Chapter 5
Emotion Modelling via Speech Content and Prosody: In Computer Games and Elsewhere

Björn Schuller

Abstract The chapter describes a typical modern speech emotion recognition engine as can be used to enhance computer games' or other technical systems' emotional intelligence. Acquisition of human affect via the spoken content and its prosody and further acoustic features is highlighted. Features for both of these information streams are shortly discussed along chunking of the stream. Decision making with and without training data is presented, each. A particular focus is then laid on autonomous learning and adaptation methods as well as the required calculation of confidence measures. Practical aspects include the encoding of the information, distribution of the processing, and available toolkits. Benchmark performances are given by typical competitive challenges in the field.

Introduction

The automatic recognition of emotion in speech dates back some twenty years by now looking back at the very first attempts, cf. e.g., [9]. It is the aim of this chapter to give a general glance 'under the hud' how today's engines work. First, a very brief overview on modelling of emotion is given. A special focus is then laid on speech emotion recognition in computer games owing to the context of this book. Finally, the structure of the remaining chapter is provided aiming at familiarising the reader with the general principles of current engines and their abilities, principles, and necessities.

Emotion Modelling

A number of different representation forms have been evaluated, with the most popular ones being discrete *emotion classes* such as 'anger', 'joy', or 'neutral' – usually reaching from two to roughly a dozen [51] depending on the

B. Schuller (✉)
Imperial College London, 180 Queen's Gate, SW7 2AZ London, UK
e-mail: bjoern.schuller@imperial.ac.uk

K. Karpouzis, G.N. Yannakakis (eds.), *Emotion in Games*,
Socio-Affective Computing 4, DOI 10.1007/978-3-319-41316-7_5

application of interest –, and a representation by continuous emotion 'primitives' in the sense of a number of (quasi-)value-continuous *dimensions* such as arousal/activation, valence/positivity/sentiment, dominance/power/potency, expectation/surprise/novelty, or intensity [43]. In a space spanned by these axes, the classes can be assigned as points or regions, thus allowing for a 'translation' between these two representation forms. Other popular approaches include *tagging* by allowing several class labels per instance of analysis (in case of two, the name *complex emotions* has been used), and calculating scores per each emotion class leading to '*soft emotion profiles*' [32] – potentially with a minimum threshold to be exceeded.

Besides choosing such a representation of emotion, one has to choose a temporal segmentation from, as the speech needs to be segmented into units of analysis. This analysis itself can be based on the spoken content or the 'way of speaking' it in the sense of prosody, articulation, and voice quality as the acoustic properties. As these two information streams tend to benefit from different temporal levels, one can choose different lengths of these units, accordingly.

Emotion and Games

It is generally believed that taking players' emotion in computer games into account possesses great potential for improved game-play and human behaviour feedback, in particular in 'serious games' [20, 46]. However, few games up to this point make use of the ability to recognise emotion from speech. This is likely owing to the fact that (1) games are not often controlled by speech, yet – however, a number of games already use this option and notably the Microsoft Xbox One console is offering speech control in many games via its Kinect interface; and (2) if no headset is used, games' and environmental audio can pose additional challenges to the use of this modality as compared, e.g., the use of bio-sensors.

Besides the commercial game "Truth or Lies", some examples of games that consider affect and emotion recognition in speech in research prototypes include [24, 29], and in particular several games for children [62, 63], especially with autism spectrum condition [35, 48] – often centred around emotion as a concept itself. Related to games, affect and emotion recognition has also been considered in educational software [26, 27]. Interestingly, also the emotion of the voice of the virtual characters of a game has been recognised automatically [38] showing that this is feasible as well. Games have, however, also been used to elicit emotion when collecting speech data 'coloured' by affect for training and evaluation of models [22, 23, 41, 53]. This makes it evident that, there is indeed 'a lot of emotion' in computer gaming that should best be captured by the game engine.

Further, a number of games exploit other modalities such as physiological measurement for emotion recognition, e.g., the "Journey to the Wild Divine" as a commercial example; other examples include mostly research prototypes [5, 19, 24, 34]. Further modalities used for emotion recognition in games include touch interaction [17], and facial expression and body gestures [48].

The remainder of this chapter is structured as follows: In section "Speech Content", we will be dealing with the recognition of emotion from the speech content, i.e., from the spoken words; then, in section "Prosodic and Acoustic Modelling" we will be considering the prosody and acoustics of the spoken words to the same end. In both of these sections, we will follow the typical chain of processing in pattern recognition systems by going from the denoising preprocessing to the feature extraction – both aiming at reaching a canonical representation of the problem of interest, usually coming at a reduction of information. Then, based on the reached representation, a decision can be made either without 'training' of a model from data, e.g., by rules such as "IF the words 'fun', 'sun', (...) are used" or "IF pitch increases and speed increases, THEN assume joy". Alternatively, as decisions are often much more complex, machine learning algorithms such as support vector machines (SVM) or neural networks (NNs) are the typically seen solutions. In section "Integration and Embedding", we will deal with practical issues when it comes to embedding 'speech emotion recognition' (SER) in an application, software system, or more specifically computer game. This includes the fusion of textual and acoustic cues, available tools ready to be used ready to 'plug and play', benchmarks on popular databases, distributed processing, confidence measures, adaptation and self-learning, and encoding of the emotion information. Then, to conclude, a summary will be given naming some existing 'white spots' in the literature that should best be filled sooner rather than later if SER is to be used widely in computer games and other real-world products.

