Chapter 9
Voice and Speech Analysis in Search of States and Traits

Björn Schuller

9.1 Vocal Behavior Analysis—An Introduction

It is the aim of this chapter to introduce the analysis of vocal behaviour and more general paralinguistics in speech and language. By 'voice' we refer to the acoustic properties of a speakers' voice—this will be dealt with in Section 9.2. By 'speech' we refer more generally to spoken language in the sense of added linguistics—dealt with in Section 9.3. Obviously, the introduced methods of linguistic analysis can also be applied to written text, albeit with slightly different pre-processing. Also, models trained on written text may differ insofar as spoken language is often grammatically different and possesses more fragments of words, etc.

9.1.1 A Short Motivation

Paralinguistic speech and language analysis, i. e., the analysis of consciously or unconsciously expressed non-verbal elements of communication, is constantly developing into a major field of speech analysis, as new human-machine interaction and media retrieval systems advance over sheer speech recognition.

The additional information over 'what' is being said bears high potential for improved interaction or retrieval of speech files. By such information, social competence is provided to systems that can react more human-like or provide more human-like information. In addition, this information can also help to better recognise 'what' is being said, as acoustic and linguistic models can be adapted to different speaker states and traits and non-verbal outbursts are not confused with linguistic entities [52]. A number of such paralinguistic phenomena is next given.

Björn Schuller
Institute for Human-Machine Communication, Technische Universität München, D-80290 Munich, Germany, e-mail: schuller@tum.de

9.1.2 From Affection to Zest

One can broadly divide the multifaceted field of paralinguistics into speaker states and speaker traits and vocal behaviour. Speaker *states* thereby deal with states changing over time, such as affection and intimacy [3], deception [14], emotion [7], interest [48], intoxication [35], sleepiness [24], health state [19], and stress [21] or zest, while the speaker *traits* identify permanent speaker characteristics such as age and gender [44], height [32], likeability [57], or personality [31]. Vocal behaviour additionally comprises non-linguistic vocal outbursts like sighs and yawns [34], laughs [10], cries [33], hesitations and consent [48], and coughs [30]. We next deal with the principle of how to computationally analyse any of these automatically.

9.1.3 Principle

Here we share a unified perspective on the computationally 'intelligent' analysis of speech as a general pattern recognition paradigm.

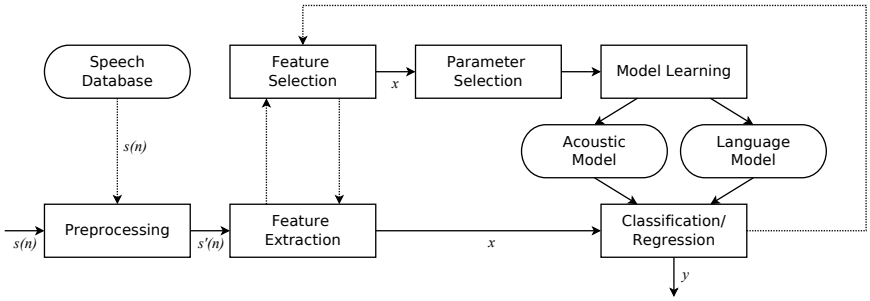

Fig. 9.1 Analysis of voice and speech—an overview. Dotted lines indicate the training phase.

Fig. 9.1 gives an overview of the typical steps in such a system. The dotted lines indicate the training or learning phase that is usually carried out once before using such a system in practice. It can, however, re-occur during application in the case of online or unsupervised and semi-supervised adaptation. Interestingly, the information is partly well suited for online learning based on user feedback, as user (dis-)satisfaction or similar states and affirmative vocalisations can be used to adapt models accordingly.

The building blocks of a voice and speech analysis system are:

Pre-processing usually deals with enhancement of signal properties of interest from input speech. Such speech may be coming from a capture device like an A/D converter in a live setting, or from offline databases of stored audio files for training and evaluation purposes. Such enhancement includes de-reverberation and noise

suppression, e.g., by exploitation of multiple microphones, or separation of multiple speakers by blind source separation.

Feature Extraction deals with the reduction of information to the relevant characteristics of the problem to be investigated in the sense of a canonical representation and will be dealt with in more detail—separately for acoustic and linguistic features.

Classification / Regression assigns the actual label to an unknown test instance. In the case of classification, discrete labels such as Ekman's 'big six' emotion classes (anger, disgust, fear, happiness, sadness, and surprise) or, e. g., binary low/high labels per each of the 'big five' personality dimensions (openness, conscientiousness, extraversion, agreeableness, and neuroticism – "OCEAN") are decided for. In the case of regression, the output is a continuous value like a speaker's height in centimetre or age in years, or—in the case of emotion—dimensions like potency, arousal, and valence, typically ranging from -1 to +1. We will discuss the frequently encountered machine learning algorithms in the field later on.

Speech Databases comprise the stored audio of exemplary speech for model learning and testing. In addition, a transcription of the spoken content may be given and the labelling of the problem at hand, such as speaker emotion, age, or personality. Usually, one wishes for adequate data in the sense of natural data rather than elicited or acted in ideal conditions, excluding disruptive influence or well-described and targeted noise or reverberation, a high total amount—which is rarely given. Further, data should ideally include a large number of speakers, a meaningful categorisation, which is usually non-trivial in this field (cf. the emotion categories vs. dimensions), a reliable annotation either by the speaker herself or a higher number of annotators to avoid skewness, additional perception tests by independent labellers to provide a comparison of human performance on the task, balanced distribution of instances among classes or the dimensional continuum, knowledge of the prior distribution, high diversity of speakers' ages, gender, ethnicity, language, etc., and high spoken content variation. Finally, one wishes for well defined test, development, and training partitions without prototypical selection of 'friendly cases' for classification [49], free availability of the data, and well-documented meta-data.

Model Learning is the actual training phase in which the classifier or regressor model is built, based on labelled data. There are classifiers or regressors that do not need this phase—so called lazy learners—as they only decide at run-time by training instances' properties which class to choose, e.g., by the training instance with shortest distance in the feature space to test instances [20]. However, these are seldom used, as they typically do not lead to sufficient accuracy in the rather complex task of speech analysis.

Feature Selection decides which features actually to keep in the feature space. This may be of interest if a new task, e.g., estimation of a speaker's weight from acoustic properties, is not well known. In such a case, a multiplicity of features can be 'brute-forced', as will be shown. From these, the ones well suited for the task at hand can be kept.

Parameter Selection fine 'tunes' the learning algorithm. Indeed, the performance of a machine learning algorithm can be significantly influenced by optimal or sub-

optimal parametrisation. As for the feature selection, it is crucial not to 'tune' on speech instances used for evaluation as obviously this would lead to overestimation of performance.

Acoustic Models consist of the learnt dependencies between acoustic observations and classes, or continuous values in the case of regression, stored as binary or text files.

Language Models resemble acoustic models—yet, they store the learnt dependencies of linguistic observations and according assignments.

9.2 'Voice' — The Acoustic Analysis

In this section we will be dealing with the acoustic properties of the voice ignoring 'what' is being said and entirely focusing on 'how' it is said (cf. also Chapter 10). For this analysis, we will first need to chunk the audio stream (for an example see Fig. 9.2.a) before extracting features for these chunks and then proceed with the selection of relevant features before the classification/regression, and 'fine tuning'.

9.2.1 Chunking

Already for the annotation of human behaviour that changes over time, one mostly needs to 'chunk' the speech, which is often stored as a single file ranging over several seconds up to hours, into 'units of analysis'. These chunks may be based on the 'quasi-stationarity' of the signal, as given by single frames obtained by applying a window function to the signal—typically having a length of some 10-30 ms and applied every 10 ms as the window often has a softening character at its ends, or larger units of constant duration. Most frequently, though, 'turns' are analysed that are based on speech onset until offset of one speaker in conversations. Onset and offset of speech are thereby often determined by a simple signal energy-based hysteresis, i. e., for a given minimum time, the speech pause energy level has to be exceeded to determine a speech onset and vice versa. While being an objective measure which is somewhat easy to obtain automatically, such turns may highly vary in length.

