The Automatic Recognition of Emotions in Speech

Anton Batliner, Björn Schuller, Dino Seppi, Stefan Steidl, Laurence Devillers, Laurence Vidrascu, Thurid Vogt, Vered Aharonson, and Noam Amir

Abstract In this chapter, we focus on the automatic recognition of emotional states using acoustic and linguistic parameters as features and classifiers as tools to predict the 'correct' emotional states. We first sketch history and state of the art in this field; then we describe the process of 'corpus engineering', i.e. the design and the recording of databases, the annotation of emotional states, and further processing such as manual or automatic segmentation. Next, we present an overview of acoustic and linguistic features that are extracted automatically or manually. In the section on classifiers, we deal with topics such as the curse of dimensionality and the sparse data problem, classifiers, and evaluation. At the end of each section, we point out important aspects that should be taken into account for the planning or the assessment of studies. The subject area of this chapter is not emotions in some narrow sense but in a wider sense encompassing emotion-related states such as moods, attitudes, or interpersonal stances as well. We do not aim at an in-depth treatise of some specific aspects or algorithms but at an overview of approaches and strategies that have been used or should be used.

1 Introduction

The study of speech and emotion can be traced back to the first decades of the last century, cf. Scripture (1921), Skinner (1935), and Fairbanks and Pronovost (1939). Whereas such studies were not very frequent during the following decades – one of the exceptions being Williams and Stevens (1972) – the topic began to attract researchers more and more during the eighties. Until the nineties most of these studies could be subsumed under the heading 'basic research in psychology and phonetics/linguistics'; an overview is given, for example, in Scherer (2003). In the nineties, the automatic processing of speech started to address topics beyond pure word recognition. First, higher linguistic levels, for instance, dialogue acts, and

A. Batliner (✉)
Lehrstuhl für Mustererkennung, Friedrich-Alexander-Universität Erlangen, Erlangen, Germany
e-mail: batliner@informatik.uni-erlangen.de

P. Petta et al. (eds.), *Emotion-Oriented Systems*, Cognitive Technologies,
DOI 10.1007/978-3-642-15184-2_6, © Springer-Verlag Berlin Heidelberg 2011

then topics beyond pure information transmission, that is, paralinguistic phenomena, e.g. emotions and attitudes conveyed via the speech channel, were addressed in studies such as Dellaert et al. (1996). At that time, however, almost all data used were 'prompted' and acted, cf. below, modelling the prototypical 'big' *n* emotions, *n* being a figure greater or equal 2 and up to 4, 6, or even more classes. Maybe the first paper dealing with 'natural(istic)' speech and emotions was Slaney and McRoberts (1998). At the turn of the century, researchers began to use non-acted databases from, generally speaking, interactions of humans with information offices/systems, i.e. human–human or human–machine interaction – the role of the machine sometimes played by a human Wizard-of-Oz (WoZ) – such as appointment scheduling or call centre dialogues, cf. Batliner et al. (2000a), Lee et al. (2001), and Ang et al. (2002).

Nowadays, it is widely acknowledged that acted data cannot model naturalistic data sufficiently, as demonstrated by Batliner et al. (2000a) and Wilting et al. (2006), especially because the emotions produced that way are too pronounced and will rather seldom be encountered as such in more realistic data. Thus a (direct) transfer from acted data onto data encountered in realistic applications is not feasible. Acted data are still used to a large extent, e.g. in Ververidis and Kotropoulos (2006); possible applications can be found within the entertainment business, e.g. for data mining in movie archives or for computer games. The main reason for this preponderance, however, is simply that non-acted data are still sparse and most often not available freely. In this chapter, we will concentrate on the genuine approach of automatically recognising/classifying emotional user states signalled in naturalistic (spontaneous) speech. We will deal with acted speech only in order to illustrate specific approaches or methodologies. Nonetheless, the basic requirements of automatic processing are the same for both acted and naturalistic data: large enough size of the database, balanced distribution of classes, large number of speakers, recording quality, class assignment as unequivocal as possible, etc. However, using realistic data requires us to face some more challenges: sparse and very unbalanced data, less pronounced emotions, and definitely the need to explicitly annotate the data, assigning emotion classes. Moreover, the data should be representative for the envisioned application; actually, this is the most important requirement: if we are interested in emotional film scenes, film actors as speakers are adequate – but not necessarily speakers prompted for emotions in the laboratory.

In the field of emotion in speech, two lines of research came together with their own standards and methods which have not converged yet: basic (psychological, clinic, phonetic) research, dealing mostly with acted data, and applied engineering – so far, too often dealing with acted data as well. Naïve conceptualisations of the respective other line of research should be replaced by a mutual understanding of innate constraints and benefits. However, it is beneficial to conceive the study and especially the automatic processing of non-acted, non-prompted emotional states as a topic sui generis.

2 Corpus Engineering

We conceive the term 'corpus engineering' as encompassing all the steps necessary before feature extraction and automatic classification can take place:

1. the design of an application-oriented scenario
2. the recruiting of the necessary personnel such as subjects, supervisors (Wizard-of-Oz), and the experimental setting or the real-life scenario
3. the recordings and – if necessary – subsequent transfer onto storage media with/without resampling of the audio signal
4. the transliteration, i.e. the orthographic transcription of the data, sometimes including the annotation of extra- or non-linguistic events such as breathing or noise
5. the definition and extraction of appropriate units of analysis such as words, chunks, turns, dialogue moves with appropriate criteria (intuitive or based on prosodic, linguistic, or pragmatic criteria)
6. the annotation of emotional states, possibly with subsequent mapping onto fewer main classes
7. evaluating the quality of these annotations by applying some measures of correlation/correspondence
8. some other pre-processing steps like manual processing or correction of automatically processed feature values
9. defining and applying exchange formats

We will sketch (1), (2), and (4) skipping the technical aspects of (3), mention (8)–(9), and concentrate on (5)–(7).

2.1 Databases

A common breakdown of emotion databases is the one into acted/non-acted, induced, and naturalistic databases, cf. Douglas-Cowie et al. (2007). This is a gross taxonomy which does not yet capture pertinent differences: the settings, i.e. the scenarios, are defined and created by the researcher; the outcome is the data that we have to deal with. Here we want to tell apart acted/non-acted and prompted/non-prompted (Schiel, 1999) settings: if the subject acts, he/she is doing as if they were in this specific situation – no matter whether it is about being emotional or not. If emotions are prompted themselves, the subjects have been told that they should produce specific emotions. The subjects can be volunteering or recorded in real-life situations. Inducing emotions means to arrange situations where the subjects are more likely to produce the desired emotional states. Strictly speaking, all these different conditions do not tell us whether our subjects will produce 'natural', realistic emotion-related states or not. It is just more likely that the outcome, i.e. the emotional database, is less natural if acted; induced data, for instance, can be more or less spontaneous, or fully spontaneous. All these differences can be evaluated

by applying a perceptive evaluation – either with naïve listeners in a perception experiment or with a more intuitive assessment.

This is a representative but not necessarily exhaustive list of scenarios where non-acted, non-prompted data have been collected, recorded, and used for the automatic classification of emotions in speech in the last decade: mother–child interaction (Slaney and McRoberts, 1998), human–robot interaction (Batliner et al., 2008b), tutoring dialogues (Ai et al., 2006), stress detection in a driving scenario (Fernandez and Picard, 2003), human–human (multi-party) conversation and interaction (Neiberg et al., 2006; Grimm et al. 2008; Schuller et al., 2009a), interaction human-information kiosk (Batliner et al., 2003b), appointment scheduling dialogues (Batliner et al., 2000a, 2003a), surgeons' speech during operations (Schuller et al., 2008), call centre applications using volunteering or real users, WoZ or real systems (Lee et al. 2001; Ang et al. 2002; Batliner et al. 2004; Steidl et al. 2004; Devillers et al. 2005). Some more references to databases mostly with acted data can be found in Cowie et al. (2005). Multi-modal databases and approaches are dealt with in Zeng et al. (2009) with the focus on other modalities, and in Cowie et al., this volume.

2.2 Annotations

Annotations can be automatic or manual or both (first automatic and then edited manually). The first annotation pass is normally the transliteration of what has been said. Even if automatic speech recognition (ASR) can be applied, a manual editing of its results is mandatory if correct transliterations are aimed at. Transliteration conventions are either implicit or following standards put forth, e.g. by LDC (http://www.ldc.upenn.edu/), cf. Devillers et al. (2005), or within the Verbmobil project in Burger et al. (2000). Apart from the 'normal' linguistic events, i.e. the words produced by the speakers, several other para-/extra-linguistic (breathing, sighing, laughter) or non-linguistic (technical noise) events can be annotated. Moreover, there are specific conventions for the annotation of typical spontaneous phenomena such as hesitations, filled or unfilled pauses, false starts, repetitions.

