al-Hames and Rigoll, 2006; Batliner et al., 2006b; Wimmer et al., 2008; Wöllmer et al., 2009) in combination with higher availability of multimodal data, the future might be expected to be dominated by multimodal approaches, potentially also integrating more contextual information (Forbes-Riley and Litman, 2004; Devillers et al., 2005a; Wöllmer et al., 2008; Schuller et al., 2009b). It is hoped that multimodality will help achieving higher recognition performance.

1.4. Modelling and annotation: categories vs. dimensions

A core issue in the psychological and theoretical discourse is the question whether emotions should be modelled as categories or dimensions, and how many categories or dimensions are needed for an adequate modelling. In the related field of phonetic processing, a strong focus on categorical perception (Harnad, 1987) has been replaced by some weaker assumptions. The same might happen to emotion processing (Said et al., 2010): Most likely, both categorical and continuous perception and processing do co-exist. Yet, it is fully open whether processing

by machine should be modelled along the lines of human processing, or simply output some correlate to the results of higher cortical processes.

The engineering approach has been a rather straightforward *we take what we get*, and so far, performance has been the decisive criterion: A reduction of complexity normally goes along with improved performance. In retrospect, the concentration on a few acted emotions at the beginning of the whole endeavour resulted in a sort of reality shock when non-prompted, realistic, and sparse events were addressed. Of course, the requirements of potential applications could be decisive in choosing amongst alternatives; so far, however, it is not yet a question of how detailed emotion modelling should be for successful processing in some application (i.e., in vivo) but how reliable and robust detection and classification can be in vitro.

Thus, it seems safe to resort to some surface taxonomy; in the case of categories, there is some hierarchy from *main categories* – such as positive, neutral, negative, or the 'big *n*' emotions, e.g., anger, fear, sadness, joy, etc. – to *sub-categories* modelling different shades of the main categories. All these categories can be thought of as *fixed* or *graded* (e.g., weak, medium, strong), or as *pure* or *mixed* – sometimes even antagonistic, if for instance a mixture of anger/joy/irony is being observed. In the case of *dimensions*, it is foremost the question how many (and which) dimensions we should assume: Traditionally, arousal and valence are modelled, with or without a third category power/dominance/control. Of course, dimensional modelling can be more or less continuous, and if we assume more than one dimension, it automatically results in 'mixed' representations in the *n*-dimensional space. If emotion is conceptualised in a broader meaning, most likely, some other dimensions representing social, interactive behaviour will be modelled as well (Batliner et al., 2008b; Vinciarelli et al., 2008).

Irrespective of strong beliefs in the one or the other type of modelling, in practice, categories can always be mapped onto dimensions and vice versa – albeit not necessarily lossless. For doing that, however, we have to produce either a detailed, fine-grained annotation by a few experts, or some coarse-grained annotation by several labellers, which normally are less experienced; all this is simply due to restricted financial resources. Moreover, the choice of recording and annotation tools such as FEELTRACE (Cowie et al., 2000) inevitably restricts the choice of phenomena that can be taken care of.

Normally, annotations are taken as representing the ground truth. This might not be (fully) true, for several reasons. First, there can be inconsistencies in the annotation process itself. Second, the effort needed often prevents a sufficiently high number of annotators; we might need more if 'naive' subjects are employed – for representing the 'wisdom of the crowd' – and a bit less if experts are employed. A number close to ten might be optimal; often, it is much less. Third, annotations rather model perception processes, not production processes, i.e., perceived, not felt emotions. In this respect, physiological signals might come closer to the ground truth; however, these are not available, most of the time. For want of something better, taking what is available – be this annotations or physiological signals or both – has always been the technological approach towards ground truth.

1.5. Unit of analysis

In our context, we define the original 'unit of analysis' as a speech episode which has been segmented – and often, but not necessarily, stored as a single speech file – according to more or less clear criteria. Often, it is called a 'turn' which starts when a speaker starts speaking, and ends when she/he stops speaking – in a conversation, when the dialogue parter takes the turn ('turn-taking'). This is an objective measure, easily obtainable for acted, read data and for interactions/conversations containing short dialogue moves. However, turns can often be very long. In such cases, either some intuitive notion of 'emotional unit' is used as criterion, or some objective measure such as pauses longer than, e.g., 0.5, or 1 s. Such prosodic units are highly correlated but not necessarily co-extensive with syntactic-semantic units and dialogue moves. Often, they span over more than one smaller unit; thus, they might be too long and the marking of specific and shorter emotional episodes might be smeared, resulting in sub-optimal classification performance. Two strategies, coping with these problems, can be observed. One defines 'technical units' such as frames, time slices, or proportions of longer units (for instance, the unit is subdivided into three parts of equal length) (Schuller and Rigoll, 2006; Schuller et al., 2008e; Shami and Verhelst, 2007; Vlasenko et al., 2008). The other one defines meaningful units with varying length such as syllables, words, phrases, i.e., chunks that are linguistically and by that, semantically, well-defined. In (Batliner et al., 2010), we opted for the word as the smallest possible, meaningful emotional unit (named *eme* along the same lines as *phonemes* and *sememes*) which can constitute larger meaningful emotional units; in (Seppi et al., 2010) we showed that stressed syllables alone are on par with words as far as classification performance is concerned.

For three reasons, finding an appropriate unit of analysis will become mandatory in the future: First, 'emotionally consistent' units will favour an optimal classification performance. Second, incremental processing, i.e., providing an estimate of the emotion before a longer utterance is finished, will often be necessary in real-life conversational systems, to enable reasonably fast system reactions. Third, for multimodal processing, an alignment of the different streams of information – speech, facial gestures, bio-signal, etc. – has to be found anyway; in the long run, ad hoc segmentation strategies will be replaced by the alignment of meaningful units for each modality, taking into account their relevance for the specific setting.

1.6. Prototypicality vs. open microphone setting

A prototype is a salient, central member of a category and typically most often associated with this category (Rosch, 1975). If no external criterion is available, real-life data have to be annotated manually for obtaining a 'ground truth' as reference: A straightforward approach is to speak of 'prototypical' cases if the labellers agree. On the other hand, the well-known, big *n* emotions might be conceived as prototypes (cf. Fehr and Russel, 1984; Shaver et al., 1992). Non-prototypical weak and/or mixed emotions (Martin et al., 2006) can be found when labellers annotate more than one emotion per item, or when we preserve the disagreement of several labellers in some sort of graded/mixed annotation. In (Steidl et al., 2009), non-prototypical emotions are conceptualised under the notion of *hinterland*: In contrast to the big *n* emotions anger, joy, sadness, despair, etc., emotion-related, affective states such as interest, tiredness, etc., and mixed/weak emotions, are conceived as constituting the hinterland of prototypical emotions.

The effects of non-prototypicallity on automatic emotion recognition have been studied only recently (cf. Batliner et al., 2005; Seppi et al., 2008a; Mower et al., 2009; Steidl, 2009; Steidl et al., 2009). If we do not select but deal with the whole database, classes might be 'noisy' and classification performance rather low. This is, however, a realistic setting. If we select most prototypical cases, classification performance might be considerably higher but we will not deal with a fully realistic scenario. In (Seppi et al., 2008a), for instance, the degree of prototypicality of the training material has been varied while keeping everything else constant. The idea behind is, of course that, using only more prototypical cases could yield higher classification performance. This approach is sometimes used in machine learning and is known as data cleaning or pruning: However, suspicious patterns may not always be garbage patterns as noisy data too might be needed to make the classifier learn those difficult patterns. Although ambiguous, many patterns proved to be still important and do characterise spontaneous emotional speech. Similar results obtained on different datasets and with different classifiers are confirmed by Mower et al. (2009) where models created for prototypical emotional speech proved to be ineffective for the classification of non-prototypical utterances.

Other experiments focus on prototypical data of both training and test material. In that case, performance rapidly improves by increasing the degree of prototypicality. Although results across increments are not directly comparable, we can see that prototypical patterns clearly allow to boost classification performance.

Finally, in (Steidl et al., 2009) a slightly different approach is taken leading to identical conclusions: Prototypical examples are selected exclusively among test patterns. This is of course the best-case scenario with respect to performance improvement that could be boosted up to an impressive performance of 85% for a two-class problem.

In an 'open microphone setting', everything that comes in is recorded and processed. This is of course a more realistic setting, mandatory for real-life applications, but yielding lower classification performance (Mower et al., 2009; Schuller et al., 2009d; Steidl et al., 2009, 2010).

2. Automatic processing

The last section addressed the necessary prerequisites – conceptualisations, models, data and their degree of prototypicality, and units of analysis. In this section, we deal with the parts and pieces necessary for automatic processing itself: features, classifiers, robustness considerations, evaluation, and system integration and implementation.

2.1. Features

Arguably the most important step in the automated recognition of emotions is the extraction of a reasonably limited, meaningful, and informative set of features. So far there has not been a large-scale, comprehensive comparison of different feature types; as for preliminary efforts in this direction (cf. Vogt and André, 2005; Batliner et al., 2006b, 2008a, 2011; Schuller et al., 2007a, 2009f; Steidl, 2009; Eyben et al., 2009; Eyben et al., 2010a).

Presenting a comprehensive overview of feature types and feature extraction methods requires some kind of division of features into classes, though there is no unique, preferable way towards such a taxonomy. The most basic distinction to be made is technology driven: Acoustic vs. linguistic features are usually considered separately, as extraction methods for these two types are extremely different. Their relative contribution can also vary greatly, depending on the database being analysed: For instance, for data based on scripted speech, linguistic features are normally of no value, apart from some specific applications such as data mining in movie archives. On the other hand, as the register comes closer to spontaneous/real-life speech, these features can gain considerably in importance.

In the past, the common focus was put on prosodic features, more specifically on pitch, duration and intensity, and less frequently on voice quality features as harmonics-to-noise ratio (HNR), jitter, or shimmer. Segmental, spectral features modelling formants, or cepstral features (MFCC) (Sato and Obuchi, 2007) are also often found in the literature.

Few years ago, a comparably small features set (few tens) has usually been employed. The recent success of systematically generated static feature vectors (derived by projection of the low-level descriptors (LLD), such as pitch or energy, on single values by descriptive statistical functionals, such as lower order moments or extrema) is probably justified by the supra-segmental nature of the phenomena found in emotional speech. As an alternative, a few studies modelled low-level features directly, mainly by Gaussian Mixture Models (GMM, i.e., continuous HMM with a single state that use weighted mixtures of Gaussian densities)

(Neiberg et al., 2006; Steidl, 2009) and HMM (Nogueiras et al., 2001; Nwe et al., 2003; Schuller et al., 2003; Yu et al., 2004; Inanoglu and Caneel, 2005; Nose et al., 2007; Wagner et al., 2007).

The large number of LLD and functionals has recently promoted the extraction of very large feature vectors (brute-force extraction), up to many thousands of features obtained either by analytical feature generation (Vogt and André, 2005; Schuller et al., 2008e; Pachet and Roy, 2009) or, in a few studies, by evolutionary generation (Schuller et al., 2006c). Such brute-forcing also often includes hierarchical functional application (e.g., mean of maxima) to better cope with statistical outliers. The brute-force extraction of acoustic parameters must be followed by feature selection methods: Lower computational costs and, sometimes, even feasibility, is guaranteed by the elimination of uninformative and redundant features.

However, also expert-based hand-crafted features will still play their role, as these are lately often crafted with more emphasis put on details hard to find by sheer brute-forcing such as perceptually more adequate ones (such as Teager energy), long-term temporal envelopes of the outputs of a gammatone filterbank (Wu et al., 2008b), or more complex features such as articulatory ones, for instance, (de-)centralisation of vowels. This can thus also be expected as a trend in future acoustic feature computation.

Fig. 1 provides an overview of the commonly used features and the principle of their brute-forcing in several layers that will be explained in more detail in the following.

2.1.1. Acoustic low-level descriptors

Pitch, intensity, duration, voice quality, and long term spectra are gathered under the umbrella of supra-segmental or prosodic features (Frick, 1985; Hess et al., 1996; Kießling, 1997; Johnstone and Scherer, 2000; Nöth et al., 2002; Scherer et al., 2003; Hui et al., 2006; Wu et al., 2008a).

Duration features model temporal aspects. Different types of normalisation can be applied. Relative positions on the time axis of base contours like energy and pitch such as maxima or on-/off-set positions do not strictly represent energy and pitch but duration – simply because they are measured in seconds, and because they are often highly correlated with duration features (Batliner et al., 2001). In other words, duration attributes can be distinguished according to their extraction nature: Those that represent temporal aspects of other acoustic base contours, and those that exclusively represent the parameter 'duration' of higher phonological units, like phonemes, syllables, words, pauses, or utterances. Duration values are usually correlated with the linguistic features described below: For instance, function words are shorter on average, content words are longer: This information can be used for classification, no matter whether it is encoded in linguistic or acoustic (i.e., duration) features (Batliner et al., 2001).

Intensity features usually model the loudness (energy) of a sound as perceived by the human ear, based on the amplitude in different intervals; different types of normalisation are applied (Zwicker and Fastl, 1999). Energy features can model intervals or characterising points. As the intensity of a stimulus increases, the hearing sensation grows logarithmically (decibel scale). It is further well-known that sound perception also depends on the spectral distribution and on its duration as well. The loudness contour is the sequence of short-term loudness values extracted on a frame base. So-called energy features are finally obtained from the loudness contour by applying functionals.

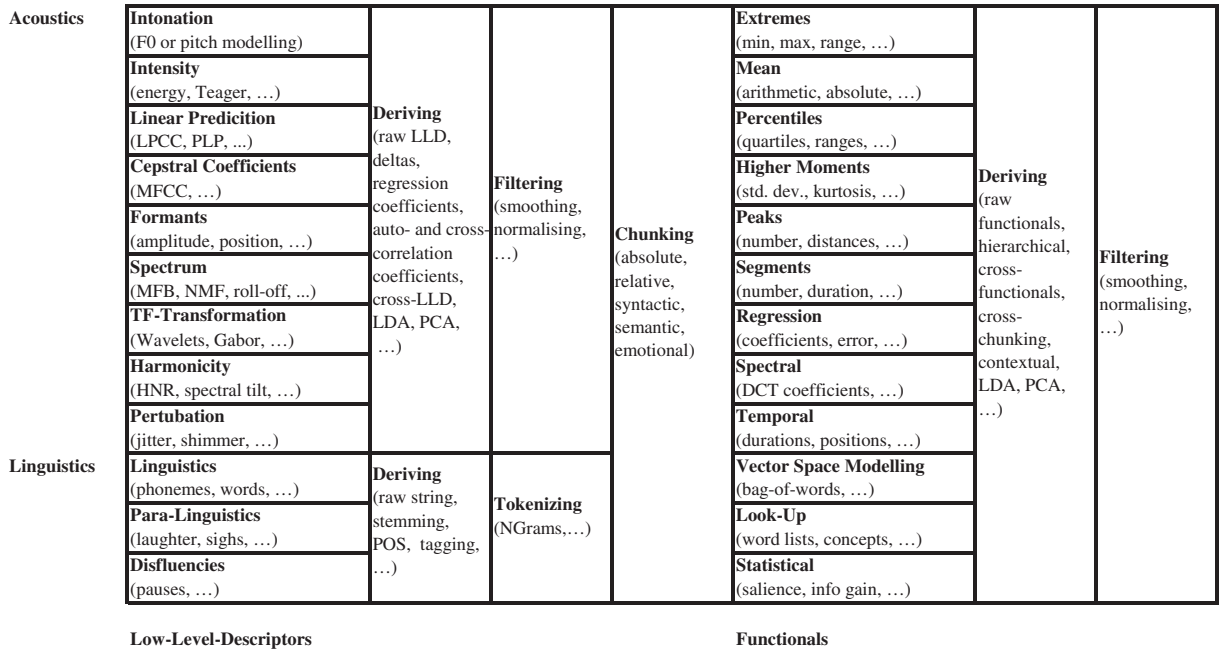


Fig. 1. Taxonomy of features commonly used for acoustic and linguistic emotion recognition. Abbreviations: Discrete Cosine Transform (DCT), Linear Prediction Cepstral Coefficients (LPCC), Mel Frequency Bands (MFB).

The basics of ***pitch*** extraction have largely remained the same; nearly all Pitch Detection Algorithms (PDAs) are built using frame-based analysis: The speech signal is broken into overlapping frames and a pitch value is inferred from each segment mostly by autocorrelation (Rabiner, 1977) in its manifold variants and derivatives (Boersma, 1993; Cheveigne and Kawahara, 2002; Sood et al., 2004; Ding et al., 2006). Often, the Linear Predictive Coding (LPC) residual or a low pass filtered version is used over the original signal. Other approaches use the cepstral representation (Noll, 1967) or exploit harmonic information by spectral compression. Pitch features are often made perceptually more adequate by logarithmic/semitone transformation, or normalisation with respect to some (speaker-specific) baseline. Pitch extraction is error-prone itself, which may influence emotion recognition performance (Batliner, 2007). However, the experimental results presented in (Steidl et al, 2008) on the FAU Aibo Emotion Corpus indicate that the influence is rather small, at least for the current state-of-the-art in modelling pitch features and for this type of data.

