Chapter 15
Affective Conversational Interfaces

Abstract In order to build artificial conversational interfaces that display behaviors that are credible and expressive, we should endow them with the capability to recognize, adapt to, and render emotion. In this chapter, we explain how the recognition of emotional aspects is managed within conversational interfaces, including modeling and representation, emotion recognition from physiological signals, acoustics, text, facial expressions, and gestures and how emotion synthesis is managed through expressive speech and multimodal embodied agents. We also cover the main open tools and databases available for developers wishing to incorporate emotion into their conversational interfaces.

15.1 Introduction

Building on the overview of approaches to affect, emotion, and personality presented in Chap. 14, this chapter discusses how these features can be incorporated into conversational interfaces to make them more believable and more expressive. The first section looks at the Emotion Markup Language (EmotionML), a recommendation of the W3C for annotating features of emotion. Next, we provide a detailed discussion of the processes of emotion recognition, looking at the phases of data collection and annotation, learning, and optimization, and at the behavioral signals that are used to recognize emotion, including physiological signals, paralinguistic features in speech and text, facial expressions, and gestures. This is followed by an overview of the synthesis of emotion. For each of the different aspects discussed, we provide a list of various tools that are available for developers.

M. McTear et al., *The Conversational Interface*,
DOI 10.1007/978-3-319-32967-3_15

15.2 Representing Emotion with EmotionML

The Emotion Markup Language (EmotionML)[1] is a recommendation of the W3C published by the Multimodal Interaction Working Group. At the time of writing, the latest version is 1.0, published in 2014. Emotion Markup Language 1.0 is designed to be practically applicable and based on concepts from research in affect and emotion. EmotionML can be used to represent emotions and related concepts following any of the approaches described in Chap. 14.

The root element of an EmotionML file is `<emotionml>`, which may contain different `<emotion>` elements that represent the annotated emotions. Depending on the model used to represent emotions, the `<emotion>` element may include a `dimension set`, `category set`, or `appraisal set` attribute indicating the dimensional space, list of categories, or appraisal model, respectively. For each of these representations, different tags may be nested inside the `<emotion>` element.

The `<dimension>` element can be used for dimensional models. The attributes required are a name for the dimension and a value, and optionally, a confidence can be assigned to the annotation. Each <emotion> that is annotated may have as children as many <dimension> elements as are required by that space. For example, in the tridimensional space of pleasure (valence), arousal, and dominance (PAD), there would be three dimension elements. As anger can be characterized by low pleasure, high arousal, and high dominance, it would be represented as shown in Code 15.1.

It is possible to define the dimension set manually or to use previously defined models that can be found in the document Vocabularies for EmotionML.[2] In this document, there are lists of emotion vocabularies that can be used with EmotionML that follow scientifically valid inventories corresponding to categories (e.g., Ekman's Big Six), dimensions (e.g., Mehrabian's PAD), appraisals (e.g., the Ortony, Clore, and Collins (OCC) model of emotion), and action tendencies (e.g., Frijda's action tendencies).

```
<emotionml version="1.0"
xmlns="http://www.w3.org/2009/10/emotionml">
<emotion dimension-set="http://www.w3.org/TR/emotion-
voc/xml#pad-dimensions">
  <dimension name="pleasure" value="0.2/>
  <dimension name="arousal" value="0.8"/>
  <dimension name="dominance" value="0.8/>
</emotion>
</emotionml>
```

Code 15.1 Representing anger in a dimensional space with EmotionXML

[1] https://www.w3.org/TR/emotionml/. Accessed February 27, 2016.

[2] http://www.w3.org/TR/emotion-voc/. Accessed March 1, 2016.

The <category> element can be used for discrete models in which emotion is assigned a name attribute. As shown in Code 15.2, an item has been tagged as "anger," but in this case, it is not represented as a point in a space but as a category chosen from a catalog called "big6."

Here, we can see how to define the vocabulary manually, though we could have used a predefined catalog as in the previous example using `<emotion category-set = "http://www.w3.org/TR/emotion-voc/xml#big6">`. Optionally, it is possible to include a confidence attribute indicating the annotator's confidence that the annotation for the category is correct.