The next step should be to define the units of emotion annotation; these, in turn, are constitutive for the units of analysis used in the classification phase. So far, this has been done mostly on a trivial or on an intuitive basis: the unit is given trivially if simply utterances/dialogue moves/turns are taken – this can be an easy endeavour in a dialogue where the partners alternate as speakers/listeners. If the turns are longer, however, chances are that there is not one and the same emotion throughout this turn. This is of course descriptively less adequate and diminishes the discriminative power of automatic classification. Sometimes, longer turns are segmented on an intuitive notion (de Rosis et al., 2007; Devillers et al. 2005) of prosodic, syntactic, or pragmatic segmentation. In Batliner et al. (2003a) an objective approach towards defining units based on syntactic–prosodic segmentation has been put forth. Another possibility is to segment automatically at prosodic boundaries, using either only pause information or more complex information on intonational/prosodic units. Although there is a high correspondence between such prosodic units and higher

syntactic/pragmatic units as shown in Batliner et al. (1998), it is not perfect and thus sub-optimal if it comes to the processing of emotion recognition in a full end-to-end system (Batliner et al., 2000b) because there will be the additional task to time-align the syntactically/semantically 'blind' prosodic units with the units processed by the higher module.

The impact of choosing the appropriate unit of analysis has been underestimated so far. However, the most important initial step is, of course, to find the adequate (number of) emotion labels. To start with, this can be done top-down or data driven: in the first case, the basis is normally a catalogue of theoretically derived or empirically obtained categories, cf. the terms used by Devillers et al. (2005) or the scheme proposed by Craggs and Wood (2004). Theoretically derived dimensional terms can be more or less elaborated (Russel, 1997). In the data-driven approach that has often been employed by more 'application-minded' studies, cf. below, only those categories are used that can be observed (often enough) in the data and are, at the same time, relevant for the intended applications.

The biggest issue in this phase concerns the two questions 'What to annotate' and 'How to annotate'. In the case of naturalistic data, a catalogue of prototypical (basic) emotion categories or dimensions falls short of the phenomena one can find; and what cannot be found cannot be annotated. Of course, different granularities can be chosen for a first annotation pass. In the short history of annotating naturalistic databases, the first studies were normally restricted to modelling a mapping onto a two-way distinction *negative* (encompassing user states such as anger, annoyance, or frustration) vs. the complement, i.e. *neutral*, even if at the beginning, more classes were annotated such as in Ang et al. (2002) neutral, annoyed, frustrated, tired, amused, other, not applicable. The minor reason for this mapping onto negative valence vs. neutral/positive valence was that in the intended application, it is most important to detect 'trouble in communication'. The major reason is simply that for statistical modelling, enough items per class are needed. The default, 'neutral', unmarked state dominates and accounts for up to $>90\%$ of the cases. The situation has not changed much recently, cf. Devillers et al. (2005). Neiberg et al. (2006) model, label, and recognize a three-way distinction neutral, emphatic, and negative for one database (voice-controlled telephone service) and for another (multi-party meetings), a three-way emotional valence negative, neutral, and positive. Ai et al. (2006) use a three-way distinction for student emotion in spoken tutoring dialogs: mixed/uncertain, certain, and neutral. Devillers et al. (2005) established an annotation scheme with the possibility to have a mixture of emotions (two labels per segment) and to use a coarse level (8 classes) and a fine-grained level (20 classes) plus neutral for annotation; a coarse label is, for example, anger with the fine-grained sub-classes anger, annoyance, impatience, cold anger, and hot anger. Mower et al. (2009) elaborate on prototypical/consensus vs. non-prototypical/no consensus for the following labels: angry, happy, sad, neutral, frustrated, or excited (audiovisual data, 10 actors). In some few studies, up to seven different emotional user states are classified as in Batliner et al. (2003b, 2008b); however, this 7-class problem cannot be used for real applications because classification performance is simply too low.

There are basically two different strategies answering the question 'How to annotate': we can start with a detailed catalogue of labels and reduce them in a more or less systematic manner to fewer labels to be used in annotation – those that really denote states that can be observed in the data – and to an even smaller set of labels to be used in automatic classification. The catalogue can be obtained from other basic studies or be based on free annotation, cf. Devillers et al. (2005). Alternatively, we can skip this step and establish in a data-driven way a set of labels suited for the intended application; for instance, in a call centre application, we might only want to find out whether the user is getting angry/annoyed, etc., i.e. whether something is going wrong. This would be a task-dependent emotion annotation with the goal of emotion detection in a real system. In the studies conducted so far, the set of labels chosen was mostly intended to be suited for the data, although aiming at the general issue of emotional behaviour annotation. However, emotional states that cannot be observed often enough were skipped in an earlier or later stage of the annotation process. Moreover, there is a certain trade-off between the number of the labellers, their expertise, and the effort to be spent: from methodological reasons, it might be desirable to employ something like >10 naïve labellers or >5 expert labellers to annotate on a fine-grained scale; by that, any 'central tendency' is not corrupted even if one expert or two naïve labellers might go astray. To follow this rule of thumb is, however, almost never feasible. Normally, more than one labellers are employed. This makes it possible to establish measures of agreement, cf. below, and to establish different levels of agreement: apart from the method to allow each labeller to give more than one label per unit, cf. the major and minor label in Devillers et al. (2005), for more labellers, either a correspondence or a majority decision can be defined or a soft vector with percentages can be created (Steidl et al., 2005; Devillers et al., 2005). Further, continuous labelling can be performed over time and space by dimensions as arousal or valence (Cowie et al., 2000). In any case, standards as, e.g. EmotionML in Schröder et al. (2007) can be of use, which allows for all of these labellings. For some scenarios, there can be some 'external ground truth', e.g. the intensity of stress-inducing tasks, a worse performance of the system, physiological measures as indicators of stress (levels). Such an external evidence can be taken as either means for assigning labels or later on as additional feature in the classification phase.

There are two classic criteria for assessing the quality of such labels: validity and reliability. Ecological validity is most important but not easy to measure; thus normally, reliability measures are aimed at such as measures of correlation, correspondence, (weighted) kappa, or (weighted) alpha (Fleiss et al., 1969; Rosenberg and Binkowski, 2004). The use of 'quantised' score ranges, based on such measures, e.g., for kappa, < 0.2 'bad', between 0.2 and 0.4 'moderate', between 0.4, and 0.6 'good', between 0.6 and 0.8 'very good', > 0.8 'excellent' (there are other scalings), seems to be a convenient way of assessing the quality of annotations. As far as we can see, however, it has almost never been used for any decision to be made – for some reasons: a lower kappa score can – apart from being caused by deficiencies in the very score itself – mean that inter- or intra-rater reliability is low because of spurious factors or because there simply are different – and valid – criteria and

thresholds for annotation, and/or simply that the task is difficult, etc. Too high scores can be rather suspicious because it can be doubted that they can be obtained when dealing with naturalistic data. Moreover, the ultimate measure (of validity) is on the one hand the performance of the classifier – which itself can be compared with the performance of the annotators by using measures such as proposed in Steidl et al. (2005), and on the other hand, the impact on the users of such systems, cf. Sect. 5.

2.3 Further Processing

The ultimate goal in ASR is fully automatic processing although important steps such as building a lexicon or transliterating the training data are still mostly done manually. Matters are different in the research of emotion in speech: here it is not yet considered to be very important whether processing is manual or not; thus we often observe a mixture of manual and automatic processing. A typical approach is, e.g., to extract acoustic features automatically and linguistic features such as non-verbals or part-of-speech classes semi-automatically or fully based on manual processing. Sometimes, automatically extracted acoustic features are corrected manually, cf. Batliner et al. (2007b) where the manual correction of word segmentation and pitch values is described. Segmentation of higher units into lower ones can be 'blind', i.e. automatic, e.g. by defining fixed length segments or by partitioning each turn into a fixed number of segments, or it can be 'intelligent', e.g. by segmenting into words or other smaller units using other higher level information. A 'blind' segmentation is normally automatic, an 'intelligent' one so far mostly manual. The choice of segmentation strategies is of course conditioned by the type of data used and by the effort needed: turns produced by one speaker taking part in a bi-directional dialogue can be segmented by hand, whereas the effort needed for a more fine-grained (word- or syllable-based) segmentation is considerably higher.

A last and decisive step is the selection of units out of the whole database for feature extraction and classification. Two easy and automatic strategies are almost never employed: simply using all the data or using a randomly chosen sub-sample. This is due to the sparse data problem: the overwhelming majority of the cases belong to the 'uninteresting' default class neutral, cf. Sect. 4.1. Non-neutral cases can often not unequivocally be attributed to one of the 'interesting' classes because they are mixed; often, more prototypical cases are chosen. This is permissible – after all, we can imagine an application looking only for very pronounced cases – but the selection criteria have to be documented clearly: simply to select more prototypical cases by sharpening the threshold criterion can yield a marked performance improvement, cf. Batliner et al. (2005) and Seppi et al. (2008b).

It should be mandatory for writing a paper on recognising emotions in speech, and it is advisable for readers of such papers, to point out explicitly and to find out the strategies used at different stages: what is automatic, what manual,

which criteria were intuitive, which objective, and which criteria for selecting the final sample were applied. Intuitive and/or selection criteria as such should not necessarily be forbidden, if stated explicitly. They simply introduce some fuzziness at a certain stage of processing. Their impact on the final results – and it is mostly recognition performance that is remembered by the readers of such studies – can be decisive or small. It would be good practice if the authors themselves pointed out the presumable impact.

