The automatic recognition of emotions in speech

Anton Batliner, Björn Schuller, Dino Seppi, Stefan Steidl, Laurence Devillers, Laurence Vidrascu, Thurid Vogt

Angaben zur Veröffentlichung / Publication details:

Batliner, Anton, Björn Schuller, Dino Seppi, Stefan Steidl, Laurence Devillers, Laurence Vidrascu, and Thurid Vogt. 2011. "The automatic recognition of emotions in speech." In *Emotion oriented systems*, edited by Roddy Cowie, Catherine Pelachaud, and Paolo Petta, 71–99. Berlin: Springer. https://doi.org/10.1007/978-3-642-15184-2_6.

The Automatic Recognition of Emotions in Speech

Anton Batliner, Björn Schuller, Dino Seppi, Stefan Steidl, Laurence Devillers, Laurence Vidrascu, Thurid Vogt, Vered Aharonson, 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 recording of databases, the annotation of emotional states, and further processing such as manual or automatic segmentation. Next we present an overview of acous-

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

Björn Schuller
Institute for Human-Machine Communication, Technische Universität München, Germany e-mail: schuller@tum.de

Dino Seppi
Fondazione Bruno Kessler - irst, Trento, Italy e-mail: seppi@fbk.eu

Stefan Steidl
Lehrstuhl für Mustererkennung, Friedrich-Alexander-Universität Erlangen, Germany e-mail: steidl@informatik.uni-erlangen.de

Laurence Devillers
Spoken Language Processing Group, LIMSI-CNRS, Orsay Cedex, France e-mail: devil@limsi.fr

Laurence Vidrascu
Spoken Language Processing Group, LIMSI-CNRS, Orsay Cedex, France e-mail: vidrascu@limsi.fr

Thurid Vogt
Multimedia Concepts and their Applications, University of Augsburg, Germany e-mail: thurid.vogt@informatik.uni-augsburg.de

Vered Aharonson
Tel Aviv academic college of engineering, Tel Aviv, Israel e-mail: vered@afeka.ac.il

Noam Amir
Department of communication disorders, Tel-Aviv University, Tel Aviv, Israel e-mail: noama@post.tau.ac.il

tic 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. [100, 103, 41]. Whereas such studies were not very frequent during the following decades, but cf. [116], 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 [91]. 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 then topics beyond pure information transmission, i.e. paralinguistic phenomena, e.g. emotions and attitudes conveyed via the speech channel, were addressed in studies such as [36]. 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 [104]. 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-center dialogues, cf. [10, 62, 4].

Nowadays, it is widely acknowledged that acted data cannot model naturalistic data sufficiently [10, 117, 110], 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. However, acted data is still used to a large extent, e.g. in [111], because non-acted data is still sparse, and most often not available freely. In this chapter, we will concentrate on the genuine approach of automatically recognizing/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: 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 un-balanced 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.

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 conceptualizations 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 esp. 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 re-sampling 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 cover 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, and (9) defining and applying exchange formats. We will sketch (1) to (4), mention (8) to (9), and concentrate on (5) to (7).

2.1 Databases

A common breakdown of emotion databases is the one into acted/non-acted, induced, and naturalistic databases [39]. 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 [93] 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 produced '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 [104], human-robot interaction [16], tutoring dialogues [2], stress detection in a driving scenario [42], human-human multi-party interaction [71], interaction human-information kiosk [21], appointment scheduling dialogues [10, 11], call-center applications (volunteering or real users, WoZ or real systems) [62, 4, 13, 106, 38]. Some more references to databases, mostly with acted data can be found in [32]. Multi-modal databases are dealt with 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. [38], or within the Verbmobil project [89], cf. [13]. 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, etc.

The next step should be to define the units of emotion annotation — which, in turn, is 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 — which can be an easy endeavour in a dialogue where the partners alternate as speakers/listeners. If the turns are longer, however, chances are that it 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 [35, 38] of prosodic, syntactic or pragmatic seg-

mentation. In [11] 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 [15] it is not perfect and thus sub-optimal if it comes to the processing of emotion recognition in a full end-to-end system [14] 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 55 terms used by [38] or the scheme proposed by [33]. Theoretically derived dimensional terms can be more or less elaborated [87]. The data-driven approach has often been employed by more 'application-minded' studies, cf. below.

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 [4] 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', un-marked state dominates and accounts for up to $> 90\%$ of the cases. The situation has not changed much recently, cf. [38]. [71] 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. [2] use a three-way distinction for student emotion in spoken tutoring dialogs: mixed/uncertain, certain, and neutral. [38] 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. In some few studies, up to seven different emotional user states are classified [21, 16]; 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. the 176 classes, reduced to 14 classes in experiments by [1]. 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-center 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 theoretical-methodological reasons, it might be desirable to employ > 10 naïve labellers or > 5 expert labellers to annotate on a fine-grained scale. This 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 [38], for more labellers, either a correspondence or a majority decision can be defined [105, 16], or a soft vector with percentages can be created. 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), etc. Such an external evidence can either be taken as 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 [44, 85]. The use of 'quantized' score ranges, based on such measures, e.g., for kappa, $< .2$ 'bad', between .2 and .4 'moderate', between .4, and .6 'good', between .6, and .8 'very good', $> .8$ '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 can itself be compared with the performance of the annotators by using measures such as proposed in [105] - within a running system, and on the other hand, the impact on the users of such systems, cf. Sec. 5.