Alternatives are either pragmatic units like time slices, or proportions of longer units obtained by subdivision into parts of relative or absolute equal length, or 'meaningful' units with varying lengths, such as syllables, words, phrases. In [4], the word as the smallest possible, meaningful unit, is favoured for the analysis of emotion in speech, and in [50] it is shown that stressed syllables alone can be on a par with words as far as classification performance is concerned.

One may assume that units that are more connected to the task of analysis will become important in future research. In addition, incremental processing will be of increasing interest. Such incremental processing means providing an online estimate after the onset, updated continuously until the offset—this is often referred

to as 'gating'. Additionally, one may want to decide for the optimal unit in a multimodal context if for example also video or physiological information is analysed that typically investigates different units, but shall be fused in a synergistic manner. In fact, this problem can already arise to a certain extent when we want to fuse acoustic and linguistic information. Even for processing exclusively acoustic information, consideration of several temporal units at the same time may be interesting, to benefit from shorter frames in the case of spectral characteristics, but larger 'supra-segmental' units in the case of prosodic features, i. e., features dealing with intonation, stress, and rhythm, such as speaker's pitch.

9.2.2 Acoustic Feature Extraction

Arguably the most important step in the automated recognition of speaker states, traits and vocal behaviour is the extraction of features that are relevant for the task at hand and providing a compact representation of the problem.

Let us divide features into groups in the following to provide a comprehensive overview. While there is no unique classification into such groups, the most basic distinction is technology driven: The main groups are at first *acoustic* and *linguistic* features.

Depending on the type of affective state of vocal behaviour one aims to analyse, different weights will be given to these. To give an obvious example, linguistic features are of limited interest when assessing non-verbal vocal outbursts such as laughter, sighs, etc. However, investigating a speaker's emotion or personality, they bear high potential.

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

Until recently, a comparably small feature set (around 20 to 30 features) has usually been employed. The recent success of systematically generated static feature vectors is probably justified by the supra-segmental nature of most paralinguistic phenomena. These features are derived by projection of the low-level descriptors (LLD, for examples see Fig. 9.2.b–f) on single scalar values by descriptive statistical functionals, such as lower order moments or extrema. As an alternative, LLD features can be modelled directly. In general, these LLD calculate a value per speech 'frame' with a typical frame rate of 100 frames per second (fps, cf. Section 9.2).

The large number of LLD and functionals has recently promoted the extraction of very large feature vectors (brute-force extraction), up to many thousands of features obtained either by analytical feature generation or, in a few studies, by evolutionary generation (note that a similar development can be found in vision analysis, where large amounts of features are produced and then reduced). Such brute-forcing also

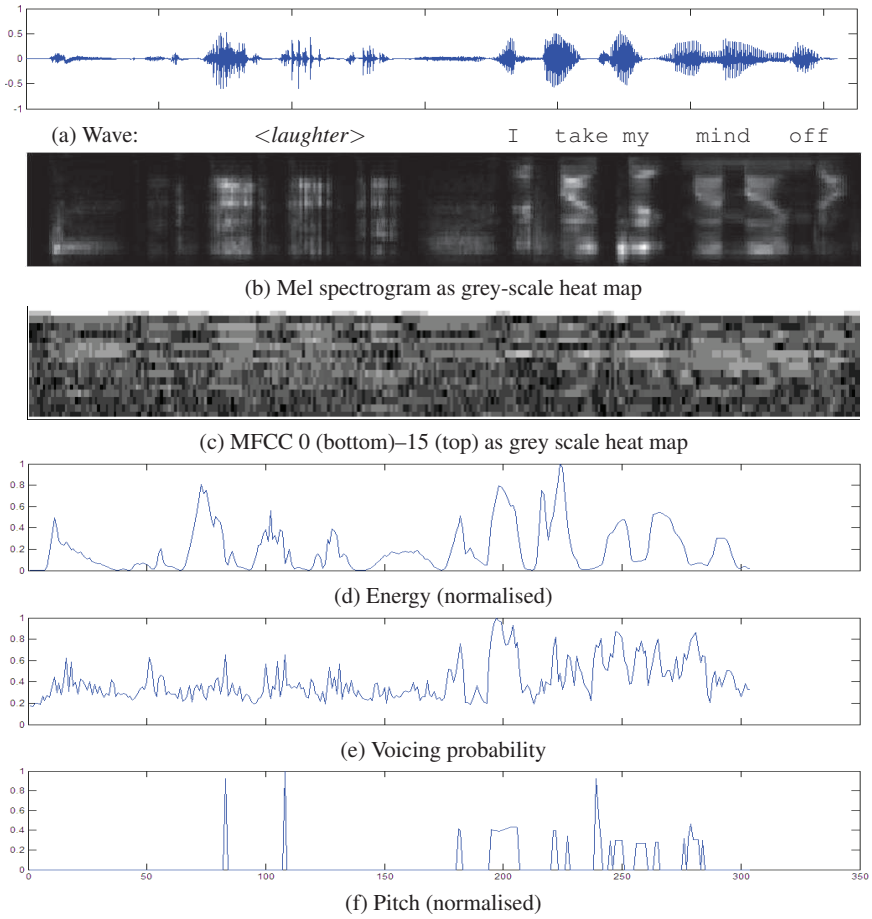

Fig. 9.2 Exemplary speech wave form over time in ms: laughter (0.0–150 ms) followed by "I take my mind off" taken from the SAL database (male speaker) and selected low-level descriptors.

often includes hierarchical functional application (e.g., mean of maxima) to better cope with statistical outliers.

However, also expert-based hand-crafted features still play their role, as these are lately often crafted with more emphasis put on details hard to find by sheer brute-forcing such as perceptually more adequate ones, or more complex features such as articulatory ones, for instance, (de-)centralisation of vowels (i. e., how exact and constant are vowels articulated). This can thus also be expected as a trend in future acoustic feature computation.

Let us now introduce the groups of features:

Intensity features usually model the loudness of a sound as perceived by the human ear, based on the amplitude, whereby different types of normalisation are

applied. Often, however, simply the frame energy is calculated for simplification, as human loudness perception requires a complex model respecting effects of duration and pitch of sound. 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. The loudness contour is thus the sequence of short-term loudness values extracted on a frame-by-frame basis.

The basics of ***pitch*** extraction have largely remained the same over the years; nearly all Pitch Detection Algorithms (PDA) are built using frame-based analysis: The speech signal is broken into overlapping frames and a pitch value is inferred from each segment mostly by the maximum in the autocorrelation function (ACF) in its manifold variants and derivatives such as Average Magnitude Difference Function (AMDF). AMDF substitutes the search of a maximum by a minimum search, as instead of multiplication of the signal with itself, a subtraction is considered for improved efficiency. Often, the Linear Predictive Coding (LPC) residual or a band pass filtered version is used over the original signal to exclude other influences from the vocal tract position. Pitch can also be determined in the time signal which allows for analysis of micro-perturbations, but is usually more error-prone. Pitch features are often made perceptually more adequate by logarithmic/semitone transformation, or normalisation with respect to some (speaker-specific) baseline. Pitch extraction is error-prone itself, which may influence recognition performance of the actual target problem [6]. However, the influence is rather small, at least for the current state-of-the-art in modelling pitch features.

Voice quality is a complicated issue in itself, since there are many different measures of voice quality [28], mostly clinical in origin and mostly evaluated for constant vowels only. Other, less well-known voice quality features were intended towards normal speech from the outset, e. g., those modelling 'irregular phonation', cf. [5]. Noise-to-Harmonic Ratio, jitter (micro-perturbation of pitch), shimmer (micro-perturbation of energy), and further micro-prosodic events are measures of the quality of the speech signal. Although they depend in part on other LLDs such as pitch and energy, they reflect peculiar voice quality properties such as breathiness or harshness.

The ***spectrum*** is characterised by formants (spectral maxima depending on the vocal tract position) 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 LPC or on cepstral analysis. A number of further spectral features can be computed either directly from a spectral transform such as by Fast Fourier Transform or the LPC spectrum, such as centroid, flux, and roll-off. Furthermore, the long term average spectrum over a unit can be employed: this averages out formant information, giving general spectral trends.