3 Features

Feature extraction is a crucial phase in automated emotion recognition. As yet there has not been a large-scale, comprehensive comparison of different feature types; as for preliminary efforts in this direction, cf. Batliner et al. (2006b) and Schuller et al. (2007a). Presenting a comprehensive overview of feature types and feature extraction methods requires some kind of division of features into classes, though there is more than one way to do so. We will present several – alternative and complementing – approaches to grouping features. The most basic distinction to be made is between acoustic vs. linguistic features, as extraction methods for these two types are extremely different. Their relative contribution can also vary greatly, depending on the database being analysed: for acted 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 we come closer to spontaneous real-life speech, these features can gain considerably in importance. Acoustic features are the more 'classic' features which have been in use since the inception of studies in this field, though researchers are far from agreeing which are most important, or whether this can even be determined. In the following subsections we discuss these two feature types separately.

3.1 Acoustic Features

Segmental features are mainly short-term spectra and derived features: MFCC (Lee et al., 2004), LPC, PLP (Perceptual Linear Prediction), etc. (Hermansky, 1990), and Wavelets (Fernandez and Picard, 2003 and Schuller et al., 2007a), TEO (Teager Energy operator) (Fernandez and Picard, 2003 and Zhou et al., 2001), LFPC, LPPC (Nwe et al., 2003). MFCCs are classically used for ASR, normally for modelling segments such as phones and, by that, words. In emotion recognition, they are rather used for modelling longer units of analysis such as utterances/turns, dialogue moves. To this aim, the features are extracted frame-wise and combined by appropriate measures such as averaging or by resorting to dynamic classification such as hidden Markov models. Although originally intended to model segments, these features have been used successfully for supra-segmental units.

Supra-segmental features model the classic prosodic types: pitch, intensity, duration, then voice quality, and long-term spectra. Prosodic features involve two steps: extracting raw prosodic *basic* features, then calculating *structured* features based on this data (Kießling, 1997; Hess et al., 1996). The raw prosodic data are the F0 contour, the intensity contour, and durational data on different levels (lengths of chunks, words, voiced segments, syllables, phonemes). Various errors can creep into the calculations at this stage. The second step involves extracting structured features from the basic prosodic features using various statistics such as mean, standard deviation, percentiles, ranges, peaks, slopes, regressions. Voice quality is a complicated issue in itself, since there are many different measures of voice quality (e.g. Lugger et al., 2006), mostly clinical in origin, 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. (2007a). There are several survey papers on prosodic features in automatic speech processing such as Hess et al. (1996) and Nöth et al. (2002) and on their use in emotion modelling, cf. Frick (1985), Scherer et al. (2003), and Johnstone and Scherer (2000).

Features can be low level or high level, i.e. statistic features or those based on pitch models such as MoMel (Hirst et al., 2000) or the Fujisaki model (Fujisaki, 1992). Features can be represented by raw values, i.e. they can be non-perceptual or they can be based on perception models. Normalisation and standardisation of pitch range, pitch mean, speech tempo, etc. are used for modelling perception as well and for making successive measurements coherent with respect to a common scale.

Using another terminology, we can speak about *Low Level Descriptors* (LLDs), i.e. basic measures of feature types, and *functionals* such as mean, percentiles. LLDs account for base contours that usually are extracted by processing a fixed number of samples contained in a sliding window. For example, pitch attributes derive from the F0 contour. Subsequently to the LLD extraction, a number of operators and functionals are applied to obtain a certain feature vector out of each contour. Functionals provide a normalisation over time: base contours associated to words have different lengths, depending on the duration of the words and on the magnitude of the window step; with the usage of functionals, we obtain one feature vector per word, with a constant number of elements.

To reduce the influence of noise and to model temporal variations of LLDs, base contours are usually filtered, and first- and second-order derivatives are extracted. These functionals that can be applied to raw contours range from simple statistics 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 kurtosis). Other functionals are positions of extremes values within a certain temporal context, quartiles, amplitude ranges, zero-crossing rates, roll-on/-off, on-/off-set, and higher level analysis. Curve fitting methods produce regression coefficients, such as slope of polynomial regressions, and regression errors (such as the mean square error between the regression and the original contour). Maybe the most comprehensive list of functionals is given in Schuller et al. (2007a) and Eyben et al. (2009).

We now characterise shortly the different types of acoustic features:

- *Duration* features model temporal aspects; the basic unit is milliseconds (ms) for the 'raw' values. Different types of normalisation can be applied. Note that 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 milliseconds 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, utterances. Duration values are usually correlated with the linguistic features described below: for instance, function words are shorter on average, content words are longer. These two main word classes are not equally distributed across emotion types; this information can be used for classification, no matter whether it is encoded in linguistic or acoustic (i.e. duration) features.
- *Energy (intensity)* features usually model the loudness of a sound as perceived by the human ear, based on the amplitude in different intervals; different types of normalisation are applied. 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 too. 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.
- 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 by either autocorrelation (Rabiner, 1977) in its manifold variants and derivatives. Often, the 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. However, also PDAs in the time domain exist that have the advantage of being able to detect changes per fundamental period, though generally being less reliable. The acoustic equivalent to the perceptual unit pitch is measured in Hertz and often made perceptually more adequate, e.g. by logarithmic/semitone transformation. Intervals, characteristic points, or contours are often modelled.
- The *spectrum* is characterised by formants (spectral maxima) modelling spoken content, especially the lower ones. 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 Linear Prediction Coding (LPC) (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 spectral features can be computed such as formant band-energies, roll-off, centroid, and

flux. 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 or secondary 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 which is related to frequency. 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. PLP coefficients (Hermansky, 1990) and the MFCCs are extremely similar, as they both correspond to a short-term spectrum smoothing – the former through an autoregressive model, the latter trough 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.
- *Voice quality* features model jitter, shimmer, and further micro-prosodic events. Noise to harmonic ratio (NHR) or harmonic-to-noise ratio (HNR) is another measure 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. Therefore, they are usually dealt with within a separate feature class. Some of these have several variants and even when their definitions are agreed upon, different software can give different values, due, for example, to differences in pitch extraction methods.
- *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.
- *Non-linguistic Vocalisations* identify non verbal phenomena such as breathing and laughter. 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.

Other acoustic features that have been used or can be used are TRAPs (Hermansky and Sharma, 1998) or Teager operator (especially for stress detection) (Zhou et al., 2001). The standard acoustic feature types used in many emotion classification studies might be – probably in this order of frequency but not necessarily of importance – pitch, energy, spectrum, cepstrum, voice quality, duration. Traditionally, pitch has been conceived as being most important – this is not backed up by empirical results; note that the reason might not be extraction errors, cf. Batliner et al. (2007b).

3.2 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 exist for this analysis: keyword 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., 2009b), 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 (Litman and Forbes, 2003, Batliner et al., 2003a), 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 et al., 2003) and *Bag-of-Words* (*vector space modelling*), cf. Schuller et al. (2005) and Batliner et al. (2006b); these will be dealt with in the following.

A first step will always be the pre-processing of the text. This seems an easy task for written text, yet, soft string matching (e.g. by Levenshtein Distance) is reported to be advantageous to overcome misspelling, or spelling variations, dialects, etc. Considering analysis from spoken text, only few results for emotion recognition rely on ASR output (Schuller et al., 2005, 2009b) rather than on manual annotation of the data (Batliner et al., 2006b). This comes, as ASR of emotional speech itself is a challenge (Athanaselis et al., 2005; Schuller et al., 2006a, 2007b, 2009b) and might be error prone.

Second, an inventory of term entities, known as vocabulary, needs to be constructed which initially consists of all different words observed in the training corpus; this usually amounts to several thousands. (Note that for instance the balanced affective wordlist (Siegle, 1995) consists of only roughly 300 words.) Eventually, the vocabulary has to be reduced somehow, by stopping or by stemming.

Stopping resembles elimination of irrelevant words. The traditional approach to stopping is an expert-based list of words as 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 such as salience or information gain-based reduction (see below) are popular. The easiest, yet often effective way, is also stopping by the general minimum frequency of occurrence within a training corpus.

Stemming stands for clustering of morphological variants, i.e. flexions (e.g. by declination or conjugation), 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 simple (character) N-gram stemming, cf. below, or by (Iterated) Lovins or Porter stemmers that base on suffix lists and rules for their application (Lovins, 1968; Porter, 1980). A very compact approach to stemming is the use of so-called part-of-speech (POS) classes, such as nouns, verbs, adjectives, particles (Batliner et al., 2006b). 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). In addition, non-linguistic

vocalisations like sighs and yawns (Russell et al., 2003), laughs (Campbell et al., 2005; Truong and van Leeuwen, 2005), cries (Pal et al., 2006), and coughs (Matos et al., 2006) can easily be integrated into the vocabulary (Batliner et al., 2006b; Schuller et al., 2006b, 2009a).

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. Following Zipf's principle of least effort stating that irrelevant function words occur very frequently opposing terms of interest, the number of considered words is reduced to N in order to prevent over-modelling. In addition, word class-based N-grams can be used as well, to better cope with data sparseness.

Nonetheless, in emotion recognition, mostly uni-grams (N=1) have been applied so far (Lee et al., 2002; Devillers et al., 2003), 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.