Voice quality is a complicated issue in itself, since there are many different measures of voice quality (Lugger et al., 2006; Lugger and Yang, 2008), mostly clinical in origin and mostly evaluated for constant vowels only, though once again standardisation in this area is lacking. Other, less well-known voice quality features were intended towards normal speech from the outset, e.g., those modelling 'irregular phonation' (cf. Batliner et al., 2007). Noise-to-Harmonic Ratio, jitter, shimmer, and further micro-prosodic events are measures of the quality of the speech signal. Although they depend in part on other LLDs such as pitch (jitter) and energy (shimmer), they reflect peculiar voice quality properties such as breathiness or harshness.

The ***spectrum*** is characterised by formants (spectral maxima) modelling spoken content, especially the lower ones (Fattah et al., 2008). Higher formants also represent speaker characteristics. Each one is fully represented by position, amplitude and bandwidth. The estimation of formant frequencies and bandwidths can be based on LPC (Atal and Hanauer, 1971; Makhoul, 1975) or on cepstral analysis (Davis and Mermelstein, 1980). LPC enables one to model the human vocal tract. Once the spectral envelope is estimated by using the LPC method, a number of further spectral features can be computed such as centroid, flux and roll-off or more specific as ratio of spectral flatness measure to spectral centre (Kim et al., 2007). Furthermore, the long term average spectrum over a unit can be employed: this averages out formant information, giving general spectral trends.

The ***cepstrum***, i.e., the inverse spectral transform of the logarithm of the spectrum (Bogert et al., 1963), emphasises changes or periodicity in the spectrum, while being relatively robust against noise. Its basic unit is quefrency. Mel-Frequency-Cepstral-Coefficients (MFCCs) – as homomorphic transform with equidistant band-pass-filters on the Mel-scale – tend to strongly depend on the spoken content. Yet, they have been proven beneficial in practically any speech processing task. Perceptual Linear Predictive (PLP) coefficients (Hermansky, 1990) and MFCCs are extremely similar, as they both correspond to a short-term spectrum smoothing – the former by an auto-regressive model, the latter by the cepstrum – and to an approximation of the auditory system by filter-bank-based methods. At the same time, PLP coefficients are also an improvement of LPC by using the perceptually based Bark filter bank.

Wavelets give a short-term multi-resolution analysis of time, energy and frequencies in a speech signal (Daubechies, 1990). Compared to similar parametric representations such as MFCCs, they are superior in the modelling of temporal aspects (cf. Fernandez and Picard, 2003; Schuller et al., 2007a).

Non-linguistic vocalisations are non-verbal phenomena such as breathing and laughter (Laskowski, 2009). Automatic detection of disfluencies and non-verbals normally requires that the vocabulary used by the ASR engine includes both these entities. Thus they could be subsumed under linguistic features as well.

2.1.2. Functionals

Subsequently to the LLD extraction, a number of operators and functionals are applied to obtain feature vectors of equal size out of each base contour. Functionals provide a sort of normalisation over time: Base contours associated, e.g., to words have different lengths, depending on the duration of each word and on the dimension of the window step; with the usage of functionals, we obtain one feature vector per word, with a constant number of elements, ready to be modelled by a static classifier. This cascade procedure, namely LLD extraction followed by functional application, has two major advantages: Features deriving from longer time intervals can be used to normalise local ones, and the overall number of features might be opportunely shrunk or expanded with respect to the number of initial LLD.

Before functionals are applied, LLD can be filtered or (perceptually) transformed, and first or second derivatives are often calculated and end up as additional LLDs. Functionals can range from statistical ones to curve fitting methods, or even methods based on perceptual criteria. The most popular statistical functionals cover the first four moments (mean, standard deviation, skewness and curtosis), order statistics (extremes value and their temporal information), quartiles, amplitude ranges, zero-crossing rates, roll-on/-off, on-/off-sets and higher level analysis. Curve (mainly linear) fitting methods produce regression coefficients, such as the slope of linear regression, and regression errors (such as the mean square errors between the regression curve and the original LLD). Maybe the most comprehensive list of functionals adopted so far in this field can be found in (Schuller et al., 2007a; Batliner et al., 2011).

Note however that functionals are not always necessary. Dynamic classifiers, such as HMMs, provide implicit time normalisation (cf. Section 2.2). An alternative option to functionals is to use a fixed number of windows per LLD; in other words, LLD can be down-sampled to vectors of equal size, by windowing the LLD a fixed number of times. In this case the window choice and displacement plays an important role: Windows can overlap or gaps can be left in-between, or even a non-uniform spacing of the windows can be opted for. Finally, a smoothed window shape, like a Hamming or Gaussian window, can help improving the robustness of the estimation.

Also, multi-instance learning techniques have been used to process features without the application of functionals, but of unknown frame sequence length (Shami and Verhelst, 2007; Schuller and Rigoll, 2009).

From the linguistic point of view, functionals are not really necessary. Features are already often extracted on a word basis. When longer segments of speech are used, time normalisation is implicitly provided by features like bag-of-words, where frequency histograms are constructed.

2.1.3. Linguistic features

Spoken or written text also carries information about the underlying affective state (Arunachalam et al., 2001). This is usually reflected in the usage of certain words or grammatical alterations, which means in turn, in the usage of specific higher semantic and pragmatic entities. A number of approaches exists for this analysis: key-word spotting (Elliott, 1992; Cowie et al., 1999), rule-based modelling (Litman and Forbes, 2003), Semantic Trees (Zhe and Boucouvalas, 2002), Latent Semantic Analysis (Goertzel et al., 2000), Transformation-based Learning (Wu et al., 2005), World-knowledge-Modelling (Liu et al., 2003), Key-Phrase-Spotting (Schuller et al., 2004), String Kernels (Schuller et al., 2009a), and Bayesian Networks (Breese and Ball, 1998). Context/pragmatic information has been modelled as well, e.g., type of system prompt (Steidl et al., 2004), dialogue acts (Batliner et al., 2003a; Litman and Forbes, 2003), or system and user performance (Ai et al., 2006). Two methods seem to be predominant, presumably because they are shallow representations of linguistic knowledge and have already been frequently employed in automatic speech processing: *(class-based) N-Grams* (Polzin and Waibel, 2000; Ang et al., 2002; Lee et al., 2002; Devillers, 2003) and *Bag-of-Words* (*vector space modelling*) (cf. Schuller et al., 2005b; Batliner et al., 2006b). In addition, exploitation of on-line knowledge sources without domain specific model training has recently become an interesting alternative or addition (Schuller et al., 2009c) – e.g., to cope with out-of-vocabulary events.

Although we are considering the analysis from spoken text, only few results for emotion recognition rely on ASR output (Schuller et al., 2005b, 2009b,a, 2010b) rather than on manual annotation of the data (Batliner et al., 2006b; Steidl, 2009). As ASR of emotional speech itself is a challenge (Athanaselis et al., 2005; Schuller et al., 2006d, 2007c, 2008a, 2009a; Wöllmer et al., 2009; Steidl et al., 2010) this step is likely to introduce errors. To some extent errors deriving from ASR and human transcription can be eliminated by soft-string-matching. In addition, the large related fields of opinion mining (Wilson et al., 2004; Popescu and Etzioni, 2005; Kim and Hovy, 2005; Missen and Boughanem, 2009) and sentiment analysis (Yi et al., 2003; Fei et al., 2006; Godbole, 2007; Missen and Boughanem, 2009) in text bear interesting alternatives and variants of methods.

To reduce complexity, stopping is usually used. This resembles elimination of irrelevant words. The traditional approach towards stopping is an expert-based list of words, e.g., of function words. Yet, even for an expert it seems hard to judge which words can be of importance in view of the affective context. Data-driven approaches as salience or information gain based reduction are popular. Another often highly effective way is stopping words that do not exceed a general minimum frequency of occurrence in the training corpus.

Tokenisation can be obtained by mapping the text onto word classes: Popular choices are lexemes and part-of-speech. The former choice is called Stemming, i.e., the clustering of morphological variants of a word by its stem into a *lexeme*. This reduces the number of entries in the vocabulary while at the same time providing more training instances per class. Thereby also words that were not seen in the training can be mapped upon lexemes, for instance by (Iterated) Lovins or Porter stemmers that base on suffix lists and rules (Lovins, 1968; Porter, 1980). Part-of-Speech (POS) tagging is a very compact approach where classes such as nouns, verbs, adjectives, particles, or more detailed sub-classes are modelled (Batliner et al., 2006b; Steidl, 2009). POS tagging and stemming have been studied thoroughly (Schuller et al., 2009a). In (Seppi et al., 2008b) the relation of ASR errors with errors based in automatic tokenisation are analysed.

Also sememes, i.e., semantic units represented by lexemes, can be clustered into higher semantic concepts such as generally positive or negative terms (Batliner et al., 2006b; Steidl, 2009). In addition, non-linguistic vocalisations like sighs and yawns (Russell et al., 2003), laughs (Campbell et al., 2005; Truong and van Leeuwen, 2005; Laskowski, 2009), cries (Pal et al., 2006), hesitations and consent (Schuller et al., 2008b; Schuller and Weninger, 2010), and coughs (Matos et al., 2006) can easily be integrated into the vocabulary (Batliner et al., 2006b; Schuller et al., 2006b, 2009b).

N-grams and class-based *N*-grams are commonly used for general language modelling. Thereby the posterior probability of a (class of a) word is given by its predecessors from left to right within a sequence of words. For emotion recognition, the probability of each emotion is determined per *N*-gram of an utterance. In addition, word-class based *N*-grams can be used as well, to better cope with data sparseness. For emotion recognition, due

to data sparseness mostly uni-grams ($N = 1$) have been applied so far (Lee et al., 2002; Devillers, 2003; Steidl, 2009), besides bi-grams ($N = 2$) and trigrams ($N = 3$) (Ang et al., 2002). The actual emotion is calculated by the posterior probability of the emotion given the actual word(s) by maximum likelihood or a-posteriori estimation. An extension of *N*-grams which copes with data sparseness even better is *Character N-grams*; in this case larger histories can be used.

Bag-of-words is a well-known numerical representation form of texts in automatic document categorisation (Joachims, 1998). It has been successfully ported to recognise sentiments (Pang et al., 2002) or emotion (Schuller et al., 2005b, 2006b; Steidl, 2009). Thereby each word in the vocabulary adds a dimension to a linguistic vector representing the term frequency within the actual utterance. Note that easily, very large feature spaces may occur, which usually require stopping and stemming. The logarithm of frequency is often used; this value is further better normalised by the length of the utterance and by the overall (log)frequency within the training corpus. It is possible not to refer to words, but to sequences of them, i.e., Bags-of-*N*-grams, to overcome the lack of word order modelling (Schuller et al., 2009a).

2.1.4. Feature selection

One of the ever-lasting questions in the field is which and how many features to choose. The answer to this question is important to improve reliability and performance but also to obtain more efficient models, also in terms of processing speed and memory requirements. A multiplicity of feature selection strategies has been employed.

Ideally, feature selection methods should not only reveal single (or groups of) most relevant attributes, but also decorrelate the feature space. In early studies, the features were designed following heuristic methods, mainly relying on human experience in this or similar research fields, such as ASR or the use of prosody for the automatic recognition of accents and boundaries (Kießling, 1997). Later and more recent studies benefit from the increase in computational power: Wrapper based selection – that is employing a target classifier's accuracy as optimisation criterion in 'closed loop' – has been used widely. However, even for relatively small data-sets, exhaustive selection is still not affordable. Therefore, the search in the feature space must employ some more restrictive, and thus less optimal, strategies. Probably the most common procedure chosen is the *sequential forward search* (Pudil et al., 1994) – a hill climbing selection starting with an empty set and sequentially adding best features; as this search function is prone to nesting, an additional floating option should be added: At each step one or more features are deleted and it is checked if others are more suited. It is noteworthy that the need for efficient feature selection in this field has also led to contributions of optimised solutions for this type of search (Ververidis and Kotropoulos, 2006; Brendel et al., 2010).

Apart from wrappers, less computationally expensive 'filter' or 'open loop' methods have attracted interest, such as information theoretic filters and correlation-based analysis (Hall, 1998).

On a different basis, *hierarchical* approaches (Lee et al., 2009) to feature selection try to optimise the feature set not globally for all emotion classes but for groups of them, mainly couples. In that way, more intelligent brute-forcing can be obtained, e.g., by search masks and by a broader selection of functionals to choose from, and of parameters to alter. In this way, an expert's experience can be combined with the freedom of exploration taken by an automatic generation.

Apart from genuine selection of features, the *reduction* of the feature space is often considered. It consists of the mapping of the input space onto a less dimensional one, while keeping as much information as possible. Principal Component Analysis (PCA) (Jolliffe, 2002) and Linear or Heteroscedastic Discriminant Analysis (LDA) (Fukunaga, 1990; Ayadi et al., 2007) are the most common techniques, used, e.g., in (Batliner et al., 2000; Kwon et al., 2003; Lee and Narayanan, 2005;Steidl, 2009). While PCA is an unsupervised feature reduction method (and thus maybe suboptimal for specific problems), LDA is a supervised feature reduction method which searches for the linear transformation that maximises the ratio of the determinants of the between-class covariance matrix and the within-class covariance matrix.

Note that all these techniques are well-known to present some drawbacks: PCA requires the guess of the dimensionality of the target space, which is not always easy; LDA demands at least some degree of Gaussian distribution and linear separability of the input space; both methods are, however, not very appropriate for feature mining, as the original features are not retained after the transformation.

Finally, it is worth to mention Independent Component Analysis (ICA) (Hyvärinen et al., 2001) and Non-negative Matrix Factorization (NMF) (Lee and Seung, 1999). ICA maps the feature space onto an orthogonal space; furthermore, the target features have the attractive property of being statistically independent. NMF, a method strictly related to probabilistic latent semantic analysis (Gaussier and Goutte, 2005), is a recent alternative to PCA in which the data and components are assumed to be non-negative: NMF learns to represent emotional classes with a set of basic emotional subclasses. There are already some studies adopting ICA (Rahurkar and Hansen, 2003), where both the input space and the output space are kept small; on the other hand, NMF is mainly employed for large linguistic feature sets. Another related factorizations is Singular Value Decomposition (SVD) that has also been used in this field (Kharat and Dudul, 2008).

Future studies will very likely address feature importance across databases (Eyben et al., 2010a) and further types of efficient feature selection (Rong et al, 2007; Altun and Polata, 2009).

2.2. Classification

The choice for an appropriate classification paradigm is dictated by a number of factors. Particularly crucial for this field are (1) the tolerance of high dimensionality, (2) the capability of exploiting a small data-set, and (3) the handling of skewed classes. Less crucial, though still important are other more general considerations such as the ability to solve non-linear problems, adaptation, missing data, efficiency and computational and memory costs.

The problem of a high dimensional feature set is usually better addressed by feature selection and elimination before actual classification takes place. Popular classifiers for emotion recognition such as Linear Discriminant Classifiers (LDCs) and k-Nearest Neighbour (kNN) classifiers have been used since the very first studies (Dellaert et al., 1996; Petrushin, 1999) and turned out to be quite successful for non-acted emotional speech as well (Batliner et al., 2000; Kwon et al., 2003; Litman and Forbes, 2003; Rahurkar and Hansen, 2003; Lee and Narayanan, 2005; Shami and Verhelst, 2007). However, they suffer from the increasing number of features that leads to regions of the feature space where data is very sparse ('curse of dimensionality' (Bellman, 1961)). Classifiers such as kNN that divide the feature space into cells are affected by the curse of dimensionality and are sensitive to outliers. A natural extension of LDCs are Support Vector Machines (SVM): If the input data have not been previously (implicitly) linearly transformed, which may have increased or decreased the number of features, and if the linear classifier obeys a maximum-margin fitting criterion, then we obtain an SVM. Although SVM are not necessarily the best classifiers for every constellation (cf. Meyer et al, 2002), they provide good generalisation properties (McGilloway et al., 2000; Lee et al., 2002; Chuang and Wu, 2004; You et al., 2006; Morrison et al., 2007b), and are nowadays conceived as a sort of state-of-the-art classifier.

Small data sets are, in general, best handled by discriminative classifiers. The most used non-linear discriminative classifiers are likely to be Artificial Neural Networks (ANNs) and decision trees. Decision hyperplanes learned with ANN might become very complex and depend on the topology of the network (number of neurons), on the learning algorithm (usually a derivation of the well-known Backpropagation algorithm) and on the activity rules. For this reason, ANNs are less robust to overfitting, and require greater amounts of data to be trained on. They are therefore rarely used for acted data (Petrushin, 1999; Martinez and Cruz, 2005), and even less for non-acted (but cf. Batliner et al., 2000, 2006b; Steidl, 2009). However, the recent incorporation of a long-short-term memory function seems to be a promising future directive (Eyben et al., 2010b; Wöllmer et al, 2010).