The `<appraisal>` element allows the use of appraisal models. The only required element is the name, but again we can also specify a value and a confidence attribute. For example, according to Scherer, anger entails appraising an event as incongruent with one's goals and values and intentionally caused. Note that, as discussed earlier in Chap. 14, with the appraisal model we are not interested in representing anger, but rather a situation that could be appraised as anger (Code 15.3).

```
<emotionml version="1.0"
xmlns="http://www.w3.org/2009/10/emotionml">

<!-- Vocabulary definition -->
<vocabulary type="category" id="big6">
  <item name="anger"/>
  <item name="disgust"/>
  <item name="fear"/>
  <item name="happiness"/>
  <item name="sadness"/>
  <item name="surprise"/>
</vocabulary>

<!-- Sample annotation of an item -->
<emotion category-set="#big6">
    <category name="anger"/>
</emotion>

</emotionml>
```

Code 15.2 Representing anger as an emotion category with EmotionXML

```
<emotionml version="1.0"
xmlns="http://www.w3.org/2009/10/emotionml">
<emotion appraisal-set="http://www.w3.org/TR/emotion-
voc/xml#scherer-appraisals">
  <appraisal name="self-compatibility" value="0.1"/>
  <appraisal name="cause-intentional" value="0.9"/>
</emotion>
</emotionml>
```

Code 15.3 Representing appraisal of anger with EmotionXML

Usually, it is necessary to tag more than one emotional item, and EmotionML facilitates different mechanisms to include time stamps and durations. For example, it is possible to indicate the absolute starting and finishing time or a starting time and duration. In the following example, we see that surprise starts at moment 1268647334000 and ends at 1268647336000, while anger starts at 1268647400000 and lasts 130 ms (Code 15.4).

To indicate relative durations, it is also possible to include an identifier for the <emotion> element. For example, we can say that anger starts 66,000 ms after surprise (Code 15.5).

Also, the <trace> element can be used to represent a periodic sampling of the value of an emotion (Code 15.6).

```
<emotion category-set="http://www.w3.org/TR/emotion-
voc/xml#big6" start="1268647334000" end="1268647336000">
  <category name="surprise"/>
</emotion>

<emotion category-set="http://www.w3.org/TR/emotion-
voc/xml#big6" start="1268647400000" duration="130">
  <category name="anger"/>
</emotion>
```

Code 15.4 Including time stamps and durations in EmotionXML

```
<emotion id="referenceSurprise" category-
set="http://www.w3.org/TR/emotion-voc/xml#big6"
start="1268647334000" end="1268647336000">
  <category name="surprise"/>
</emotion>

<emotion category-set="http://www.w3.org/TR/emotion-
voc/xml#big6" time-ref-uri="#referenceSurprise" offset-
to-start="66000">
  <category name="surprise"/>
</emotion>
```

Code 15.5 Including relative durations in EmotionXML

```
<emotion category-set="http://www.w3.org/TR/emotion-
voc/xml#big6">
  <category name="anger">
  <trace freq="10Hz" samples="0.1 0.1 0.15 0.2 0.2 0.25
0.25 0.25 0.3 0.3 0.35 0.5 0.7 0.8 0.85 0.85"/>
</category>
</emotion>
```

Code 15.6 Representing a sampled emotion in EmotionXML

According to its specification, EmotionML is conceived as a multipurpose language that can be used for manual annotation of corpora, as a standard language for the output of emotion recognizers, or to specify the emotional behavior generated by automated systems. In its specification, there are also different examples of how it can be used in combination with other compatible languages such as Extensible Multimodal Annotation Markup Language (EMMA), Synchronized Multimedia Integration Language (SMIL), or Speech Synthesis Markup Language (SSML).

15.3 Emotion Recognition

The processes involved in building an emotion recognizer are shown in Fig. 15.1. In this section, we focus on the following phases: data collection and annotation, learning, and optimization.

During the *data collection* phase different signals are recorded from the user and preprocessed to eliminate noise and other phenomena that may degrade them. The question here is which information can be obtained from the user that is a reliable source of emotional information and how to acquire it. Emotion recognition can be performed using any of the input modalities of the conversational interface (e.g., detecting emotion in the user's voice or facial expression) or using a combination of them. It can also take into account the appraisal mechanisms in users and the effect that the interaction itself may have on their emotional responses.