2.3 Further Processing

State-of-the-art and 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. [20] 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 'un-interesting' default class neutral, cf. Sec. 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, e.g. in [17] from 59.2% recognition rate up to 77.5% for four classes.

It should be mandatory for writing a paper on recognizing 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. [19, 94]. 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 analyzed: For acted data, based on scripted speech, linguistic features are of no value. 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. There are several survey papers on prosodic features in automatic speech processing [53, 73] and on their use in emotion modelling [45, 92, 57].

3.1 Acoustic Features

Segmental features are mainly short term spectra and derived features: MFCC, LPC, PLP, etc. [51], and Wavelets [42, 94], TEO (Teager Energy operator) [42, 123], LFPC, LPPC [74]. These features are classically used for ASR where they are normally used for modelling segments such as phones and by that, words, rather than in emotion recognition where they are used for modelling longer units of analysis such as utterances/turns, dialogue moves, etc. To this aim the features are extracted framewise and combined by appropriate measures such as averaging and computing delta coefficients. Although originally intended to model segments, these features have been used successfully for supra-segmental units or for dynamic classification such as HMMs.

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 [53]. The raw prosodic data is 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 etc. Voice quality is a complicated issue in itself, since there are many different measures of voice quality, mostly clinical in origin, though once again standardization 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' [18].

Features can be low level vs. high level, i.e. statistic features vs. those based on pitch models such as MoMel [54], the Fujisaki model [46], and others. Features can be represented by raw values, i.e. they can be non-perceptual, or they can be based on perception models; *normalization*, taking into account pitch range, speech tempo, etc., — a straightforward way is to subtract some reference value - is used for modelling perception as well.

Using another terminology, we can speak about *Low Level Descriptors* (LLDs), i.e. basic measures of feature types, and *functionals* such as mean, percentiles, etc. LLDs account for base contours that usually are extracted by elaborating 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 contours. Functionals provide a normalization 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 derivation 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 curtosis). 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 [94].

We now characterize 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 normalization 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 ms, and because they are often highly correlated with duration features [8]. 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 normalization 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 also 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 time or spectral analysis: the speech signal is broken into overlapping frames and a pitch value is inferred from each segment by either autocorrelation [83] or spectral analysis [72]. The acoustic equivalent to the perceptual unit pitch is measured in Hz and often made perceptually more adequate by logarithmic transformation etc. Intervals, characterising points, or contours are often modelled.

The *spectrum* is characterized by formants (spectral maxima) modelling spoken content, esp. the lower ones. Higher ones 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) [65] or on cepstral analysis [34]. LPC is a model of 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*, the spectrum of the spectrum, emphasises changes or periodicity in the spectrum, while being relatively robust against noise. Its basic unit is quefrency which is related to time. 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 [51] and the MFCCs are extremely similar, as they both correspond to a short-term spectrum smoothing - the former through an ASR 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 other 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 (NHR), they reflect peculiar voice quality properties such as breathiness or harshness. Therefore they are usually dealt 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 difference in pitch extraction methods.

Wavelets give a short-term multi-resolution analysis of time, energy and frequencies in a speech signal. Compared to similar parametric representations such as MFCCs, they are superior in the modelling of temporal aspects.

Non-verbals 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 [52] Teager operator (esp. for stress detection) [123], and dynamic features for HMM, cf. below. The standard acoustic feature types used in many emotion classification studies might be — probably in this order of frequency but not necessarily 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. [20].

3.2 Linguistic Features

Spoken or written text also carries information about the underlying affective state [5]. 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: key-word spotting [40, 31], rule-based modelling [63], Semantic Trees [122], Latent Semantic Analysis [48], Transformation-based Learning [120], World-knowledge-Modelling [64], Key-Phrase-Spotting [98], and Bayesian Networks [24]. Context/pragmatic information has been modelled as well, e.g. type of system prompt [106], dialogue acts [63, 11], or system and user performance [2]. 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* [80, 4, 61, 37] and *vector space modelling* [96, 19]; 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 advantageous to overcome misspelling, or spelling variations, dialects, etc. Considering analysis from spoken text, only few results for emotion recognition rely on ASR output [96] rather than on manual annotation of data [19]. Secondly, an inventory of term entities, known as vocabulary, needs to be constructed which initially consists of all different words observed in the training corpus - usually several thousand (as opposed to this, e.g. the Balanced Affective Wordlist [102] consists of only roughly 300 words) — and 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 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 in 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, as e.g. by simple N-Gram Stemming, cf. below, or by (Iterated) Lovins, Snowball, Dawson, Porter, Paice and Husk, and Krovetz stemmers that base on suffix lists and rules for their application. A very compact approach to stemming is the use of so called Part-of-Speech (POS) classes, such as nouns, verbs, adjectives, particles [7, 19]. Also *sememes*, i.e. semantic units represented by lexemes, can be clustered into higher semantic concepts such as generally positive or negative terms [19]. In addition, non-linguistic vocalizations like sighs and yawns [88], laughs [28, 108], cries [76], and coughs [67] can easily be integrated into the vocabulary [19, 95].