Decision trees are also characterised by the property of handling non-linearly separable data; moreover, they are less of a 'black box' compared to SVM or neural networks, since they are based on simple recursive splits (i.e., questions) of the data. These binary questions are very readable, especially if the tree has been adequately pruned. As accuracy degrades in case of irrelevant features or noisy patterns, Random Forests (RF) (Schuller et al., 2007a) can be employed: They consist of an ensemble of trees, each one accounting for random, small subsets of the input features obtained by sampling with replacement. They are practically insensitive to the curse of dimensionality, while, at the same time, still providing all the benefits of classification trees.

As spontaneous emotion classes are seldom evenly distributed, balancing of the training instances with respect to instances per emotion class is often a necessary step before classification (Schuller et al., 2009d,b; Steidl, 2009). Normally, most cases belong to the neutral class. The balancing of the output space can be addressed either by considering proper class weights (e.g., priors), or by resampling, i.e., (random) up- or down-sampling. Class priors are implicitly taken into account by discriminative classifiers. Yet another widely adopted option is the introduction of main classes by clustering of instances, such as the couples 'neutral vs. non-neutral' and 'positive vs. negative'.

As explained above, applying functionals to LLD is done for obtaining the same number of features for different lengths of units such as turns or words. Dynamic classifiers like Hidden Markov Models, Dynamic Bayesian Networks or simple Dynamic Time Warp allow to skip this step in the computation by implicitly warping observed feature sequences over time. Among dynamic classifiers, HMM have been used widely. The reason why is probably that elaborated tools such as HTK have been previously developed for similar tasks such as speech and speaker recognition. For acted emotions, there are numerous references (tenBosch, 2003; Schuller et al., 2003; Zeng et al., 2009); for non-acted emotions, fewer are known (Kwon et al., 2003; Vlasenko et al., 2007b; Wagner et al., 2007; Schuller et al., 2009d): The performance of static modelling through functionals is often reported as superior (Schuller et al., 2003, 2009d; Vlasenko et al., 2007b) as emotion is apparently better modelled on a time-scale above frame-level; note that a combination of static features such as minimum, maximum, onset, offset, duration, regression, etc. implicitly shape contour dynamics as well. A possibility to use static classifiers for frame-level feature processing is further given by multi-instance learning techniques, where a time series of unknown length is handled as one by SVM or similar techniques (Shami and Verhelst, 2007; Schuller and Rigoll, 2009). Still, when the spoken content is fixed, the combination of static and dynamic processing may help improve overall accuracy (Vlasenko et al., 2007a; Schuller et al., 2008c).

Ensembles of classifiers (Schuller et al., 2005b,c,a; Morrison et al., 2007a,) combine their individual strengths, and might improve training stability. There exists a number of different approaches to combine classifiers. Popular are methods based on majority voting such as *Bagging*,

Boosting and other variants (e.g., *MultiBoosting*). More powerful, however, is the combination of diverse classifiers by the introduction of a meta-classifier that learns 'which classifier to trust when' and is trained only on the output of 'base-level' classifiers, known as *Stacking* (Wolpert, 1992). If confidences are provided on lower level, they can be exploited as well. Still, the gain over single strong classifiers such as SVM may not justify the extra computational costs (Schuller et al., 2005a).

In line with the different models of emotion (cf. Section 1.4), also different approaches towards classification are needed. Historically, classification into a limited number of (few) discrete classes came first (Liscombe et al., 2003). With the advent of databases annotated in the dimensional space, *regression* was found as an alternative (Grimm et al., 2007). As real-life application is not limited to prototypical cases (cf. Section 1.6), also *detection* as opposed to classification can be expected as alternative paradigm: 'Out-of-vocabulary' emotions need to be handled as well, and apart from the easiest solution of introducing a garbage class (Schuller et al., 2009d), detection allows for more flexibility. Detection is thereby defined by inheriting a rejection threshold. Likewise, one model per emotion detects 'its' emotion, e.g., by log-likelihood ratios as in a typical speaker verification setup. At the same time, complex emotions demand for a shift in paradigm, as not one emotion is recognised, but an intensity score per emotion is provided. In this respect, *confidence measurements* should be mentioned, which is also a 'black spot' in the literature yet.

2.3. Robustness

First attempts to cope with noise were seen rather recently (Lugger et al., 2006; Schuller et al., 2006a, 2007c; You et al., 2006; Grimm et al., 2007) and are characterised by the simplification of additive stationary noise. This does of course not take into account the alteration of voice characteristics of the speaker herself, speaking in noisy environment – the well-known Lombard effect, which is, however, covered to some extent by parts of the SUSAS database. Also, methods applied to cope with noise are at present rather basic in this discipline: matched conditions training and feature spaces. Surprisingly, these seem to suffice to restore disturbances in the stationary case. However, a considerable amount of well elaborated speech and feature enhancement techniques (Schuller et al., 2009g) such as Switching Linear Dynamic Models, two-stage Wiener filtering, Histogram Equalisation, etc. are ready to be tested to better cope with noisy speech for emotion recognition – in particular when it comes to first attempts to deal with non-stationary and complex noise.

Apart from noise also reverberation impact has so far been considered only with stationary impulse responses (Schuller et al., 2007c). Effects of dynamic reverberation while, e.g., moving through a church or when entering or leaving a car, or more simply turning the head from the wall to the open still need to be investigated, though, and might profit from well-elaborated de-reverberation methods.

Interestingly, coding and transmission errors as package loss in Voice over IP or mobile telephony have also been neglected so far. In this respect also distributed emotion recognition could be a future topic, because with new technical requirements, a new issue will get more and more important: Real-time ability, which has been addressed only rarely so far, e.g., in (Vogt et al., 2009). However, some small, e.g., wearable devices might not provide sufficient computational power.

Apart from noise-based model adaptation, speaker adaptation seems to be one of the major factors to enhance robustness of general models. In speech processing, there is a long tradition to this end reflected in methods such as Vocal Tract Length Normalisation (VTLN), speaker clustering, or acoustic model adaptation. At present, first methods are successfully being employed such as VTLN (Vlasenko et al., 2008) or speaker normalisation (Schuller et al., 2006a, 2008d; Sethu et al., 2007). However, all these are usually not performed as true adaptation, as all processing is carried out in a static way, and not gradually with a truly unknown speaker. Also, acoustic models are not truly adapted dynamically yet. Thus, genuine dynamic adaptation and its effects have to be investigated. Another interesting possibility could be emotional user profiling and adaptation, e.g., in the form of emotional 'language models', i.e., *N*-grams or priors, or as full probabilistic state transition graphs.

2.4. Evaluation

To understand the uses and abuses of evaluation strategies adopted in the automatic recognition of emotion, a historical perspective is appropriate. Early studies started with speaker dependent recognition of emotion, just as in the recognition of speech. Therefore, even today, the lion's share of research presented relies on either subject dependent percentage split or subject dependent cross-validated test-runs. However, only Leave-One-Subject-Out (cf. Seppi et al., 2008b;Steidl, 2009) or Leave-One-Subject-Group-Out (Schuller et al., 2009d) (cross-)validation would ensure true speaker independence.

Apart from this, an even bigger problem is the simple lack of exact reproducibility: Clearly defined experiments where training and test partitioning are explicitly stated are rare. Ideally, such definition should be provided with, but not restricted to, future database releases. This kind of documentation should be employed in oncoming studies and allow a less chaotic development and improvement of the methods and the models adopted in this research field.

When doing comparative evaluations, most important is that everything that is done to modify or prepare the classifier must be done in advance before looking at the test data (Salzberg, 1997). To our knowledge, only few studies in emotion recognition clearly explain what – if any – part of the data has been used for the tuning of the parameters.

This is not reassuring because it might be that most of the studies might have actually optimised the classifier on the test set rather than on a proper, separate validation set.

The use of accuracy as performance measure in most of the studies nowadays might be based on the former studies using acted/read emotional databases, where there usually is no skewness at all in the emotional classes. In such cases, accuracy is an appropriate measure of performance but it rapidly fails to convey an adequate quantitative understanding of the performance of the models as soon as one or more classes become rare. A way around this problem is chosen in recent studies in emotion recognition that report the unweighted mean of the recall of each class (Schuller et al., 2009d) – this better reflects the imbalance among classes (usually a high percentage of neutral speech, but sparse instances of diverse non-neutral examples). In other studies, the very same metric has been called class-wise recognition rate (CL) (Batliner et al., 2003b).

Naturally, one will find also other measures such as precision or F1 – the harmonic mean of recall and precision. In the end, all these may be provided weighted or unweighted, and a very popular and at the same time complete representation in the sense of the above named measures is the confusion matrix, yet coming at the cost of 'several numbers' rather than 'one measure'.

In a binary class problem, from an application point of view, the benefit of the Receiver Operating Characteristic (ROC) curve was proved in (Steidl et al., 2009). The ROC plots the true positive rate (TPR) over the false positive rate (FPR) given different tunings of the classifier: The goal consists of the parallel maximisation of TPR while keeping a low FPR, thus leading to producing points in the upper triangle, close to the upper left corner of the graph. This methods is useful when, for instance, low false alarm rate is critical: In that case we should aim at a low FPR. If, on the other hand, we are interested in off-line analyses of the felicity of call centre interactions, we can live with a high amount of false instances as long as we get a reliable estimate for the quality of each interaction in total (Batliner et al., 2006a). Aiming at a single measure for the ROC, one can use the area under the ROC (AUC), or the equal error rate (EER), i.e., the point of equality of TPR and FPR. Naturally, these measures can also be calculated for multiple classes – again weighted or unweighted (Steidl et al., 2009). This can be done by modelling each class against all other classes.

While measures for ground truth evaluation such as kappa (Fleiss et al., 1969; Rosenberg and Binkowski, 2004) are often found, significance tests (Nickerson, 2000; Armstrong, 2007) are practically ignored to the present day, but (cf. Steidl, 2009; Schuller et al., 2009b; Seppi et al., 2010). This comes with some worry: Due to the ever-present sparseness of training and testing material, and due to the general limited size of the databases, the significance of results and improvements could be seriously compromised and should therefore be investigated. Another important issue are repeated measurements: The same databases are used over and over again in subsequent studies; further, in some research topics such as feature selection, a large number of models are trained on the same data, typically leading to the so-called multiplicity effect. In such circumstances a correction of the significance threshold (e.g., the Bonferroni adjustment (Pernegger, 1998)) would be required; unfortunately, in most of the cases, such adjustment would clearly invalidate any significant difference. On the other hand, already in (Eysenck, 1960; Rozeboom, 1960) and more recently again in (Nickerson, 2000), it was suggested to use significance not in the inferential meaning but as a sort of descriptive device. Moreover, there are strong arguments against 'statistical rituals' such as significance testing (Gigerenzer, 2004); effect size/power seem to be more adequate measures (Cohen, 1988; Ferguson, 2009) but have been largely neglected so far.

As should be clear now, comparability between research results in this field is considerably low. Differences are inflated by the diversity of corpora collected under different acoustic conditions, focussing on specific sub-populations such as children or students, or being filtered through a number of disparate channels such as close vs. far microphone, or fixed-line vs. mobile phone, just to mention a few. A severe issue in cross/multi-corpora studies is the inhomogeneous labelling process, which often leads to inconsistent, incompatible or even distinct emotional classes (Eyben et al., 2010a). Moreover, many studies report results on proprietary, mostly unavailable corpora (Chuang and Wu, 2004; Schuller et al., 2005a; Vidrascu and Devillers, 2007).

Other differences can be traced back to a lack of standard in processing and in jargon. Examples range from definition of low-level contours and functionals to an accurate description of the classifier chosen. This again is in contrast to the more or less settled and clearly defined feature types as, e.g., MFCC that allow for higher comparability in speech recognition.

In addition, differences exist with respect to the degree of automation: Pre-segmentation into turns is mostly provided beforehand by semi-automatic chopping. In some cases it is also done based on forced alignment (Schuller et al., 2007c), which requires (manual) transcripts of the spoken content. Also speaker adaptation – often reported as beneficial (Sethu et al., 2007) – has apparently not been investigated fully automatically so far; instead, the whole speaker content is utilised. Finally, linguistic features need the spoken content transcriptions. The reported success of linguistic feature inclusion, however, often relies on manual transcripts (Batliner et al., 2011). Only few studies compare manual vs. fully automatic processing in this respect (Schuller et al., 2006b, 2009a; Seppi et al., 2008b), or contrast manually corrected error-prone acoustic descriptors as pitch with exclusively automatically derived pitch (cf. Batliner, 2007; Steidl et al, 2008).

An attempt to overcome most of the problems listed in this section was obtained by cooperations and

competitions. Probably the first comparative experiments can be found in the CEICES initiative (Batliner et al., 2006b), where seven sites compared their classification results under exactly the same conditions. However, the primary goal was cooperation, not competition, in order to pool the different features together for one combined, unified selection process (Batliner et al., 2011). More comparisons and challenges, along the lines of the NIST evaluations in the field of speech processing, or the MIREX challenges in the field of music retrieval, will be needed. Thereby, detailed information on feature extraction and classification procedures will have to be made available to the organisers or, even better, to the community. First attempts in this direction have been the INTERSPEECH 2009 Emotion challenge (Schuller et al., 2009d) (cf. Section 3) and the following INTERSPEECH 2010 Paralinguistic Challenge (Schuller et al., 2010c).

Yet, we might ask whether test runs on databases are the optimal choice for measuring the true performance that is to be expected in a system operating in 'real life'. Thus, it seems desirable to test systems within their application framework, e.g., by usability analysis. However, due to the fact that there are only few systems operating to date, there are rather few studies of this kind. These, however, seem to indicate that there is a gain over not having automatic recognition of emotion (Burkhardt et al., 2009), while a human wizard would perform better (Devillers, 2003, 2007; Schuller et al., 2009b) – meaning that there definitely is room for improvement.

2.5. Implementation and system integration

Toolkits and freely available codes have always considerably influenced research, cf. the Hidden Markov Model Toolkit (HTK)[2] (Young et al., 2006) in the field of automatic speech recognition. In this respect, two types of available packages influenced the development in the recognition of emotion recognition: Firstly tools for feature extraction such as PRAAT[3] (Boersma and Weenink, 2005) or the SNACK sound toolkit,[4] and secondly classification software and learning environments such as the WEKA[5] (Witten and Frank, 2005) or Rapid Miner[6] packages. The frequent choice of HTK resulted in an almost exclusive use of MFCC and energy features as implemented in the HTK, if Hidden Markov Models were used for emotion processing. Of course, many individual feature extractions and classification algorithms are found as well.

However, only few dedicated open source projects are found that aim at a free platform, in particular for emotion recognition from speech. The Munich open-source Emotion and Affect Recognition Toolkit (openEAR) (Eyben et al., 2009) is the first of its kind to provide a free open source toolkit that integrates all three necessary components: feature extraction (by the fast openSMILE[7] backend (Eyben et al., 2010c)), classifiers, and pre-trained models. An interesting aspect of the openEAR toolkit is that it provides pre-defined sets of derived functionals for comparison such as the sets provided for the INTERSPEECH 2009 Emotion Challenge (Schuller et al., 2009d) and the following INTERSPEECH 2010 Paralinguistic Challenge (Schuller et al., 2010c). This is beneficial, as apart from some publicly available scripts for extractors such as PRAAT, highly individual implementations and solutions are found, also for the functionals. A similar initiative, yet not open source, is the EmoVoice[8] (Vogt et al., 2008) toolkit.

Once a recognition result is produced, it will have to be communicated to the application. A standard was not needed in the early days of emotion recognition from speech because the performances were not sufficiently mature, and no broad application basis existed; the need for standardised coding of information was dealt with only at a later stage. The HUMAINE EARL (Schröder et al., 2006) can be named as one of the first standards tailored to provide a well-defined description of recognised or to be synthesised emotions. Released as a working draft, the W3C EmotionML[9] (Schröder et al., 2006, 2007) followed EARL providing more flexibility and a broader coverage including action tendencies, appraisals, meta context or a basis to encode regulation, acting, meta-data, and ontologies. Particularly for the encoding of acoustic and linguistic features, a standard was proposed and used within the CEICES initiative (Batliner et al., 2011). Further, a number of standards exists which do not exclusively deal with emotion but contain definitions for suited tags, as EMMA[10] (Baggia et al., 2007). However, these standards have not been used frequently so far. This will most likely change when more and more applications will be addressed. At the same time, it remains to hope that either their number stays limited, or well-defined translations among standards will be provided, to avoid incompatibilities.