However, raw signals are not appropriate inputs for a classifier. They must be sampled and different features and statistics are usually computed in order to make

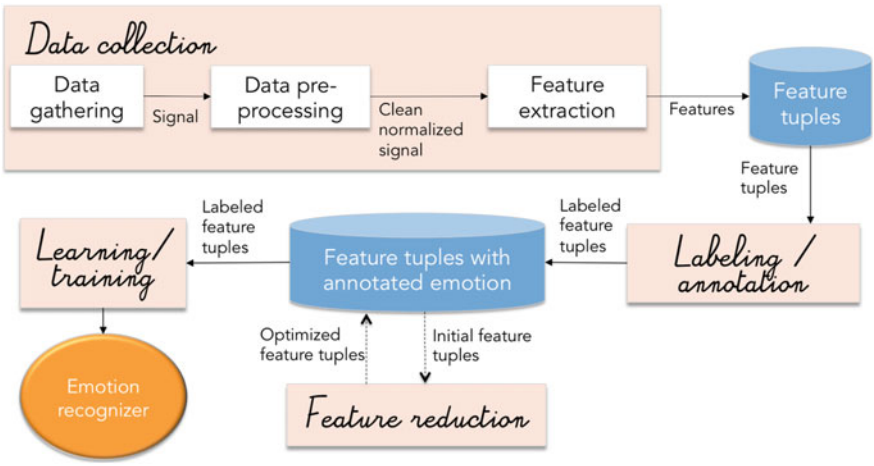

Fig. 15.1 Processes involved in building an emotion recognizer

them suitable for processing. Thus, the features are not only the raw values computed from the input signal (e.g., heart rate or voice volume), but statistical measures (e.g., heart rate variance and average volume). Sometimes there is a classification process in order to obtain meaningful features to be entered into the emotion recognizer. For example, from the recorded video, we may just focus on the mouth and from the mouth apply a classifier to determine whether the user is smiling or not. The unit being used for classification can also have an important impact on the performance of the classifier. Schuller and Batliner (2013) present a discussion of the advantages and disadvantages of different units.

Once we have obtained a database with all the recognition units (e.g., system utterances) represented as feature vectors, the database must be annotated to assign an emotion to each of the units. The *annotation* procedure depends on how the data was collected. The data can be obtained from acted emotions, elicited emotions, or spontaneous naturally occurring emotions. Acted data by professionals can be appropriate for some modalities although they miss some of the subtleties of emotional response production that cannot be consciously produced. For other signals, such as physiological data, databases of acted emotions are not suitable. With respect to elicited emotions, it is important to avoid inducing emotions different from the target emotion and eliminating the chances of inducing several emotions. There are some widespread emotion elicitation methods using pictures, films, music, and personal images (Calvo et al. 2014). Some authors have tuned video games or produced faulty versions of systems to induce negative emotions in users. Spontaneous emotions are the most natural, but they demand a complex emotion annotation process in order to obtain a reliable database. Also, they may have the drawback that not all emotions are frequent in all application domains and usually databases of spontaneous emotions are unbalanced.

Once annotated, the emotional database is used to *train a pattern recognition algorithm* that from a feature vector generates a classification hypothesis for an emotion. Different algorithms have been compared to check their suitability for this task, as will be described in the following sections.

As the overall idea is to recognize emotional states automatically from patterns of input features, there must be a process of feature design and optimization to keep only the relevant features. If too many features are considered, they may mislead the recognition process and slow it down to the point that it cannot be computable online (while the user is interacting with the system). On the other hand, the feature set must contain all the relevant features so that recognition is reliable. Sometimes *feature selection* is done offline, and on other occasions, algorithms are used that automatically select the best features while the system is operating.

The process used to recognize the user's emotion while interacting with a system is shown in Fig. 15.2.

Emotion recognition is a vast research area. In the following sections, we will describe the different sources of information that can be used in the recognition of emotion along with some discussion of the challenges involved.