Class-based back-off N-Grams are commonly used for general language modelling. Thereby the posterior probability of a word is given by its predecessors from left to right within an utterance. For emotion recognition, class-based N-grams are needed: given an emotion, this leads to the according posterior probability for an emotion under the condition of the words 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. Due to the typical data sparseness in emotion recognition, mostly uni-grams (N=1) have been applied so far [61, 37], besides bi-grams (N=2) and tri-grams (N=3) [4]. The actual emotion is calculated by the posterior probability of the emotion given the actual word(s).

Bag-of-Words, also known as vector space-modelling, is a well-known numerical representation form of text in automatic document categorization [56]. It has been successfully ported to recognize sentiments [77] or emotion [96, 95]. Thereby each word in the vocabulary adds a dimension to a linguistic vector representing the term frequency within the actual utterance. Note that thereby easily very large feature spaces may occur, which usually require stopping and stemming. To overcome linearities, the logarithmic term frequency is often used, and the term frequency is also often normalized by the utterance length, and with respect to the overall term frequency of occurrence within the training corpus. 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 (c.f. below) show high performance for this task. The possibility of early fusion with acoustic features helped make this technique very popular [95, 19].

The preponderance of acoustics in emotion modelling so far is conditioned by the traditional focus on segmentally identical, acted utterances. For natu-

ralistic 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 have 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. A common claim in machine learning is: 'any method is as good provided a good feature vector'. Unfortunately emotion recognition from speech has to deal with noisy, redundant, 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, (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 is very sparse. This problem is known as 'curse of dimensionality' [22], 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, without loosing as much information as possible. Common reduction techniques used in the field of emotion recognition are Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Independent Component Analysis (ICA), and more sophisticated derivations like Heteroscedastic Discriminant Analysis [6]. PCA is the feature transformation that minimises the sum of square error. Furthermore, the base of the new space is ortho-normal, which means that PCA de-correlates the original features: new features are constructed 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 suboptimal for specific problems), LDA is a supervised feature reduction method which searches for the linear transformation that maximises the ratio of the determinants of the between-class covariance matrix and the within-class covariance matrix [47]. LDA is less used as feature reduction, but it is widely adopted for direct classification [60, 84, 58, 10]. Finally ICA 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, 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 optimization 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 []], or inter-feature and feature-class correlation []], etc. 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 conservative 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 [62, 60, 58]. Sequential floating forward selection SFFS [81, 55] 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 [55]. Note that a good feature selection should de-correlate the feature space to optimize 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 [19].

With the growing interest in spontaneous data, class skewness or the 'sparse data' problem in the output (classes) 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 cover classes (clustering similar classes under the same hat). The most frequent couples of cover 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 speaks for considering diverse classifiers for different tasks: mostly high recognition rates (e.g. ability to solve non-linear problems, learn discriminatively, online adapt, generalize, 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, handling of input confidence). These considerations, and the simple availability of implementations such as WEKA ([118]) or HTK led to a considerable band-width of variants being used in the recognition of emotion from speech.

Very popular classifiers for emotion recognition are Linear Discriminant Classifiers (LDCs) [47] and k-Nearest Neighbour (kNN) classifiers [30]: 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 look-up method:

the training data is 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 [36, 79], with good results for non acted emotional speech as well [60, 101]. LDC (as a natural extension of LDA, see [47]) is basically a classifier with straight line decision surfaces (hyperplanes). LDA is one possible method of estimating LDC hyperplane parameters by maximization of class separability (see above). They have often been used [60, 84, 58, 63, 10], with a competitive performance [19] in spite of some limitations: the data should be linearly separable, and the method is sensitive to outliers. A natural extension of LDCs are Support Vector Machines (SVMs): if the input data have previously undergone a nonlinear 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 [109]. SVM provide very good generalization properties [68, 61, 29, 121, 70], which positioned them among the number one choices in recent works; note, however, that their performance is not always (way) better than the one obtained by using alternative classifiers [69].

The most used non linear discriminative classifiers are Artificial Neural Networks (ANNs) and decision trees. Feed Forward ANNs, also known as Multi Layered 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 [86]), 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 [79, 66], and even less for non-acted, but cf. [10, 19]. Although they are also characterized 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. cut off according to the ranking. Popular decision tree algorithms are C4.5 [82], and CART [27]. Note, however, that accuracy degrades in case of irrelevant features or noisy patterns. A solution are Random Forests (RF) [26], 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 [94].

Stochastic classifiers are very popular. This probably also derives from their general popularity in speech analysis tasks. Apart from the already named kNN, which can be seen as very basic statistical classifier, one also basic representative of this group is the Naive Bayes classifier [59, 49]. 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, usually by Expectation Maximization.