Whereas, for instance, EmotionML already provides a definition for communication of confidence of the results a system reports, this non-trivial goal has not really been targeted in studies on emotion recognition. Naturally, one can think of distances to hyperplanes of SVM or posterior probabilities, etc., yet evidence from independent sources as often followed in speech recognition, e.g., acoustic stability, have not been, to our knowledge, considered in this field so far. One peculiarity of the field appears promising in this respect, though: The fact that usually several labellers annotate the data allows for training of labeller specific emotion models. As a consequence, one can predict

[2] http://htk.eng.cam.ac.uk/.
[3] http://www.fon.hum.uva.nl/praat.
[4] http://www.speech.kth.se/snack/.
[5] http://www.cs.waikato.ac.nz/ml/weka.
[6] http://rapid-i.com/.
[7] http://www.openaudio.eu/.
[8] http://mm-werkstatt.informatik.uni-augsburg.de/EmoVoice.html.
[9] http://www.w3.org/2005/Incubator/emotion/XGR-emotionml/.
[10] http://www.w3.org/TR/emma/.

the inter-labeller agreement in addition to the emotion, which may serve as a confidence: In a pioneering study, for example, an error of only one labeller on average out of five is reported on the FAU AIBO corpus (Steidl et al., 2009). However, employing the error or confidences when it comes to the temporal alignment, or a change to a spotting paradigm, may be further promising strategies.

Another question that usually arises when dealing with system integration are real-time issues (Lefter et al., 2010) and incremental processing or 'gating' – i.e., how fast after the beginning of a speech turn one can provide a reasonable prediction of the likely emotion. This is for example of importance in dialogue systems such as the SEMAINE system (Schröder et al., 2008) where complex multimodal system outputs have to be generated in advance to 'fire' the right alternative just when it is needed. For that, an early prediction in the right direction is of course meaningful. Few studies deal with this topic; Schuller and Rigoll (2006) indicates that one second of speech might be sufficient. An incremental update of the prediction is also possible by classifiers suited for this task that may at the same time learn how much past context information is beneficial (Wöllmer et al., 2008); some toolkits, e.g., openEAR, already allow for this type of processing (Eyben et al., 2009).

3. Lessons learnt from the first emotion challenge

The INTERSPEECH 2009 Emotion Challenge (Schuller et al., 2009d), organised by Schuller et al., was held in conjunction with INTERSPEECH 2009 in Brighton, UK. This challenge was the first open public evaluation of speech-based emotion recognition systems with a strict comparability where all participants were using the same corpus and identical settings.

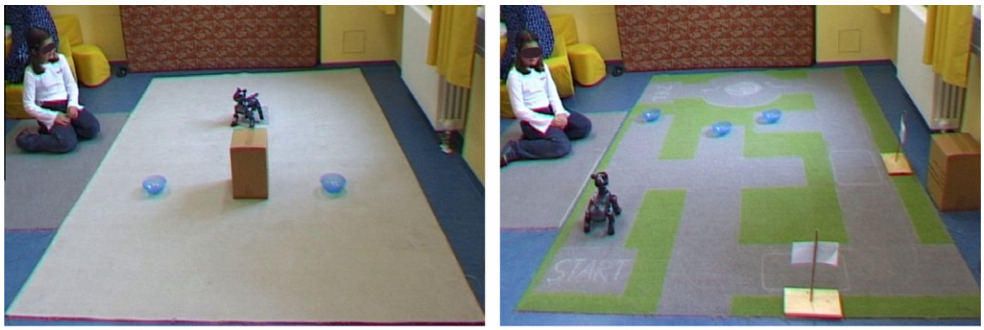
3.1. Database

The German FAU Aibo Emotion Corpus (Steidl, 2009) with 8.9 h of spontaneous, emotionally coloured children's speech comprises recordings of 51 German children at the age of 10–13 years from two different schools. The children were given five different tasks where they had to direct Sony's dog-like robot Aibo to certain objects (Fig. 2(a)) and through a given 'parcours' (Fig. 2(b)). The children were told that they could talk to Aibo the same way as to a real dog. However, Aibo was remote-controlled and followed a fixed, pre-determined course of actions, which was independent of what the child was actually saying. At certain positions Aibo disobeyed in order to elicit negative forms of emotions. The corpus is annotated by five human labellers on the word level using 11 emotion categories that have been chosen prior to the labelling process by iteratively inspecting the data. In the INTERSPEECH 2009 Emotion Challenge, the unit of analysis are not single words, but semantically and syntactically meaningful, manually defined chunks (18 216 chunks, 2.66 words per chunk on average). Heuristic algorithms are used to map the decisions of the five human labellers on the word level onto a single emotion label for the whole chunk. The emotional states that can be observed in the corpus are rather non-prototypical, emotion-related states than 'pure' emotions. Mostly, they are characterised by low emotional intensity. For the Emotion Challenge, the complete corpus is used, i.e., no balanced subsets were defined, no rare states and no ambiguous states are removed – all data had to be processed and classified (cf. Steidl et al., 2009). Two different classification problems were designed: a 2-class problem with the two main classes *negative valence* and the default state *idle*, and a 5-class problem with the emotion classes *anger*, *emphatic* (the child speaks in a pronounced, accentuated, sometimes hyper-articulated way, but without showing any emotion), *neutral*, *positive valence*, and an explicit *rest* class representing all other categories as well as ambiguous states. Test and training partitions were defined. As it is typical for realistic data, the various emotion classes are highly unbalanced. The number of instances for the 2- and the 5-class problem are given in Table 1(a) and (b). As the children of one school were used for training and the children of the other school for testing, the partitions feature speaker independence as needed in most real-life settings.

3.2. Sub-Challenges

The challenge consists of three Sub-Challenges:

1. The *Open Performance Sub-Challenge* allowed contributors to find their own features with their own classification algorithms. However, they had to stick to the definition of test and training sets.
2. In the *Classifier Sub-Challenge*, participants designed their own classifiers and had to use a selection of 384 standard acoustic features, computed with the open SMILE toolkit (Eyben et al., 2009, 2010c) provided by the organisers. These features are based on 16 base contours (MFCC 1–12, RMS energy, F0, zero crossing rate, and HNR) and their first derivatives. Features for the whole chunk are obtained by applying 12 functionals (mean, standard deviation, kurtosis, skewness, maxima and minima by value, range and relative position, and the linear regression coefficients offset and slope and the according mean squared error). Participants had an option to subsample, alter, and combine features (e.g., by standardisation or analytical functions). The training could be bootstrapped, and several classifiers could be combined by tools such as Ensemble Learning, or side tasks learnt as gender, etc. However, the audio files could not be used for additional feature extraction in this task.
3. In the *Feature Sub-Challenge*, participants were encouraged to design 100 best features for emotion classification to be tested by the organisers in equivalent setting. In particular, novel, high-level, or perceptually adequate features were sought-after.



(a) Object localisation task (b) Parcours Experiment

Fig. 2. Experimental setup of the FAU Aibo Emotion Corpus (Steidl, 2009).

Table 1
Number of instances in the FAU Aibo Emotion Corpus.

(a) 2-class problem: **NEG**ative vs. **IDL**e

#	NEG	IDL	∑
Train	3358	6601	9959
Test	2465	5792	8257
∑	5823	12393	**18216**

(b) 5-class problem: **A**nger, **E**mphatic, **N**eutral, **P**ositive, **R**est

#	A	E	N	P	R	∑
Train	881	2093	5590	674	721	9959
Test	611	1508	5377	215	546	8257
∑	1492	3601	10967	889	1267	**18216**

Participants did not have access to the labels of the test data, and all learning and optimisations were based only on the training data. However, each participant could upload instance predictions to receive the confusion matrix and results from the test data set up to 25 times; the best result was counted. This opportunity was used to full capacity by most participants.

The number of 25 trials was decided for to start such evaluations gently allowing for multiple uploads, as this was the first official challenge in this field. In addition, this allowed for reporting results for different configurations, for instance for using acoustics or linguistics or both in the same paper submission. A limit was necessary in order to make sure no sampling of the test labels was possible. Note that we changed this rule by limiting to two uploads in the Affect Sub-Challenge of the successive INTERSPEECH 2010 Paralinguistic Challenge (Schuller et al., 2010c).

The format contained instance and prediction, and optionally additional probabilities per class. This later allowed a final fusion by majority vote of predicted classes of all participants' results to demonstrate the best possible performance of the combined efforts. As classes were unbalanced, the primary measure to optimise was firstly unweighted average (UA) recall, and secondly the accuracy. The choice of unweighted average recall was a necessary step to better reflect imbalance of instances among classes in real-world emotion recognition, where an emotionally 'idle' state usually dominates (cf. Section 2.4). Other well-suited and interesting measures such as the area under the Receiver Operating Characteristic curve were also considered (cf. Section 2.4); however, they were not used as they are no common measures in this field, yet. Furthermore, the ROC curve can be plotted only for 2-class problems. Classifiers for multi-class problems can be evaluated more easily if only a single performance figure is used as measure. The organisers did not take part in the sub-challenges but provided baselines using the two most popular approaches: static and dynamic modelling. In the first approach, chunk-level feature vectors of constant length are obtained by applying functionals to the sequence of frame-level features. These chunk-level features are then classified with the WEKA toolkit (Witten and Frank, 2005) and Support Vector Machines. In the second approach, the frame-level features are classified directly using the implementation of hidden Markov models from the HTK toolkit (Young et al., 2006). Intentionally standard-tools were used in order to obtain reproducible results (cf. Section 2.5). The static approach outperformed slightly the dynamic one. The best baseline results were 67.7% UA for the 2-class problem and 38.2% UA for the 5-class problem. Compared to classification rates for emotion portrayals, these rates are rather low and demonstrate the difficulty of this challenge: classifying real-life emotions without the possibility to focus on prototypes only. As it turned out, exceeding the baseline results was possible, but certainly not an easy task.

3.3. Participants and their approaches

Although a non-English speech corpus was used in this challenge, many participants from all over the world were interested in the challenge, and 33 research groups registered to get access to the data. Seventeen groups were from Europe (5 from Germany), 8 from Middle and North America, 6 from Asia, and 2 from Australia. Seventeen research groups actually uploaded classification results on our server, and 14 groups finally submitted a paper to the INTERSPEECH 2009 Emotion Challenge Special Session. One additional paper was not submitted to the special session but to the general conference. Ten of these

Table 2
Participants of the three sub-challenges of the INTERSPEECH 2009 Emotion Challenge.

Open performance Sub-Challenge		Classifier Sub-Challenge		Feature Sub-Challenge		Number of participants
2 Classes	5 Classes	2 Classes	5 Classes	2 Classes	5 Classes	
✓	✓	–	–	–	–	7
✓	–	–	–	–	–	2
–	–	✓	✓	–	–	2
–	–	–	✓	–	–	1
–	–	–	✓	✓	✓	1
–	–	–	–	✓	✓	1

15 submissions could be accepted for publication in the conference proceedings of INTERSPEECH 2009.

All participants were encouraged to compete in multiple sub-challenges. However, most participants focused on only one sub-challenge (cf. Table 2). The most popular sub-challenge was the *Open Performance Sub-Challenge*, where 9 research groups took part. Seven of these 9 groups addressed both the 2-class and the 5-class problem. The other two groups addressed only the 2-class problem. Furthermore, 4 research groups took part in the *Classifier Sub-Challenge*. Two of them addressed both classification tasks, the other two addressed only the 5-class problem. One of these groups as well as one extra group submitted their own set of features to the *Feature Sub-Challenge*.

In the *Classifier Sub-Challenge*, Planet et al. (2009) evaluated different WEKA classifiers as well as a hierarchical waterfall classification scheme for the 5-class classification problem. However, the best result, which is significantly better than the baseline, was obtained with the simple Naïve Bayes classifier. Lee et al. (2009) also evaluated a hierarchical classification framework, where "the top level classification is performed on the *easiest* emotion recognition task". Bayesian Logistic Regression was used as binary classifier. They outperformed the other three participants of the *Classifier Sub-Challenge*, although the results were not significantly higher than those of Planet et al.

In the *Open Performance Sub-Challenge*, the participants evaluated their own set of features using their own classification techniques. The submitted contributions deliver insight into the performance of current state-of-the-art systems if they are used in real-life settings. Furthermore, these contributions exemplify the kind of features and classification techniques currently used and the focus of current activities in the area of emotion recognition from speech. In general, the feature sets can be roughly categorised into two groups: The first group are small sets of features that are carefully designed by experts based on their knowledge on emotion, speech production, etc. The second group are large sets of features that are obtained by calculating 'everything that is possible'. In these 'brute-force' approaches, various base contours on the frame-level such as F0, energy, HNR, MFCC, etc. are computed. Often, the number of base contours is multiplied by taking the first, second, and sometimes even the third derivative of these base contours. In order to obtain features on the chunk level, numerous functionals are applied to each base contour. By that, an impressive number of features (in general several thousands) can be generated including not only meaningful features but also many features that are not suited for distinguishing the different emotional states. Thus, a data-driven feature selection step is absolutely essential to reduce the number of features again. The features that were provided for the *Classifier Sub-Challenge* and that were used to produce the baseline results belong to the second group of brute-force features, though only a small number of base contours (16 contours plus their first derivatives) and functionals (12) has been selected, resulting in 384 standard acoustic features. This brute-force approach has become quite popular over the last years. Thus, it is surprising that only two participants of the *Open Performance Sub-Challenge* used such large feature sets.

Vogt and André (2009) used a set of 1451 features that was then reduced drastically to only 79 and 154 features for the 2-class problem and the 5-class problem, respectively. The classification result did not reach the baseline, mainly due to the focus on real-time processing.

Polzehl et al. (2009) also defined a set of about 1500 features. They were the only research group that also used linguistic features based on the output of an ASR system. Linguistic and acoustic features were classified separately and the decisions were then fused (late or decision-level fusion). The authors presented classification rates for the 2-class problem only. In terms of the unweighted average recall (UA), the achieved result was very close to the baseline. However, the reported accuracy was clearly higher than those of the baseline system indicating that there would have been space for improving the UA values – of course, at the expense of a lower weighted average recall.

All the other participants of the *Open Performance Sub-Challenge* used quite small feature sets. Surprisingly, most of them were mainly based on MFCC features, which are well-known from speech recognition and have proven to be very useful in emotion recognition as well, although their design goal was clearly to remove all the information about *how* something was produced and to keep only information about *what* was said. In the majority of cases, the MFCC features are complemented by a few prosodic features based on F0 and energy contours.

Vlasenko (2009) achieved good results by using only 13 MFCC features and their first and second derivative with a simple GMM classifier.

Kockmann et al. (2009) used the same type of features (13 MFCC features, Δ, ΔΔ) and evaluated different GMM training methods: standard Maximum Likelihood (ML) training, training of a Universal Background Model (UBM) and Maximum a Posteriori (MAP) adaptation to the emotion classes, discriminative GMM training based on Maximum Mutual Information (MMI), and GMM training based on Joint Factor Analysis (JFA). The latter technique yielded the best results for the 2- and the 5-class problem. Slight improvements are obtained for the 5-class problem by fusing the decisions of these four classifiers.

Bozkurt et al. (2009) also trained different GMM classifiers and fused their results. One classifier was based on MFCC features (13 MFCCs, Δ, ΔΔ), one on Line Spectral Features (LSF), which are closely related to formant frequencies, and one on a posteriori scores from a parallel multi-branch HMM that is trained on MFCC, F0 and energy features.

Dumouchel et al. (2009) used also decision-level fusion to fuse the results of three different GMM classifiers. The first classifier was a GMM that is trained on MFCC features (13 MFCC, Δ, ΔΔ) using ML training. The second classifier used the same set of features but employed the GMM supervector approach, where a UBM is adapted to the current utterance (or chunk as in the case of the FAU Aibo Emotion Corpus). The mean vectors of the adapted GMM then form the GMM supervector. This static feature vector was finally classified with Support Vector Machines (SVM). The third classifier was a GMM (UBM + MAP) that is trained on prosodic features modelling F0, energy, and the first two formants for pseudo-syllable units.

Luengo et al. (2009) used 18 Mel-scale short-term log-frequency power coefficients (LFPC) and their first and second derivative instead of the MFCC features. This first set of frame-level features was also classified with GMMs. From the second set of 56 chunk-level prosodic features, the best 10 features were selected and classified with SVM. The results of both classifiers were finally fused.

Barra-Chicote et al. (2009) used again MFCC features (complemented with the logarithm of the energy and F0) and their first and second derivative but fed these features into a Dynamic Bayesian Network classifier.

Although the participants of the *Open Performance Sub-Challenge* had the free choice of features, most of them focused strongly on MFCC features. With respect to prosodic features, most participants only added energy and the fundamental frequency to the set of base contours. Other prosodic features (for example, features based on durations of words or vowels) as well as other types of features such as voice quality features did not play any role in most cases. Furthermore, not only the features, but also the chosen classifiers differed only slightly in most cases in terms of classification techniques as well as in terms of resulting recognition rates. Consequently, the results of the participants of the Emotion Challenge are very close.