Speech Content

We tend to express our emotion in the words we choose. Not only is this true for the words themselves, but also for their 'part of speech' (POS) classes such as noun, verb or adjective. The latter means that depending on the emotion, different frequencies of the POS classes can be observed. In fact, also semantically higher 'tagging' of words such as 'first person pronoun' or 'about religion' (such as 'Christ', 'Jesus', 'holy', 'Lord', etc.) may reveal the affect 'of the moment'. At the same time, the level of control by an individual for regulation or masking of this type of affective information is usually high, i.e., we are trained to choose these words different from our emotion in case we want to hide our feelings. In the following, let us take a look on how to model emotion in such a way that the described speech content can be exploited.

Speech Recognition and Emotion

One bottleneck in exploiting spoken or written words for the recognition of emotion of players or in general is that these need to be recognised in the first place. Even if they should be typed rather than spoken, such as by a keybord or a game console, some preprocessing is usually necessary to ‘correct’ misspelling, dialects or remove or handle special characters, while ensuring not to lose relevant information such as ‘smileys’ as “:)” or “:-(” and alike. This is often handled by allowing for a certain Levenshtein distance (the minimum number of operations out of insertion, deletion, and substitution to match the observed word with one known by the computer – usually calculated by dynamic programming) or look-up dictionaries for different spellings. Luckily, automatic speech recognition (ASR) has matured impressively in recent time to the degree where it is available on even very low resource devices, smart phones, game consoles such as the XBox One via its Kinect, or even modern TV sets such as the Samsung UE46F7090 and alike. These ASR engines are, however, mostly focused on the recognition of (a few) keywords at a time depending on the current context. This comes, as robustness is particularly required as when gaming at home, the sounds and music as well as non-player-character (NPC) voices interfere; when playing on mobile devices, ambient noises and ever-changing reverberation patterns have to be coped with. This allows to either use an additional continuous speech larg(er) vocabulary ASR engine alongside – even if at lower requirements in terms of robustness – or to limit the emotion analysis to a few affective keywords. Surprisingly, it seems as if even lower accuracies of the ASR engine can be sufficient to be able to capture the emotion of the words [30] as long as the affective keywords are not lost or ‘changed’ to another affective context.

Besides the recognition of verbal units, also non-verbal units such as laughter or sighing and hesitations bear information on the emotion of a speaker. Their recognition can be handled within the ASR process, but is often executed individually, such as in [6]. Independent of that, the information of verbal and non-verbal tokens can be handled commonly such as in “*this is really <laughs> funny!*”.

Textual Features

In order to represent text in a numerical representation form, there is a need for segmentation of the word or text stream at first, potentially followed by re-tagging of the words, and an optional representation in a vector space. The last choice depends on the type of decision making.

Tokenisation and Tagging

A first step is – similar to the processing of an acoustic speech signal as described later – the *tokenisation* in the sense of a segmentation. Typical approaches to this end include sequences of N characters or words, the latter usually delimited by spaces or special characters when considering written text. Such sequences are known as N-grams, or character N-grams and word N-grams, respectively. Typical lengths of sequences are roughly three to eight characters or one to three words [47].

Either before or after this tokenisation, a (re-)*tagging* of the tokens can take place. Above, POS classes and higher semantical classes were named already. Besides, stemming is a popular way of re-tagging tokens in order to cluster these by their word stem. An example would be “stems”, “stemming”, or “stemmed” re-tagged as “stemX”. This increases the number of observations of words of the same stem and likewise allows to train more meaningful models as long as the different morphological variants indeed do not represent different emotional ‘colourings’. Differently tagged variants of the text can be combined, to combine different levels of information and highlight different aspects, such as the original ‘fine grained’ words (usually in the order of some thousand different entries) alongside their semantic classes (rather in the order of a few hundreds) alongside their POS classes (rather some tenths).

Vector Space Modelling

In order to reach a feature representation in terms of a vector of numerical variables that allows, e.g., for the calculation of distances between such vectors belonging to different phrases, a popular approach is to count the frequency of occurrence of each token in the text of analysis. This is known as term frequency modelling [21]. Then, a feature vector is constructed by using one such feature per entry in the ‘vocabulary’ of known different tokens. As most texts of analysis will consist of very few words when dealing with analysis of spoken language in search of affective cues, the feature vector will accordingly contain mostly zero values. Different variants exist to normalise such a term frequency feature vector, such as using the logarithmic term frequency, dividing by the number of tokens in the text of analysis or by the number of occurrences as seen in the training material. However, it has repeatedly been observed that, given a large database for training of a machine learning algorithm, the effect of different such ‘normalisation approaches’ is comparably minor [47], and even a simple boolean representation of the term frequency can be efficient and sufficient when dealing with short phrases.

Zero-Resource Modelling

Without data to train from, emotion can be recognised from text thanks to a richer number of available affective dictionaries and other related knowledge sources such as ConceptNet, General Inquirer, or WordNet [31, 47]. Usually, one ‘looks up’ the words in the text of analysis in the according knowledge resource and computes a score per emotion or a value for an affect dimension according to the entries in these knowledge sources. Alternatively, one computes the distance in the resource to the affective tag(s) of interest, e.g., the distance in a word relation database between “sun” and “happniess” or alike.

Learning

Alternatively, one can train standard machine learning algorithms on annotated text. Most popular solutions include SVMs due to their ability to handle large feature spaces. Depending on the feature representation of the text, these can have several thousand entries.

Prosodic and Acoustic Modelling

Let us now have a look at the speech signal by exploiting ‘how’ something is being said.