Bag-of-Words is a well-known numerical representation form of text in automatic document categorisation (Joachims, 1998). It has been successfully ported to recognise sentiments (Pang et al., 2002) or emotion (Schuller et al., 2005, 2006b). 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. Also, it is possible not to refer to words, but sequences of them, i.e. Bags-of-N-grams, to overcome the lack of word order modelling (Schuller et al., 2009b).

Note that most vector elements will resemble zero, as feature vectors are constructed for short utterances rather than for longer texts, as in document retrieval, and only few words of the vocabulary will be seen. Support vector machines (cf. below) show high performance for this task. The possibility of early fusion with acoustic features helped make this technique very popular (Schuller et al., 2006b; Batliner et al., 2006b).

The preponderance of acoustics in emotion modelling so far is conditioned by the traditional focus on segmentally identical, acted utterances. For naturalistic data, both acoustic and linguistic features should be employed, both for a deeper understanding and a better classification performance. Basic feature extraction and subsequent computation of structured features employing (combinations of) functionals will certainly be the subject of much research in the future, examined in different contexts. We are far from knowing which feature (type) models best which emotional states in which context. Thus we have to resort to the general advise to use a representative set of features of different types rather than only one type of feature.

4 Classification

The data-driven way to evaluate extracted features and classification performance is to rely on machine learning and/or pattern recognition techniques: we let the machine find and learn regularities in the data. In the past decades, a prolific amount of methods has emerged for automatic modelling and extraction of informative patterns of the data. The number of successive refinements and slight variations of each machine learning algorithm is even bigger. One challenge to address in emotion classification is how to prune into this depth of options and find a good method for this specific task.

Emotion recognition from speech has to deal with noisy, redundant, and correlated features. Furthermore speech feature vectors are often complex and large, contaminated with interferences, background noise, and overlapping signals; this is especially true for naturalistic emotional speech. Thus different studies have shown that the same feature vector can yield very different classification results using different algorithms.

4.1 The Curse of Dimensionality and the Sparse Data Problem

Realistic emotional speech databases are characterised by the following problems: (1) small number of patterns, (2) potentially high number of features, and (3) skewed classes. Typically such databases comprise some hundreds of labelled utterances, while the features for classifying them can be chosen within a high-dimensional space, usually up to some hundreds as well. As the amount of available data is usually fixed, any increase in the feature space rapidly (exponentially in the number of features) leads to regions of the feature space where data are very sparse. This problem is known as 'curse of dimensionality' (Bellman, 1961), and it affects classifiers that divide the feature space into cells. A good rule of thumb requires that the number of patterns should never be lower than twice the number of features. Although some classifiers implicitly and successfully cope with the curse of dimensionality, pre-processing methods such as 'feature selection' and 'feature reduction' are generally applied to the input space. A favourable by-product of reducing the feature space is the reduction of the computational burden and implementation complexity while training the classifier. Both should not be underestimated: the former may lead to no solution at all (in reasonable time), the latter can yield wrong results due to numerical instabilities and overflows. Furthermore, feature reduction and selection methods selectively proceed to discard correlated and non-relevant features, resulting in higher reliability of the results.

Feature reduction consists in the mapping of the input space onto a less-dimensional one, while keeping as much information as possible. Common reduction techniques used in the field of emotion recognition are principal component analysis (PCA), linear discriminant analysis (LDA) and more sophisticated derivations like heteroscedastic discriminant analysis (Ayadi et al., 2007) and independent

component analysis (ICA). PCA is the feature transformation that minimises the sum of square error (Jolliffe, 2002). Furthermore, the base of the new space is orthonormal, which means that PCA de-correlates the original features: new features are constructed as linear superpositions so that the first one explains the largest amount of total variance of the data while each subsequent component explains the largest amount of the remaining variance while remaining uncorrelated with previously constructed features. The use of PCA requires the guess of the dimensionality of the target space. This can be done by the Kaiser–Guttman test, Log-Eigenvalue (LEV) diagram, Cattell's scree test (broken stick model), cross-validation, etc.

While PCA is an unsupervised feature reduction method (and thus maybe sub-optimal 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 (Fukunaga, 1990). LDA is less used as feature reduction, but it is widely adopted for direct classification (Lee and Narayanan, 2005; Kwon et al., 2003; Batliner et al., 2000a). Finally ICA (Hyvärinen et al., 2001) is the transformation that maps the feature space into an orthogonal space; furthermore, the target features are independent. Both theoretical and practical assumptions must hold, like the non-Gaussianity of the input features and the low dimensionality of the transformed space. There are already some studies adopting ICA (Rahurkar and Hansen, 2003), where both the input space and the output space are kept small.

Feature reduction is not appropriate for feature mining, as the original features are not retained after the transformation. *Feature selection* denotes a set of techniques that remove features which are irrelevant for modelling. This is a combinatorial optimisation problem: the feature space is traversed and at each step of the search, a different feature combination is evaluated. Evaluation is usually done following two possible strategies: the closed-loop 'wrapper' method, which trains and re-evaluates a given classifier at each search step using accuracy as objective function and the open-loop 'filter' method, which maximises simpler objective functions. While a wrapper can consist of any classifier, filter objective functions are usually measures such as information gain ratio (Witten and Frank, 2005) or inter-feature and feature-class correlation (Hall, 1998). As an exhaustive search through all possible feature combinations is unfeasible, faster but sub-optimal search functions are chosen. Most popular thereby is hill-climbing search or random injection as within random or genetic search. Typical hill-climbing procedures are sequential forward (SFS) and backward (SBS) selection by adding (deleting) at each search step the feature reporting the best performance according to the chosen wrapper or filter. SFS and SBS are commonly used (Lee et al., 2001; Lee and Narayanan, 2005; Kwon et al., 2003). Sequential floating forward selection (SFFS) (Pudil et al., 1994; Jain and Zongker, 1997) is an improved SFS method in the sense that at each step, previously selected features are considered for being discarded from the optimal group (SBS steps) to overcome nesting effects. Experiments show SFFS to dominate over other methods (Jain and Zongker, 1997). Note that a good feature selection should de-correlate the feature space to optimise a set of features as opposed to sheer ranking of features. This is in particular the case for wrapper search, which at the

same time usually demands considerably higher computational effort. Some studies combine feature selection with feature generation to find better representations and combinations of features by simple mathematical operations such as addition, multiplication, or reciprocal value of features (Batliner et al., 2006b). Also, balancing of the training instances with respect to instances per emotion class may be done before feature selection if these are highly skewed (Schuller et al., 2009c).

With the growing interest in spontaneous data, class skewness or the 'sparse data' problem in the output space came to the fore: many classes are characterised by few observations only. Normally, most cases belong to the neutral class. The skewness of the output space can be addressed by considering proper class weights, by resampling, i.e. (random) up- or down-sampling, or by introducing main classes (clustering similar classes under the same hat). The most frequent couples of main classes are 'neutral vs. non-neutral' and 'positive vs. negative' emotions modelling the 'valence' dimension, where neutral generally encompasses the absence of any emotion while 'positive' emotions span from neutral to happiness.

4.2 Classifiers

A number of reasons speak for considering diverse classifiers for different tasks: mostly high recognition rates (e.g. ability to solve non-linear problems, learn discriminatively, adapt online, generalise, tolerate high dimensionality), adequate modelling (static or dynamic, data- or knowledge-based, model or instance-based, handling of missing feature values and uncertainty, training stability), efficiency and economical factors (real-time capability, low computational cost for training and recognition, low memory requirement, need of only few exemplary instances, easy implementation), and optimal integration in a system context, e.g. (class-wise) provision of confidences or handling of input confidence. These considerations, and the simple availability of implementations in toolboxes such as WEKA (Witten and Frank, 2005) or HTK (Young et al., 2006), led to a considerable bandwidth of variants being used in the recognition of emotion from speech.

Very popular classifiers for emotion recognition are linear discriminant classifiers (LDCs) (Fukunaga, 1990) and k-nearest neighbour (kNN) classifiers (Cover and Hart, 1967): their implementation is easy, the time needed for training is short, unbalanced classes can be handled, and the sensitivity to lack of data in general is small. kNN is a lookup method: the training data are simply stored ('lazy' or instance-based learning, as opposed to model building classifiers) and each new pattern is assigned by averaging its nearest neighbour classes. They are widely used (Dellaert et al., 1996; Petrushin, 1999), with good results for non-acted emotional speech as well (Lee and Narayanan, 2005; Shami and Verhelst, 2007). LDC – as a natural extension of LDA, see Fukunaga (1990) – is basically a classifier with straight-line decision surfaces (hyperplanes). LDA is one possible method of estimating LDC hyperplane parameters by maximisation of class separability (see above). They have often been used (Lee and Narayanan, 2005; Rahurkar and

Hansen, 2003; Kwon et al., 2003; Litman and Forbes, 2003; Batliner et al., 2000a), with a competitive performance (Batliner et al., 2006b) in spite of some limitations: the data should be linearly separable and the method is sensitive to outliers. A natural extension of LDCs is support vector machines (SVM): if the input data have previously implicitly undergone a non-linear transformation, 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 (Vapnik, 1995). SVM provide very good generalisation properties (McGilloway et al., 2000; Lee et al., 2002; Chuang and Wu, 2004; You et al., 2006; Morrison et al., 2007); thus, they became increasingly popular. Note, however, that their performance is not always (way) better than the one obtained by using alternative classifiers (Meyer et al., 2002).