The ***cepstrum***, i. e., the inverse spectral transform of the logarithm of the spectrum, emphasises changes or periodicity in the spectrum, while being relatively robust against noise. Its basic unit is quefrency. Mel-Frequency Cepstral Coefficients (MFCCs)—as homomorphic transform with equidistant band-pass-filters on the Mel-scale—tend to strongly depend on the spoken content. Yet, they have been

proven beneficial in practically any speech processing task. MFCC are calculated based on the Fourier transform of a speech frame. Next, overlapping windows—usually of triangular shape and equidistant on the Mel scale—are used for mapping the powers of the obtained spectrogram onto the Mel scale to model human frequency resolution. Next, the logarithms of the powers are taken per such Mel frequency filter band—the idea at this point is to decouple the vocal tract transfer function from the excitation signal of human sound production. Then, the Discrete Cosine Transform (DCT) of the list of mel log powers is taken for de-correlation (other transforms are often used as well) to finally obtain the MFCCs as the amplitudes of the resulting DCT spectrum.

Perceptual Linear Predictive (PLP) coefficients and MFCCs are extremely similar, as they both correspond to a short-term spectrum smoothing—the former by an autoregressive model, the latter by the cepstrum—and to an approximation of the auditory system by filter-bank-based methods. At the same time, PLP coefficients are also an improvement of LPC by using the perceptually based Bark filter bank. Variants such as Mel Frequency Bands (MFB) that do not decorrelate features as a final step are also found in this particular field.

Wavelets give a short-term multi-resolution analysis of time, energy and frequencies in a speech signal. Compared to similar parametric representations they are able to minimise the time-frequency uncertainty.

Duration features model temporal aspects. 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—because they are measured in seconds, and because they are often highly correlated with duration features. By that, they can be distinguished according to the way they are extracted: Those that represent temporal aspects of other acoustic base contours, and those that exclusively represent the 'duration' of higher phonological units, like phonemes, syllables, words, pauses, or utterances. Duration values are usually correlated with linguistic features: For instance, function words are shorter on average, content words are longer: This information can be used for classification, no matter whether the signal is encoded in linguistic, or acoustic (i. e., duration) features.

Subsequent to the LLD extraction, a number of operators and functionals can be applied to obtain feature vectors of equal size from each LLD. Functionals provide a sort of normalisation over time: LLD associated with words (and other units) have different lengths, depending on the duration of each word and on the dimension of the window step; with the usage of functionals, we obtain one feature vector per chunk, with a constant number of elements that can be modelled by a static classifier or regressor. This cascade procedure, namely LLD extraction followed by functional application, has two major advantages: Features derived from longer time intervals can be used to normalise local ones, and the overall number of features might be opportunely shrunk or expanded with respect to the number of initial LLDs [38].

More intelligent brute-forcing can be obtained by search masks and by a broader selection of functionals and parameters. In this way, an expert's experience can be combined with the freedom of exploration taken by an automatic generation.

Before functionals are applied, LLDs can be filtered or (perceptually) transformed, and first or second derivatives are often calculated and end up as additional LLDs. Functionals can range from statistical ones to curve fitting methods. The most popular statistical functionals cover the first four moments (mean, standard deviation, skewness and kurtosis), higher order statistics (extreme values and their temporal information), quartiles, amplitude ranges, zero-crossing rates, roll-on/-off, on-/off-sets and higher level analysis. Curve fitting methods (mainly linear) produce regression coefficients, such as the slope of linear regression, and regression errors (such as the mean square errors between the regression curve and the original LLD). A comprehensive list of functionals adopted so far in this field can be found in [7].

Fig.9.3 provides an overview of the commonly used features and the principle of their brute-forcing in several layers.

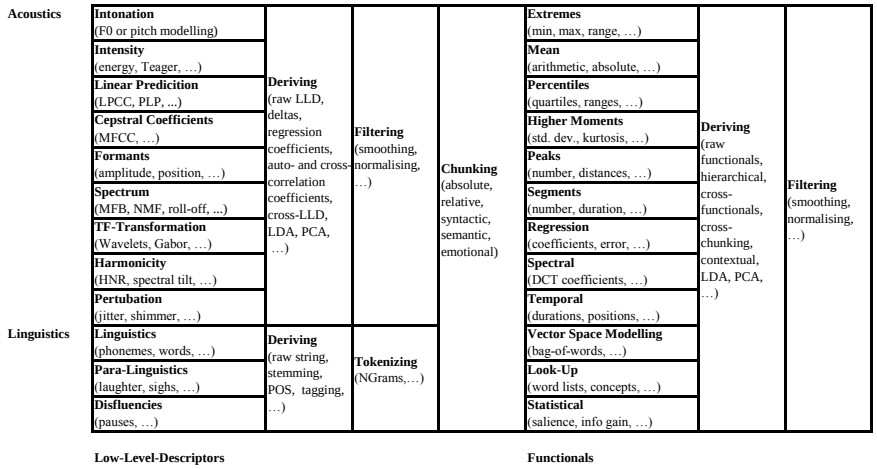

Fig. 9.3 Overview on features commonly used for acoustic and linguistic emotion recognition. Abbreviations: Linear Prediction Cepstral Coefficients (LPCC), Mel Frequency Bands (MFB).

As a typical example, we can have a look at the 'large' feature set of the public open source toolkit openSMILE [16] that is frequently used in the field: Acoustic feature vectors of 6.552 dimensions are extracted as 39 functionals of 56 acoustic LLDs, including first and second order delta regression coefficients: Table 9.2 summarizes the statistical functionals which were applied to the LLDs shown in Table 9.1 to map a time series of variable length onto a static feature vector as described above.

Table 9.1 33 exemplary typical Low-Level Descriptors (LLD).

Feature Group	Features in Group
Raw Signal	Zero-crossing-rate
Signal energy	Logarithmic
Pitch	Fundamental frequency F_0 in Hz via Cepstrum and Autocorrelation (ACF). Exponentially smoothed F_0 envelope.
Voice Quality	Probability of voicing ($\frac{ACF(T_0)}{ACF(0)}$)
Spectral	Energy in bands 0–250 Hz, 0–650 Hz, 250–650 Hz, 1–4 kHz 25 %, 50 %, 75 %, 90 % roll-off point, centroid, flux, and rel. pos. max. / min.
Mel-spectrum	Band 1–26
Cepstral	MFCC 0–12

Table 9.2 39 exemplary functionals as typically applied to LLD contours.

Functionals	#
Respective rel. position of max./min. value	2
Range (max.-min.)	1
Max. and min. value - arithmetic mean	2
Arithmetic mean, Quadratic mean, Centroid	3
Number of non-zero values	1
Geometric, and quadratic mean of non-zero values	2
Mean of absolute values, Mean of non-zero abs. values	2
Quartiles and inter-quartile ranges	6
95 % and 98 % percentile	2
Std. deviation, variance, kurtosis, skewness	4
Zero-crossing rate	1
# of peaks, mean dist. btwn. peaks, arth. mean of peaks, arth. mean of peaks - overall arth. mean	4
Linear regression coefficients and error	4
Quadratic regression coefficients and error	5

9.2.3 Feature Selection

To improve reliability and performance, but also to obtain more efficient models in terms of processing speed and memory requirements, one usually has to select a subset of features that best describe the audio analysis task. A multiplicity of feature selection strategies have been employed, e.g., for recognition of emotion or personality, but even for non-linguistic vocalisations, different types of features are often considered and selected.

Ideally, feature selection methods should not only reveal single (or groups of) most relevant attributes, but also de-correlate the feature space. Wrapper-based selection—that is employing a target classifier's accuracy or regressor's cross-

correlation as optimisation criterion in 'closed loop'—is widely used to tailor the feature set in match with the machine learning algorithm. However, even for relatively small data-sets, exhaustive selection considering any permutation of features is still not affordable. Therefore, the search in the feature space must employ some more restrictive, and thus less optimal, strategies. Probably the most common procedure chosen is the *sequential forward search*—a hill climbing selection starting with an empty set and sequentially adding best features; as this search function is prone to nesting, an additional floating option should be added: At each step one or more features are deleted and it is checked if others are more suited.