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 [23, 97]; for non-acted emotion fewer are known [58, 114, 113]. The performance of static modelling is usually not reached [23, 97], as emotion apparently is better modelled on a time-scale above frame-level; note that a combination of static features such as minimum, maximum, onset, offset, duration, regression, etc. implicitly shape contour dynamics as well. Still, when the spoken content is fixed, the combination of static and dynamic processing may help improve overall accuracy [112]. However, it is not clear whether emotion can be satisfyingly modelled using the simplifying Markov assumption that underlies HMM modelling [23].

Ensembles of classifiers [96] combine their individual strengths, or overcome training instability deriving from the sparseness of data. In the highly popular *Bagging* [25] 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 by their sequential application. Most powerful however is the combination of diverse classifiers by either simple *Voting* 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* [119]. 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.

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 as Support Vector Regression [50] or less complex solutions 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 [43] used in [19].

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 is

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 unfeasible. 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. 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, it is not speaker-independent, and recognition performance will thus be too optimistic. Both these downsides can be overcome by leave-one-speaker-out, meaning training with all but one speaker in each cycle, or leave a known group of speakers out to spare computational effort.

Most of the studies report performance measures expressed by accuracy, i.e. Recognition Rate (RR), the number of correctly recognised patterns divided by the total number of patterns. Given the skewness of spontaneous emotional databases, this is not always 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). Note that RR and CL for a balanced multi-class recognition problem are always identical; the more the class distribution is un-balanced, the higher the difference between RR and CL. 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 negative). 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.

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, esp. 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 as harmonic mean of recall and precision.

Studies eventually end up with the conclusion that a specific classifier is better than another one - which is a conclusion that must not be generalized. Most of the time no significance of the differences is reported. Actually, there are some reasons to handle significance tests with care: the more experiments we do on a certain dataset, 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; in our field, this is not the rule but the exception. The Bonferroni adjustment is a possible choice of a correction factor. For a cookbook on multi-experiment studies see [90]. There are some drawbacks to the Bonferroni correction as it is usually too conservative; these are outlined in [78].

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 [90]. To our knowledge, only few studies in emotion recognition clearly explain what - if any - part of the data has been used for parameter tuning: they describe how the data has been divided into test and training but nothing is said about held-out data for classifier tuning, which 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 generalized. 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 standardized yet.

5 Applications

Apart from some 'off-line' applications such as data mining in movie archives or screening call-center 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 in — generated system reaction such as synthesised speech, produced by the system out, and vice versa'. Several systems have been envisaged so far [9]. 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 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 [12, 99]. In real-life settings, chances are that a worse signal-to-noise ratio will deteriorate ASR and by that, emotion classification; esp. 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, cf. [99], but we don't know yet. 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 [107] where an implementation of the OCC model [75] 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 process-

ing, from conceptualization 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 break-through 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 recognize. 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 addressed different databases using the same approaches, cf. [101]. Initiatives such as CEICES [19], 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 standardization', i.e., making comparisons across classifiers and features easier and more reliable.

References

1. S. Abrilian, L. Devillers, S. Buisine, and J.-C Martin. EmoTV1: Annotation of Real-life Emotions for the Specification of Multimodal Affective Interfaces. In *Proceedings of Human-Computer Interaction International*, Las Vegas, 2005.
2. H. Ai, D. Litman, K. Forbes-Riley, M. Rotaru, J. Tetreault, and A. Purandare. Using system and user performance features to improve emotion detection in spoken tutoring dialogs. In *Proceedings of ICSLP*.
3. E. André, L. Dybkiaer, W. Minker, and P. Heisterkamp, editors. *Affective Dialogue Systems, Proc. of a Tutorial and Research Workshop*, volume 3068 of *Lecture Notes in Artificial Intelligence*, Berlin, 2004. Springer-Verlag.
4. J. Ang, R. Dhillon, E. Shriberg, and A. Stolcke. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proceedings of the International Conference on Spoken Language Processing (Interspeech 2002 – ICSLP)*, pages 2037–2040, 2002.
5. S. Arunachalam, D. Gould, E. Anderson, D. Byrd, and S.S. Narayanan. Politeness and frustration language in child-machine interactions. pages 2675–2678, Aalborg, Denmark, 2001.
6. M. M. H. El Ayadi, M. S. Kamel, and F. Karray. Speech emotion recognition using gaussian mixture vector autoregressive models. pages 957–960, 2007.
7. A. Batliner, J. Buckow, R. Huber, V. Warnke, E. Nöth, and H. Niemann. Prosodic Feature Evaluation: Brute Force or Well Designed? In *Proc. of the 14th Int. Congress of Phonetic Sciences*, volume 3, pages 2315–2318, San Francisco, 1999.
8. A. Batliner, J. Buckow, R. Huber, V. Warnke, E. Nöth, and H. Niemann. Boiling down Prosody for the Classification of Boundaries and Accents in German and English. In *In Proc. 7th Eurospeech*, pages 2781–2784, Aalborg, 2001.