Among the most used non-linear discriminative classifiers are artificial neural networks (ANNs) and decision trees. Feedforward ANNs, also known as multilayered perceptrons, are equivalent to fitting pre-defined non-linear functions to some given data. Decision surfaces 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 (Rumelhart et al., 1986), and on the activity rules (how the input patterns and the ANN weights are combined to obtain a decision output class). ANNs are therefore not robust to overfitting and require greater amounts of data to be trained on. Therefore, ANNs are rarely used for acted data (Petrushin, 1999; Martinez and Cruz, 2005) and even less for non-acted, but cf. (Batliner et al., 2000a, 2006b). Recurrent networks can further be complemented by long short-term Memory to integrate emotional context (Wöllmer et al., 2008). Although they are also characterised by the property of handling non-linearly separable data, decision trees are less of a 'black box' compared to SVMs or neural networks, since they are based on simple recursive splits of the data. These splits (yes/no questions usually ranked by information gain) are very readable, especially if the tree has been adequately pruned, i.e. cutoff according to the ranking. Popular decision tree algorithms are C4.5 (Quinlan, 1993) and CART (Breiman et al., 1984). Note, however, that accuracy degrades in case of irrelevant features or noisy patterns. A solution is random forests (RF) (Breiman, 2001), an ensemble of trees each one accounting for a random subset of the input features and learned on variants of the training set by sampling with replacement. They are practically insensitive to the curse of dimensionality (Schuller et al., 2007a).

Apart from the already named kNN, which can be seen as a very basic statistical classifier, one also basic representative of this group is the Naive Bayes classifier (Langley et al., 1992; Good, 1965). It is robust with respect to irrelevant features but its performance may degrade quickly if correlated – even relevant – features are added. Less 'naïve' are Gaussian mixture models (GMM) that employ a number of multivariate Gaussians to model the original densities in the feature space. However, this of course also requires more training data.

Dynamic classifiers like hidden Markov models (HMM), dynamic Bayesian networks (DBN), or simple dynamic time warp (DTW) implicitly warp observed

feature sequences over time. No further processing of the raw feature contours on a per-frame basis as pitch or energy is needed (like the application of functionals, to obtain the same number of features for different lengths of units such as turns or words). Among dynamic classifiers, apparently only HMM were studied yet, probably mostly because of the presence of well-elaborated tools such as HTK. For acted emotion there are numerous references as given by ten Bosch (2003); Schuller et al. (2003), and Zeng et al. (2009); for non-acted emotion, fewer are known (Kwon et al., 2003; Wagner et al., 2007; Vlasenko et al., 2007b; Schuller et al., 2009c). The performance of static modelling is usually not reached (ten Bosch, 2003; Schuller et al., 2003), as emotion apparently is better modelled on a timescale above frame level; note that a combination of static features such as minimum, maximum, onset, offset, duration, regression implicitly shape contour dynamics as well. Still, when the spoken content is fixed, the combination of static and dynamic processing may help improve overall accuracy (Vlasenko et al., 2007a). However, it is not clear whether emotion can be satisfyingly modelled using the simplifying Markov assumption that underlies HMM modelling (ten Bosch, 2003).

Ensembles of classifiers (Schuller et al., 2005) combine their individual strengths, and overcome training instability deriving from the sparseness of data. In the highly popular *Bagging* (Breiman, 1996) method, several instances of the same classifier are trained on sub-samples of the data set, usually of the same size, obtained by sampling with replacement. The final decision is then made by majority voting. *Boosting* decides by weighted majority voting after iteratively assigning (high) weights for hardly separable instances throughout learning. Next, *MultiBoosting* combines bias and variance reduction of these two methods by their sequential application. Most powerful, however, is the combination of diverse classifiers by either simple *Voting* (Morrison and Silva, 2007) or 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, one speaks of *StackingC*. Still, the gain over single strong classifiers as SVM may not justify the extra computational need.

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).

Regression – that is mapping on a continuum rather than on discrete classes – is also used in emotion recognition to handle the dimensional approach. Usually each axis, such as arousal, valence, or dominance, is thereby taken care of by one regression model such as support vector regression (Grimm et al., 2007) or less complex solutions such as multiple linear regression.

Features belonging to different types, e.g. acoustic and linguistic features, can be combined in *early fusion* within the same classifier or the class assignment with or without confidence measures obtained with different classifiers using different features can be combined in *late fusion*, cf. the ROVER approach (Fiscus, 1997) used in Batliner et al. (2006b).

4.3 Evaluation

To assess the performance of a classifier, we have to split the data into train and test. The easiest approach is a percentage split. However, data in emotion recognition are usually sparse, as mentioned. Therefore, it seems desirable to test on all instances: the training set is thereby usually kept as large as possible, the limit being a single pattern at a time for testing; this is repeated j times changing the tested pattern each time. Such a high number of trainings can be infeasible. Splitting the data into $j = 10$ parts, training on 9 parts, and testing on the remaining data is a good, popular compromise, called j-fold cross validation, cf. Salzberg (1997). Throughout partitioning of the data the distribution among classes should be kept, known as stratification. However, the partitioning is usually not explicitly stated, thus not easily allowing for comparative studies. Also, if it is not speaker independent, recognition performance will be too optimistic. Both these downsides can be overcome by leave-one-speaker-out, meaning training with all but one speaker in each cycle (Steidl, 2009), or leave a known group of speakers out to spare computational effort.

Most of the studies report performance measures expressed by accuracy, i.e. the recognition rate (RR), also known as weighted average (WA): the number of correctly recognised patterns divided by the total number of patterns. Given the skewness of spontaneous emotional databases, this is rarely appropriate. A possibility is to measure both, Precision (P, the number of true positives over all positive patterns) and Recall (R, the number of true positives over the number of all reference patterns). When there are more than two classes, it is useful to give a P- and an R-value for each class separately. In this sense R of a class corresponds to the RR of this class. As a general measure over the entire data is useful, we can introduce the mean of the accuracies (RR) over all classes, i.e. the class-wisely averaged classification rate (CL), also known as unweighted average (UA) recall in contrast to weighted average (WA) recall resembling RR (Schuller et al., 2009c). Note that RR and CL for a balanced multi-class recognition problem are always identical; the more the class distribution is unbalanced, the higher the difference between RR and CL. To have a unique measure of the goodness of classification – for comparison aims – the F-measure can be used; it is the harmonic mean of P and R. A similar score can be obtained by averaging UA and WA. The receiver operating characteristic (ROC) curve is independent of the data distribution but has the disadvantage that curves are not easy to compare. It is the plot of R over 1-Specificity (S, the false negative over all negatives). ROC curves are constructed by modifying a threshold during the training of the classifier. Different thresholds correspond to different performance of the classifier (in terms of Recall and Specificity) and thus to different points on the ROC curve, cf. Steidl et al., (2009).

The complete source of information is the confusion matrix. The figures described above all derive from it and try to highlight or smooth some aspects, especially for multiple classes when it might be difficult to interpret or during the training of a classifier when optimisation is achievable only w.r.t. few or one single parameter such as accuracy or F-measure.

Studies eventually end up with the conclusion that a specific classifier is better than another one – which is a conclusion that must not be generalised. Most of the

time no significance of the differences is reported. Actually, there are some reasons to handle significance tests with care, for general reasons (Nickerson, 2000) and because of repeated measurements: the more experiments we do on a certain data set, the more probable it is that we accidentally run into some significant results. Significance thresholds should be augmented whenever we increment the number of experiments; however, this is not done very often. The Bonferroni adjustment is a possible choice of a correction factor. For a cookbook on multi-experiment studies, see Salzberg (1997). There are some drawbacks of the Bonferroni correction as it is usually too conservative; these are outlined in Pernegger (1998).

Also, when doing comparative evaluations, 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 have been used for parameter tuning: they describe how the data have been divided into test and training but nothing is said about held-out data for classifier tuning, i.e. a development set; this should be part of future investigations.

Finding, fine-tuning, and evaluating classifiers is a broad topic in its own; although there might be preferences to use one or the other approach in specific fields – such as emotion recognition, it generally suffers from too many degrees of freedom: a strict comparison across studies is practically never possible. Statements such as 'it has been proved that classifier X is superior to classifier Y', should never be generalised. Often it only means that there has been more fine-tuning for X than for Y. In the long run, it might turn out that specific models and classifiers based on them are – on the average – better suited for emotion recognition. However, searching for an optimal classifier alone will not be a panacea; it will not improve unsatisfying recognition rates to such an extent that the intended application will be successful. Anyway, it should be mandatory to document the steps explicitly, e.g. whether a cross-validation has been done speaker-independently or in a speaker-dependent way. This statement holds similarly for comparison across whole studies: what never should be done is simply to compare recognition rates between two studies. Such performance depends crucially on too many factors which have not been standardised yet.

5 Applications

Apart from some 'offline' applications such as data mining in movie archives or screening call-centre agents as for their behaviour against customers, the ultimate goal of the whole endeavour described in this chapter is employing classified emotional user states in an end-to-end system; by end-to-end system we mean 'spontaneous speech, produced by human users as input – generated system reaction such as synthesised speech, produced by the system, as output, and vice versa'.