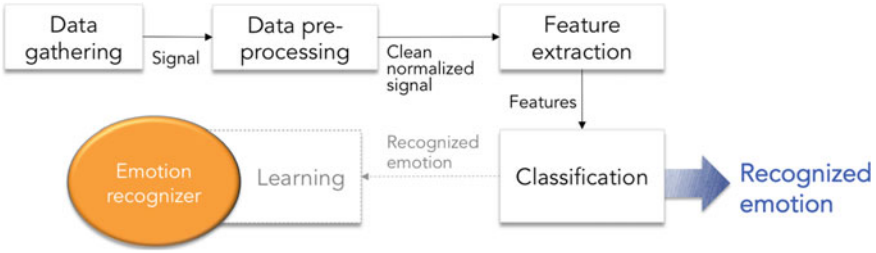

Fig. 15.2 The process of emotion recognition

15.3.1 Emotion Recognition from Physiological Signals

Different emotional expressions result in changes in autonomic activity, producing physiological signals that can be measured and used for emotion recognition. This activation affects different areas. Jerritta et al. (2011) classify them into cardiovascular system, electrodermal activity (EDA), respiratory system, muscular system, and brain activity and present a detailed table of measures. In this section, we will describe some of the most relevant of these.

The advantage of physiological signals is that they can be collected continuously to check for changes or patterns. The signals are robust against social artifacts that can hide emotions (e.g., a polite smile when the person is tense). As explained in Jerritta et al. (2011), even if a person does not overtly express his/her emotion through speech, gestures, or facial expression, a change in physiological patterns is inevitable and detectable because the sympathetic nerves of the autonomous nervous system become activated when a person is positively or negatively excited.

On the other hand, recording physiological signals requires sensors that, although in most cases they are not invasive, can be intrusive and uncomfortable for users. As discussed in Chap. 13, there are an increasing number of devices that can be connected to computers and smartphones and that provide sensing capabilities. However, developers will have to find a balance between the wearability and price of the equipment and its reliability as a source of information for recognizing emotion, since the cheap sensing devices that most users would find in stores are not sufficiently reliable in most cases and sensors that are reliable are expensive and mostly only used in laboratory conditions.

15.3.1.1 The Cardiovascular System

A heartbeat is a series of electrical impulses, involving depolarization and repolarization of the muscle whose electrical waveforms can be recorded. For example, electrocardiography (ECG) detects the electrical activity of the heart through electrodes attached to the outer surface of the skin and reflects emotional states such as tension or stress.

With every beat, the volume of blood is pushed and the generated pulse wave travels from the heart to all regions of the body. Blood volume pulse (BVP), also called heart photoplethysmography (PPG), sensors detect blood flow by using infrared light through the tip of a finger and measuring how much light is reflected.

The typical features for emotion classification obtained with the techniques described are heart rate variability (HRV), respiratory sinus arrhythmia (RSA), cardiac output, interbeat interval (IBI), and blood pressure (BP). Using these features, it is possible to differentiate mainly valence (distinguishing between positive and negative emotions), but it is also possible to recognize mental stress and effort. In this way, decreasing heart rate is a cue of relaxation and happiness, and increasing HRV is a sign of stress, anger, fear, and frustration.

Anger increases diastolic BP to the greatest degree, followed by fear, sadness, and happiness, and can be distinguished from fear by larger increases in blood pulse volume, while an increased interbeat interval can be used to detect amusement and sadness.

15.3.1.2 Electrodermal Activity

EDA measures the skin's ability to conduct electricity, which reflects changes in sympathetic nervous systems due to emotional responses and is specifically correlated with arousal (e.g., it usually increases for amusement and sadness as found in emotion elicitation using films). It is usually measured in terms of skin conductance (SC) and galvanic skin response (GSR).

Sensors for SC and GSR, which are usually placed on fingers, are based on applying a small voltage to the skin to measure its conductance or resistance. As SC depends on the activity of the sweat glands, it is important to consider the perspiration component when measuring these features. Sometimes skin temperature is used to measure the temperature at the surface of the skin, usually also on the fingers. It is important to take into account that the temperature varies depending on where the sensor is placed on the user's body and even on the time of day or the activity.

15.3.1.3 The Respiratory System

Typical features related to the respiratory system are breaths per minute, respiration volume, and relative breath amplitude. These features measure how deep and fast the user breathes and can be gathered with a sensor incorporated in a band fixed to the user's chest that accounts for chest expansions. An increasing respiration rate is related to anger and joy and a decreased respiration rate with relaxation and bliss. Respiratory cues can then be used to recognize arousal. However, it is important to consider that shocking events may cause the user's respiration to cease for a

moment. Also, this can be indicative of negative valence as negative emotions may cause irregular respiratory patterns.