Apart from wrappers, less computationally expensive 'filter' or 'open loop' methods are frequently used if repeated selection is necessary, such as information theoretic filters and correlation-based analysis.

There are, however, also classifiers and regressors with 'embedded' selection, such as Decision Trees or Ridge Regression.

As a refinement, *hierarchical* approaches to feature selection try to optimise the feature set not globally for all target classes, but for groups of them, mainly couples.

Apart from genuine selection of features, the *reduction* (i.e. feature extraction) of the feature space is often considered to reduce the complexity and number of free parameters to be learnt for the machine learning algorithms while benefiting from all original feature information. This is achieved by mapping of the input space onto a less dimensional target space, while keeping as much information as possible. Principal Component Analysis (PCA) and Linear or Heteroscedastic Discriminant Analysis (LDA) are the most common techniques.

While PCA is an unsupervised feature reduction method and thus is often suboptimal for more complex 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, i. e., it is a discriminative method as the name indicates.

In fact, none of these methods is optimal: There is no straight forward way of knowing the optimal target space size—typically the variance covered is a decisive measure. Further, a certain degree of normal distribution is expected, and LDA additionally demands linear separability of the input space. PCA and LDA are also not very appropriate for feature mining, as the original features are not retained after the transformation.

Finally, Independent Component Analysis (ICA) and Non-negative Matrix Factorization (NMF) [25] can be named. ICA maps the feature space onto an orthogonal space and the target features have the attractive property of being statistically independent. NMF is a recent alternative to PCA in which the data and components have to be non-negative. NMF is at present mainly employed for large linguistic feature sets.

Also, it seems important to mention that there is a high danger of over-adaptation to the data that features are selected upon. As a counter-measure, it seems wise to address feature importance across databases [15].

9.2.4 Classification and Regression

A number of factors motivate consideration of diverse machine learning algorithms, the most important being tolerance to high dimensionality, capability of exploiting sparse data, and handling of skewed classes. In addition, more general considerations such as the ability to solve non-linear problems, discriminative learning, self-learning of relevant features, high generalisation, on-line adaptation, handling of missing data, efficiency with respect to computational and memory costs in training and recognition, etc. can play a decisive role. Further, one may wish for human-readable learnt models, provision of meaningful confidence measures and handling of input uncertainties (features like pitch are not determined flawlessly—here an algorithm may also consider a certainty measure in addition to the predicted pitch value) for optimal integration in a system context.

As previously mentioned, we can basically differentiate between classifiers that decide for discrete classes and regressors that estimate a continuous value in the sense of a function learner. However, practically any classifier can be turned into a regressor and vice versa, although the result would not necessarily be as efficient for this task as for its 'native' task. Classification using regression methods can for example be obtained by having each class binarised and one regression model built for each class value. The other way round, a regression scheme can be realised by using any classifier on a copy of the data where the continuous 'class' is discretised. The predicted value is the expected value of the mean class value for each discretised interval, based on the predicted probabilities for each interval [58] (also see 'squashing' in Chapter 1).

The problem of a high dimensional feature set is usually better addressed by feature selection and elimination before actual classification takes place. Popular classifiers such as Linear Discriminant Classifiers (LDCs) and k-Nearest Neighbor (kNN) classifiers have been used since the very first studies. However, they suffer from the increasing number of features that leads to regions of the feature space where data are very sparse ('curse of dimensionality'). Classifiers such as kNN that divide the feature space into cells are affected by the curse of dimensionality and are sensitive to outliers. A natural extension of LDCs are Support Vector Machines (SVM): they combine discriminative learning and solving of non-linear problems by a Kernel-based transformation of the feature space. While they may not always lead to the best result, they provide good generalisation properties, and can be seen as a sort of state-of-the-art classifier (or regressor, as the related Support Vector Regression allows for handling of continuous problem descriptions).

Small data sets are, in general, best handled by discriminative classifiers. The most used non-linear discriminative classifiers apart from SVM are likely to be Artificial Neural Networks (ANNs) and decision trees. Decision hyperplanes learnt with ANN might become very complex and depend on the topology of the network (number of neurons), on the learning algorithm (usually a derivation of the well-known Backpropagation algorithm) and on the activity rules. For this reason, ANNs are less robust to over-fitting, and require greater amounts of data to be trained on. The recent incorporation of a long-short-term memory function seems to be a promising

future direction [60] that may raise their popularity. Also, multi-task learning is well established, which may be of particular interest in this field to, e.g., assess emotion and personality in one pass, benefiting from mutual dependencies.

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

As many paralinguistic tasks (such as emotion) are not evenly distributed among classes in databases, balancing of the training instances with respect to instances per class is often a necessary step before classification [43]. The balancing of the output space can be addressed either by considering proper class weights (e. g., priors), or by resampling, i. e., (random) up- or down-sampling. Class priors are implicitly taken into account by discriminative classifiers.

As explained above, applying functionals to LLD is done for obtaining the same number of features for different lengths of units such as turns or words. Dynamic classifiers like Hidden Markov Models, Dynamic Bayesian Networks or simple Dynamic Time Warp allow to skip this step in the computation by implicitly warping observed feature sequences over time. Among dynamic classifiers, Hidden Markov Models (HMM) have been used widely. The performance of static modelling through functionals is often reported as superior [43], as paralinguistic tasks are apparently better modelled on a time-scale above frame-level; note that a combination of static features such as minimum, maximum, onset, offset, duration, regression, etc. implicitly shape contour dynamics as well. A possibility to use static classifiers for frame-level feature processing is further given by multi-instance learning techniques, where a time series of unknown length is handled by SVM or similar techniques. Also, a combination of static and dynamic processing may help improve overall accuracy [55].

Ensembles of classifiers combine their individual strengths, and might improve training stability. There exists a number of different approaches to combine classifiers. Popular are methods based on majority voting such as *Bagging*, *Boosting* and other variants (e. g., *MultiBoosting*). More powerful, however, is the combination of diverse classifiers by the introduction of a meta-classifier that learns 'which classifier to trust when' and is trained only on the output of 'base-level' classifiers, known as *Stacking*. If confidences are provided on lower level, they can be exploited as well. Still, the gain over single strong classifiers such as SVM may not justify the extra computational costs [39].

In line with the different models to describe the named problems, e.g., by classes or continuous dimensions, also different approaches towards classification are needed: As real-life application is not limited to prototypical cases, also *detection* as opposed to classification can be expected as an alternative paradigm: 'Out-of-

vocabulary' classes need to be handled as well (as an example, imagine the emotions anger and neutral having been trained, but in the recognition phase joy appears), and apart from the easiest solution of introducing a garbage class [43], detection allows for more flexibility. Detection is thereby defined by inheriting a rejection threshold. In this respect, *confidence measurements* should be mentioned, which are, however, not sufficiently explored, yet.

9.2.5 Parameter Tuning

Apart from the selection of features, a crucial factor in optimisation of performance is the 'fine tuning' of classifiers' parameters on a development partition of the training data. Typically such parameters comprise the exponent of polynomial Kernels for Support Vector Machines or the number of nearest neighbors in k nearest neighbor classification, etc. While these can be optimised by equidistant scanning of the parameter space, more efficient methods exist, of which grid search is the most frequently encountered in the field (e. g., [23]). Grid search is a greedy algorithm that first performs a rough search over the values and then narrows down on promising areas in terms of best accuracy for a classifier or cross-correlation for a regressor in a recursive manner. Obviously, just as for the selection of features, such searches do not necessarily lead to the global optimum, if the search is not exhaustive. In addition, they also may differ drastically for different databases, depending on their size and complexity. Thus, again cross-corpus parameter tuning may help find more generally valid sets than considering just intra-database variation.

9.3 'Speech' — The (Non-) Linguistic Analysis

As said, in this chapter speech stands for spoken text, i. e., the analysis of textual cues. Apart from the analysis of linguistic content, non-linguistic vocal outbursts as laughter are dealt with.