3.4. Winners of the INTERSPEECH 2009 Emotion Challenge

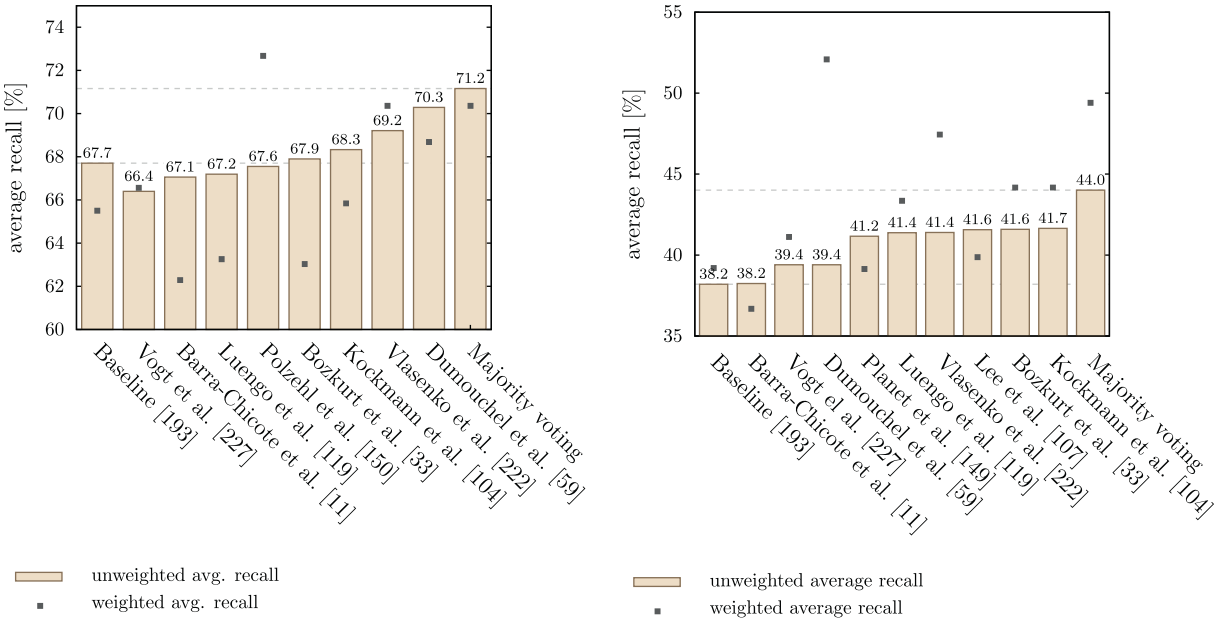
The Open Performance Sub-Challenge Prize was awarded to Pierre Dumouchel et al. (University de Quebec, Canada (Dumouchel et al., 2009)): They obtained the best result (70.29% UA recall) in the two-class task, significantly ahead of their eight competitors. The best result in the five-class task (41.65% UA recall) was achieved by Marcel Kockmann et al. (Brno University of Technology, Czech Republic, (Kockmann et al., 2009)) who surpassed six further results and were awarded the Best Special Session's Paper Prize as they had received the highest reviewers' score for their paper at the same time. The Classifier Sub-Challenge Prize was given to Chi-Chun Lee et al. (University of Southern California, USA (Lee et al., 2009)) for their best result in the five-class task in advance of three further participants. In the two-class task, the baseline was not exceeded by any of two participants.

Regrettably, no award could be given in the Feature Sub-Challenge. Neither of the feature sets provided by three participants in this sub-challenge exceeded the baseline feature set provided by the organisers. Overall, the results of all 17 participating sites were often very close to each other, and significant differences were as seldom as one might expect in such a close competition.

3.5. Fusion of the individual contributions

Fig. 3 summarises the results of the participants, who took part in either the *Classifier* or the *Open Performance Sub-Challenge* and whose papers were accepted for publication at INTERSPEECH 2009. Fig. 3(a) shows the results for the 2-class problem, Fig. 3(b) the ones for the 5-class problem. In both cases, the best baseline result from (Schuller et al., 2009d) is given as well. The bars indicate the primary measure – unweighted average recall UA – that had to be optimised. The dots show the corresponding accuracy values (weighted average recall). Fig. 3 also shows the best result obtained by majority voting of the best n participants.

Fig. 4 shows which absolute improvements over a given experiment are significantly better for the four levels of significance $\alpha = .050$, .010, .005, and .001. The null hypothesis H0 assumes that the accuracies of both experiments are identical. We apply a one-tailored significance test since we are interested in whether the second experiment is better than the first one. We assume that H0 is true and disprove it at various levels of significance. It depends on the accuracy of the first experiment which absolute improvements are necessary for the second one to be significantly better. Compared to the baseline of the 5-class problem (38.2%), accuracies $\geqslant$40.2% are significantly better at a significance level of $\alpha = .005$. Compared to the baseline of the 2-class



(a) 2-class classification problem (b) 5-class classification problem

Fig. 3. Results of the participants of the INTERSPEECH 2009 Emotion Challenge.

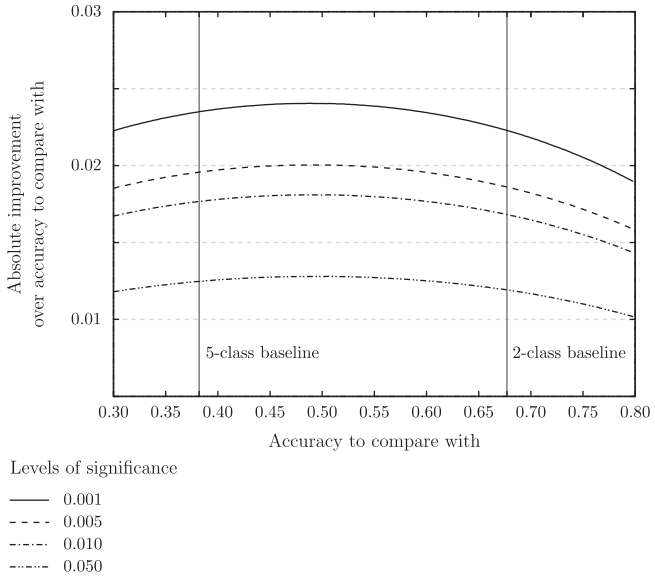


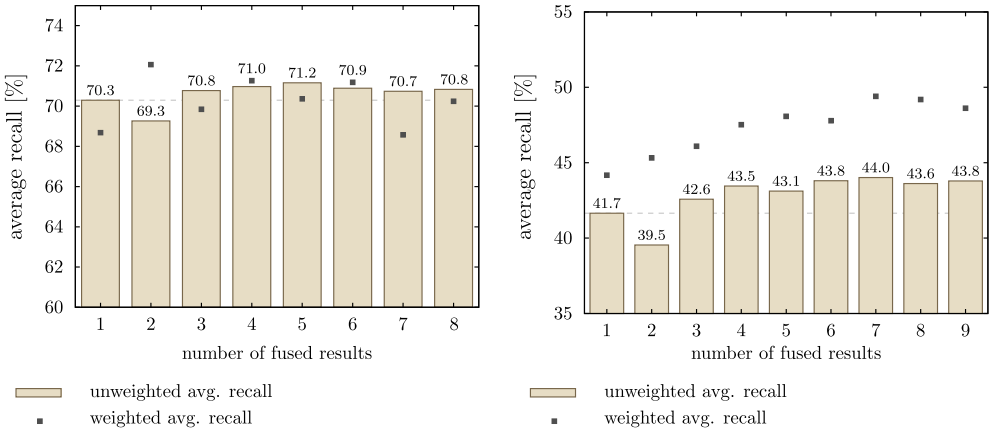
Fig. 4. Lines of significant absolute improvements for different levels of significance.

problem (67.6%), accuracies $\geqslant$69.5% are significantly better ($\alpha = .005$).

Majority voting has been chosen to fuse the single contributions since the participants only uploaded predictions for the test set. More sophisticated fusion techniques could not be used since their parameters must not be optimised on the test set. Some participants also uploaded a posteriori scores. However, better results were obtained if the majority voting was based on hard decisions. Fig. 5(a) shows the results of the majority vote for the 2-class problem if the best n ($1 \leqslant n \leqslant 8$) contributions were fused. The result of the best single contribution is outperformed if at least three contributions are fused. The best result is obtained for the fusion of $n = 5$ contributions. Along the same lines, Fig. 5(b) shows the results of the majority vote for the 5-class problem ($n \leqslant 1 \leqslant 9$). Again, better results than the best single contribution are obtained if at least $n = 3$ contributions are fused. For the 5-class problem, the best results were obtained by considering the best 7 contributions. Here, the improvement compared to the best single contribution is significant: The unweighted average recall of the majority voting is 44.0% compared to the result of Kockmann et al., who achieved 41.7% UA. For the 2-class problem, the majority voting resulted also in better – albeit not significantly better – recognition rates. Note: The evaluation of different numbers of contributions used in the majority vote is also some sort of optimisation on the test set. However, this optimisation is in agreement with the rules of the INTERSPEECH 2009 Emotion Challenge, where the participants could upload their results up to 25 times and finally submit their best system. It is pointed out again that the best single contribution is always outperformed if at least three contributions are fused.

3.6. Lessons learnt

The INTERSPEECH 2009 Emotion Challenge was well attended (17 participants) and finally 10 papers could be accepted for publication in the conference proceedings of INTERSPEECH 2009. The Emotion Challenge clearly benefited from the large popularity of a big international conference such as INTERSPEECH 2009. However, the



(a) 2-class classification problem (b) 5-class classification problem

Fig. 5. Combination of the results of the n best participants of the INTERSPEECH 2009 Emotion Challenge.

time constraints given by INTERSPEECH 2009 led to a very short period of time where interested research groups could set up and optimise their systems. As a consequence, many research groups participated only in one of the three sub-challenges. Furthermore, the people who are able to take part in such a challenge are those who are able to adapt their existing systems quickly to the new tasks of the challenge. Most likely, these groups are used to take part in challenges – even if the challenges are in a different area of research. Hence, it is not astonishing that many participants came from a speech recognition background and focused mainly on MFCC features, which are the standard features in speech recognition.

Yet, emotion recognition is a very interdisciplinary field of research and researchers from other disciplines (e.g., psychologists) are excluded if they are not able to process a new corpus automatically in such a short time. Thus, many interesting approaches and other types of features are probably not covered. An event as the Emotion Challenge is certainly suited to foster progress in emotion recognition, however, other ways should be considered as well in order to integrate people and knowledge from other disciplines. One of these possibilities can be along the lines of CEICES, an initiative within the European network of excellence HUMAINE, where the goal was to foster co-operation and not competition (for more details see (Batliner et al., 2011)).

The final fusion of the contributions of the single participants demonstrated that "together we are best". In general, this is not surprising, but in this challenge, many groups focused on the same type of features as well as on the same type of classifiers. Still, a significant relative improvement of 5.5% in terms of the unweighted average recall was obtained for the 5-class problem by a simple majority vote. By that, an unweighted average recall of 44.0% was reached. This result clearly demonstrates the difficulty of dealing with a real-life non-prototypical emotion recognition scenario – this challenge remains.

3.7. Release of the FAU Aibo Emotion Corpus

After the INTERSPEECH 2009 Emotion Challenge, the owners of the FAU Aibo Emotion Corpus – the Pattern Recognition Lab of the Friedrich-Alexander University Erlangen-Nuremberg (FAU) – released the data. The corpus can be obtained for a small service charge for sole scientific, non-commercial use. It includes the audio files, the transcription, and the original emotion labels of the five human labellers who annotated 11 different emotion categories on the word level. More details can be found on the web page http://www5.cs.fau.de/FAUAiboEmotionCorpus/.

4. Where to Go from Here

Obtaining more realistic data will still be the most important issue in the foreseeable future. This holds for basic categories such as main classes, and even more for sub-categories or mixed emotions: We simply do need enough data for modelling. New databases with acted data, even for minority languages, will not really help promoting the state-of-the-art. Establishing realistic databases, however, requires high efforts, and progress will therefore be not too fast.

Paradigms such as unsupervised learning may reduce the problem if engines are sufficiently robust to find emotional data by themselves. Another alternative might be the further exploitation or addition of synthesized speech data (Schuller and Burkhardt, 2010); yet, these two alternatives will most likely be confined to providing training data.

An important side topic is genericity of the data although it will not always go together with appropriateness within specific scenarios. Some of the topics pursued quite often such as searching for new features, trying out new and more sophisticated classification procedures, and more complex emotion representations, might only come second in relevance because as is, they too often clash with

the data requirements addressed above: It is convenient to use publicly available, acted databases – or the very few publicly available, realistic databases such as VAM or FAU Aibo (Grimm et al., 2008; Steidl, 2009) – for benchmarking. We do not know yet, however, whether and up to which extent the results can be transferred onto other (types of) data. Moreover, more complex representations inevitably lead to smaller number of cases representing one of the classes to be modelled. Comparisons of performances – be this of feature sets or of classifiers – within one and the same study are susceptible to the 'doubt of unequal edge conditions': We do not know whether a better performing feature set or classifier really is better or whether less effort has been spent on developing and tuning. Cross-corpora, cross-cultural, and multilingual experiments will further be needed to broaden the scope and to come closer to genericity. Commonalities of related fields as speaker or language identification and other speaker state and trait analysis as speaker role and health state or speaker age, gender or height can be further exploited for mutual exchange of methods – the results of the first challenge clearly showed that, e.g., borrowing from such methods as different GMM training techniques and introduction of a universal background model can pay off. For all these problems, challenges as the one described above can help. As we pointed out in Section 3.6, strict challenges, however, have to be supplemented by initiatives such as CEICES in order to really bridge the gap between subcultures such as ASR on the one hand, and basic phonetic/linguistic/psychological approaches on the other hand.

Another topic, largely untouched so far, is user and usability studies, especially tailored for specific applications (Batliner et al., 2006a). Marked by the lack of actual real-life application, no user studies were carried out in the beginning. Technology was driven without actual proof whether the user really liked a machine judging her feelings. While techniques such as Wizard-of-Oz could have easily been employed to simulate and evaluate the acceptance of social machine competence, it was mainly used to record data. Moreover, asking volunteering participants to evaluate the system not necessarily mirrors the acceptance of the same system if used 'in the wild'. Even today almost no such studies exist – apparently the community is highly convinced of the need for automatic emotion recognition. An example of such a studies can be found in (Burkhardt et al., 2009), where 200 paid test subjects calling a voice portal with anger feedback strategies: 20% noticed such a strategy of which 70% of these judged it as helpful. In (Schuller et al., 2009b) a further study with 40 participants is reported that interacted with a virtual product and company tour that took participants' interest into account to change topic in case of boredom. Three variants were used: topic change after a fixed time, with fully automatic interest recognition or by a human Wizard-of-Oz. The question "Did you think the system was taking into account your interest?" was positively answered by 35% in the first case (no interest recognition), by 63% in the second case (fully automatic interest recognition) and by 84% in the last case (human interest recognition) nicely demonstrating that the technology seems to be generally working, but that there is also still headroom for improvement to reach human-alike performance.

In the beginning, performance simply was not sufficient for first real-world applications though the goals were often set as high as language independent affect recognition in brand slogans (Cheng et al., 2006). Not surprisingly, the first were found in the entertainment area: lie and stress level detectors for the mobile phone, or a software intended to detect the level of intimacy of a caller. It lies in the nature of industrial production that no details on implementation, features, classifiers, training material, or true performances are known for these. Recent use-cases contain first test runs in call-centres to switch to human operators in case of annoyed customers (T-Labs) or server-based sending of intimacy information via SMS (Chinese Academy of Science) among first video games (Row the boat). All these pilot applications are, however, still rather unnoticed by the public.

Coming back to more general questions, a related topic is whether for specific applications, a dimensional representation is more adequate, or a sub-division into a fine-graded category system, or whether we can do most successfully with a coarse dichotomy into, for instance, negative/positive vs. neutral. These questions relate to topics that can be addressed in more basic research: Should we try and model human processing (in the case of emotions conveyed via speech, from peripheral auditory processing to higher cortical processes), or will it rather be sufficient to model the more categorical outcome of this processing? Normally, we remember the core semantics or illocution of an utterance, and not, whether the verb form was passive or active (Fillenbaum, 1966; Sachs, 1967); the same might hold for emotions.

Acknowledgement

This work was partly funded by the European Union under Grant Agreement No. 211486 (FP7/2007-2013, SEMAINE), IST-2001-37599 (PF-STAR), IST-2002-50742 (HUMAINE), and RTN-CT-2006-035561 (S2S). The authors would further like to thank the sponsors of the challenge, the HUMAINE Association and Deutsche Telekom Laboratories. The responsibility lies with the authors.

References

Ai, H., Litman, D., Forbes-Riley, K., Rotaru, M., Tetreault, J., Purandare, A., 2006. Using system and user performance features to improve emotion detection in spoken tutoring dialogs. In: Proc. Interspeech, Pittsburgh, PA, USA, pp. 797–800.

Al-Hames, M., Rigoll, G., 2006. Reduced complexity and scaling for asynchronous HMMs in a bimodal input fusion application. In: Proc. ICASSP, Toulouse, France, pp. 757–760.