9. A. Batliner, F. Burkhardt, M. van Ballegooy, and E. Nöth. A Taxonomy of Applications that Utilize Emotional Awareness. In *Proceedings of IS-LTC 2006*, pages 246–250, Ljubliana, 2006.
10. A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. Desperately Seeking Emotions: Actors, Wizards, and Human Beings. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 195–200, Newcastle, Northern Ireland, 2000.
11. A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. How to find trouble in communication. *Speech Communication*, 40:117–143, 2003.
12. A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D'Arcy, M. Russell, and M. Wong. "You stupid tin box" - children interacting with the AIBO robot: A cross-linguistic emotional speech corpus. In *Proc. of the 4th International Conference of Language Resources and Evaluation LREC 2004*, pages 171–174, Lisbon, 2004.
13. A. Batliner, C. Hacker, S. Steidl, E. Nöth, and J. Haas. From Emotion to Interaction: Lessons from Real Human-Machine-Dialogues. In André et al. [3], pages 1–12.
14. A. Batliner, R. Huber, H. Niemann, E. Nöth, J. Spilker, and K. Fischer. The Recognition of Emotion. In Wahlster [115], pages 122–130.
15. A. Batliner, R. Kompe, A. Kießling, M. Mast, H. Niemann, and E. Nöth. M = Syntax + Prosody: A syntactic–prosodic labelling scheme for large spontaneous speech databases. *Speech Communication*, 25(4):193–222, September 1998.
16. A. Batliner, S. Steidl, C. Hacker, and E. Nöth. Private Emotions vs. Social Interaction — a Data-Driven Approach Towards Analysing Emotions in Speech. *User Modeling and User-Adapted Interaction, The Journal of Personalization Research*, 2008. to appear.
17. A. Batliner, S. Steidl, C. Hacker, E. Nöth, and H. Niemann. Tales of Tuning – Prototyping for Automatic Classification of Emotional User States. In *Proc. 9th Eurospeech - Interspeech 2005*, pages 489–492, Lisbon, 2005.
18. A. Batliner, S. Steidl, and E. Nöth. Laryngealizations and Emotions: How Many Babushkas? In *Proceedings of the International Workshop on Paralinguistic Speech — between Models and Data (ParaLing'07)*, pages 17–22, Saarbrücken, 2007.
19. A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson. Combining Efforts for Improving Automatic Classification of Emotional User States. In *Proceedings of IS-LTC 2006*, pages 240–245, Ljubliana, 2006.
20. A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson. The Impact of F0 Extraction Errors on the Classification of Prominence and Emotion. In *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS 2007)*, pages 2201–2204, Saarbrücken, 2007.
21. A. Batliner, V. Zeissler, C. Frank, J. Adelhardt, R. Shi, and E. Nöth. We are not amused - but how do you know? user states in a multi-modal dialogue system. pages 733–736, 2003.
22. R.E. Bellman. *Adaptive Control Processes*. Princeton University Press, 1961.
23. L.T. Bosch. Emotions: what is possible in the asr framework? In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 189–194, Newcastle, Northern Ireland, 2000.
24. J. Breese and G. Ball. Modeling emotional state and personality for conversational agents. Technical Report MS-TR-98-41, Microsoft, 1998.
25. L. Breiman. Bagging predictors. *Machine Learning*, 26:123–140, 1996.
26. L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.
27. L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Pacific Grove, CA, USA, 1984.
28. N. Campbell and R. Ohara. No laughing matter. In *EUROSPEECH*, Lisbon, Portugal, 2005.
29. Z.-J. Chuang and Chung-Hsien Wu. Emotion recognition using acoustic features and textual content. In *Proc. of ICME*, pages 53–56, 2004.
30. T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, 1967.
31. R. Cowie, E. Douglas-Cowie, B. Apolloni, J. Taylor, A. Romano, and W. Fellenz. What a neural net needs to know about emotion words. *Journal of Computational Intelligence and Applications*, pages 109–114, 1999.