Several systems have been envisaged so far (Batliner et al., 2006a); example applications are depicted in Burkhardt et al., (2009) and Vogt et al. (2009). The contribution of automatic classification is rather straightforward: each speech unit such as words/chunks/turns/dialogue moves is attributed one out of a rather reduced set of emotion labels, maybe with some probability or confidence measure. This attribution can be correct or wrong – basically the same way as human beings can be right or wrong or disagreeing when estimating the emotions of other human beings. In both cases, some cost function has to be established – is it costly, or does it not matter at all, whether I attribute the wrong emotion or the right one? But it is not only an erroneous classification of emotion which can cause erroneous results: ASR is not perfect. We do not fully know yet whether emotional speech causes more speech recognition errors because it is more difficult than 'normal' speech, or because we simply do not have enough data of this variety to train an ASR engine successfully (Athanaselis et al., 2005; Schuller et al., 2007b, 2009b). In real-life settings, chances are that a worse signal-to-noise ratio will deteriorate ASR and by that, emotion classification; especially using linguistic features might not yield good recognition performance. If ASR is erroneous, this will result in erroneous words and erroneous segmentation, so both acoustic and linguistic features might be computed in a sub-optimal way, resulting in lower classification performance. The impact of erroneous extraction might not be too high if acoustic features are used, cf. Schuller et al., (2007b), but might be problematic if only linguistic information is exploited (Seppi et al., 2008a). Moreover, erroneous ASR is of course not really helpful for processing the user's semantic/pragmatic intentions within the whole system.

ASR normally aims at speaker-independent modelling and recognition; this is state of the art in our field as well. Speaker-dependent processing yields better recognition performance; we want to point out that even if speaker independency is, of course, the ultimate goal, we can imagine applications where speaker-dependent modelling is possible and makes sense. This will always be the case when the speaker can be identified and is a frequent user of the system.

The exchange format with other modules within a full end-to-end system is nowadays normally some XML dialect, cf. Schröder et al. in this volume. However, we do not know yet of any system where really speech and not written language has been used as input into such a representation and subsequent use within a full system – apart from the SmartKom system (Streit et al., 2006) where an implementation of the OCC model (Ortony et al., 1988) had to be restricted to some few so-called use cases. It could be shown that the module was functional on a principled basis in the whole end-to-end system; however, it has to await much more testing and more robust recognition modules to be functional in any practical application.

In this section we want to point out that even if we solved somehow the problems we addressed in this chapter, this is not the end of the story because most of the time, we will have to use ASR output within a 'real system' – and this output inevitably can be erroneous which in turn can cause erroneous processing of not only emotion attributions.

6 Concluding Remarks

In this chapter, we gave an overview of the state of the art in the automatic recognition of real life, natural emotional user states, pointing out problems, pitfalls, and to-do's and not-to-do's. We deliberately refrained from comparing classification performance across studies in terms of recognition rates – this cannot be done in a serious way and would be misleading. We dealt with the full sequence of processing, from conceptualisation to recognition rates, although mostly not in an in-depth manner. We hope to have introduced almost all of the pertinent topics; the references can be used for more detailed information.

As for the future of our topic, the pivotal desideratum is databases; a comparable albeit way easier problem that somehow has been 'solved' – i.e. a satisfying recognition performance has been obtained – in recent time is the performance of automatic dictation systems. Here, the breakthrough came with the use of training material larger by some order of magnitude. However, already the basic unit is not comparable: whereas there can be a fair agreement on what a word is and which word has been produced, there is neither full agreement on what an emotion is nor on the way how to obtain the ground truth, i.e. the types and tokens we want to recognise. Moreover, the creation of databases is expensive, and progress will be slow. Even if the field is emerging – which can be seen from the growing number of contributions to conferences and journal papers – the methodological problem is that practically always, results cannot be compared across studies because too many factors are not kept constant. A few studies have begun to address different databases using the same approaches, cf. Shami and Verhelst (2007) and Batliner et al. (2008a). Initiatives such as CEICES, cf. Batliner et al. (2006b), combining thoroughly annotated data with the fusion of a plethora of different feature types, generated at different sites, might be one way of establishing 'islands of standardisation', i.e. making comparisons across classifiers and features easier and more reliable. The Interspeech 2009 Emotion Challenge, cf. Schuller et al. (2009c), has been the first attempt towards strict comparability and reproducibility of emotion classification results. However, further steps in this direction will be needed to provide comparability among researchers for a multiplicity of remaining challenges.

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: Proceedings of the Interspeech, Pittsburgh, PA, September 17–21, pp 797–800

Ang J, Dhillon R, Shriberg E, Stolcke A (2002) Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In: Proceedings of the Interspeech, Denver, September 16–20, pp 2037–2040

Arunachalam S, Gould D, Anderson E, Byrd D, Narayanan S (2001) Politeness and frustration language in child-machine interactions. In: Proceedings of the Eurospeech, Aalborg, September 3–7, pp 2675–2678

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 Netw. 18:437–444

Ayadi MMHE, Kamel MS, Karray F (2007) Speech emotion recognition using gaussian mixture vector autoregressive models. In: Proceedings of ICASSP, Honolulu, April 15–20, pp 957–960

Batliner A, Kompe R, Kießling A, Mast M, Niemann H, Nöth E (1998) M = Syntax + Prosody: a syntactic–prosodic labelling scheme for large spontaneous speech databases. Speech Communi 25(4):193–222

Batliner A, Fischer K, Huber R, Spilker J, Nöth E (2000a) Desperately Seeking Emotions: Actors, Wizards, and Human Beings. In: Proceedings of the ISCA workshop on speech and emotion, Newcastle, Northern Ireland, September 5–7, pp 195–200

Batliner A, Huber R, Niemann H, Nöth E, Spilker J, Fischer K (2000b) The recognition of emotion. In: Wahlster W. (ed) Verbmobil: Foundations of speech-to-speech translations. Springer, Berlin, pp 122–130.

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: Proceedings of the Eurospeech, Aalborg, September 3–7, pp 2781–2784

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

Batliner A, Zeissler V, Frank C, Adelhardt J, Shi RP, Nöth E (2003b) We are not amused - but how do you know? User states in a multi-modal dialogue system. In: Proceedings of the Interspeech, Geneva, September 1–4, 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: Affective dialogue systems, proceedings of a tutorial and research workshop, Kloster Irsee, June 14–16, 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: Proceedings of the Interspeech, Lisbon, September 4–8, pp 489–492

Batliner A, Burkhardt F, van Ballegooy M, Nöth E (2006a) A taxonomy of applications that utilize emotional awareness. In: Proceedings of IS-LTC 2006, Ljubliana, October 9–10, 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: Proceedings of IS-LTC 2006, Ljubliana, October 9–10, pp 240–245

Batliner A, Steidl S, Nöth E (2007a) Laryngealizations and Emotions: How Many Babushkas? In: Proceedings of the international workshop on paralinguistic speech – between models and data (ParaLing'07), Saarbrücken, August 3, pp 17–22

Batliner A, Steidl S, Schuller B, Seppi D, Vogt T, Devillers L, Vidrascu L, Amir N, Kessous L, Aharonson V (2007b) The impact of F0 extraction errors on the classification of prominence and emotion. In: Proceedings of the ICPhS, Saarbrücken, August 6–10, pp 2201–2204

Batliner A, Schuller B, Schaeffler S, Steidl S (2008a) Mothers, adults, children, pets — towards the acoustics of intimacy. In: Proceedings of the ICASSP 2008, Las Vegas, NV, March 30–April 04, 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-Adap Interact 18:175–206

Bellman R (1961) Adaptive control processes. Princeton University Press, Princeton, NJ

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. Wiley, New York, NY, pp 209–243

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

Breiman L (1996) Bagging predictors. Mach Learn 26:123–140

Breiman L (2001) Random forests. Mach Learn 45:5–32

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth and Brooks, Pacific Grove, CA

Burger S, Weilhammer K, Schiel F, Tillman HG (2000) Verbmobil data collection and annotation. In: Wahlster W. (ed) Verbmobil: foundations of speech-to-speech translations. Springer, Berlin, pp 537–549

Burkhardt F, van Ballegooy M, Engelbrecht K-P, Polzehl T, Stegmann J (2009) Emotion detection in dialog systems: applications, strategies and challenges. In: Proceedings of the ACII, Amsterdam, September 10–12, pp 684–689

Campbell N, Kashioka H, Ohara R (2005) No laughing matter. In: Proceedings of the Interspeech, Lisbon, September 12–14, pp 465–468

Chuang Z-J, Wu C-H (2004) Emotion recognition using acoustic features and textual content. In: Proceedings of ICME, Taipei, June 27–30, pp 53–56

Cover T, Hart P (1967) Nearest neighbor pattern classification. IEEE Trans Info Theoy 13:21–27

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. World Scientific Engineering 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: Proceedings of the ISCA Workshop on Speech and Emotion, Newcastle, Northern Ireland, September 5–7, pp 19–24