Speaker Separation and Denoising

Considerable progress had been made over the last years in (blind) audio source separation allowing for high(er) quality separation of a speaker even in adverse acoustic conditions or even in the presence of other overlapping speakers. Some interesting approaches include a range of variants of Non-Negative Matrix Factorisation (NMF) [59] or the usage of (recurrent) neural networks (preferably with memory [60]) to estimate clean from noisy speech or at least ‘clean’ features derived from noisy speech. In the luxurious case of availability of several microphones such as in a microphone array, e.g., four in the case of Microsoft’s Kinect sensor, Independent Component Analysis (ICA) or derivatives can separate as many sources as microphones with a very high quality. One can also mix single- and multi-microphone approaches, e.g., by ICA-NMF hybrids. Most of these approaches target additive noise, however, also convolutional artifacts such as by reverberation due to a changing room-impulse response in a mobile setting can be handled – again, data-driven approaches have lately become very powerful in this respect [60]. Surprisingly few works have investigated the usage of speaker separation for the recognition of emotion (e.g., [59]) from speech as compared to those works considering it for ASR. Interestingly, such separation also allows for recognition of group emotion.

Prosodic and Acoustic Features

Prosody is usually summarised by three feature groups: intensity, intonation, and ‘rhythm’. Whereas in emotional speech synthesis, these are mostly focussed upon, in the recognition, also further acoustic features are of crucial relevance. In terms of being most descriptive, voice quality is to mention next. This includes the harmonics to noise ratio, the jitter and shimmer – micro pertubations of pitch and energy, and other aspects describing the quality of the voice such as breathy, whispering, shouting, etc. Finally, a rich selection of further spectral and cepstral features is usually considered, such as spectral band energies and ratios – often presented in a manner closer to human hearing sensation such as mel frequency and auditory spectrum coefficients – or formants, i.e., spectral maxima described by position, amplitude, and bandwidth – usually in the order of the first five to seven.

The above named features are usually computed per ‘frame’ after windowing the speech signal with a suited function such as a rectangular function for analysis in the time domain and a smooth function such as Hamming or Hanning (basically, the positive half wave of a cosine signal) or Gaussian (allowing to come closest to the Heisenberg-alike time-frequency uncertainty resolution optimum). A typical frame length then is around 20–50 ms, with a frame shift of around 10 ms resembling a feature sampling frequency of 100 frames per second (fps) at this so called ‘low level’ – overall one thus speaks of low-level descriptors (LLDs) at this point. To add to the feature basis, further LLDs can be derived from the original contours such as by calculating delta regression or correlation coefficients. Based on the LLDs, statistical functionals are then mostly applied to lead to a supra-segmental view – emotion is a usually better modelled at a feature-level for a window of analysis of around 1 s [45] depending on the LLD type. These functionals project a LLD onto a single value per ‘macro’ time frame (fixed length or semantically or syntactically meaningful such as per word or phrase). Typical functionals include the lower order moments such as mean, standard deviation, kurtosis, skewness, or extrema, such as minimum, maximum, range, percentiles, segments, or even spectral functionals. This process can be carried out systematically leading to a brute-force feature extraction: First, LLDs are extracted – then, per LLD derivations are produced. Next, for each of the overall LLDs – be it original or derived – the same functionals are computed – potentially even hierarchically such as the mean of maxima or the maxima of means. This can also be considered across different window lengths or cross LLD, thus easily leading to a feature basis of several thousand features. Usually, one then carries out a data-driven feature selection to reduce the space to a few hundred or even less features. Over the years, sufficient experience was gained to formulate recommendations for such compact feature sets. However, as the emotion representation and type targeted can vary considerably, it may be worth to adapt the feature set to a specific task or domain and culture (cf. also below). Also, one can execute a controlled brute-forcing, thus not necessarily producing all possible LLD/functional/(functional) combinations (which can sometimes be meaningless as the minimum pitch value which is often simply zero), but rather searching in the space of possible combinations in a targeted way by combining expert knowledge and brute-forcing efficiently [33].

More recently, unsupervised learning of features has been considered as an alternative of expert-crafted or brute-forced descriptors. This comes often in combination with ‘deep learning’ approaches [25, 57] based on (sparse) auto-encoders or similar. An interesting ‘conventional’ method for unsupervised acoustic feature generation is the Bag-of-Audio-Words variant. It is similar to the Bag-of-Words modelling described above for linguistic content features. However, it produces the ‘audio words’ in unsupervised manner by vector quantisation over some (often raw) acoustic representation such as in the spectrum, e.g., by k-means clustering or similar. Then, frequency of occurrences of these audio words are counted.

Zero-Resource Modelling

As opposed to the linguistic exploitation of the spoken content, there exists practically no acoustic SER engine that is based on (human-readable) rules that have not been learnt from data by a machine learning algorithm. This does, however, not mean that it is not possible to implement what is known from the literature on how speech changes depending on the emotion, such as in [40]. It rather means that usually, the task is of non-linear and complex nature and a few rules will not suffice. However, the advantage of not needing training data seems obvious: such a model will not be biased by the chracteristics of the speakers contained in the training data – may it be their gender, age, social class, origin, culture, or simply their co-influencing states and traits at the time of the recording of the training data. For example, in serious gaming for specific target groups such as young individuals with Autism Spectrum Condition [48], training data may be so rare and the variation of the deviation from the ‘typical’ so large that it can be more promising to rely on a few basic rules.

An interesting alternative version in case of limited availability of (labelled) human speech samples is the synthesis of training material [52]. Given a high quality speech synthesiser that can produce speech in different emotions, one can produce sheer infinite amounts of training material for an emotion recogniser. As such a speech synthesiser is often given in computer games, this may be a promising road in affect recognition for gaming or dialogue systems. Obviously, one can combine synthesised speech with human speech in training of classifiers often leading to an additional gain in performance.