Altun, H., Polata, G., 2009. Boosting selection of speech related features to improve performance of multi-class SVMs in emotion detection. Expert Systems Appl. 36 (4), 8197–8203.

Ang, J., Dhillon, R., Shriberg, E., Stolcke, A., 2002. Prosody-based automatic detection of annoyance and frustration in human–computer dialog. In: Proc. Interspeech, Denver, CO, USA, pp. 2037–2040.

Armstrong, J., 2007. Significance tests harm progress in forecasting. Internat. J. Forecast. 23, 321–327.

Arunachalam, S., Gould, D., Anderson, E., Byrd, D., Narayanan, S., 2001. Politeness and frustration language in child–machine interactions. In: Proc. Eurospeech, Aalborg, Denmark, pp. 2675–2678.

Atal, B., Hanauer, S.L., 1971. Speech analysis and synthesis by linear prediction of the speech wave. J. Acoust. Soc. Amer. 50, 637–655.

Athanaselis, T., Bakamidis, S., Dologlu, I., Cowie, R., Douglas-Cowie, E., Cox, C., 2005. ASR for emotional speech: clarifying the issues and enhancing performance. Neural Networks 18, 437–444.

Ayadi, M.M.H.E., Kamel, M.S., Karray, F., 2007. Speech emotion recognition using gaussian mixture vector autoregressive models. In: Proc. ICASSP, Honolulu, HY, pp. 957–960.

Baggia, P., Burnett, D.C., Carter, J., Dahl, D.A., McCobb, G., Raggett, D., 2007. EMMA: Extensible MultiModal Annotation markup language.

Barra-Chicote, R., Fernandez, F., Lutfi, S., Lucas-Cuesta, J.M., Macias-Guarasa, J., Montero, J.M., San-Segundo, R., Pardo, J.M., 2009. Acoustic emotion recognition using dynamic bayesian networks and multi-space distributions. In: Proc. Interspeech, Brighton, pp. 336–339.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., Aharonson, V., 2007. The impact of F0 extraction errors on the classification of prominence and emotion. In: Proc. ICPhS, Saarbrücken, Germany, pp. 2201–2204.

Batliner, A., Fischer, K., Huber, R., Spilker, J., Nöth, E., 2000. Desperately seeking emotions: actors, wizards, and human beings. In: Proc. ISCA Workshop on Speech and Emotion, Newcastle, Northern Ireland, pp. 195–200.

Batliner, A., Buckow, J., Huber, R., Warnke, V., Nöth, E., Niemann, H., 2001. Boiling down prosody for the classification of boundaries and accents in German and English. In: Proc. Eurospeech, Aalborg, Denmark, pp. 2781–2784.

Batliner, A., Fischer, K., Huber, R., Spilker, J., Nöth, E., 2003a. How to find trouble in communication. Speech Comm. 40, 117–143.

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

Batliner, A., Hacker, C., Steidl, S., Nöth, E., Haas, J., 2004. From emotion to interaction: lessons from real human–machine-dialogues. In: Proc. Tutorial and Research Workshop on Affective Dialogue Systems, Kloster Irsee, Germany, pp. 1–12.

Batliner, A., Steidl, S., Hacker, C., Nöth, E., Niemann, H., 2005. Tales of tuning – prototyping for automatic classification of emotional user states. In: Proc. Interspeech, Lisbon, Portugal, pp. 489–492.

Batliner, A., Burkhardt, F., van Ballegooy, M., Nöth, E., 2006a. A taxonomy of applications that utilize emotional awareness. In: Proc. IS-LTC 2006, Ljubliana, Slovenia, pp. 246–250.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Laskowski, K., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., Aharonson, V., 2006b. Combining efforts for improving automatic classification of emotional user states. In: Proc. IS-LTC 2006, Ljubliana, Slovenia, pp. 240–245.

Batliner, A., Steidl, S., Nöth, E., 2007a. Laryngealizations and emotions: How many Babushkas? In: Proc. Internat. Workshop on Paralinguistic Speech – between Models and Data (ParaLing'07), Saarbrücken, Germany, pp. 17–22.

Batliner, A., Schuller, B., Schaeffler, S., Steidl, S., 2008a. Mothers, adults, children, pets — towards the acoustics of intimacy. In: Proc. ICASSP 2008, Las Vegas, NV, pp. 4497–4500.

Batliner, A., Steidl, S., Hacker, C., Nöth, E., 2008b. Private emotions vs. social interaction — a data-driven approach towards analysing emotions in speech. User Model. User-Adapted Interact. 18, 175–206.

Batliner, A., Seppi, D., Steidl, S., Schuller, B., 2010. Segmenting into adequate units for automatic recognition of emotion-related episodes: a speech-based approach. Advances in Human–Computer Interaction, Vol. 2010. Article ID 782802, 15 pages.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Aharonson, V., Amir, N., 2011. Whodunnit – searching for the most important feature types signalling emotional user states in speech. Comput. Speech Lang. 25, 4–28.

Bellman, R., 1961. Adaptive Control Processes. Princeton University Press.

Bengio, S., 2003. An asynchronous hidden markov model for audio-visual speech recognition. Advances in NIPS 15.

Bengio, Y., Frasconi, P., 1995. An input output HMM architecture. Adv. Neural Inform. Process. Systems 7, 427–434.

Boda, P.P., 2004. Multimodal integration in a wider sense. In: Proc. COLING 2004 Satellite Workshop on Robust and Adaptive Information Processing for Mobile Speech Interfaces, Geneva, Switzerland, pp. 22–30.

Boersma, P., 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. Proc. Inst. Phonetic Sci. (Univ. Amsterdam) 17, 97–110.

Boersma, P., Weenink, D., 2005. Praat: doing phonetics by computer (version 4.3.14). <http://www.praat.org/>.

Bogert, B., Healy, M., Tukey, J., 1963. The quefrency analysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. In: Rosenblatt, M. (Ed.), Symposium on Time Series Analysis. John Wiley & Sons, New York, pp. 209–243.

Bozkurt, E., Erzin, E., Erdem, Ç.E., Erdem, A.T., 2009. Improving automatic emotion recognition from speech signals. In: Proc. Interspeech, Brighton, pp. 324–327.

Breese, J., Ball, G., 1998. Modeling emotional state and personality for conversational agents. Technical Report MS-TR-98-41, Microsoft.

Brendel, M., Zaccarelli, R., Schuller, B., Devillers, L., 2010. Towards measuring similarity between emotional corpora. In: Proc. 3rd ELR0A Internat. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect, Valetta, pp. 58–64.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B., 2005. A database of german emotional speech. In: Proc. Interspeech, Lisbon, Portugal, pp. 1517–1520.

Burkhardt, F., van Ballegooy, M., Engelbrecht, K.-P., Polzehl, T., Stegmann, J., 2009. Emotion detection in dialog systems: applications, strategies and challenges. In: Proc. ACII, Amsterdam, Netherlands, pp. 1–6.

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S., 2004. Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Proc. ICMI '04: Proc. 6th Internat. Conf. on Multimodal interfaces, New York, USA, pp. 205–211.

Campbell, N., Kashioka, H., Ohara, R., 2005. No laughing matter. In: Proc. Interspeech, Lisbon, Portugal, pp. 465–468.

Chen, L.S., Tao, H., Huang, T.S., Miyasato, T., Nakatsu, R., 1998. Emotion recognition from audiovisual information. In: Proc. IEEE Workshop on Multimedia Signal Processing, pp. 83–88.

Cheng, Y.-M., Kuo, Y.-S., Yeh, J.-H., Chen, Y.-T., Chien, C., 2006. Using recognition of emotions in speech to better understand brand slogans. In: Proc. IEEE 8th Workshop on Multimedia Signal Processing, Victoria, BC, pp. 238–242.

Cheveigne, A.D., Kawahara, H., 2002. Yin: a fundamental frequency estimator for speech and music. J. Acoust. Soc. Amer. 111 (4), 1917–1930.

Chuang, Z.-J., Wu, C.-H., 2004. Emotion recognition using acoustic features and textual content. In: Proc. ICME, Taipei, Taiwan, pp. 53–56.

Cohen, J., 1988. Statistical Power Analysis for the Behavioural Sciences, second ed. Erlbaum, Hillsdale, NJ.

Cowie, R., Douglas-cowie, E., Apolloni, B., Taylor, J., Romano, A., Fellenz, W., 1999. What a neural net needs to know about emotion words. In: Mastorakis, N. (Ed.), Computational Intelligence and Applications. Word Scientific Engineering Society, Society Press, pp. 109–114.

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

Cowie, R., Douglas-Cowie, E., Cox, C., 2005. Beyond emotion archetypes: databases for emotion modelling using neural networks. Neural Networks 18, 3388.

Cowie, R., Sussman, N., Ben-Ze'ev, A., 2010. Emotions: concepts and definitions. In: Cowie, R., Petta, P., Pelachaud, C. (Eds.), Emotion-Oriented Systems: The HUMAINE Handbook. Springer.

Daubechies, I., 1990. The wavelet transform, time–frequency localization and signal analysis. TransIT 36 (5), 961–1005.

Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. 29, 917–919.

de Gelder, B., Vroomen, J., 2000. The perception of emotions by ear and eye. Cognition Emotion 14 (3), 289–311.

de Gelder, B., Böcker, K.B.E., Tuomainen, J., Hensen, M., Vroomen, J., 1999. The combined perception of emotion from voice and face: early interaction revealed by human electric brain responses. Neurosci. Lett. 260 (2), 133–136.

Dellaert, F., Polzin, T., Waibel, A., 1996. Recognizing emotion in speech. In: Proc. ICSLP, Philadelphia, PA, USA, pp. 1970–1973.

Devillers, L., Vasilescu, I., Lamel, L., 2003. Emotion detection in task-oriented spoken dialogs. In: Proc. ICME 2003, IEEE, Multimedia Human–Machine Interface and Interaction, Baltimore, MD, USA, pp. 549–552.

Devillers, L., Vidrascu, L., 2007. Real-life emotion recognition in speech. In: Müller, C. (Ed.), Speaker Classification II, . In: Lecture Notes in Computer Science, Vol. 4441/2007. Springer, Berlin/Heidelberg, pp. 34–42.

Devillers, L., Abrilian, S., Martin, J.-C., 2005a. Representing real-life emotions in audiovisual data with non basic emotional patterns and context features. In: Proc. ACII, Beijing, China, pp. 519–526.

Devillers, L., Vidrascu, L., Lamel, L., 2005b. Challenges in real-life emotion annotation and machine learning based detection. Neural Networks 18, 407–422.

Ding, H., Qian, B., Li, Y., Tang, Z., 2006. A method combining lpc-based cepstrum and harmonic product spectrum for pitch detection. In: Proc. 2006 Internat. Conf. on Intelligent Information Hiding and Multimedia, IEEE, pp. 537–540.

Dumouchel, P., Dehak, N., Attabi, Y., Dehak, R., Boufaden, N., 2009. Cepstral and long-term features for emotion recognition. In: Proc. Interspeech, Brighton, pp. 344–347.

Elliott, C., 1992. The affective reasoner: a process model of emotions in a multi-agent system. Ph.D. Thesis, Dissertation, Northwestern University.

Engberg, I.S., Hansen, A.V., Andersen, O., Dalsgaard, P., 1997. Design, recording and verification of a Danish emotional speech database. In: Proc. Eurospeech, Rhodes, Greece, pp. 1695–1698.

Erickson, D., Yoshida, K., Menezes, C., Fujino, A., Mochida, T., Shibuya, Y., 2004. Exploratory study of some acoustic and articulatory characteristics of sad speech. Phonetica 63, 1–25.

Eyben, F., Wöllmer, M., Schuller, B., 2009. openEAR – Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. In: Proc. ACII, Amsterdam, Netherlands, pp. 576–581.

Eyben, F., Batliner, A., Schuller, B., Seppi, D., Steidl, S., 2010a. Cross-corpus classification of realistic emotions some pilot experiments. In: Proc. 3rd Internat. Workshop on EMOTION (Satellite of LREC): Corpora for Research on Emotion and Affect, Valetta, pp. 77–82.

Eyben, F., Wöllmer, M., Graves, A., Schuller, B., Douglas-Cowie, E., Cowie, R., 2010b. On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues. J. Multimodal User Interfaces 3 (1–2), 7–12, Special Issue on "Real-Time Affect Analysis and Interpretation: Closing the Affective Loop in Virtual Agents and Robots".

Eyben, F., Wöllmer, M., Schuller, B., 2010c. openSMILE – the munich versatile and fast open-source audio feature Extractor. In: Proc. ACM Multimedia, Florence, Italy, pp. 1459–1462.

Eysenck, H., 1960. The concept of statistical significance and the controversy about one-tailed tests. Psychol. Rev. 67, 269–271.

Fattah, S.A., Zhu, W.P., Ahmad, M.O., 2008. A cepstral domain algorithm for formant frequency estimation from noise-corrupted speech. In: Internat. Conf. on Neural Networks and Signal Processing 2008, Zhenjiang, China, pp. 114–119.

Fehr, B., Russel, J.A., 1984. Concept of emotion viewed from a prototype perspective. J. Exp. Psychol.: Gen. 113, 464–486.

Fei, Z., Huang, X., Wu, L., 2006. Mining the relation between sentiment expression and target using dependency of words. In: Proc. 20th Pacific Asia Conf. on Language, Information and Computation (PACLIC20), Wuhan, China, pp. 257–264.

Ferguson, C.J., 2009. An effect size primer: a guide for clinicians and researchers. Prof. Psychol.: Res. Practice 40, 532–538.

Fernandez, R., Picard, R.W., 2003. Modeling drivers' speech under stress. Speech Comm. 40, 145–159.

Fillenbaum, S., 1966. Memory for gist: some relevant variables. Lang. Speech 9, 217–227.

Fiscus, J., 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In: Proc. ASRU, Santa Barbara, CA, USA, pp. 347–352.

Fleiss, J., Cohen, J., Everitt, B., 1969. Large sample standard errors of kappa and weighted kappa. Psychol. Bull. 72 (5), 323–327.

Forbes-Riley, K., Litman, D., 2004. Predicting emotion in spoken dialogue fromm multiple knowledge sources. In: Proc. Human Language Technology Conf. of the North American Chap. of the Assoc. for Computational Linguistics, Boston, MA, USA, no pagination.

Frick, R., 1985. Communicating emotion: the role of prosodic features. Psychol. Bull. 97, 412–429.

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition. Academic Press.

Gaussier, E., Goutte, C., 2005. Relation between PLSA and NMF and implications. In: Proc. 28th Internat. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR-05), Salvador, Brazil, pp. 601–602.

Gigerenzer, G., 2004. Mindless statistics. J. Socio-Econ. 33, 587–606.

Godbole, N., Srinivasaiah, M., Skiena, S., 2007. Large-scale sentiment analysis for news and blogs. In: Proc. Internat. Conf. on Weblogs and Social Media (ICWSM), Boulder, CO, no pagination.

Goertzel, B., Silverman, K., Hartley, C., Bugaj, S., Ross, M., 2000. The baby webmind project. In: Proc. Annual Conf. of The Society for the Study of Artificial Intelligence and the Simulation of Behaviour (AISB), no pagination.

Grimm, M., Kroschel, K., Harris, H., Nass, C., Schuller, B., Rigoll, G., Moosmayr, T., 2007. On the necessity and feasibility of detecting a driver's emotional state while driving. In: Paiva, A., Prada, R., Picard, R.W. (Eds.), Affective Computing and Intelligent Interaction. Springer, Berlin–Heidelberg, pp. 126–138.

Grimm, M., Kroschel, K., Narayanan, S., 2008. The Vera am Mittag German audio–visual emotional speech database. In: Proc. IEEE Internat. Conf. on Multimedia and Expo (ICME), Hannover, Germany, pp. 865–868.

Gunes, H., Piccardi, M., 2005. Affect recognition from face and body: early fusion vs. late fusion. In: IEEE Internat. Conf. Systems, Man and Cybernetics, Vol. 4, pp. 3437–3443.

Hall, M.A., 1998. Correlation-based feature selection for machine learning. Ph.D. Thesis, Hamilton, NZ: Waikato University, Department of Computer Science.

Hansen, J., Bou-Ghazale, S., 1997. Getting started with susas: a speech under simulated and actual stress database. In: Proc. EUROSPEECH-97, Vol. 4, Rhodes, Greece, pp. 1743–1746.

Harnad, S., 1987. Categorical Perception. In: The Groundwork of Cognition. University Press, Cambridge.

Hermansky, H., 1990. Perceptual linear predictive (plp) analysis for speech. J. Acoust. Soc. Amer. (JASA) 87, 1738–1752.

Hess, W., Batliner, A., Kießling, A., Kompe, R., Nöth, E., Petzold, A., Reyelt, M., Strom, V., 1996. Prosodic modules for speech recognition and understanding in verbmobil. In: Sagisaka, Y., Campell, N., Higuchi, N. (Eds.), Computing Prosody. Approaches to a Computational Analysis and Modelling of the Prosody of Spontaneous Speech. Springer-Verlag, New York, pp. 363–383.