9.3.1 Analysis of linguistic content

Spoken or written text provides cues on emotion, personality or further states and traits. This is usually reflected in the usage of certain words or grammatical alterations, which means in turn, in the usage of specific higher semantic and pragmatic entities. A number of approaches exists for this analysis: key-word spotting [12], rule-based modelling [26], Semantic Trees [61], Latent Semantic Analysis [17], World-knowledge-Modelling, Key-Phrase-Spotting, String Kernels [37], and Bayesian Networks [8]. Contextual and pragmatic information has been mod-

elled as well, e.g., dialogue acts [26], or system and user performance [1]. 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* and *Bag of Words* (*vector space modelling*), cf. [41].

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 N words. For recognition of a target problem such as emotion or personality, the probability of each target class is determined per N-gram of an utterance. In addition, word-class based N-grams can be used as well, to better cope with data sparseness. For the example of emotion recognition, due to data sparseness mostly uni-grams (N=1) have been applied so far, besides bi-grams (N=2) and trigrams (N=3) [2]. The actual target class is calculated by the posterior probability of the class given the actual word(s) by maximum likelihood or a-posteriori estimation. An extension of N-Grams which copes with data sparseness even better is *Character N-Grams*; in this case larger histories can be used.

Bag of Words is a well-known numerical representation form of texts in automatic document categorisation [22]. It has been successfully ported to recognise sentiments or emotion [41] and can equivalently be used for other target problems. In this approach 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 intelligent reduction. 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.

In addition, exploitation of on-line knowledge sources without domain specific model training has recently become an interesting alternative or addition [42]—e.g., to cope with out-of-vocabulary events. The largely related fields of opinion mining and sentiment analysis in text bear interesting alternatives and variants of methods.

Although we are considering the analysis from spoken text, only few results for paralinguistic speaker state and trait recognition rely on automatic speech recognition (ASR) output [40] rather than on manual annotation of the data. As ASR of affective speech itself is a challenge [52], this step is likely to introduce errors. To some extent errors deriving from ASR and human transcription can be eliminated by soft-string-matching such as tolerating a number of deletions, insertions, or substitutions of characters.

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

Tokenisation, i. e., chunking of the continuous text string similar to chunking of the acoustic stream above, can be obtained by mapping the text onto word classes:

Stemming is the clustering of morphological variants of a word (such as “fight”, “fights”, “fought”, “fighting”, etc.) 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 their representative morphological variant, for instance by (Iterated) Lovins or Porter stemmers that are based on suffix lists and rules. Part-of-Speech (POS) tagging is a very compact approach where classes such as nouns, verbs, adjectives, particles, or more detailed sub-classes are modelled [51]. POS tagging and stemming have been studied thoroughly [37].

Also *sememes*, i. e., semantic units represented by lexemes, can be clustered into higher semantic concepts such as generally positive or negative terms [7]. In addition, non-linguistic vocalisations can easily be integrated into the vocabulary [48].

9.3.2 Analysis of non-linguistic content

While non-linguistic events such as laughter can be modelled as an extra type of feature stream or information, a very simple way is to include them in the string of linguistic events. On the positive side, this can put events like laughter in direct relation with the words. This may, however, disrupt linguistically meaningful sequences of words. Alternatively, frequencies of occurrences normalised to time or even functionals applied to occurrences are alternative solutions.

9.3.3 (Non-)Linguistic Feature Extraction

While non-linguistic events can be recognised directly in-line with speech as by an Automatic Speech Recogniser, it seems noteworthy to mention that one can also use brute-forced features as described above. Interestingly, little to no difference is reported for these two types of representation [48]. The incorporation into a speech recogniser has the advantage that speech is recognised with integration of higher-level knowledge as coming from the language model. However, if non-linguistic vocalisations are modelled on their own, a richer feature representation can be used that may unnecessarily increase space complexity for speech recognition. Furthermore, in case of non-linguistic vocalisations such as laughter, these may also appear ‘blended’ with speech, as in the case of ‘speech-laughter’, i. e., laughter while actually speaking words. This cannot easily be handled in-line with ASR, as the ASR engine typically would have to decide for phonetically meaningful units or laughter.

9.3.4 Classification and Regression

In principle, any of the formerly discussed learning algorithms can be used for linguistic analysis, as well. However, different ones may be typically preferred owing to the slightly different characteristics of linguistic features. In particular, statistical algorithms and Kernel machines such as Support Vector Machines are popular. Noteworthy, there are also specific algorithms that may operate directly on string input such as the String Kernels [27] for Support Vector Machines.

9.4 Data, Benchmarks, and Application Examples

In this section, let us first have a look at some typical databases focusing on affective speaker states. Next, two examples of systems that analyse vocal behaviour on different levels will shortly be described.

9.4.1 Frequently Encountered Data-Sets and their Benchmarks

As benchmark databases, nine most frequently encountered databases that span a range from acted over induced to spontaneous affect portrayals are presented, focusing in particular on affective speaker states. For better comparability of obtained performances among corpora, the diverse affect groups are additionally mapped onto the two most popular axes in the dimensional emotion model as in [46]: arousal (i. e., passive ("-") vs. active ("+")) and valence (i. e., negative ("-") vs. positive ("+")). Note that these mappings are not straight forward—here we will favour better balance among target classes. Let us further discretize into the four quadrants (q) 1–4 of the arousal-valence plane for continuous labelled corpora. In the following, each set is shortly introduced, including the mapping to binary arousal/valence by "+" and "-" per emotion and its number of instances in parentheses. Note that the emotions are referred to as in the original database descriptions.

The *Danish Emotional Speech* (DES) database [13] is professionally acted and contains nine sentences, two isolated words, and chunks that are located between two silent segments of two passages of fluent text. Affective states contain angry (+/-, 85), happy (+/+, 86), neutral (-/+, 85), sadness (-/-, 84), and surprise (+/+, 79).

The *Berlin Emotional Speech Database* (EMOD) [9] features professional actors speaking ten emotionally undefined sentences. 494 phrases are commonly used: angry (+/-, 127), boredom (-/-, 79), disgust (-/-, 38), fear (+/-, 55), happy (+/+, 64), neutral (-/+, 78), and sadness (-/-, 53).

The *eNTERFACE* (eNTER) [29] corpus consists of recordings of subjects from 14 nations speaking pre-defined spoken content in English. The subjects listened to six successive short stories eliciting a particular emotion out of angry (+/-, 215),

disgust (-/-, 215), fear (+/-, 215), happy (+/+, 207), sadness (-/-, 210), and surprise (+/+, 215).

The *Airplane Behavior Corpus* (ABC) [47] is based on induced mood by pre-recorded announcements of a vacation (return) flight, consisting of 13 and 10 scenes. It contains aggressive (+/-, 95), cheerful (+/+, 105), intoxicated (+/-, 33), nervous (+/-, 93), neutral (-/+, 79), and tired (-/-, 25) speech.

The *Speech Under Simulated and Actual Stress* (SUSAS) database [21] serves as a first reference for spontaneous recordings. Speech is additionally partly masked by field noise in the chosen speech samples of actual stress. SUSAS content is restricted to 35 English air-commands in the speaker states of high stress (+/-, 1 202), medium stress (+/-, 1 276), neutral (-/+, 701), and scream (+/-, 414).

The *Audiovisual Interest Corpus* (AVIC) [48] consists of spontaneous speech and natural emotion. In its scenario setup, a product presenter leads subjects through a commercial presentation. AVIC is labelled in "levels of interest" (loi) 1–3 having loi1 (-/-, 553), loi2 (+/+, 2279), and loi3 (+/+, 170).

The *Belfast Sensitive Artificial Listener* (SAL) data are part of the HUMAINE database. The subset used—as in [59]—has an average length of 20 minutes per speaker of natural human-SAL conversations. The data have been labelled continuously in real time with respect to valence and activation, using a system based on FEELtrace [11]. The annotations were normalized to zero-mean globally and scaled so that 98 % of all values are in the range from -1 to +1. The 25 recordings have been split into turns using energy based Voice Activity Detection. Labels for each obtained turn are computed by averaging over the complete turn. Per quadrant the samples are: q1 (+/+, 459), q2 (-/+, 320), q3 (-/-, 564), and q4 (+/-, 349).