Learning

There are at best trends visible which machine learning algorithms are preferred in the literature, as in principle, both, ‘static’ algorithms (with a fixed feature dimension and stream/time series length) as well as ‘dynamic’ algorithms (that can cope with time series of unknown length) can be used [50]. This depends on the feature representation of choice: the functional representation is usually modelled by static approaches, whereas the ‘raw’ LLDs are best modelled by the dynamic alternative [50]. Popular static approaches include SVMs, NNs, and decision trees, but also more simple solutions such as k-Nearest Neighbours (kNN) and related distance-based solutions, or naive Bayes and related simple statistical considerations. As for static approaches, Hidden Markov Models (HMMs) and more general Dynamic Bayesian Networks or Graphical Models prevail over distance-based alternatives such as Dynamic Time Warping (DTW). Noteworthy, recent trends include the usage of deep learning [25, 57] and specifically of long-short-term memory-enhanced (LSTM) recurrent NNs [61]. This comes, as the first offer the ability to learn feature representations in an unsupervised manner (as outlined above) which seems attractive in a field where the optimal feature representation is highly disputed. The latter, i.e., the LSTM paradigm, has the charm of providing the additional learning of the optimal temporal context. Again, a valuable asset, as the optimal temporal context for emotion recognition from speech and the best 'unit of analysis' such as fixed length frames, voiced or unvoiced sounds, syllables, words, sequences of words or whole speaker turns is also highly disputed [44, 49].

Integration and Embedding

In this section, we will deal with practical aspects of 'putting SER in there' (i.e., embedding an emotion recogniser in a software system) where 'there' can be a computer game, a (software) application or in principal various kinds of technical, computer, and software systems.

Fusion

Obviously, it will be of interest to exploit both – the spoken content and the way it is acoustically expressed – rather than only any individual aspect of these two. In fact, they provide different strengths when considering emotion in the dimensional model: The words usually better reflect valence aspects, the prosody and acoustics arousal.

The integration of these two cues can be executed either by combining the features before the decision ('early fusion') or after the individual decisions for content and prosody/acoustics ('late fusion'), e.g., by a weighted sum of scores or another classifier/regressor or similar. Further approaches have been investigated such as fusion by prediction, but remained a fringe phenomenon in this domain up to this point.

Available Tools

A number of tools exist that are free for research purposes and often are even open-source.

For modelling emotion and affect via prosody and acoustic features, the openSMILE [15] package provides a ready to use solution of voice activity detection and speech signal segmentation, feature extraction, and classification/regression including pre-trained models for an instant setup of an online emotion recogniser. However, it is mostly focused on acoustic/prosodic analysis. A similar tool is described in [58].

For modelling via spoken content analysis, one needs to 'put together' one of the manifold automatic speech recognition engines available these dates such as HTK, Julius, Sphinx, etc. and then plug it into openSMILE or Weka [18], etc., for feature computation and machine learning.

Obviously, further tools exist – in particular for labelling of emotional speech data, e.g., [7] – that could not be mentioned here due to space restrictions.

Data and Benchmarks

Luckily, a richer number of affective speech and language resources is available these days, albeit only few are 'sufficiently' large covering different age groups, and realistic spontaneous emotion. For many languages and cultures, emotionally labelled speech data is unfortunately still missing. Here, only those data that have been featured in a research competition will be named as examples, as they provide a well-defined test-bed and benchmark results are available. These are the FAU Aibo Emotion Corpus (AEC) as was featured in the first comparative public emotion challenge ever – the Interspeech 2009 Emotion Challenge [50]. AEC contains children speech in two or five emotion categories in realistic settings – the best results reached in the Challenge were 71.2 % (two classes) and 44.0 % (five classes) by fusion of the best participants' engines. In 2013, emotion was revisited by the series of challenges in the computational paralinguistic domain by the GEMEP corpus [51]. While it contains enacted data in an artificial syllable 'language', this time 12 classes were targeted with the best result reaching 46.1 %. Another series is the Audio/Visual Emotion Challenge (AVEC) series that has in its so far five editions seen three different databases which are available for research and testing: the AVEC 2011/2012 corpus was based on the SEMAINE database; AVEC 2013/2014's featured AVDLC depression data, and the RECOLA database [37] is featured in AVEC 2015. These are usually larger by an order of magnitude and all labelled in time and value-continuous arousal and valence, and partially further dimensions or classes. A number of further related challenges includes, e.g., [12].

Distribution

The recognition of emotion from speech can usually be done (1) in real-time, and (2) at 'reasonable' computational effort thus (3) allowing for processing on the same platform that is used for a computer game or similar realisation. However, distributing the recognition between a end-user based client and a server bears a huge advantage: if the audio, the features, or compressed version of the features are collected centrally, learning models can be improved for all players that are making use of such a distributed service. In order to increase the level of privacy of users in this case, and reduce the bandwidth needed for transmission to the server, sub-band feature vector quantisation can be used. In fact, this bears similarity with the generation of the 'audio words' as were described above: Rather than a large feature vector of several thousand features, only one reference ID number of a spatially close reference feature vector is submitted to the server per feature-sub-band and unit of analysis in time [65].

Confidence Measures

The provision of confidence measures gives extra information to a game or application alongside the recognition result on its certainty. This seems needed looking at the current benchmarks in the field as were outlined above. Further to that, confidence measures are an important factor of the self-training as will be described next, as it helps a machine to decide whether it can label data by itself or needs human help prior to using such data to (re-)train itself. Few works have investigated calculation of meaningful confidence measures in this field up to this point. Most notably, these include some aspects tailored to the characteristics of emotion recognition, as they exploit the uncertainty of labels and imbalance of observations per class both typical for the field. The first approach tries to predict the degree of human agreement such as n raters out of N on a data point independent of the emotion assumed as an indication on how reliable the assumption is likely to be. This can be done, as usually several raters (usually some three given expert labellers up to around several tenths, e.g., when crowd sourcing the information) label a data point in this field, and the percentage of agreeing raters can be used as a learning target either instead of the emotion or – even better – alongside the emotion in a multitarget learning approach [16]. The second approach is based on engines trained on additional data with the learning target whether the actual emotion recognition engine is correct or incorrect or – more fancily – correctly positive, correctly negative, incorrectly positive or incorrectly negative [10]. In fact, such confidence measures could be shown to provide reasonable indication on the reliability of a recognised emotion [10] thus providing a genuine extra value that can be exploited, e.g., by a game engine.