32. R. Cowie, E. Douglas-Cowie, and C. Cox. Beyond emotion archetypes: databases for emotion modelling using neural networks. *Neural Networks*, 18:371–388, 2005.
33. R. Craggs and M. McGee Wood. A categorical annotation scheme for emotion in the linguistic content of dialogue. In André et al. [3], pages 89–100.
34. S.B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. on Acoust., Speech and Signal Processing*, 29:917–919, 1980.
35. F. de Rosis, A. Batliner, N. Novielli, and S. Steidl. 'You are Sooo Cool, Valentina!' Recognizing Social Attitude in Speech-Based Dialogues with an ECA. In Ana Paiva, Rui Prada, and Rosalind W. Picard, editors, *Affective Computing and Intelligent Interaction*, pages 179–190, Berlin-Heidelberg, 2007. Springer.
36. F. Dellaert, T. Polzin, and A. Waibel. Recognizing emotion in speech. In *Proceedings of 4th International Conference on Spoken Language Processing*, pages 1970–1973, 1996.
37. L. Devillers, L. Lamel, and I. Vasilescu. Emotion Detection in Task-Oriented Spoken Dialogs. In *ICME*, Baltimore, July 2003.
38. L. Devillers, L. Vidrascu, and L. Lamel. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18:407–422, 2005.
39. E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilan, A. Batliner, N. Amir, and K. Karpousis. The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data. In Ana Paiva, Rui Prada, and Rosalind W. Picard, editors, *Affective Computing and Intelligent Interaction*, pages 488–500, Berlin-Heidelberg, 2007. Springer.
40. C. Elliott. The affective reasoner: A process model of emotions in a multi-agent system. 1992.
41. G. Fairbanks and W. Pronovost. An experimental study of the pitch characteristics of the voice during the expression of emotion. *Speech Monograph*, 6:87–104, 1939.
42. R. Fernandez and R. W. Picard. Modeling drivers' speech under stress. *Speech Communication*, 40:145–159, 2003.
43. J. Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *ASRU*, Santa Barbara, USA, 1997.
44. J.L. Fleiss, J. Cohen, and B.S. Everitt. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72(5):323–327, 1969.
45. R.W. Frick. Communicating emotion: the role of prosodic features. *Psychological Bulletin*, 97:412–429, 1985.
46. H. Fujisaki. Modelling the process of fundamental frequency contour generation. In Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisasaka, editors, *Speech Perception, Production and Linguistic Structure*, pages 313–328. IOS Press, 1992.
47. K. Fukunaga. Introduction to statistical pattern recognition. 1990.
48. B. Goertzel, K. Silverman, C. Hartley, S. Bugaj, and M. Ross. 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)*, 2000.
49. I.J. Good. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. MIT Press, 1965.
50. M. Grimm, K. Kroschel, H. Harris, C. Nass, B. Schuller, G. Rigoll, and T. Moosmayr. On the Necessity and Feasibility of Detecting a Driver's Emotional State While Driving. In Ana Paiva, Rui Prada, and Rosalind W. Picard, editors, *Affective Computing and Intelligent Interaction*, pages 126–138, Berlin-Heidelberg, 2007. Springer.
51. H. Hermansky. Perceptual linear predictive (plp) analysis for speech. *The Journal of The Acoustical Society of America (JASA)*, 87:1738–1752, 1990.
52. H. Hermansky and S. Sharma. Traps - classifiers of temporal patterns. In *Proc. of ICSLP 98*, pages 1003–1006, Sydney, 1998.
53. W. Hess, A. Batliner, A. Kießling, R. Kompe, E. Nöth, A. Petzold, M. Reyelt, and V. Strom. Prosodic Modules for Speech Recognition and Understanding in Verbmobil. In Yoshinori Sagisaka, Nick Campell, and Norio Higuchi, editors, *Computing Prosody. Approaches to a Computational Analysis and Modelling of the Prosody of Spontaneous Speech*, pages 363–383. Springer–Verlag, New York, 1996.

54. D. Hirst, A. Di Cristo, and R. Espesser. Levels of representation and levels of analysis for intonation. In M. Horne, editor, *Prosody : Theory and Experiment*, pages 51–87. Kluwer Academic Publishers, Dordrecht, 2000.
55. A. Jain and D. Zongker. Feature selection: evaluation, application and small sample performance. *PAMI*, 19(2):153–158, 1997.
56. T. Joachims. Text categorization with support vector machines: learning with many relevant features. In Claire Nédellec and Céline Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, pages 137–142, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE.
57. T. Johnstone and K. R. Scherer. Vocal communication of emotion. In M. Lewis and J. M. Haviland-Jones, editors, *Handbook of Emotions*, chapter 14, pages 220–235. Guilford Press, New York, London, 2nd edition, 2000.
58. O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee. Emotion recognition by speech signals. pages 125–128, 2003.
59. P. Langley, W. Iba, and K. Thompson. An analysis of bayesian classiers. In *Proceedings of the National Conference on Articial Intelligence*, pages 223–228, San Jose, CA, USA, 1992.
60. C. M. Lee and S. S. Narayanan. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13(2):293–303, 2005.
61. C. M. Lee, S. S. Narayanan, and R. Pieraccini. Combining acoustic and language information for emotion recognition. pages 873–376, 2002.
62. C.M. Lee, S. Narayanan, and R. Pieraccini. Recognition of Negative Emotions from the Speech Signal. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU'01)*, 2001. no pagination.
63. D. Litman and K. Forbes. Recognizing emotions from student speech in tutoring dialogues. In *Proceedings of ASRU*, pages 25–30, Virgin Island, 2003.
64. H. Liu, H. Liebermann, and T. Selker. A model of textual affect sensing using real-world knowledge. In *Proc. 7th International Conference on Intelligent User Interfaces (IUI 2003)*, pages 125–132, 2003.
65. J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63:561–580, 1975.
66. C. A. Martinez and A.B. Cruz. Emotion recognition in non-structured utterances for human-robot interaction. In *IEEE International Workshop on Robot and Human Interactive Communication*, pages 19–23, 2005.
67. S. Matos, S.S. Birring, I.D. Pavord, and D.H. Evans. Detection of cough signals in continuous audio recordings using hmm. *IEEE Trans. Biomedical Engineering*, pages 1078–108, 2006.
68. S. McGilloway, R. Cowie, E. Doulas-Cowie, S. Gielen, M. Westerdijk, and S. Stroeve. Approaching automatic recognition of emotion from voice: A rough benchmark. In *Proceedings of the ISCA workshop on Speech and Emotion*, pages 207–212, Newcastle, 2000.
69. D. Meyer, F. Leisch, and K. Hornik. Benchmarking Support Vector Machines. Report Series No. 78, Adaptive Informations Systems and Management in Economics and Management Science , 2002.
70. D. Morrison, R. Wang, W.L. Xu, and L. C. De Silva. Incremental learning for spoken affect classification and its application in call-centres. *International Journal of Intelligent Systems Technologies and Applications*, 2:242–254, 2007.
71. D. Neiberg, K. Elenius, and K. Laskowski. Emotion Recognition in Spontaneous Speech Using GMMs. In *Proceedings of The International Conference on Spoken Language Processing (Interspeech 2006 – ICSLP)*, pages 809–812, Pittsburgh, 2006.
72. A. M. Noll. Cepstrum pitch determination. *JASA*, 14:293–309, 1967.
73. E. Nöth, A. Batliner, V. Warnke, J. Haas, M. Boros, J. Buckow, R. Huber, F. Gallwitz, M. Nutt, and H. Niemann. On the Use of Prosody in Automatic Dialogue Understanding. 36(1-2), January 2002.
74. T. Nwe, S. Foo, , and L. De Silva. Speech emotion recognition using hidden markov models. *Speech Communication*, 41:603–623, 2003.
75. A. Ortony, G. L. Clore, and A. Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, Cambridge, New York, 1988.