The *SmartKom* (Smart) [53] corpus consists of Wizard-Of-Oz dialogues. For evaluations, the dialogues recorded during a public environment technical scenario are used. It is structured into sessions which contain one recording of approximately 4.5 min length with one person, and labelled as anger/irritation (+/-, 220), helplessness (+/-, 161), joy/gratification (+/+, 284), neutral (-/+, 2179), pondering/reflection (-/+, 643), surprise (+/+, 70), and unidentifiable episodes (-/+, 266).

Finally, the *Vera-Am-Mittag* (VAM) corpus [18] consists of recordings taken from a German TV talk show. The audio recordings were manually segmented to the utterance level, whereas each utterance contains at least one phrase. The labelling is based on a discrete five point scale for each of the valence, activation, and dominance dimensions. Samples among quadrants are q1 (+/+, 21), q2 (-/+, 50), q3 (-/-, 451), and q4 (+/-, 424).

Further details on the corpora are summarized in Table 9.3 and found in [45]. Looking at the table, some striking facts become evident: most notably, the high sparseness of data with these sets typically providing only one hour of speech from only around 10 subjects. In related fields as ASR, several hundreds of hours of speech and subjects are typically contained. This is one of the major problems in this field at the moment. In addition, one sees that often such data are rather acted as opposed to natural and that the linguistic content is often restricted to pre-defined phrases or words. Obviously, this is rather an annotation challenge, as emotional speech data per se would be available.

Table 9.3 Overview on the selected corpora (E/D/G: English/German/Danish, act/ind/nat: acted/induced/natural, Lab: labellers, Rec: recording environment, f/m: (fe-)male subjects). Speaker-independent recognition performance benchmarks are provided by weighted (WA) and unweighted (UA) average accuracy. * indicates results obtained by Support Vector Machines if these had outperformed Deep Neural Networks as taken in all other cases.

Corpus	Speech	# All	h:mm	# m	# f	# Lab	Rec	kHz	# All		# Arousal		# Valence	
									UA	WA	UA	WA	UA	WA
ABC	G fixed act	430	1:15	4	4	3	studio	16	56.1	61.5	69.3	80.6	79.6	79.0
AVIC	E free nat	3002	1:47	11	10	4	studio	44	59.9	79.1	75.6	85.3	75.2	85.5
DES	D fixed act	419	0:28	2	2	–	studio	20	59.9*	60.1*	90.0	90.3	71.7	73.7
EMOD	G fixed act	494	0:22	5	5	–	studio	16	84.6*	85.6*	97.6	97.4	82.2	87.5
eNTER	E fixed ind	1277	1:00	34	8	2	studio	16	72.5*	72.4*	78.1	79.3*	78.6*	80.2*
SAL	E free nat	1692	1:41	2	2	4	studio	16	35.9	34.3	65.1	66.4	57.7	53.0
Smart	G free nat	3823	7:08	32	47	3	noisy	16	25.0	59.5	55.2	79.2	52.2	89.4
SUSAS	E fixed nat	3593	1:01	4	3	–	noisy	8	61.4*	56.5*	68.2	83.3	74.4	75.0
VAM	G free nat	946	0:47	15	32	6/17	noisy	16	39.3	68.0	78.4	77.1	52.4	92.3

To provide an impression on typical performances in the field, the last columns of Table 9.3 provide weighted (WA) and unweighted (UA) accuracy of speaker independent recognition by feature reduction with Deep Neural Networks and subsequent distance classification (DNN) or Support Vector Machines on the original space (SVM), with the 'large' standard feature set of openSMILE introduced in Section 9.2. Such speaker independence is obtained by partitioning the data in a 'leave-one-speaker-out', or—for databases with many speakers, here starting at more than 10—'leave-one-speaker-group-out' cross-validation manner. This cross-validation is very popular in this field, as it allows to test on all instances of the very limited resources. The accuracy of the classifier that produced the higher result on development data is presented, each, in the table. Balancing of the training partition is used to cope with the imbalance of instances in the training set among affective states. More details are found in [54]. If we now look at these numbers, it seems clear that acted data are considerably easier to recognise automatically owing to their often exaggerated display. Naturally, this is more true in the case where the verbal content is limited. Another interesting but typical fact is that arousal is usually recognised more reliably. To better handle valence, one would best integrate linguistic feature information.

To conclude this chapter, let us now have a look at two examples of voice and speech analysis systems that are currently used in practice and that investigate a number of different issues in contrast to the above named problems.

9.4.2 Human-to-Human Conversation Analysis

The AVIC corpus as introduced above and as used in the INTERSPEECH Paralinguistic Challenge [44] provides a good example of vocal behaviour analysis in natural human conversational speech: In [48] an analysis of non-linguistic vocalisations and speaker's interest is shown, based on these non-linguistic vocalisations and further acoustic and linguistic features as introduced above. The acoustic feature space consists of a brute-force large space with subsequent feature selection with the classifier in the loop and SVM for classification. Linguistic features are the described Bag of Words, integrated directly into the feature vector. Non-linguistics are recognised in a separate recognition pass by the same basis of acoustic features but optimised for this task. The occurrence of non-linguistics is simply added to the linguistic feature string.

To demonstrate efficiency over weighted and unweighted accuracies like we presented in Table 9.3, 40 participants interacted with a virtual product and company tour that took participants' interest into account to change topic in case of their boredom. Three variants were used: topic change after a fixed time, with fully automatic interest recognition with this system or by a human Wizard-of-Oz. The question "Did you think the system was taking into account your interest?" was positively answered by 35 % in the first case (no interest recognition), by 63 % in the second case (fully automatic interest recognition) and by 84 % in the last case (human interest recognition) nicely demonstrating that the technology seems to be generally working, but that there is also still headroom for improvement to reach human-like performance.

In our next example, let us switch to human-computer conversation.

9.4.3 Human-to-Agent Conversation Analysis

In the European SEMAINE project, a Sensitive Artificial Listener (SAL)—a multimodal dialogue system with the social interaction skills needed for a sustained conversation with a human user—was built [36]. Such a system demands for on-line incremental emotion recognition, in order to select responses as early as possible. In SEMAINE, the user's affective state and non-verbal behaviour are the major factors for deciding upon agent actions. Therefore it is essential to obtain a fast estimate of the user's affective state as soon as the user starts speaking, and refine the estimate as more data are available (for example, in [56] 350 ms are suggested for human-like back-channelling in certain situations). Moreover, the system needs to know how reliable the affect dimension predictions are, in order to identify salient parts of highly affective speech reliably, in order to choose appropriate actions. The verbal dialogue capabilities of the system are very limited on purpose. They are basically limited to agreement/disagreement, emotionally relevant keywords, and changing characters (see below for more information on the four different SEMAINE characters/personalities).

In the SEMAINE system, which is freely available as release for research and tutoring[1], *Feature extractors* analyse low-level audio and video signals, and provide feature vectors periodically (10 ms) to the *analysers*, which process the low-level features and produce a representation of the current user state, in terms of epistemic-affective states (emotion, interest, etc.). Since, automatic speech recognition or emotion recognition might benefit from the dialogue context or user profiles at a higher level, *interpreter* components are contained in the system to address this issue. A typical and obvious example is the 'turn-taking interpreter', which decides when it is time for the agent to take the turn. These are examples—the SEMAINE API goes beyond these capabilities [36].

The next group of components is a set of *action proposers* which produce agent action candidates independently from one another. The action proposers take their input mainly from the user, dialog, and agent state. As for the voice and speech analysis, the free open source openSMILE[2] [16] module extracts state of the art features stemming from the large feature set described in Section 9.2 for voice activity detection, prosody analysis, keyword spotting, non-linguistic vocalisation detection, and an acoustic emotion recognition module. Prosodic features, which are used by other SEMAINE components (e.g., for turn taking decisions), include pitch contour, energy/loudness, and per pseudo-syllable pitch direction estimates. Classification and regression are based on on-line Long Short-Term Memory (LSTM) Recurrent Neural Networks.