Adaptation and Self-Training

In this section, methods are presented on how to exploit existing labelled data similar to the labelled or unlabelled target data or unlabelled data in utmost efficient manners.

In the first case, i.e., the existence of similar data such as adult emotional speech when in need of a child emotion recogniser, e.g., for a child computer game, transfer learning comes in handy. Transfer learning per se is a rather loosely formulated approach in the sense that many different variants and starting points exist. As an example, in [11] a sparse autoencoder is trained in both – the source and the target domain, thus reaching compact representations of both domains. Then, a neural network is trained to learn to map across domains, thus allowing according transfer. In comparison with other standard transfer learning approaches, this has shown the best results in the study for SER. In such a way, a quick adaptation to a new individual, e.g., a new player can be reached.

Further to that, games usually are characterised by repeated interaction with a player. This fact can be exploited to give a game the chance to better 'learn' the user. Ideally, this is done without being noticed or requiring help from the user by weakly supervised learning approaches.

Semi-supervised learning does not require any human help – a system labels the data for its training by itself once it 'hears' (or 'reads') new data and is sufficiently confident that it 'does it right'. This can best be done in co-training exploiting different 'views' on the data such as acoustic and linguistic feature information. This approach has been shown to actually improve recognition performance in the recognition of emotion from speech [28, 67], or text [8] thus enabling systems to also make use of unlabelled data in addition to labelled data.

Unfortunately, it usually requires large amounts of unlabelled data to obtain similar gains in performance as one would see when using (human) labelled data. This can be eased by adding as a promising aid *active learning*. The idea of active learning is to ask a human for help whenever new data appears to be 'promising'. The art here is to quantise and identify what is 'promising'. Approaches that have been shown to work well for emotion recognition are based, e.g., on the likely sparseness of a new unlabelled instance in terms of the expected emotion class. This means that likely to be 'neutral' samples are considered less worthy labelling as there are usually plenty of these available. However, should the system believe a new speech sample is emotional or even potentially from a hardly or never (requiring some additional novelty detection) seen before emotion, it will (only then) ask the player or user about this state, thus considerably reducing the amount of required human labels. Further approaches to measure the interest in a data instance include the expected change in model parameters, i.e., if a new speech sample is not likely to change the parameters of the trained classifier or regressor, it is not important to inquire to which emotion it belongs. Overall, active learning could be shown to reduce the labelling effort by up to 95 % at the same or even higher recognition accuracy of accordingly trained models in the study [66].

Most efficiently, *cooperative learning* combines the strengths and increases the efficiency of active and semi-supervised learning in the sense that a computer system first decides if it can label the data itself, and only if not, evaluates if it is worth to ask for human help [64].

Encoding and Standards

A number of standards exist to encode the information on emotion to be used by a game or application. To name but a few, the Humaine EARL and the W3C Emotion Markup Language recommendation (or EmotionML [42] for short) have been designed in particular for encoding affective information. A number of further standards do, however, provide tags for emotion as well, and may likely be used in a video game or similar context, anyhow, such as W3C EMMA [1].

Further, in particular for the speech modality, a standard has been developed to encode the feature information used [4]. This allows to exchange collected feature/label pairs even if different systems or games use different types of feature extraction. In addition, standards exist for the encoding of individual feature types, e.g., [55].

Finally, standardised feature sets are offered and used in this field by the openSMILE toolkit (cf. above).

Summary and White Spots

Towards the end of this chapter, let us quickly summarise and identify white spots left for future research efforts.

Summary

In this chapter, we took a look 'under the hood' of a modern speech emotion recognition engine. Ideally, it exploits both, acoustic and linguistic cues and combines these efficiently. The features can either be extracted based on known 'expert' feature types or learnt unsupervised from data, e.g., in deep learning. Then, either a rule-based 'zero-resource' decision takes place, or a machine learning algorithm is trained based on labelled speech data. The recognition result for a previously unseen new speech instance will then usually be one or several emotion labels or one or several continuous values of emotion primitives such as arousal or valence. What's more, an engine should ideally provide confidence measures alongside to give a feeling for the reliability of its prediction. It can, based on this confidence measure, also decide to retrain itself after having heard new data points and having been able to predict them likely correctly. If it is unsure about the emotion of a new sample but feels it would benefit from knowing it, it can ask a user for help in labelling in a very targeted and efficient way. If it has 'similar' labelled data to the target data to exploit or start with, transfer learning allows to adapt the data to the domain. For practical implementation, encoding and feature standards can be used, and a number of 'out of the box' ready tools exist which are free for research purposes. There have repeatedly been evaluation campaigns showing that recognition of emotion from speech 'works', but leaves head room for improvement at this moment in time.