Breathing is linked to cardiac features and is also related to talking. These correlations must be considered in multimodal conversational interfaces in order to avoid interdependencies.

15.3.1.4 The Muscular System

A frequent method for gathering information from the muscular system is electromyography (EMG), which measures the electric impulses generated by muscles during their activity: the higher the amplitude of the electric signal, the higher the power of muscular contraction, although the signal can be very subtle, and thus, it is mainly used for strong emotions.

Although it can be useful to distinguish facial expressions, facial EMG requires fixing electrodes on the user's face, which can be intrusive, and thus, other approaches based on video are usually employed to recognize emotions from facial expressions and gestures (see further Sect. 15.3.3).

15.3.1.5 Brain Activity

Brain activity is usually measured by means of electroencephalography (EEG) and brain imaging methods such as positron emission tomography. EEG measures the electrical voltages generated by neurons when they fire. There are different frequency subsets: high beta (20–40 Hz), beta (15–20 Hz), sensorimotor rhythm (13–15 Hz), alpha (8–13 Hz), theta (4–8 Hz), and delta (2–4 Hz). The meaning of the signals gathered and their relations to emotional states are described in Bos (2006).

EEG is still not very well suited for practical implementations because of the high sensitivity to physiological artifacts such as eye blinks and electrostatic artifacts.

15.3.1.6 Classification Based on Physiological Cues

Usually, classification based on physiological cues makes use of a combination of the features described in order to obtain patterns for a certain emotion. For example, Jang et al. (2014) report that the responses of the autonomous system for fear comprise broad sympathetic activation including cardiac acceleration, increased myocardial contractility, vasoconstriction, and electrodermal activity (EDA).

It is interesting to note that the accuracy of arousal discrimination is usually higher than that for valence. The reason might be that the change in the arousal

level corresponds directly to the intensity of activities such as sweat glands and BP, which is straightforward to measure with single features, while the valence differentiation of emotion requires a multifactor analysis (Kim and André 2008).

With respect to classification accuracy, Kim and André (2008) report results from early studies that already attained over 80 % accuracy. Picard et al. (2001) obtained a recognition accuracy of over 80 % on average with a linear approach; Nasoz et al. (2003) achieved an emotion classification accuracy of 83 %; and Haag et al. (2004) classified arousal and valence separately using a neural network classifier and obtained recognition accuracy rates of 96.6 and 89.9 %, respectively. Kim et al. (2004) obtained a classification ratio of over 78 % for three emotions (sadness, stress, and anger) and over 61 % for four emotions (sadness, stress, anger, and surprise) by adopting support vector machines as a pattern classifier. In all these approaches, the emotion database used was acted or elicited, which may make the results more difficult to replicate in spontaneous settings.

A more recent study by Jerritta et al. (2011) presents a detailed survey of more than 40 studies on physiological emotion recognition with accuracies ranging from 66 to 95 % using support vector machines, linear discriminant analysis, Bayesian networks, Hidden Markov models, k-nearest neighbors, and neural networks, with databases containing from 1 to 154 subjects.

15.3.1.7 Open Tools for the Analysis of Physiological Signals

The Augsburg Biosignal Toolbox (AuBT)[3] is a toolbox written in MATLAB and developed at the University of Augsburg. The toolbox can be used to analyze physiological signals by extracting their features, automatically selecting the relevant features, and using these features to train and evaluate a classifier. AuBT includes two corpora: a corpus containing physiological data of a single user in four different emotional states (Augsburg database of biosignals (AuDB)) and a corpus containing physiological data recorded from a single user under varying stress DRIving under VArying WORKload (DRIVAWORK).

AuDB was collected by recording ECG, EMG, SC, and respiratory features of one participant while listening to music in order to induce one of the emotions of joy, anger, sadness, and pleasure. There were 25 separate sessions on different days with a total of 200 min of data.

DRIVAWORK contains recordings of ECG, EMG, SC, temperature, BVP, and respiratory features along with audio and video recordings of participants in a simulated car drive. It contains a total of 15 h from 24 participants where relaxed and stressed states were elicited by giving the participants different tasks on top of a driving task.