The SEMAINE keyword spotter detects a set of 176 keywords (including the non-linguistic vocalisations 'breathing', 'laughing', and 'sighing' handled in-line) which are relevant for the dialogue management and for linguistic emotion recognition. As system responses have to be prepared already before the user has finished speaking, the keyword spotter operates incrementally. The acoustic feature extractor extracts large sets of acoustic features used for recognition of the user"s affective state (5 continuous dimensions: arousal, expectation, intensity, power, and valence, and 3 'levels of interest': bored, neutral, interested) incrementally in real-time with regression models trained on the SEMAINE database. High dimensional acoustic feature vectors are concatenated with linguistic Bag of Words vectors, which are computed from the keyword spotter output. An incremental segmentation scheme is applied to the continuous audio input: analysis is conducted over windows of up to five seconds length, which are shifted forward in time with a step of two seconds, thus producing an estimate of the user's affective state every two seconds. The same acoustic feature set as for the 5 dimensional affect recognition is used in models trained on the AVIC corpus, as described in Section 9.4.

[1] http://semaine.sourceforge.net/

[2] http://www.openaudio.eu

9.5 Summary

This chapter introduces the principles of analysis of acoustic and linguistic properties of the voice and speech for the assessment of speaker states, traits, and vocal behaviour such as laughter. While voice and speech analysis follows the general pattern recognition paradigm, one of its major peculiarities might be the choice of features. In particular, brute-forcing of rather large feature spaces and subsequent selection are common procedure. Further, the type of features—either low-level descriptors that provide a value per short frames of speech (usually around 100 per second), or functionals per larger units of time—decide on the type of classifier or regressor. Owing to the diversity of tasks reaching from emotion to personality or laughter, different machine learning algorithms are preferred and used. Features and parameters of these learning algorithms can be fine tuned to the problem and data at hand, yet this comes at the risk of over-adaptation.

Another main peculiarity is the ambiguity of ground truth due to the often very subjective nature of labelling and to the fact that models for the description of tasks like emotion or personality prediction are usually non-trivial. Finally, one of the most decisive limiting factors is typically the ever-present lack of data—in particular of natural data of a multiplicity of speakers and languages and cultures. However, reasonably functioning accuracies independent of speakers can already be provided allowing for first systems to be operated 'in the wild'. At the same time, further research will usually be needed to achieve human-like performance for cross-database and task operation potentially in the presence of noise and reverberation.

9.6 Questions

1. Discuss the difference between speaker states and traits and list at least three examples for each of these two.
2. Name at least five ideal conditions for a collection of speech and voice data.
3. Describe the chain of processing for the analysis of speech and voice including each block and its function.
4. Explain the difference between Low-Level Descriptors and functionals and name at least five examples for each of these two.
5. Which units for chunking exist and why is chunking needed?
6. Name at least five ideal conditions for a classification or regression algorithm.
7. How can linguistic information be incorporated in the analysis process? Name at least two alternative strategies and describe their principle.

9.7 Glossary

- *Chunking* Segmentation of the audio stream into units of analysis.

- *Low-Level Descriptor* Time series of extracted feature values—typically on frame level.
- *Functional* Projection of a function onto a scalar value by statistical or other functions.
- *Regression* Mapping of a feature input vector onto a real-valued output instead of discrete classes as in classification.
- *Prosody* Rhythm, stress, and intonation of speech.
- *N-Gram* Subsequence (e. g., words or characters) from a given sequence (e. g., turns or words) with n consecutive items.
- *Bag of Words* Representation of text (e. g., of a speaker turn) as numerical feature representation (e. g., per word or N-Gram of words) without modelling of order of units.
- *Wizard-of-Oz (experiment)* The Wizard-of-Oz simulates an autonomous system by a human response during an experiment, for example to test new technology and its acceptance before it actually exists or to allow for data collection.
- *Arousal* Physiological/psychological state of being (re-)active.
- *Valence* Here used to categorize emotion as 'positive' (e. g., joy) or 'negative' (e. g., anger).
- *Non-Linguistic (event)* Describes vocal outbursts of non-linguistic character such as laughter or sigh.
- *Pitch* Perceived frequency of sound (here speech) as opposed to the fundamental frequency—perception can vary according to the intensity, duration, and frequency of the stimuli.
- *Keyword Spotter* Automatic Speech Recogniser that focuses on the highly robust detection of selected words within a speech or general audio stream.
- *Lexeme* In linguistics, this roughly subsumes a number of forms (such as flexions) of a single word (such as *speak*, *speaks*, *spoken* as forms of the lexeme SPEAK).

References

1. 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 *Proc. Interspeech*, pages 797–800, Pittsburgh, PA, USA, 2006.
2. J. Ang, R. Dhillon, E. Shriberg, and A. Stolcke. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proc. Interspeech*, pages 2037–2040, Denver, CO, USA, 2002.
3. A. Batliner, B. Schuller, S. Schaeffler, and S. Steidl. Mothers, Adults, Children, Pets — Towards the Acoustics of Intimacy. In *Proc. ICASSP 2008*, pages 4497–4500, Las Vegas, NV, 2008.
4. A. Batliner, D. Seppi, S. Steidl, and B. Schuller. Segmenting into adequate units for automatic recognition of emotion-related episodes: a speech-based approach. *Advances in Human-Computer Interaction*, 2010. vol. 2010, Article ID 782802, 15 pages.
5. A. Batliner, S. Steidl, and E. Nöth. Laryngealizations and Emotions: How Many Babushkas? In *Proc. of the International Workshop on Paralinguistic Speech – between Models and Data (ParaLing'07)*, pages 17–22, Saarbrücken, Germany, 2007.

6. 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 *Proc. ICPhS*, pages 2201–2204, Saarbrücken, Germany, 2007.
7. A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, and N. Amir. Whodunnit – Searching for the Most Important Feature Types Signalling Emotional User States in Speech. *Computer Speech and Language*, 25:4–28, 2011.
8. J. Breese and G. Ball. Modeling emotional state and personality for conversational agents. Technical Report MS-TR-98-41, Microsoft, 1998.
9. F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. A Database of German Emotional Speech. In *Proc. Interspeech*, pages 1517–1520, Lisbon, Portugal, 2005.
10. Nick Campbell, Hideki Kashioka, and Ryo Ohara. No laughing matter. In *Proc. Interspeech*, pages 465–468, Lisbon, Portugal, 2005.
11. R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder. Feeltrace: An instrument for recording perceived emotion in real time. In *Proc. of the ISCA Workshop on Speech and Emotion*, pages 19–24, Newcastle, Northern Ireland, 2000.
12. C. Elliott. *The Affective Reasoner: A Process Model of Emotions in a Multi-agent System*. PhD thesis, Dissertation, Northwestern University, 1992.
13. I. S. Engbert and A. V. Hansen. Documentation of the Danish Emotional Speech Database DES. Technical report, Center for PersonKommunikation, Aalborg University, Denmark, 2007. `http://cpk.auc.dk/~tb/speech/Emotions/`, last visited 11/13/2007.
14. F. Enos, E. Shriberg, M. Graciarena, J. Hirschberg, and A. Stolcke. Detecting deception using critical segments. pages 2281–2284.
15. F. Eyben, A. Batliner, B. Schuller, D. Seppi, and S. Steidl. Cross-Corpus Classification of Realistic Emotions • Some Pilot Experiments. In *Proc. 3rd International Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, pages 77–82, Valetta, 2010.
16. F. Eyben, M. Wöllmer, and B. Schuller. openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proc. ACM Multimedia*, pages 1459–1462, Florence, Italy, 2010.
17. B. Goertzel, K. Silverman, C. Hartley, S. Bugaj, and M. Ross. The baby webmind project. In *Proc. of The Annual Conference of The Society for the Study of Artificial Intelligence and the Simulation of Behavior (AISB)*, 2000.
18. M. Grimm, Kristian Kroschel, and Shrikanth Narayanan. The Vera am Mittag German Audio-Visual Emotional Speech Database. In *Proc. of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 865–868, Hannover, Germany, 2008.
19. T. Haderlein, E. Nöth E, H. Toy, A. Batliner, M. Schuster, U. Eysholdt, J. Hornegger, and F. Rosanowski. *Automatic Evaluation of Tracheoesophageal Substitute Voices*, volume 264. 2007.
20. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11, 2009.
21. J.H.L. Hansen and S. Bou-Ghazale. Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database. In *Proc. EUROSPEECH-97*, volume 4, pages 1743–1746, Rhodes, Greece, 1997.
22. T. Joachims. Text categorization with support vector machines: learning with many relevant features. In Claire Nédellec and Céline Rouveirol, editors, *Proc. of ECML-98, 10th European Conference on Machine Learning*, pages 137–142, Chemnitz, Germany, 1998. Springer, Heidelberg.
23. Y.-H. Kao and L.-S. Lee. Feature Analysis for Emotion Recognition from Mandarin Speech Considering the Special Characteristics of Chinese Language. In *Proc. ICLSP*, pages 1814–1817, 2006.
24. J. Krajewski and B. Kröger. Using prosodic and spectral characteristics for sleepiness detection. In *Eighth Annual Conf. Int. Speech Communication Association*, pages 1841–1844, 2007.
25. Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