76. P. Pal, A.N. Iyer, and R.E. Yantorno. Emotion detection from infant facial expressions and cries. *ICASSP*, 2006.
77. B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. pages 79–86, 2002.
78. T. V. Pernegger. What's wrong with Bonferroni adjustment. *British Medical Journal*, 316:1236–1238, 1998.
79. V. Petrushin. Emotion in speech: Recognition and application to call centers. In *Proc. of Artificial Neural Networks in Engineering (ANNIE '99)*, pages 7–10, 1999.
80. T. S. Polzin and A. Waibel. Emotion-sensitive human-computer interfaces. pages 201–206, 2002.
81. P. Pudil, J. Novovicova, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15:1119–1125, 1994.
82. J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.
83. L. R. Rabiner. On the use of autocorrelation analysis for pitch detection. *IEEE_J_ASSP*, 25:24–33, 1977.
84. M. A. Rahurkar and J.H.L.Hansen. Towards affect recognition: an ica approach. pages 1017–1022, 2003.
85. A. Rosenberg and E. Binkowski. Augmenting the kappa statistic to determine interannotator reliability for multiply labeled data points. In D. Marcu S. Dumais and S. Roukos, editors, *HLT-NAACL 2004: Short Papers*, pages 77–80, Boston, Massachusetts, USA, 2004. Association for Computational Linguistics.
86. D.E. Rumelhart, G.E. Hinton, and R.J. Williams. *Learning internal representations by error propagation*, volume 1, pages 318–362. MIT Press, 1986.
87. J. A. Russel. How shall an emotion be called? In *Circumplex Models of Personality and Emotions*, chapter 9, pages 205–220. American Psychological Association, Washington D.C., 1997.
88. J.A. Russell, J.A. Bachorowski, and J.M. Fernandez-Dols. Facial and vocal expressions of emotion. *Annual Review of Psychology*, pages 329–349, 2003.
89. F. Schiel H. G. Tillman S. Burger, K. Weilhammer. Verbmobil Data Collection and Annotation. In Wahlster [115], pages 537–549.
90. S. Salzberg. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1(3):317–328, 1997.
91. K. R. Scherer. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40:227–256, 2003.
92. K. R. Scherer, T. Johnstone, and G. Klasmeyer. Vocal expression of emotion. chapter 23, pages 433–456. 2003.
93. F. Schiel. Automatic phonetic transcription of non-prompted speech. In *Proc. of ICPhS 1999*, pages 607–610, San Francisco, 1999.
94. B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson. The Relevance of Feature Type for the Automatic Classification of Emotional User States: Low Level Descriptors and Functionals. In *Proceedings of Interspeech*, pages 2253–2256, Antwerp, 2007.
95. B. Schuller, N. Köhler, R. Müller, and G. Rigoll. Recognition of Interest in Human Conversational Speech. In *Proc. INTERSPEECH*, pages 793–796, Pittsburgh, 2006.
96. B. Schuller, R. Müller, M. Lang, and G. Rigoll. Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensembles. In *Proc. 9th Eurospeech - Interspeech 2005*, pages 805–809, Lisbon, 2005.
97. B. Schuller, G. Rigoll, and M. Lang. Hidden markov model-based speech emotion recognition. In *Proc. ICASSP*, pages II: 1–4, 2003.
98. B. Schuller, G. Rigoll, and M. Lang. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. pages I:577–580, 2004.
99. B. Schuller, D. Seppi, A. Batliner, A. Meier, and S. Steidl. Towards more Reality in the Recognition of Emotional Speech. In *Proc. of ICASSP 2007*, pages 941–944, Honolulu, 2007.