Hirschberg, J., Liscombe, J., Venditti, J., 2003. Experiments in emotional speech. In: Proc. ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo, Japan, pp. 1–7.

Hui, L., Dai, B.-Q., Wei, L., 2006. A pitch detection algorithm based on amdf and acf. In: Proc. ICASSP, Toulouse, France, p. I.

Hyvärinen, A., Karhunen, J., Oja, E., 2001. Independent Component Analysis. Wiley & Sons, New York.

Inanoglu, Z., Caneel, R., 2005. Emotive alert: HMM-based emotion detection in voicemail messages. In: Proc. 10th Internat. Conf. on Intelligent User Interfaces, San Diego, CA, USA, pp. 251–253.

Joachims, T., 1998. Text categorization with support vector machines: learning with many relevant features. In: Nédellec, C., Rouveirol, C. (Eds.), Proc. ECML-98, 10th European Conf. on Machine Learning. Springer, Heidelberg, Chemnitz, Germany, pp. 137–142.

Johnstone, T., Scherer, K.R., 2000. Vocal communication of emotion. In: Lewis, M., Haviland-Jones, J.M. (Eds.), Handbook of Emotions. Guilford Press, New York, London, pp. 220–235, second ed. (Chapter 14).

Jolliffe, I.T., 2002. Principal Component Analysis. Springer, Berlin, Germany.

Kharat, G.U., Dudul, S.V., 2008. Human emotion recognition system using optimally designed SVM with different facial feature extraction techniques. WSEAS Trans. Comput. 7 (6).

Kießling, A., 1997. Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung. Berichte aus der Informatik. Shaker, Aachen, Germany.

Kim, S.-M., Hovy, E., 2005. Automatic detection of opinion bearing words and sentences. In: Companion Volume to the Proc. Internat. Joint Conf. on Natural Language Processing (IJCNLP), Jeju Island, Korea, pp. 61–66.

Kim, K.H., Bang, S.W., Kim, S.R., 2004. Emotion recognition system using short-term monitoring of physiological signals. Medical Biological Eng. Comput. 42 (3), 419–427.

Kim, J., André, E., Rehm, M., Vogt, T., Wagner, J., 2005. Integrating information from speech and physiological signals to achieve emotional sensitivity. In: Proc. Interspeech, Lisbon, Portugal, pp. 809–812.

Kim, E., Hyun, K., Kim, S., Kwak, Y., 2007. Emotion interactive robot focus on speaker independently emotion recognition. In: Proc. IEEE/ASME Internat. Conf. on Advanced Intelligent Mechatronics, Zurich, Switzerland, pp. 1–6.

Kockmann, M., Burget, L., Černocký, J., 2009. Brno University of Technology System for Interspeech 2009 Emotion Challenge. In: Proc. Interspeech, Brighton, pp. 348–351.

Kwon, O.-W., Chan, K., Hao, J., Lee, T.-W., 2003. Emotion recognition by speech signals. In: Proc. Interspeech, pp. 125–128.

Laskowski, K., 2009. Contrasting emotion-bearing laughter types in multiparticipant vocal activity detection for meetings. In: Proc. ICASSP, IEEE, Taipei, Taiwan, pp. 4765–4768.

Lee, C.M., Narayanan, S.S., 2005. Toward detecting emotions in spoken dialogs. IEEE Trans. Speech Audio Process. 13 (2), 293–303.

Lee, D.D., Seung, H.S., 1999. Learning the parts of objects by non-negative matrix factorization. Nature 401 (6755), 788–791.

Lee, C.M., Narayanan, S.S., Pieraccini, R., 2002. Combining acoustic and language information for emotion recognition. In: Proc. Interspeech, Denver, CO, USA, pp. 873–376.

Lee, C., Mower, E., Busso, C., Lee, S., Narayanan, S., 2009. Emotion recognition using a hierarchical binary decision tree approach. In: Proc. Interspeech, Brighton, pp. 320–323.

Lefter, I., Wiggers, P., Rothkrantz, L., 2010. EmoReSp: an online emotion recognizer based on speech. In: Proc. 11th Internat. Conf. on Computer Systems and Technologies (CompSysTech), Sofia, Bulgaria, pp. 287–292.

Liscombe, J., Hirschberg, J., Venditti, J., 2005. Detecting certainness in spoken tutorial dialogues. In: Proc. INTERSPEECH, Lisbon, Portugal, pp. 1837–1840.

Liscombe, J., Riccardi, G., Hakkani-Tür, D., 2005b. Using context to improve emotion detection in spoken dialog systems. In: Proc. Interspeech, Lisbon, Portugal, pp. 1845–1848.

Liscombe, J., Venditti, J., Hirschberg, J., 2003. Classifying subject ratings of emotional speech using acoustic features. In: Proc. Eurospeech, Geneva, Switzerland, pp. 725–728.

Litman, D., Forbes, K., 2003. Recognizing emotions from student speech in tutoring dialogues. In: Proc. ASRU, Virgin Island, USA, pp. 25–30.

Liu, H., Liebermann, H., Selker, T., 2003. A model of textual affect sensing using real-world knowledge. In: Proc. 7th Internat. Conf. on Intelligent User Interfaces (IUI 2003), pp. 125–132.

Lizhong, W., Oviatt, S., Cohen, P.R., 1999. Multimodal integration – a statistical view. IEEE Trans. Multimedia 1, 334–341.

Lovins, J.B., 1968. Development of a stemming algorithm. Mech. Transl. Comput. Linguist. 11, 22–31.

Luengo, I., Navas, E., Hernáez, I., 2009. Combining spectral and prosodic information for emotion recognition in the interspeech 2009 emotion challenge. In: Proc. Interspeech, Brighton, pp. 332–335.

Lugger, M., Yang, B., 2008. Psychological motivated multi-stage emotion classification exploiting voice quality features. In: Mihelic, F., Zibert, J. (Eds.), Speech Recognition, IN-TECH, p. 1.

Lugger, M., Yang, B., Wokurek, W., 2006. Robust estimation of voice quality parameters under real world disturbances. In: Proc. ICASSP, Toulouse, France, pp. 1097–1100.

Makhoul, J., 1975. Linear prediction: a tutorial review. Proc. IEEE 63, 561–580.

Martin, J.-C., Niewiadomski, R., Devillers, L., Buisine, S., Pelachaud, C., 2006. Multimodal complex emotions: gesture expressivity and blended facial expressions. Int. J. Human. Robot. 3 (3), 1–23.

Martinez, C.A., Cruz, A., 2005. Emotion recognition in non-structured utterances for human–robot interaction. In: IEEE Internat. Workshop on Robot and Human Interactive Communication, Nashville, TN, USA, pp. 19–23.

Matos, S., Birring, S., Pavord, I., Evans, D., 2006. Detection of cough signals in continuous audio recordings using hidden markov models. IEEE Trans. Biomed. Eng., 1078–1108.

McGilloway, S., Cowie, R., Doulas-Cowie, E., Gielen, S., Westerdijk, M., Stroeve, S., 2000. Approaching automatic recognition of emotion from voice: A rough benchmark. In: Proc. ISCA Workshop on Speech and Emotion, Newcastle, Northern Ireland, pp. 207–212.

Meyer, D., Leisch, F., Hornik, K., 2002. Benchmarking support vector machines. Report Series No. 78, Adaptive Informations Systems and Management in Economics and Management Science, 19 pages.

Missen, M., Boughanem, M., 2009. Using WordNet's semantic relations for opinion detection in blogs. In: Advances in Information Retrieval, Lecture Notes in Computer Science, Vol. 5478/2009. Springer, pp. 729–733.

Morrison, D., Silva, L.C.D., 2007. Voting ensembles for spoken affect classification. J. Network Comput. Appl. 30, 1356–1365.

Morrison, D., Wang, R., Silva, L.C.D., 2007a. Ensemble methods for spoken emotion recognition in call-centres. Speech Comm. 49 (2), 98–112.

Morrison, D., Wang, R., Xu, W., Silva, L.C.D., 2007b. Incremental learning for spoken affect classification and its application in call-centres. Int. J. Intell. Systems Technol. Appl. 2, 242–254.

Mower, E., Metallinou, A., Lee, C.-C., Kazemzadeh, A., Busso, C., Lee, S., Narayanan, S., 2009. Interpreting ambiguous emotional expressions. In: Proc. ACII, Amsterdam, Netherlands, pp. 662–669.

Nasoz, F., Alvarez, K., Lisetti, C.L., Finkelstein, N., 2004. Emotion recognition from physiological signals using wireless sensors for presence technologies. Cognition Technol. Work 6 (1), 4–14.

Nefian, A.V., Luhong, L., Xiaobo, P., Liu, X., Mao, C., Murphy, K., 2002. A coupled HMM for audio–visual speech recognition. In: Proc. ICASSP, Orlando, FL, USA, pp. 2013–2016.

Neiberg, D., Elenius, K., Laskowski, K., 2006. Emotion recognition in spontaneous speech using GMMs. In: Proc. Interspeech, Pittsburgh, PA, USA, pp. 809–812.

Nickerson, R.S., 2000. Null hypothesis significance testing: a review of an old and continuing controversy. Psychol. Methods 5, 241–301.

Nogueiras, A., Moreno, A., Bonafonte, A., Mariño, J.B., 2001. Speech emotion recognition using hidden markov models. In: Proc. Eurospeech, Aalborg, Denmark, pp. 2267–2270.

Noll, A.M., 1967. Cepstrum pitch determination. J. Acoust. Soc. Amer. (JASA) 14, 293–309.

Nose, T., Kato, Y., Kobayashi, T., 2007. Style estimation of speech based on multiple regression hidden semi-markov model. In: Proc. Interspeech, Antwerp, Belgium, pp. 2285–2288.

Nöth, E., Batliner, A., Warnke, V., Haas, J., Boros, M., Buckow, J., Huber, R., Gallwitz, F., Nutt, M., Niemann, H., 2002. On the use of prosody in automatic dialogue understanding. Speech Comm. 36 (1–2), 45–62.

Nwe, T., Foo, S., Silva, L.D., 2003. Speech emotion recognition using hidden markov models. Speech Comm. 41, 603–623.

Pachet, F., Roy, P., 2009. Analytical features: a knowledge-based approach to audio feature generation. EURASIP J. Audio Speech Music Process., 23 pages.

Pal, P., Iyer, A., Yantorno, R., 2006. Emotion detection from infant facial expressions and cries. In: Proc. ICASSP, Toulouse, France, pp. 809–812.

Pang, B., Lee, L., Vaithyanathan, S., 2002. Thumbs up? sentiment classification using machine learning techniques. In: Proc. 2002 Conf. on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, PA, USA, pp. 79–86.

Pantic, M., Rothkrantz, L., 2003. Toward an affect-sensitive multimodal human–computer interaction. Proc. IEEE 91 (9), 1370–1390.

Pernegger, T.V., 1998. What's wrong with Bonferroni adjustment. Brit. Med. J. 316, 1236–1238.

Petrushin, V., 1999. Emotion in speech: recognition and application to call centers. In: Proc. Artificial Neural Networks in Engineering (ANNIE '99), St. Louis, MO, USA, pp. 7–10.

Picard, R., Vyzas, E., Healey, J., 2001. Toward machine emotional intelligence: Analysis of affective physiological state. IEEE Trans. Pattern Anal. Machine Intell. 23 (10), 1175–1191.

Planet, S., Iriondo, I., Socoró, J.-C., Monzo, C., Adell, J., 2009. GTM-URL contribution to the INTERSPEECH 2009 Emotion Challenge. In: Proc. Interspeech, Brighton, pp. 316–319.

Polzehl, T., Sundaram, S., Ketabdar, H., Wagner, M., Metze, F., 2009. Emotion classification in children's speech using fusion of acoustic and linguistic features. In: Proc. Interspeech, Brighton, pp. 340–343.

Polzin, T.S., Waibel, A., 2000. Emotion-sensitive human–computer interfaces. In: Proc. ISCA Workshop on Speech and Emotion, Newcastle, Northern Ireland, pp. 201–206.

Popescu, A.-M., Etzioni, O., 2005. Extracting product features and opinions from reviews. In: Proc. Human Language Technology Conf. and the Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP), Vancouver, British Columbia, pp. 339–346.

Porter, M., 1980. An algorithm for suffix stripping. Program 14 (3), 130–137.

Pudil, P., Novovicova, J., Kittler, J., 1994. Floating search methods in feature selection. Pattern Recognition Lett. 15, 1119–1125.

Rabiner, L.R., 1977. On the use of autocorrelation analysis for pitch detection. IEEE Trans. Acoust. Speech Signal Process. 25, 24–33.

Rahurkar, M.A., Hansen, J.H.L., 2003. Towards affect recognition: an ICA approach. In: Proc. 4th Internat. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003), Nara, Japan, pp. 1017–1022.

Rong, J., Chen, Y.-P.P., Chowdhury, M., Li, G., 2007. Acoustic features extraction for emotion recognition. In: Proc. ACIS Internat. Conf. on Computer and Information Science. IEEE Computer Society, Los Alamitos, CA, pp. 419–424.

Rosch, E., 1975. Cognitive representations of semantic categories. J. Exp. Psychol.: Gen. 104 (3), 192–233.

Rosenberg, A., Binkowski, E., 2004. Augmenting the kappa statistic to determine interannotator reliability for multiply labeled data points. In: Dumais, D.M., Roukos, S. (Eds.), HLT-NAACL 2004: Short Papers. Association for Computational Linguistics, Boston, MA, USA, pp. 77–80.

Rozeboom, W., 1960. The fallacy of the null-hypothesis significance test. Psychol. Bull. 57, 416–428.

Russell, J., Bachorowski, J., Fernandez-Dols, J., 2003. Facial and vocal expressions of emotion. Annu. Rev. Psychol., 329–349.

Sachs, J., 1967. Recognition memory for syntactic and semantic aspects od connected discourse. Percept. Psychophys. 2, 437–442.

Said, C., Moore, C., Norman, K., Haxby, J., Todorov, A., 2010. Graded representations of emotional expressions in the left superior temporal sulcus. Front. Systems Neurosci. 4, 6, doi:10.3389/fnsys.2010.00006.

Salzberg, S., 1997. On comparing classifiers: pitfalls to avoid and a recommended approach. Data Mining Knowl. Discov. 1 (3), 317–328.

Sato, N., Obuchi, Y., 2007. Emotion recognition using mel-frequency cepstral coefficients. Inform. Media Technol. 2 (3), 835–848.

Scherer, K.R., Johnstone, T., Klasmeyer, G., 2003. Vocal expression of emotion. In: Davidson, R.J., Scherer, K.R., Goldsmith, H.H. (Eds.), Handbook of Affective Sciences. Oxford University Press, Oxford, New York, pp. 433–456 (Chapter 23).

Schiel, F., 1999. Automatic phonetic transcription of non-prompted speech. In: Proc. ICPhS, San Francisco, CA, USA, pp. 607–610.

Schröder, M., Pirker, H., Lamolle, M., 2006. First suggestions for an emotion annotation and representation language. In: Devillers, L., Martin, J.-C., Cowie, R., Douglas-Cowie, E., Batliner, A. (Eds.), Proc. Satellite Workshop of LREC 2006 on Corpora for Research on Emotion and Affect, Genoa, Italy, pp. 88–92.

Schröder, M., Devillers, L., Karpouzis, K., Martin, J.-C., Pelachaud, C., Peter, C., Pirker, H., Schuller, B., Tao, J., Wilson, I., 2007. What should a generic emotion markup language be able to represent?. In: Paiva A., Prada, R., Picard, R.W. (Eds.), Affective Computing and Intelligent Interaction. Springer, Berlin–Heidelberg, pp. 440–451.

Schröder, M., Cowie, R., Heylen, D., Pantic, M., Pelachaud, C., Schuller, B., 2008. Towards responsive sensitive artificial listeners. In: Proc. 4th Internat. Workshop on Human–Computer Conversation, Bellagio, Italy, no pagination.

Schuller, B., Rigoll, G., Lang, M., 2003. Hidden markov model-based speech emotion recognition. In: Proc. ICASSP, Hong Kong, pp. 1–4.

Schuller, B., Rigoll, G., Lang, M., 2004. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In: Proc. ICASSP, Montreal, Canada, pp. 577–580.

Schuller, B., Jiménez Villar, R., Rigoll, G., Lang, M., 2005a. Meta-classifiers in acoustic and linguistic feature fusion-based affect recognition. In: Proc. ICASSP, Philadelphia, PA, USA, pp. I:325–328.

Schuller, B., Müller, R., Lang, M., Rigoll, G., 2005b. Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensemble. In: Proc. Interspeech, Lisbon, Portugal, pp. 805–808.

Schuller, B., Reiter, S., Mller, R., Al-Hames, M., Lang, M., Rigoll, G., 2005c. Speaker independent speech emotion recognition by ensemble classification. In: Proc. ICME, Amsterdam, Netherlands, pp. 864–867.