26. D. Litman and K. Forbes. Recognizing emotions from student speech in tutoring dialogues. In *Proc. ASRU*, pages 25–30, Virgin Island, USA, 2003.
27. H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text Classification using String Kernels. *Journal of Machine Learning Research*, 2:419–444, 2002.
28. Marko Lugger and Bin Yang. Psychological motivated multi-stage emotion classification exploiting voice quality features. In F. Mihelic and J. Zibert, editors, *Speech Recognition*, page 1. IN-TECH, 2008.
29. O. Martin, I. Kotsia, B. Macq, and I. Pitas. The enterface'05 audio-visual emotion database. *IEEE Workshop on Multimedia Database Management*, 2006.
30. S. Matos, S.S. Birring, I.D. Pavord, and D.H. Evans. Detection of cough signals in continuous audio recordings using hidden Markov models. *IEEE Trans. Biomedical Engineering*, pages 1078–108, 2006.
31. G. Mohammadi, A. Vinciarelli, and M. Mortillaro. The Voice of Personality: Mapping Nonverbal Vocal Behavior into Trait Attributions. In *Proc. SSPW 2010*, pages 17–20, Firenze, Italy, 2010.
32. I. Mporas and T. Ganchev. Estimation of unknown speakers' height from speech. *International Journal of Speech Technology*, 12(4):149–160, 2009.
33. P. Pal, A.N. Iyer, and R.E. Yantorno. Emotion detection from infant facial expressions and cries. In *Proc. ICASSP*, pages 809–812, Toulouse, France, 2006.
34. 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.
35. F. Schiel and C. Heinrich. Laying the foundation for in-car alcohol detection by speech. In *Proc. INTERSPEECH 2010*, pages 983–986, Brighton, UK, 2009.
36. M. Schröder, R. Cowie, D. Heylen, M. Pantic, C. Pelachaud, and B. Schuller. Towards responsive sensitive artificial listeners. In *Proc. 4th Intern. Workshop on Human-Computer Conversation*, Bellagio, Italy, 2008.
37. B. Schuller, A. Batliner, S. Steidl, and D. Seppi. Emotion Recognition from Speech: Putting ASR in the Loop. In *Proc. ICASSP*, pages 4585–4588, Taipei, Taiwan, 2009. IEEE.
38. B. Schuller, A. Batliner, S. Steidl, and D. Seppi. Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from the First Challenge. *Speech Communication - Special Issue on Sensing Emotion and Affect •- Facing Realism in Speech Processing*, 2011.
39. B. Schuller, R. Jiménez Villar, G. Rigoll, and M. Lang. Meta-classifiers in acoustic and linguistic feature fusion-based affect recognition. In *Proc. ICASSP*, pages I:325–328, Philadelphia, PA, USA, 2005.
40. B. Schuller, F. Metze, S. Steidl, A. Batliner, F. Eyben, and T. Polzehl. Late Fusion of Individual Engines for Improved Recognition of Negative Emotions in Speech – Learning vs. Democratic Vote. In *Proc. 35th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5230–5233, Dallas, 2010.
41. B. Schuller, R. Müller, M. Lang, and G. Rigoll. Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensemble. In *Proc. Interspeech*, pages 805–808, Lisbon, Portugal, 2005.
42. B. Schuller, J. Schenk, G. Rigoll, and T. Knaup. The "Godfather" vs. "Chaos": Comparing Linguistic Analysis based on Online Knowledge Sources and Bags-of-N-Grams for Movie Review Valence Estimation. In *Proc. International Conference on Document Analysis and Recognition*, pages 858–862, Barcelona, Spain, 2009.
43. B. Schuller, S. Steidl, and A. Batliner. The INTERSPEECH 2009 Emotion Challenge. In *Proc. Interspeech*, pages 312–315, Brighton, UK, 2009.
44. B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan. The INTERSPEECH 2010 Paralinguistic Challenge. In *Proc. INTERSPEECH 2010*, pages 2794–2797, Makuhari, Japan, 2010.
45. B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth. Acoustic Emotion Recognition: A Benchmark Comparison of Performances. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 552–557, Merano, 2009.

46. B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll. Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies. *IEEE Transactions on Affective Computing*, 1:119–132, 2010.
47. B. Schuller, M. Wimmer, D. Arsic, G. Rigoll, and B. Radig. Audiovisual behaviour modeling by combined feature spaces. In *Proc. ICASSP*, pages 733–736, Honolulu, HY, 2007.
48. Björn Schuller, Ronald Müller, Florian Eyben, Jürgen Gast, Benedikt Hörnler, Martin Wöllmer, Gerhard Rigoll, Anja Höthker, and Hitoshi Konosu. Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-Life Application. *Image and Vision Computing Journal (IMAVIS), Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior*, 27:1760–1774, 2009.
49. D. Seppi, A. Batliner, B. Schuller, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, and V. Aharonson. Patterns, Prototypes, Performance: Classifying Emotional user States. In *Proc. Interspeech*, pages 601–604, Brisbane, Australia, 2008.
50. D. Seppi, A. Batliner, S. Steidl, B. Schuller, and E. Nöth. Word Accent and Emotion. In *Proc. Speech Prosody 2010*, Chicago, IL, 2010.
51. S. Steidl. *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. Logos Verlag, Berlin, Germany, 2009. (PhD thesis, FAU Erlangen-Nuremberg).
52. Stefan Steidl, Anton Batliner, Dino Seppi, and Björn Schuller. On the Impact of Children's Emotional Speech on Acoustic and Language Models. *EURASIP Journal on Audio, Speech, and Music Processing*, page 14 pages, 2010. doi:10.1155/2010/783954.
53. S. Steininger, F. Schiel, O. Dioubina, and S. Raubold. Development of user-state conventions for the multimodal corpus in smartkom. In *Proc. Workshop on Multimodal Resources and Multimodal Systems Evaluation*, pages 33–37, Las Palmas, Spain, 2002.
54. A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller. Deep Neural Networks for Acoustic Emotion Recognition: Raising the Benchmarks. In *Proc. ICASSP*, Prague, Czech Republic, 2011.
55. 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, Antwerp, Belgium, 2007.
56. N. Ward and W. Tsukahara. Prosodic features which cue backchannel responses in English and Japanese. *Journal of Pragmatics*, 32:1177–1207, 2000.
57. B. Weiss and F. Burkhardt. Voice attributes affecting likability perception. In *Proc. INTERSPEECH*, Makuhari, Japan, 2010.
58. I. H. Witten and E. Frank. *Data mining: Practical machine learning tools and techniques, 2nd Edition*. Morgan Kaufmann, San Francisco, CA, USA, 2005.
59. M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie. Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. In *Proc. Interspeech*, pages 597–600, Brisbane, Australia, 2008.
60. M. Wöllmer, B. Schuller, F. Eyben, and G. Rigoll. Combining Long Short-Term Memory and Dynamic Bayesian Networks for Incremental Emotion-Sensitive Artificial Listening. *IEEE Journal of Selected Topics in Signal Processing, Special Issue on "Speech Processing for Natural Interaction with Intelligent Environments"*, 4:867–881, 2010.
61. X. Zhe and A.C. Boucouvalas. Text-to-emotion engine for real time internet communication. In *Proc. of the International Symposium on Communication Systems, Networks, and DSPs*, pages 164–168, Staffordshire University, 2002.