[3] https://www.informatik.uni-augsburg.de/de/lehrstuehle/hcm/projects/tools/aubt/. Accessed February 27, 2016.

15.3.2 Emotion Recognition from Speech

For conversational interfaces, the user's spoken input is probably the most relevant source of emotional information in that it encodes the message being conveyed (the textual content) as well as how it is conveyed (paralinguistic features such as tone of voice).

15.3.2.1 Paralinguistic Features

Many acoustic features can be obtained from the speech signal, although there is no single approach for classifying them. Batliner et al. (2011) distinguish segmental and suprasegmental features.

Segmental features are short-term spectral and derived features, including mel-frequency cepstral coefficients (MFCCs), linear predictive coding (LPC), and wavelets. Suprasegmental features model prosodic types such as pitch, intensity duration, and voice quality. Features can be represented as raw data or they can be normalized, standardized, and presented as statistics (means, averages, etc.).

The main groups of acoustic features used for emotion recognition are listed below. Usually, for each of these groups, different features are computed, including statistics such as minimum, maximum, variance, mean, and median.

- **Intensity (energy)**. Intensity is the physical energy of the speech signal and models the loudness of a sound as perceived by the human ear.
- **Duration**. Duration models temporal aspects of voiced and unvoiced segments. It can be computed over the whole signal or on higher-order phonological units, e.g., words, to be correlated with their linguistic content.
- **Zero Crossing Rate (ZCR)**. ZCR counts the number of times the speech signal changes its sign and thus at some point equals zero. It is useful to tell whether a speech signal is voiced (low ZCR) or not.
- **Pitch/Fundamental frequency**. The fundamental frequency F0 is very representative of emotion, as human perception is very sensitive to changes in pitch.
- **Linear Prediction Cepstral Coefficients (LPCCs)**. Spectral features represent phonetic information. Their extraction can be based on LPC. The main idea of linear prediction is that the current speech sample can be predicted from its predecessors, i.e., it can be approximated by a linear combination of previous samples.
- **Mel-Frequency Cepstral Coefficients (MFCCs)**. MFCCs are among the most widely used speech features for automatic speech processing including speech and speaker recognition. They are computed by transforming the signal into a cepstral space. Coefficient 0 describes the signal energy. Coefficients 1–12 (approximately) describe mainly the phonetic content, and higher-order coefficients describe more the vocal tract and thus speaker characteristics.

Table 15.1 Common effects of emotion in speech features of Western languages (based on Väyrynen 2014)

	Anger	Fear	Joy	Sadness	Disgust	Surprise
Speech rate	>	≫	> or <	<	⋘	>
Pitch average	⋙	⋙	≫	<	⋘	>
Pitch range	≫	≫	≫	<	>	>
Pitch changes	Abrupt	Normal	Smooth up	Down	Down terminal	High
Intensity	>	=	>	<	<	>
Voice quality	Breathy	Irregular	Breathy	Resonant	Grumbled	Breathy
Articulation	Tense	Precise	Normal	Slurring	Normal	

The symbols ">," "≫," and "⋙" represent increase and symbols "<," "≪," and "⋘" decrease, while "=" indicates no perceived change

- **Formants**. Cepstral coefficients are very widespread features for speech processing, but they have a poor performance with noisy speech. Thus, to handle real-life speech, they can be supplemented with formant parameters. Formants are used to model changes in the vocal tract shape and they vary according to the spoken content, in particular formants F1 and F2 and their bandwidths.
- **Wavelets** represent a multilevel analysis of time, energy, and frequencies of a speech signal and account for its sharp transitions and drifts.
- **Voice Quality**. Voice quality features are based on acoustical models of the vocal folds. They model jitter, shimmer, and further microprosodic events.

For a review of emotionally relevant features and extraction techniques, see (Batliner et al. 2011; Cowie and Cornelius 2003; Ververidis and Kotropoulos 2006). A summary of commonly associated emotion effects in relation to normal speech is shown in Table 15.1.