White Spots

The field is still somewhat young and not all problems could be touched upon sufficiently or at all, yet. To conclude this chapter, these white spots in the literature shall be quickly named without lengthy explanation: When dealing with textual cues, *multilinguality* is an obvious problem [2, 36]. While one can benefit from the progress machine translation has made, there is still need for more experience with this challenge. Obviously, multilinguality also has an impact on acoustic features, e.g., when considering the differences between tonal or non-tonal languages such as Chinese in comparison to English or German. Similarly, *cross-cultural* aspects [14, 39, 54] need additional investigation and in particular solutions to identify the culture automatically and choose or adapt models accordingly, e.g., by suited means of transfer learning. Further, 'faking', *regulation*, and *masking* of emotions has hardly been investigated in terms of automatic recognition. Similarly, '*atypicality*' is not sufficiently explored up to this moment. This deficit includes the lack of providing an engine prepared for less typical emotion portrayal as, e.g., from individuals with ADHD, Autism Spectrum Condition, vocal disfluencies, and alike or simply of less investigated age groups such as the elderly or (very) young individuals. Obviously, a number of *ethical implications* come with such data, and in general when having a machine analysing human emotion, giving (potentially flawed) feedback on it, and potentially distributing processing or storing information in a centralised fashion [3, 13, 56]. Then, *user studies* and experiences with *real systems 'in the wild'* are still very sparse and among most urgent issues to tackle – often a careful system or game design or response pattern can cover up very elegantly for potential misrecognitions, or – the other way round – an almost perfectly functioning emotion recognition not exploited in the right way can 'ruin it all' easily. Finally, better modelling of *co-influencing states and traits* of humans of analysis is needed: All speaker states and traits impact on the same speech production mechanism and the choice of our verbal behaviour: Whether we have a cold, are sleepy, or intoxicated – emotion should be recognised reliably independent of these factors and our personality profile. A promising route to reaching this ability seems parallel modelling of the wide range of state and traits, e.g., by a neural network with multiple output nodes as targets for various speaker states and traits rather than just for emotion, yet, allowing for missing labels during training, as not all data are likely to be labelled in such a wide range of factors.

With the advent of solutions to these problems as well as in general, one can expect to see gaming and general intelligent technology soon to be recognising our emotion when having us talk (or type) to them. May the best use be made of this new technical ability considering highest ethical standards at all time.

Acknowledgements The author acknowledges the support of the European Union's Horizon 2020 Framework Programme under grant agreement no. 645378 (ARIA-VALUSPA).

References

1. Baggia P, Burnett DC, Carter J, Dahl DA, McCobb G, Raggett D (2007) Emma: extensible multimodal annotation markup language. W3C working draft
2. Banea C, Mihalcea R, Wiebe J (2011) Multilingual sentiment and subjectivity analysis. Multiling Nat Lang Process 6:1–19
3. Batliner A, Schuller B (2014) More than fifty years of speech processing – the rise of computational paralinguistics and ethical demands. In: Proceedings ETHICOMP 2014. CERNA, Paris, 11p
4. Batliner A, Steidl S, Schuller B, Seppi D, Vogt T, Wagner J, Devillers L, Vidrascu L, Aharonson V, Kessous L, Amir N (2011) Whodunnit – searching for the most important feature types signalling emotion-related user states in speech. Comput Speech Lang Spec Issue Affect Speech Real-life Interact 25(1):4–28
5. Becker C, Nakasone A, Prendinger H, Ishizuka M, Wachsmuth I (2005) Physiologically interactive gaming with the 3D agent max. In: Proceedings international workshop on conversational informatics in conjunction with JSAI-05, Kitakyushu, pp 37–42
6. Brückner R, Schuller B (2014) Social signal classification using deep BLSTM recurrent neural networks. In: Proceedings 39th IEEE international conference on acoustics, speech, and signal processing, ICASSP 2014, Florence. IEEE, pp 4856–4860
7. Cowie R, Douglas-Cowie E, Savvidou S, McMahon E, Sawey M, Schröder M (2000) Feeltrace: an instrument for recording perceived emotion in real time. In: Proceedings ISCA workshop on speech and emotion, Newcastle, pp 19–24
8. Davidov D, Tsur O, Rappoport A (2010) Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In: Proceedings CoNNL, Uppsala, pp 107–116
9. Dellaert F, Polzin T, Waibel A (1996) Recognizing emotion in speech. In: Proceedings ICSLP, Philadelphia, pp 1970–1973
10. Deng J, Schuller B (2012) Confidence measures in speech emotion recognition based on semi-supervised learning. In: Proceedings of INTERSPEECH. ISCA, Portland
11. Deng J, Zhang Z, Schuller B (2014) Linked source and target domain subspace feature transfer learning – exemplified by speech emotion recognition. In: Proceedings 22nd international conference on pattern recognition (ICPR 2014). IAPR, Stockholm, pp 761–766
12. Dhall A, Goecke R, Joshi J, Wagner M, Gedeon T (eds) (2013) Proceedings emotion recognition in the wild challenge and workshop. ACM, Sydney
13. Döring S, Goldie P, McGuinness S (2011) Principalism: a method for the ethics of emotion-oriented machines. In: Petta P, Pelachaud C, Cowie R (eds) Emotion-oriented systems: the HUMAINE handbook, cognitive technologies. Springer, Berlin/Heidelberg, pp 713–724
14. Elfenbein HA, Mandal MK, Ambady N, Harizuka S, Kumar S (2002) On the universality and cultural specificity of emotion recognition: a meta-analysis. Psychol Bull 128(2):236–242
15. Eyben F, Weninger F, Groß F, Schuller B (2013) Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In: Proceedings of the 21st ACM international conference on multimedia, MM 2013. ACM, Barcelona, pp 835–838
16. Eyben F, Wöllmer M, Schuller B (2012) A multi-task approach to continuous five-dimensional affect sensing in natural speech. ACM Trans Interact Intell Syst Spec Issue Affect Interact Nat Environ 2(1):29
17. Gao Y, Bianchi-Berthouze N, Meng H (2012) What does touch tell us about emotions in touchscreen-based gameplay? ACM Trans Comput-Hum Interact 19(4/31):1–30
18. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten I (2009) The WEKA data mining software: an update. SIGKDD Explor 11:10–18