Schuller, B., Arsić, D., Wallhoff, F., Rigoll, G., 2006. Emotion recognition in the noise applying large acoustic feature sets. In: Proc. Speech Prosody 2006, Dresden, Germany, no pagination.

Schuller, B., Rigoll, G., 2006. Timing levels in segment-based speech emotion recognition. In: Proc. Interspeech, Pittsburgh, PA, USA, pp. 1818–1821.

Schuller, B., Köhler, N., Müller, R., Rigoll, G., 2006b. Recognition of interest in human conversational speech. In: Proc. Interspeech, Pittsburgh, PA, USA, pp. 793–796.

Schuller, B., Reiter, S., Rigoll, G., 2006c. Evolutionary feature generation in speech emotion recognition. In: Proc. Internat. Conf. on Multimedia and Expo ICME 2006, Toronto, Canada, pp. 5–8.

Schuller, B., Stadermann, J., Rigoll, G., 2006d. Affect-robust speech recognition by dynamic emotional adaptation. In: Proc. Speech Prosody 2006, Dresden, Germany, no pagination.

Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., Aharonson, V., 2007a. The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals. In: Proc. Interspeech, Antwerp, Belgium, pp. 2253–2256.

Schuller, B., Müller, R., Hörnler, B., Höthker, A., Konosu, H., Rigoll, G., 2007b. Audiovisual recognition of spontaneous interest within conversations. In: Proc. 9th Internat. Conf. on Multimodal Interfaces (ICMI), Special Session on Multimodal Analysis of Human Spontaneous Behaviour. ACM SIGCHI, Nagoya, Japan, pp. 30–37.

Schuller, B., Seppi, D., Batliner, A., Meier, A., Steidl, S., 2007c. Towards more reality in the recognition of emotional speech. In: Proc. ICASSP, Honolulu, HY, USA, pp. 941–944.

Schuller, B., Batliner, A., Steidl, S., Seppi, D., 2008a. Does affect automatic recognition of children's speech? In: Proc. 1st Workshop on Child, Computer and Interaction, Chania, Greece, 4 pages, no pagination.

Schuller, B., Eyben, F., Rigoll, G., 2008b. Static and dynamic modelling for the recognition of non-verbal vocalisations in conversational speech. In: André, E. (Ed.), Proc. 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-based Systems (PIT 2008), Kloster Irsee, Germany, LNCS, Vol. 5078. Springer, pp. 99–110.

Schuller, B., Vlasenko, B., Arsic, D., Rigoll, G., Wendemuth, A., 2008c. Combining speech recognition and acoustic word emotion models for robust text-independent emotion recognition. In: Proc. ICME, Hannover, Germany, pp. 1333–1336.

Schuller, B., Wimmer, M., Arsic, D., Moosmayr, T., Rigoll, G., 2008d. Detection of security related affect and behaviour in passenger transport. In: Proc. Interspeech, Brisbane, Australia, pp. 265–268.

Schuller, B., Wimmer, M., Mösenlechner, L., Kern, C., Arsic, D., Rigoll, G., 2008e. Brute-forcing hierarchical functionals for paralinguistics: a waste of feature space? In: Proc. ICASSP, Las Vegas, NV, pp. 4501–4504.

Schuller, B., Rigoll, G., 2009. Recognising interest in conversational speech – comparing bag of frames and supra-segmental features. In: Proc. Interspeech, Brighton, UK, pp. 1999–2002.

Schuller, B., Batliner, A., Steidl, S., Seppi, D., 2009a. Emotion recognition from speech: putting ASR in the loop. In: Proc. ICASSP, IEEE, Taipei, Taiwan, pp. 4585–4588.

Schuller, B., Müller, R., Eyben, F., Gast, J., Hörnler, B., Wöllmer, M., Rigoll, G., Höthker, A., Konosu, H., 2009b. Being bored? Recognising natural interest by extensive audiovisual integration for real-life Application. Image Vision Comput. J. (IMAVIS) 27, 1760–1774 (Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior).

Schuller, B., Schenk, J., Rigoll, G., Knaup, T., 2009c. The "Godfather" vs. "Chaos": comparing linguistic analysis based on online knowledge sources and Bags-of-*N*-grams for movie review valence estimation. In: Proc. Internat. Conf. on Document Analysis and Recognition, Barcelona, Spain, pp. 858–862.

Schuller, B., Steidl, S., Batliner, A., 2009d. The INTERSPEECH 2009 Emotion Challenge. In: Proc. Interspeech, Brighton, UK, pp. 312–315.

Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., Wendemuth, A., 2009e. Acoustic emotion recognition: a benchmark comparison of performances. In: Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Merano, pp. 552–557.

Schuller, B., Wöllmer, M., Eyben, F., Rigoll, G., 2009f. Spectral or voice quality? Feature type relevance for the discrimination of emotion pairs. In: Hancil, S. (Ed.), The Role of Prosody in Affective Speech. Linguistic Insights, Studies in Language and Communication, Vol. 97. Peter Lang, pp. 285–307.

Schuller, B., Wöllmer, M., Moosmayr, T., Rigoll, G., 2009g. Recognition of noisy speech: a comparative survey of robust model architectures and feature enhancement. EURASIP J. Audio Speech Music Process. (JASMP), 17 pages, Article ID 942617.

Schuller, B., Burkhardt, F., 2010. Learning with synthesized speech for automatic emotion recognition. In: Proc. 35th IEEE Internat. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Dallas, pp. 5150–5153.

Schuller, B., Weninger, F., 2010. Discrimination of speech and non-linguistic vocalizations by non-negative matrix factorization. In: Proc. ICASSP, Dallas, pp. 5054–5057.

Schuller, B., Eyben, F., Can, S., Feussner, H., 2010a. Speech in minimal invasive surgery – towards an affective language resource of real-life medical operations. In: Proc. 3rd ELRA Internat. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect, Valetta, pp. 5–9.

Schuller, B., Metze, F., Steidl, S., Batliner, A., Eyben, F., Polzehl, T., 2010b. Late fusion of individual engines for improved recognition of negative emotions in speech – learning vs. democratic vote. In: Proc. 35th IEEE Internat. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Dallas, pp. 5230–5233.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S., 2010c. The INTERSPEECH 2010 paralinguistic challenge. In: Proc. INTERSPEECH 2010, Makuhari, Japan, pp. 2794–2797.

Seppi, D., Batliner, A., Schuller, B., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Aharonson, V., 2008a. Patterns, prototypes, performance: classifying emotional user states. In: Proc. Interspeech, Brisbane, Australia, pp. 601–604.

Seppi, D., Gerosa, M., Schuller, B., Batliner, A., Steidl, S., 2008b. Detecting problems in spoken child–computer interaction. In: Proc. 1st Workshop on Child, Computer and Interaction, Chania, Greece, no pagination.

Seppi, D., Batliner, A., Steidl, S., Schuller, B., Nöth, E., 2010. Word accent and emotion. In: Proc. Speech Prosody 2010, Chicago, IL, no pagination.

Sethu, V., Ambikairajah, E., Epps, J., 2007. Speaker normalisation for speech-based emotion detection. In: Proc. 15th Internat. Conf. on Digital Signal Processing, Cardiff, pp. 611–614.

Shami, M., Verhelst, W., 2007. Automatic classification of expressiveness in speech: a multi-corpus study. In: Müller, C. (Ed.), Speaker Classification II, Lecture Notes in Computer Science/Artificial Intelligence, Vol. 4441. Springer, Heidelberg–Berlin–New York, pp. 43–56.

Shaver, P.R., Wu, S., Schwartz, J.C., 1992. Cross-cultural similarities and differences in emotion and its representation: a prototype approach. Emotion, 175–212.

Sood, S., Krishnamurthy, A., 2004. A robust on-the-fly pitch (otfp) estimation algorithm. In: Proc. 12th Annual ACM Internat. Conf. on Multimedia (MULTIMEDIA '04). ACM, New York, NY, USA, pp. 280–283.

Steidl, S., Ruff, C., Batliner, A., Nöth, E., Haas, J., 2004. Looking at the last two turns, I'd say this dialogue is doomed — measuring dialogue success. In: Sojka, P., Kopeček, I., Pala, K. (Eds.), 7th Internat. Conf. on Text, Speech and Dialogue, TSD 2004, Berlin, Heidelberg, pp. 629–636.

Steidl, S., Batliner, A., Nöth, E., Hornegger, A., 2008. Quantification of segmentation and fo errors and their effect on emotion recognition. In: 11th Internat. Conf. on Text, Speech and Dialogue, TSD 2008, pp. 525–534.

Steidl, S., 2009. Automatic classification of emotion-related user states in spontaneous children's speech. Logos Verlag, Berlin, Germany, (Ph.D. Thesis, FAU Erlangen-Nuremberg).

Steidl, S., Schuller, B., Batliner, A., Seppi, D., 2009. The hinterland of emotions: facing the open-microphone challenge. In: Proc. ACII, Amsterdam, Netherlands, pp. 690–697.

Steidl, S., Batliner, A., Seppi, D., Schuller, B., 2010. On the impact of children's emotional speech on acoustic and language models.

EURASIP J. Audio Speech Music Process., 14. doi:10.1155/2010/783954.

Takahashi, K., 2004. Remarks on emotion recognition from bio-potential signals. In: Proc. 2nd Internat. Conf. on Autonomous Robots and Agents, pp. 186–191.

tenBosch, L., 2003. Emotions, speech and the ASR framework. Speech Comm. 40 (1–2), 213–225.

Tomlinson, M.J., Russell, M.J., Brooke, N.M., 1996. Integrating audio and visual information to provide highly robust speech recognition. In: Proc. ICASSP, Atlanta, GA, USA, pp. 812–824.

Truong, K., van Leeuwen, D., 2005. Automatic detection of laughter. In: Proc. Interspeech, Lisbon, Portugal, pp. 485–488.

Ververidis, D., Kotropoulos, C., 2003. A review of emotional speech databases. In: PCI 2003, 9th Panhellenic Conf. on Informatics, November 1–23, 2003, Thessaloniki, Greece, pp. 560–574.

Ververidis, D., Kotropoulos, C., 2006. Fast sequential floating forward selection applied to emotional speech features estimated on DES and SUSAS data collection. In: Proc. European Signal Processing Conf. (EUSIPCO 2006), Florence, Italy, no pagination.

Vidrascu, L., Devillers, L., 2007. Five emotion classes in real-world call center data: the use of various types of paralinguistic features. In: Proc. PARALING07, pp. 11–16.

Vinciarelli, A., Pantic, M., Bourlard, H., Pentland, A., 2008. Social signals, their function, and automatic analysis: a survey. In: Proc. 10th Internat. Conf. on Multimodal Interfaces, ACM, New York, USA, pp. 61–68.

Vlasenko, B., 2009. Processing affected speech within human machine interaction. In: Proc. Interspeech, Brighton, pp. 2039–2042.

Vlasenko, B., Schuller, B., Wendemuth, A., Rigoll, G., 2007a. Combining frame and turn-level information for robust recognition of emotions within speech. In: Proc. Interspeech, Antwerp, Belgium, pp. 2249–2252.

Vlasenko, B., Schuller, B., Wendemuth, A., Rigoll, G., 2007b. Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing. In: Paiva, A., Prada, R., Picard, R.W. (Eds.), Affective Computing and Intelligent Interaction. Springer, Berlin–Heidelberg, pp. 139–147.

Vlasenko, B., Schuller, B., Mengistu, T.K., Rigoll, G.A.W., 2008. Balancing spoken content adaptation and unit length in the recognition of emotion and interest. In: Proc. Interspeech, Brisbane, Australia, pp. 805–808.

Vogt, T., André, E., 2005. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In: Proc. Multimedia and Expo (ICME05), Amsterdam, Netherlands, pp. 474–477.

Vogt, T., André, E., 2009. Exploring the benefits of discretization of acoustic features for speech emotion recognition. In: Proc. Interspeech, Brighton, pp. 328–331.

Vogt, T., André, E., Bee, N., 2008. Emovoice – a framework for online recognition of emotions from voice. In: Proc. IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems (PIT 2008). Lecture Notes in Computer Science, Vol. 5078. Springer, Kloster Irsee, Germany, pp. 188–199.

Vogt, T., André, E., Wagner, J., Gilroy, S., Charles, F., Cavazza, M., 2009. Real-time vocal emotion recognition in artistic installations and interactive storytelling: Experiences and lessons learnt from CALLAS and IRIS. In: Proc. ACII, Amsterdam, Netherlands, pp. 670–677.

Wagner, J., Kim, J., André, E., 2005. From physiological signals to emotions: implementing and comparing selected methods for feature extraction and classification. In: Proc. ICME, Amsterdam, Netherlands, pp. 940–943.

Wagner, J., Vogt, T., André, 2007. A systematic comparison of different HMM designs for emotion recognition from acted and spontaneous speech. In: Paiva, A., Prada, R., Picard, R.W. (Eds.), Affective Computing and Intelligent Interaction. Springer, Berlin–Heidelberg, pp. 114–125.

Wang, Y., Guan, L., 2005. Recognizing human emotion from audiovisual information. In: Proc. ICASSP, Vol. 2. Philadelphia, PA, USA, pp. 1125–1128.

Wilson, T., Wiebe, J., Hwa, R., 2004. Just how mad are you? Finding strong and weak opinion clauses. In: Proc. Conf. American Association for Artificial Intelligence (AAAI), San Jose, CA, no pagination.

Wimmer, M., Schuller, B., Arsic, D., Radig, B., Rigoll, G., 2008. Low-level fusion of audio and video features for multi-modal emotion recognition. In: Proc. 3rd Internat. Conf. on Computer Vision Theory and Applications, Funchal, Portugal, pp. 145–151.

Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques, second ed. Morgan Kaufmann, San Francisco, CA, USA.

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

Wöllmer, M., Al-Hames, M., Eyben, F., Schuller, B., Rigoll, G., 2009. A multidimensional dynamic time warping algorithm for efficient multimodal fusion of asynchronous data streams. Neurocomputing 73, 366–380.

Wöllmer, M., Eyben, F., Keshet, J., Graves, A., Schuller, B., Rigoll, G., 2009. Robust discriminative keyword spotting for emotionally colored spontaneous speech using bidirectional LSTM networks. In: Proc. ICASSP, Taipei, Taiwan, pp. 3949–3952.

Wöllmer, M., Schuller, B., Eyben, F., Rigoll, G., 2010. Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening. IEEE J. Select. Topics Signal Process. 4, 867–881 (Special Issue on "Speech Processing for Natural Interaction with Intelligent Environments").

Wolpert, D., 1992. Stacked generalization. Neural Networks 5, 241–259.

Wu, T., Khan, F., Fisher, T., Shuler, L., Pottenger, W., 2005. Posting act tagging using transformation-based learning. In: Lin, T.Y., Ohsuga, S., Liau, C.-J., Hu, X., Tsumoto, S. (Eds.), Foundations of Data Mining and Knowledge Discovery. Springer, Berlin–Heidelberg, pp. 319–331.

Wu, C.-H., Yeh, J.-F., Chuang, Z.-J., 2008a. Emotion perception and recognition from speech. In: Affective Information Processing. II. Springer, London, pp. 93–110.

Wu, S., Falk, T., Chan, W.-Y., 2008b. Long-term spectro-temporal information for improved automatic speech emotion classification. In: Proc. INTERSPEECH, Brisbane, Australia, pp. 638–641.

Yi, J., Nasukawa, T., Bunescu, R., Niblack, W., 2003. Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. In: Proc. IEEE Internat. Conf. on Data Mining (ICDM), Melbourne, FL, pp. 427–434.

You, M., Chen, C., Bu, J., Liu, J., Tao, J., 2006. Emotion recognition from noisy speech. In: Proc. ICME, Toronto, Canada, pp. 1653–1656.

Young, S., Evermann, G., Gales, M., Hain, T., D.Kershaw, Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., 2006. The HTK Book, for htk version 3.4 Edition, Cambridge University Engineering Department.

Yu, C., Aoki, P., Woodruf, A., 2004. Detecting user engagement in everyday conversations. In: Proc. ICSLP, pp. 1329–1332.

Zeng, Z., Hu, Y., Roisman, G.I., Wen, Z., Fu, Y., Huang, T.S., 2007a. Audio–visual spontaneous emotion recognition. Artif. Intell. Human Comput., 72–90.

Zeng, Z., Tu, J., Liu, M., Huang, T.S., Pianfetti, B., Roth, D., Levinson, S., 2007b. Audio–visual affect recognition. IEEE Trans. Multimedia 9 (2), 424–428.

Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S., 2009. A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Machine Intell. 31 (1), 39–58.

Zhe, X., Boucouvalas, A., 2002. Text-to-emotion engine for real time internet communication. In: Proc. Internat. Symp. on Communication Systems, Networks, and DSPs, Staffordshire University, pp. 164–168.

Zwicker, E., Fastl, H., 1999. Psychoacoustics – Facts and Models, second ed. Springer-Verlag.