15.3.2.2 Classification of Paralinguistic Features

Since 2009, the INTERSPEECH Conference, organized by the International Speech Communication Association (ISCA), has held Computational Paralinguistics Challenges. Not all editions have focused on emotion, but they have provided an opportunity to share databases, to replicate and compare results, and to compile the best features and algorithms to be used for classification. All subchallenges allow contributors to use their own features and machine learning algorithms, although participants adhere to the definition of training, development, and test sets. Thus, results are directly comparable, and it is easy for the interested reader to check which approaches have obtained the best results in each challenge.[4]

[4]http://compare.openaudio.eu/. Accessed February 27, 2016.

15.3.2.3 Open Tools for the Analysis of Paralinguistic Features

There are several tools—most of them open source—that provide the algorithms to perform acoustic analysis and visualization, as well as tools for scripting and classification.

Praat phonetics software, developed at the University of Amsterdam, is an open-source software package for the analysis of speech (Boersma and Weenink 2016). Praat implements algorithms to perform the main phonetic measurement and analysis procedures, including working with waveforms and spectrograms, measuring pitch, pulses, harmonics, formants, intensity, and sound quality parameters. Praat also features graphic representations and statistics and allows users to create their own scripts and communicate with other programs. On its Web page[5], it is possible to find information about how to use it to compute the features described previously. There is also a Praat Users Group in yahoo[6] and conversations about Praat in other communities of programmers such as Stack Overflow.[7]

EmoVoice/Open SSI. The Open Social Signal Interpretation framework (Open SSI)[8] offers tools to record, analyze, and recognize human behavior in real time, including gestures, mimics, head nods, and emotional speech (Wagner et al. 2013). EmoVoice was developed in the Human-Centered Multimedia Lab in the University of Augsburg and has been used by several EU-funded projects related to affective interaction. EmoVoice is integrated into the Social Signal Interpretation (SSI) framework and provides modules for real-time recognition of emotions from acoustics. The modules include speech corpus creation, segmentation, feature extraction, and online classification. The phonetic analysis used by EmoVoice relies on the algorithms provided by Praat.

openSMILE. The Speech and Music Interpretation by Large-space Extraction (SMILE) tool also provides general audio signal processing, feature extraction, and statistics as in the previously described tools. Its input/output formats are compliant with other widespread tools for machine learning such as the Hidden Markov Toolkit, Waikato Environment for Knowledge Analysis (WEKA), and the Library for Support Vector Machines (LibSVM). openSMILE was started at the Technical University of Munich by Florian Eyben, Martin Wöllmer, and Björn Schuller (Eyben et al. 2013) and is now maintained by audEERING and distributed free of charge for research and personal use.[9]

Databases. In order to train emotion recognizers, there is a need for emotionally labeled corpora. There are different corpora that have been released under varying licenses. The Association for the Advancement of Affective Computing compiles the main databases and tools and is constantly updated. For example, they have the

[5] http://www.fon.hum.uva.nl/praat/. Accessed February 27, 2016.

[6] https://uk.groups.yahoo.com/neo/groups/praat-users. Accessed February 27, 2016.

[7] http://stackoverflow.com/questions/tagged/praat. Accessed February 27, 2016.

[8] http://hcm-lab.de/projects/ssi/. Accessed February 27, 2016.

[9] http://www.audeering.com/research/opensmile. Accessed February 27, 2016.

HUMAINE, Belfast Naturalistic, and Geneva Vocal Emotion Expression Stimulus databases.[10] We recommend readers to check this Web page as it contains information about projects, journals and conferences, researchers, tools, and databases related to affective computing. Also, the European Language Resources Association (ELRA) has a catalog of spoken, written, and multimodal resources, some of them related to emotion.[11]

15.3.2.4 Extracting Affective Information from Text

In conversational interfaces, the user's spoken input is translated into text by means of an automatic speech recognizer. The text is used to extract the semantics of the message conveyed and to compute the most adequate system response. However, the text also carries information about the user's emotional state. This is encoded in the words and grammatical structure. For example, saying "as you wish" is not the same as saying "do what the hell you want."

There are many techniques for extracting affective information from text. Most of these involve applying techniques that are widely used for ASR and SLU to this new classification task. For example, emotion recognition from text uses preprocessing stages that are common to techniques used in SLU, such as stemming (separating lexemes from morphemes in order to avoid a dimensionality problem, especially in highly inflective languages) and stopping (removing non-relevant words). For the processing of affect, non-linguistic vocalizations such as sighs, yawns, laughs, and cries are important and are usually included as vocabulary.