19. Holmgard C, Yannakakis G, Karstoft KI, Andersen H (2013) Stress detection for PTSD via the StartleMart game. In: Proceedings of 2013 humaine association conference on affective computing and intelligent interaction (ACII). IEEE, Memphis, pp 523–528
20. Hudlicka E (2009) Affective game engines: motivation and requirements. In: Proceedings of the 4th international conference on foundations of digital games. ACM, New York, 9p
21. Joachims T (1998) Text categorization with support vector machines: learning with many relevant features. In: Nédellec C, Rouveirol C (eds) Proceedings of ECML-98, 10th European conference on machine learning. Springer, Heidelberg/Chemnitz, pp 137–142
22. Johnstone T (1996) Emotional speech elicited using computer games. In: Proceedings ICSLP, Philadelphia, 4p
23. Johnstone T, van Reekum CM, Hird K, Kirsner K, Scherer KR (2005) Affective speech elicited with a computer game. Emotion 5:513–518
24. Kim J, Bee N, Wagner J, André E (2004) Emote to win: affective interactions with a computer game agent. In: GI Jahrestagung, Ulm, vol 1, pp 159–164
25. Kim Y, Lee H, Mower-Provost E (2013) Deep learning for robust feature generation in audiovisual emotion recognition. In: Proceedings the 2nd CHiME workshop on machine listening in multisource environments held in conjunction with ICASSP 2013, Vancouver. IEEE, pp 86–90
26. Liscombe J, Hirschberg J, Venditti JJ (2005) Detecting certainness in spoken tutorial dialogues. In: Proceedings INTERSPEECH. ISCA, Lisbon, pp 1837–1840
27. Litman D, Forbes K (2003) Recognizing emotions from student speech in tutoring dialogues. In: Proceedings ASRU, Virgin Island. IEEE, pp 25–30
28. Mahdhaoui A, Chetouani M (2009) A new approach for motherese detection using a semi-supervised algorithm. In: Machine learning for signal processing XIX – Proceedings of the 2009 IEEE signal processing society workshop, MLSP 2009, Grenoble. IEEE, pp 1–6
29. Martyn C, Sutherland JJ (2005) Creating an emotionally reactive computer game responding to affective cues in speech. In: Proceedings HCI, Las Vegas, vol 2, pp 1–2
30. Metze F, Batliner A, Eyben F, Polzehl T, Schuller B, Steidl S (2010) Emotion recognition using imperfect speech recognition. In: Proceedings INTERSPEECH. ISCA, Makuhari, pp 478–481
31. Missen M, Boughanem M (2009) Using WordNet's semantic relations for opinion detection in blogs. In: Advances in information retrieval. Lecture notes in computer science, vol 5478/2009. Springer, Berlin, pp 729–733
32. Mower E, Mataric MJ, Narayanan SS (2011) A framework for automatic human emotion classification using emotion profiles. IEEE Trans Audio Speech Lang Process 19:1057–1070
33. Pachet F, Roy P (2009) Analytical features: a knowledge-based approach to audio feature generation. EURASIP J Audio Speech Music Process 2009:1–23
34. Park S, Sim H, Lee W (2014) Dynamic game difficulty control by using EEG-based emotion recognition. Int J Control Autom 7:267–272
35. Ploog BO, Banerjee S, Brooks PJ (2009) Attention to prosody (intonation) and content in children with autism and in typical children using spoken sentences in a computer game. Res Autism Spectr Disord 3:743–758
36. Polzehl T, Schmitt A, Metze F (2010) Approaching multi-lingual emotion recognition from speech – on language dependency of acoustic/prosodic features for anger detection. In: Proceedings speech prosody, Chicago. ISCA
37. Ringeval F, Eyben F, Kroupi E, Yuce A, Thiran JP, Ebrahimi T, Lalanne D, Schuller B (2015) Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. Pattern Recognit Lett 66:10
38. Rudra T, Kavakli M, Tien D (2007) Emotion detection from female speech in computer games. In: Proceedings of TENCON 2007 – 2007 IEEE region 10 conference, Taipei. IEEE, pp 712–716
39. Sauter DA, Eisner F, Ekman P, Scott SK (2010) Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. Proc Natl Acad Sci USA 107(6):2408–2412
40. Scherer KR (2003) Vocal communication of emotion: a review of research paradigms. Speech Commun 40:227–256