Cowie R, Douglas-Cowie E, Cox C (2005) Beyond emotion archetypes: databases for emotion modelling using neural networks. Neural Netw 18:371–388

Craggs R, Wood MM (2004) A categorical annotation scheme for emotion in the linguistic content of dialogue. In: Affective dialogue systems, proceedings of a tutorial and research workshop, Kloster Irsee, June 14–16, pp 89–100

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

Dellaert F, Polzin T, Waibel A (1996) Recognizing emotion in speech. In: Proceedings of the ICSLP, Philadelphia, PA, October 3–6, pp 1970–1973

Devillers L, Vasilescu I, Lamel L (2003) Emotion detection in task-oriented spoken dialogs. In: Proceedings of ICME 2003, IEEE, multimedia human-machine interface and interaction, Baltimore, MD, July 6–9, pp 549–552

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

Douglas-Cowie E, Cowie R, Sneddon I, Cox C, Lowry O, McRorie M, Martin J-C, Devillers L, Abrilan S, Batliner A, Amir N, Karpousis K (2007) The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data. In: Paiva A, Prada R, Picard RW, (eds), Affective computing and intelligent interaction. Springer, Berlin, pp 488–500

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

Eyben F, Wöllmer M, Schuller B (2009) openear - introducing the munich open-source emotion and affect recognition toolkit. In: Proceedings of the ACII, Amsterdam, September 10–12, pp 576–581

Fairbanks G, Pronovost W (1939) An experimental study of the pitch characteristics of the voice during the expression of emotion. Speech Monogr, 6:87–104

Fernandez R, Picard RW (2003) Modeling drivers' speech under stress. Speech Commun 40: 145–159

Fiscus J (1997) A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER). In: Proceedings of the ASRU, Santa Barbara, CA, December 14–17, 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

Frick R (1985) Communicating emotion: the role of prosodic features. Psychol Bull 97:412–429
Fujisaki H (1992) Modelling the process of fundamental frequency contour generation. In: Tohkura Y, Vatikiotis-Bateson E, Sagisasaka Y, (eds), Speech perception, production and linguistic structure. IOS Press, Amsterdam, pp 313–328
Fukunaga K (1990) Introduction to statistical pattern recognition. Academic Press, London
Goertzel B, Silverman K, Hartley C, Bugaj S, Ross M (2000) The baby webmind project. In: Proceedings of the annual conference of the society for the study of artificial intelligence and the simulation of behaviour (AISB), Birmingham, April 17–20
Good I (1965) The estimation of probabilities: an essay on modern bayesian methods. MIT Press, Cambridge, MA
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 RW, (eds), Affective computing and intelligent interaction. Springer, Berlin, pp 126–138
Grimm M, Kroschel K, Narayanan S (2008) The vera am mittag german audio-visual emotional speech database. In: Proceedings of the IEEE international conference on multimedia and expo (ICME), Hannover, Germany, June 23–26, pp 865–868
Hall MA (1998) Correlation-based feature selection for machine learning. Ph.D. thesis, Department of Computer Science, Waikato University, Hamilton, NZ
Hermansky H (1990) Perceptual linear predictive (plp) analysis for speech. J Acoust Soc Am (JASA), 87:1738–1752
Hermansky H, Sharma S (1998) Traps - classifiers of temporal patterns. In: Proceedings of the ICSLP, Sydney, November 30–December 04, pp 1003–1006
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, New York, NY, pp 363–383
Hirst D, Cristo AD, Espesser R (2000) Levels of representation and levels of analysis for intonation. In: Horne M, (ed), Prosody : theory and experiment Kluwer, Dordrecht, pp 51–87
Hyvärinen A, Karhunen J, Oja E (2001) Independent component analysis. Wiley, New York, NY
Jain A, Zongker D (1997) Feature selection: evaluation, application and small sample performance. PAMI 19(2):153–158
Joachims T (1998) Text categorization with support vector machines: learning with many relevant features. In: Nédellec C, Rouveirol C, (eds), Proceedings of ECML-98, 10th European conference on machine learning. Springer, Heidelberg, pp 137–142
Johnstone T, Scherer KR (2000) Vocal communication of emotion. In: Lewis M, Haviland-Jones JM, (eds), Handbook of emotions, chapter 14. 2nd edn. Guilford Press, London
Jolliffe IT (2002) Principal component analysis. Springer, Berlin
Kießling A (1997) Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung. Berichte aus der Informatik. Shaker, Aachen
Kwon O-W, Chan K, Hao J, Lee T-W (2003) Emotion recognition by speech signals. In: Proceedings of the Interspeech, Geneva, September 1–4, pp 125–128
Langley P, Iba W, Thompson K (1992) An analysis of Bayesian classifiers. In: Proceedings of the national conference on articial intelligence, San Jose, CA, pp 223–228
Lee C, Narayanan S, Pieraccini R (2001) Recognition of negative emotions from the speech signal. In: Proceedings of the ASRU, Madonna di Campiglio, December 9–13, no pagination
Lee CM, Narayanan SS (2005) Toward detecting emotions in spoken dialogs. IEEE Trans Speech Audio Process 13(2):293–303
Lee CM, Narayanan SS, Pieraccini R (2002) Combining acoustic and language information for emotion recognition. In: Proceedings of the Interspeech, Denver, September 16–20, pp 873–376
Lee CM, Yildirim S, Bulut M, Kazemzadeh A, Busso C, Deng Z, Lee S, Narayanan SS (2004) Emotion recognition based on phoneme classes. In: Proceedings of the Interspeech, Jeju Island, Korea, October 4–8, pp 889–892

Litman D, Forbes K (2003) Recognizing emotions from student speech in tutoring dialogues. In: Proceedings of the ASRU, Virgin Island, November 30–December 3, pp 25–30

Liu H, Liebermann H, Selker T (2003) A model of textual affect sensing using real-world knowledge. In: Proceedings of the 7th International conference on intelligent user interfaces (IUI 2003), Miami, Florida, USA, January 12–15, pp 125–132

Lovins JB (1968) Development of a stemming algorithm. Mech Transl Comput Linguist 11:22–31

Lugger M, Yang B, Wokurek W (2006) Robust estimation of voice quality parameters under real world disturbances. In: Proceedings of the ICASSP, Toulouse, May 15–19, pp 1097–1100, 2006

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

Martinez CA, Cruz A (2005) Emotion recognition in non-structured utterances for human-robot interaction. In: IEEE international workshop on robot and human interactive communication, August 13–15, pp 19–23, 2005

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 pp 1078–108

McGilloway S, Cowie R, Douglas-Cowie E, Gielen S, Westerdijk M, Stroeve S (2000) Approaching automatic recognition of emotion from voice: A rough benchmark. In: Proceedings of the ISCA workshop on speech and emotion, Newcastle, Northern Ireland, September 5–7, pp 207–212

Meyer D, Leisch F, Hornik K (2002) Benchmarking support vector machines. Report series no. 78, SFB Adaptive informations systems and management in economics and management science, Wien, Austria, 19 pp

Morrison D, Silva LCD (2007) Voting ensembles for spoken affect classification. J Netw Comput Appl 30:1356–1365

Morrison D, Wang R, Xu W, Silva LCD (2007) Incremental learning for spoken affect classification and its application in call-centres. Int J Intell Syst Tech: 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: Proceedings of the ACII, Amsterdam, pp 662–669

Neiberg D, Elenius K, Laskowski K (2006) Emotion Recognition in Spontaneous Speech Using GMMs. In: Proceedings of the Interspeech, Pittsburgh, PA, September 17–21, pp 809–812

Nickerson RS (2000) Null hypothesis significance testing: a review of an old and continuing controversy. Psychol Methods 5:241–301

Noll AM (1967) Cepstrum pitch determination. J Acoust Soc Am (JASA), 14:293–309

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 Commun, 36:(1–2), pp 45–62

Nwe T, Foo S, Silva LD (2003) Speech emotion recognition using hidden Markov models. Speech Commun 41:603–623

Ortony A, Clore GL, Collins A (1988) The cognitive structure of emotions. Cambridge University Press, Cambridge

Pal P, Iyer A, Yantorno R (2006) Emotion detection from infant facial expressions and cries. In: Proceedings of ICASSP, Toulouse, May 15–19, pp 809–812

Pang B, Lee L, Vaithyanathan S (2002) Thumbs up? sentiment classification using machine learning techniques. In: Proceedings of the 2002 conference on empirical methods in natural language processing (EMNLP), Philadelphia, PA, July 6–7, pp 79–86

Pernegger T.V (1998) What's wrong with Bonferroni adjustment. Br Med J, 316:1236–1238

Petrushin V (1999) Emotion in speech: recognition and application to call centers. In: Proceedings of artificial neural networks in engineering (ANNIE '99), St. Louis, MO, November 7–10, pp 7–10

Polzin TS, Waibel A (2000) Emotion-sensitive human-computer interfaces. In: Proceedings of the ISCA workshop on speech and emotion, Newcastle, Northern Ireland, September 5–7, pp 201–206

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 Recogn Lett 15:1119–1125

Quinlan JR (1993) C4.5: Programs for machine learning. Morgan Kaufmann, San Francisco, CA

Rabiner LR (1977) On the use of autocorrelation analysis for pitch detection. IEEE Trans Acoust Speech Signal Process 25:24–33