100. E.W. Scripture. A study of emotions by speech transcription. *Vox*, 31:179–183, 1921.
101. M. Shami and W. Verhelst. Automatic Classification of Expressiveness in Speech: A Multi-corpus Study. In Christian Müller, editor, *Speaker Classification II*, volume 4441 of *Lecture Notes in Computer Science / Artificial Intelligence*, pages 43–56. Springer, Heidelberg - Berlin - New York, 2007.
102. G. Siegle. The balanced affective word list project, 1995. http://www.sci.sdsu.edu/CAL/wordlist/.
103. E.R. Skinner. A calibrated recording and analysis of the pitch, force, and quality of vocal tones expressing happiness and sadness. *Speech Monographs*, 2:81–137, 1935.
104. M. Slaney and G. McRoberts. Baby Ears: A Recognition System for Affective Vocalizations. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1998)*, pages 985–988, Seattle, 1998.
105. S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann. "Of All Things the Measure is Man": Automatic Classification of Emotions and Inter-Labeler Consistency. In *Proc. of ICASSP 2005*, pages 317–320, Philadelphia, 2005.
106. S. Steidl, C. Ruff, A. Batliner, E. Nöth, and J. Haas. Looking at the Last two Turns, I'd Say this Dialogue is Doomed — Measuring Dialogue Success. In P. Sojka, I. Kopeček, and K. Pala, editors, *Text, Speech and Dialogue, 7th International Conference, TSD 2004*, pages 629–636, Berlin, Heidelberg, 2004.
107. M. Streit, A. Batliner, and T. Portele. Emotions Analysis and Emotion-Handling Subdialogues. In W. Wahlster, editor, *SmartKom: Foundations of Multimodal Dialogue Systems*, pages 317–332. Springer, Berlin, 2006.
108. K.P. Truong and D.A. van Leeuwen. Automatic detection of laughter. pages 485–488, Lisbon, Portugal, 2005.
109. V.N. Vapnik. *The nature of statistical learning theory*. Springer, 1995.
110. E. Veleten. A laboratory task for induction of mood states. *Behavior Research & Therapy*, (6):473–482, 1968.
111. D. Ververidis and C. Kotropoulos. Fast sequential floating forward selection applied to emotional speech features estimated on des and susas data collection. In *Proc. of European Signal Processing Conf. (EUSIPCO 2006)*, Florence, 2006.
112. B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll. Combining frame and turn-level information for robust recognition of emotions within speech. In *Proc. Interspeech*, pages 2249–2252, 2007.
113. B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll. Frame vs. Turn-Level: Emotion Recognition from Speech Considering Static and Dynamic Processing. In Ana Paiva, Rui Prada, and Rosalind W. Picard, editors, *Affective Computing and Intelligent Interaction*, pages 139–147, Berlin-Heidelberg, 2007. Springer.
114. J. Wagner, T. Vogt, and André. A Systematic Comparison of different HMM designs for Emotion Recognition from Acted and Spontaneous Speech. In Ana Paiva, Rui Prada, and Rosalind W. Picard, editors, *Affective Computing and Intelligent Interaction*, pages 114–125, Berlin-Heidelberg, 2007. Springer.
115. W. Wahlster, editor. *Verbmobil: Foundations of Speech-to-Speech Translations*. Springer, Berlin, 2000.
116. C.E. Williams and K.N. Stevens. Emotions and speech: some acoustic correlates. *Journal of the Acoustical Society of America*, 52:1238–1250, 1972.
117. J. Wilting, E. Krahmer, and M. Swerts. Real vs. acted emotional speech. In *Proc. Interspeech*, 2006.
118. I. H. Witten and E. Frank. *Data mining: Practical machine learning tools and techniques, 2nd Edition*. Morgan Kaufmann, San Francisco, 2005.
119. D. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992.
120. T. Wu, F.M. Khan, T.A. Fisher, L.A. Shuler, and W.M. Pottenger. Posting act tagging using transformation-based learning. In Tsau Young Lin, Setsuo Ohsuga, Churn-Jung Liau, Xiaohua Hu, and Shusaku Tsumoto, editors, *Foundations of Data Mining and knowledge Discovery*, pages 319–331. Springer, Berlin-Heidelberg, 2005.

121. M. You, C. Chen, J. Bu, J. Liu, and J. Tao. Emotion recognition from noisy speech. In *Proc. ICME*, pages 1653–1656, 2006.
122. X. Zhe and A.C. Boucouvalas. Text-to-emotion engine for real time internet communication. In *Proceedings of the International Symposium on Communication Systems, Networks, and DSPs*, pages 164–168, Staffordshire University, 2002.
123. G. Zhou, J. H. L. Hansen, , and J. F. Kaiser. Nonlinear feature based classification of speech under stress. *IEEE transactions on speech and audio processing*, 9:201–216, 2001.