41. Scherer S, Hofmann H, Lampmann M, Pfeil M, Rhinow S, Schwenker F, Palm G (2008) Emotion recognition from speech: stress experiment. In: Proceedings of the international conference on language resources and evaluation, LREC 2008, Marrakech. ELRA, 6p
42. Schröder M, Devillers L, Karpouzis K, Martin JC, Pelachaud C, Peter C, Pirker H, Schuller B, Tao J, Wilson I (2007) What should a generic emotion markup language be able to represent? In: Paiva A, Prada R, Picard RW (eds) Proceedings of ACII. Springer, Berlin/Heidelberg, pp 440–451
43. Schuller B (2012) The computational paralinguistics challenge. IEEE Signal Process Mag 29(4):97–101
44. Schuller B, Batliner A (2013) Computational paralinguistics: emotion, affect and personality in speech and language processing. Wiley, New York
45. Schuller B, Devillers L (2010) Incremental acoustic valence recognition: an inter-corpus perspective on features, matching, and performance in a gating paradigm. In: Proceedings INTERSPEECH, Makuhari. ISCA, pp 2794–2797
46. Schuller B, Dunwell I, Weninger F, Paletta L (2013) Serious gaming for behavior change – the state of play. IEEE Pervasive Comput Mag 12(3):48–55
47. Schuller B, Knaup T (2011) Learning and knowledge-based sentiment analysis in movie review key excerpts. In: Esposito A, Esposito AM, Martone R, Müller V, Scarpetta G (eds) Toward autonomous, adaptive, and context-aware multimodal interfaces: theoretical and practical issues: third COST 2102 international training school. Lecture notes on computer science (LNCS), vol 6456/2010, 1st edn. Springer, Heidelberg, pp 448–472
48. Schuller B, Marchi E, Baron-Cohen S, Lassalle A, O'Reilly H, Pigat D, Robinson P, Davies I, Baltrusaitis T, Mahmoud M, Golan O, Friedenson S, Tal S, Newman S, Meir N, Shillo R, Camurri A, Piana S, Staglianò A, Bölte S, Lundqvist D, Berggren S, Baranger A, Sullings N, Sezgin M, Alyuz N, Rynkiewicz A, Ptaszek K, Ligmann K (2015) Recent developments and results of ASC-inclusion: an integrated internet-based environment for social inclusion of children with autism spectrum conditions. In: Proceedings of the of the 3rd international workshop on intelligent digital games for empowerment and inclusion (IDGEI 2015) as part of the 20th ACM international conference on intelligent user interfaces, IUI 2015, Atlanta. ACM, 9p
49. Schuller B, Rigoll G (2006) Timing levels in segment-based speech emotion recognition. In: Proceedings of INTERSPEECH, Pittsburgh. ISCA, pp 1818–1821
50. Schuller B, Steidl S, Batliner A (2009) The interspeech 2009 emotion challenge. In: Proceedings of INTERSPEECH, Brighton. ISCA, pp 312–315
51. Schuller B, Steidl S, Batliner A, Vinciarelli A, Scherer K, Ringeval F, Chetouani M, Weninger F, Eyben F, Marchi E, Mortillaro M, Salamin H, Polychroniou A, Valente F, Kim S (2013) The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In: Proceedings of INTERSPEECH, Lyon. ISCA, pp 148–152
52. Schuller B, Zhang Z, Weninger F, Burkhardt F (2012) Synthesized speech for model training in cross-corpus recognition of human emotion. Int J Speech Technol Spec Issue New Improv Adv Speak Recognit Technol 15(3):313–323
53. Shahid S, Krahmer E, Swerts M (2007) Audiovisual emotional speech of game playing children: effects of age and culture. In: Proceedings of INTERSPEECH, Antwerp, pp 2681–2684
54. Shaver PR, Wu S, Schwartz JC (1992) Cross-cultural similarities and differences in emotion and its representation: a prototype approach. Review of Personality and Social Psychology. Vol. XIII: Emotion, 175–212
55. Silverman K, Beckman M, Pitrelli J, Ostendorf M, Wightman C, Price P, Pierrehumbert J, Hirschberg J (1992) ToBI: a standard for labeling English prosody. In: Proceedings of ICSLP, Banff, pp 867–870
56. Sneddon I, Goldie P, Petta P (2011) Ethics in emotion-oriented systems: the challenges for an ethics committee. In: Petta P, Pelachaud C, Cowie R (eds) Emotion-oriented systems: the HUMAINE handbook, cognitive technologies. Springer, Berlin/Heidelberg, pp 753–768

57. Stuhlsatz A, Meyer C, Eyben F, Zielke T, Meier G, Schuller B (2011) Deep neural networks for acoustic emotion recognition: raising the benchmarks. In: Proceedings 36th IEEE international conference on acoustics, speech, and signal processing, ICASSP 2011, Prague. IEEE, pp 5688–5691
58. Vogt T, André E, Bee N (2008) Emovoice – a framework for online recognition of emotions from voice. In: Proceedings IEEE PIT, Kloster Irsee. Lecture notes in computer science, vol 5078. Springer, pp 188–199
59. Weninger F, Schuller B, Batliner A, Steidl S, Seppi D (2011) Recognition of non-prototypical emotions in reverberated and noisy speech by non-negative matrix factorization. EURASIP J Adv Signal Process Spec Issue Emot Ment State Recognit Speech 2011:Article ID 838790
60. Weninger FJ, Watanabe S, Tachioka Y, Schuller B (2014) Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition. In: Proceedings 39th IEEE international conference on acoustics, speech, and signal processing, ICASSP 2014, Florence. IEEE, pp 4656–4660
61. Wöllmer M, Schuller B, Eyben F, Rigoll G (2010) Combining long short-term memory and dynamic Bayesian networks for incremental emotion-sensitive artificial listening. IEEE J Select Top Signal Process Spec Issue Speech Process Nat Interact Intell Environ 4(5):867–881
62. Yildirim S, Lee C, Lee S, Potamianos A, Narayanan S (2005) Detecting politeness and Frustration state of a child in a conversational computer game. In: Proceedings of INTERSPEECH, Lisbon. ISCA, pp 2209–2212
63. Yildirim S, Narayanan S, Potamianos A (2011) Detecting emotional state of a child in a conversational computer game. Comput Speech Lang 25:29–44
64. Zhang Z, Coutinho E, Deng J, Schuller B (2015) Cooperative learning and its application to emotion recognition from speech. IEEE/ACM Trans Audio Speech Lang Process 23(1): 115–126
65. Zhang Z, Coutinho E, Deng J, Schuller B (2014) Distributing recognition in computational paralinguistics. IEEE Trans Affect Comput 5(4):406–417
66. Zhang Z, Deng J, Marchi E, Schuller B (2013) Active learning by label uncertainty for acoustic emotion recognition. In: Proceedings of INTERSPEECH, Lyon. ISCA, pp 2841–2845
67. Zhang Z, Deng J, Schuller B (2013) Co-training succeeds in computational paralinguistics. In: Proceedings 38th IEEE international conference on acoustics, speech, and signal processing, ICASSP 2013, Vancouver. IEEE, pp 8505–8509