Rahurkar MA, Hansen JHL (2003) Towards affect recognition: an ICA approach. In: Proceedings of 4th international symposium on independent component analysis and blind signal separation (ICA2003), Nara, April 1–4, pp 1017–1022

Rosenberg A, Binkowski E (2004) Augmenting the kappa statistic to determine interannotator reliability for multiply labeled data points. In: Dumais DMS, Roukos S, (eds), HLT-NAACL 2004: short papers. Association for Computational Linguistics, Boston, MA, pp 77–80

de Rosis F, Batliner A, Novielli N, Steidl S (2007) 'You are Sooo Cool, Valentina!' Recognizing social attitude in speech-based dialogues with an ECA. In: Paiva A, Prada R, Picard RW, (eds), Affective computing and intelligent interaction, Springer, Berlin, pp 179–190

Rumelhart D, Hinton G, Williams R (1986) Learning internal representations by error propagation. In: Rumelhart D, McClelland L, the PDP Research Group, (eds), Parallel distributed processes: exploration in the microstructure of cognition, vol 1. MIT Press, Cambridge, MA, pp 318–362

Russel JA (1997) How shall an emotion be called? In: Plutchik R, Conte HR (eds), Circumplex models of personality and emotions, chapter 9. American Psychological Association, Washington, DC, pp 205–220

Russell J, Bachorowski J, Fernandez-Dols J (2003) Facial and vocal expressions of emotion. Ann Rev Psychol 54:329–349

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

Scherer KR (2003) Vocal communication of emotion: a review of research paradigms. Speech Commun 40:227–256

Scherer KR, Johnstone T, Klasmeyer G (2003) Vocal expression of emotion. In: Davidson RJ, Scherer KR, Goldsmith HH, (eds), Handbook of affective sciences, chapter 23. Oxford University Press, Oxford NY, pp 433–456

Schiel F (1999) Automatic phonetic transcription of non-prompted speech. In: Proceedings of the ICPhS, San Francisco, CA, August 1–7, pp 607–610

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 RW, (eds), Affective computing and intelligent interaction. Springer, Berlin, pp 440–451

Schuller B, Rigoll G (2009) Recognising interest in conversational speech – comparing bag of frames and supra-segmental features. In: Proceedings of the Interspeech, Brighton, UK, September 6–10, pp 1999–2002

Schuller B, Rigoll G, Lang M (2003) Hidden Markov model-based speech emotion recognition. In: Proceedings of the ICASSP, Hong Kong, April 6–10, 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: Proceedings of the ICASSP, Montreal, QC, Canada, May 17–21, pp 577–580

Schuller B, Müller R, Lang M, Rigoll G (2005) Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensemble. In: Proceedings of the Interspeech, Lisbon, September 4–8, pp 805–808

Schuller B, Stadermann J, Rigoll G (2006a) Affect-robust speech recognition by dynamic emotional adaptation. In: Proceedings of speech prosody 2006, Dresden, May 2–5, no pagination

Schuller B, Köhler N, Müller R, Rigoll G (2006b) Recognition of interest in human conversational speech. In: Proceedings of the Interspeech, Pittsburgh, PA, September 17–21, pp 793–796

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: Proceedings of the Interspeech, Antwerp, Belgium, August 27–31, pp 2253–2256

Schuller B, Seppi D, Batliner A, Meier A, Steidl S (2007b) Towards more reality in the recognition of emotional speech. In: Proceedings of the ICASSP, Honolulu, April 15–20, pp 941–944

Schuller B, Rigoll G, Can S, Feussner H (2008) Emotion sensitive speech control for human-robot interaction in minimal invasive surgery. In: Proceedings of the 17th International Symposium on robot and human interactive communication, RO-MAN 2008, Munich, Germany, August 1–3, pp 453–458

Schuller B, Müller R, Eyben F, Gast J, Hörnler B, Wöllmer M, Rigoll G, Höthker A, Konosu H (2009a) Being bored? Recognising natural interest by extensive audiovisual integration for real-life application. Image Vis Comput J, Special Issue on Vis Multimodal Anal Hum Spontaneous Behav 27:1760–1774

Schuller B, Batliner A, Steidl S, Seppi D (2009b) Emotion recognition from speech: putting ASR in the loop. In: Proceedings of ICASSP, Taipei, Taiwan. IEEE, April 19–24, pp 4585–4588

Schuller B, Steidl S, Batliner A (2009c) The INTERSPEECH 2009 emotion challenge. In: Proceedings of the Interspeech, Brighton, September 6–10, pp 312–315

Scripture E (1921) A study of emotions by speech transcription. Vox 31:179–183

Seppi D, Gerosa M, Schuller B, Batliner A, Steidl S (2008a) Detecting problems in spoken child-computer interaction. In: Proceedings of the 1st workshop on child, computer and interaction, Chania, Greece, October 23, no pagination

Seppi D, Batliner A, Schuller B, Steidl S, Vogt T, Wagner J, Devillers L, Vidrascu L, Amir N, Aharonson V (2008b) Patterns, prototypes, performance: classifying emotional user states. In: Proceedings of the Interspeech, Brisbane, September 22–26, pp 601–604

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, pp 43–56

Siegle G (1995) The balanced affective word list project. http://www.sci.sdsu.edu/CAL/wordlist/ (accessed October 17, 2010)

Skinner E (1935) A calibrated recording and analysis of the pitch, force, and quality of vocal tones expressing happiness and sadness. Speech Monogr 2:81–137

Slaney M, McRoberts G (1998) Baby Ears: A Recognition System for Affective Vocalizations. In: Proceedings of the ICASSP, Seattle, WA, pp 985–988

Steidl S (2009) Automatic classification of emotion-related user states in spontaneous children's speech. Berlin. PhD thesis, Logos Verlag

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), Text, speech and dialogue, 7th international conference, TSD 2004. Berlin, Heidelberg, pp 629–636

Steidl S, Levit M, Batliner A, Nöth E, Niemann H (2005) "Of all things the measure is man": automatic classification of emotions and inter-labeler consistency. In: Proceedings of ICASSP, Philadelphia, PA, May 12–15, pp 317–320

Steidl S, Schuller B, Batliner A, Seppi D (2009) The hinterland of emotions: facing the open-microphone challenge. In: Proceedings of ACII, Amsterdam, September 10–12, pp 690–697

Streit M, Batliner A, Portele T (2006) Emotions analysis and emotion-handling subdialogues. In: Wahlster W, (ed), SmartKom: foundations of multimodal dialogue systems. Springer, Berlin, pp 317–332

ten Bosch L (2003) Emotions, speech and the ASR framework. Speech Commun 40(1–2):213–225

Truong K, van Leeuwen D (2005) Automatic detection of laughter. In: Proceedings of the interspeech, Lisbon, Portugal, September 4–8, pp 485–488

Vapnik V (1995) The nature of statistical learning theory. Springer, Berlin

Ververidis D, Kotropoulos C (2006) Fast sequential floating forward selection applied to emotional speech features estimated on DES and SUSAS data collection. In: Proceedings of european signal processing Conference (EUSIPCO 2006), Florence, September 4–8, no pagination

Vlasenko B, Schuller B, Wendemuth A, Rigoll G (2007a) Combining frame and turn-level information for robust recognition of emotions within speech. In: Proceedings of Interspeech, Antwerp, Belgium, August 27–31, 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 RW, (eds), Affective computing and intelligent interaction. Springer, Berlin, pp 139–147

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: Proceedings of the ACII, Amsterdam, September 10–12, pp 670–677

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 RW, (eds), Affective computing and intelligent interaction. Springer, Berlin, pp 114–125

Williams C, Stevens K (1972) Emotions and speech: some acoustic correlates. J Acoust Soc Am (JASA) 52:1238–1250

Wilting J, Krahmer E, Swerts M (2006) Real vs. acted emotional speech. In: Proceedings of Interspeech, Pittsburgh, PA, September 17–21, pp 805–808

Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques, 2nd Edn. Morgan Kaufmann, San Francisco, CA

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: Proceedings of Interspeech, Brisbane, September 22–26, pp 597–600

Wolpert D (1992) Stacked generalization. Neural Netw 5:241–259

Wu T, Khan F, Fisher T, Shuler L, Pottenger W (2005) Posting act tagging using transformation-based learning. In: Lin TY, Ohsuga S, Liau C-J, Hu X, Tsumoto S, (eds), Foundations of data mining and knowledge discovery. Springer, Berlin, pp 319–331

You M, Chen C, Bu J, Liu J, Tao J (2006) Emotion recognition from noisy speech. In: Proceedings of ICME, Toronto, ON, July 9–12, pp 1653–1656

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

Zeng Z, Pantic M, Roisman GI, Huang TS (2009) A Survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans Pattern Anal Mach Intell 31(1):39–58

Zhe X, Boucouvalas A (2002) Text-to-emotion engine for real time internet communication. In: Proceedings of the international symposium on communication systems, networks, and DSPs. Staffordshire University, Stoke-on-Trent, July 15–17, pp 164–168

Zhou G, Hansen JHL, Kaiser J.F (2001) Nonlinear feature based classification of speech under stress. IEEE Trans Speech Audio Process 9:201–216