The main approaches used are bag of words, n-grams, rule models, and semantic analysis. Of these, bag of words and n-grams are widely used because of their simplicity.

Bag of words. The main idea behind this approach is that words have an affective charge and some words are more frequent in expressions produced under certain emotional states than others. In this approach, the emotional salience of each word in the vocabulary to be considered is computed in a training corpus that is emotionally annotated. Then, when a new user utterance is ready to be processed, an overall emotional score is computed taking into account the most representative emotional category for each word in the utterance.

Sometimes we can also use already prepared vocabularies so that we just have to count the number of appearances of each word to compute the probability of each emotion being considered. For example, "hell" is very likely to appear under an anger setting, and thus, "do what the hell you want" can we considered as anger, while "as you wish" could be considered neutral as the vocabulary employed would have a low probability of being related to any affect category.

[10] http://emotion-research.net/toolbox/toolbox_query_view?category=Database. Accessed February 27, 2016.

[11] http://catalog.elra.info/. Accessed February 27, 2016.

N-grams. Sometimes considering words in isolation does not provide accurate results, as it may be necessary to account for the relation between the different words in the phrase. For example, if the user says "this is fun as hell!", "hell" should not be considered an indicative of anger. Using n-grams (see Chap. 8), we could account for how the preceding structure "fun as" changes the polarity of "hell" from negative to positive. In current work, mostly unigrams and bigrams (and very rarely trigrams) have been employed for emotion recognition, e.g., Polzin and Waibel (2000).

Rule-based approaches. These approaches are based on expert knowledge of the topic and have been extensively used in opinion mining, usually to determine the polarity of opinion (whether it is positive or negative). For example, saying that the battery of a wearable is rechargeable is usually categorized as a positive opinion. This is, however, a tricky method in application domains that are very variable, for example, saying that a smartphone is big might have been negative some years ago but is positive now (and may be negative again in the future).

The rules can also be applied to quantify the effects of connectors and qualifiers in combination with a bag-of-words approach. Thus, if the word "funny" indicates positive polarity, "extremely funny" should have even a more positive value, and a rule-based approach can be used to indicate explicitly how much higher. Similarly, if we say someone is "friendly and kind," this is more positive than just "friendly" or just "kind."

The rules may also change depending on user models. For example, in a tutoring system, if a student says that he or she finds a problem difficult, it may be bad for a student who is having problems with the subject and who may become frustrated, but good for a student who is doing well and who likes challenges.

Linguistic analysis. Emotion can also be obtained from text by means of a semantic analysis using the techniques described in Chap. 8.

Usually, techniques such as these are used for sentiment analysis (see the discussion on the difference between the semantic analysis and affective interaction communities in Clavel and Callejas (2016)). Sentiment analysis is becoming very popular for opinion mining (e.g., for companies to control conversations about their brands in social networks). Although the bag-of-words and n-gram approaches can also be employed in this area, it is very relevant to detect which is the object of the opinion/affect, something that is also very important for emotion recognition when using appraisal approaches.

The conversational context. Irrespective of the approach used, the information derived from the current user utterance must be interpreted in the context of the ongoing dialog. Techniques used to evaluate the contribution of the semantics of the input to the conversation and how the system response is computed are described in Chap. 10. With respect to the affect-related interpretation, it is possible to use the conversational context to compute the probability of each emotional state of the user more reliably. For example, if the system has been faulty, this may cause a negative user state (Callejas et al. 2011).

15.3.2.5 Open Tools for Extracting Emotional Information from Text

Tools for processing natural language text and that can also be used for extracting emotional information from text were described in Chap. 9, for example, the Natural Language Toolkit (NLTK)[12] and Apache OpenNLP.[13]

Specialized databases. The following are lexical resources for sentiment analysis and/or opinion mining. Sentiment lexicons are the most crucial resource for most sentiment analysis algorithms.

- **SentiWordNet**.[14] SentiWordNet assigns to each synset of WordNet three sentiment scores: positivity, negativity, and objectivity (Baccianella et al. 2010).
- **Affective Norms for English Words** (ANEW).[15] This dataset provides normative emotional ratings for a large number of words in the English language in terms of pleasure, arousal, and dominance.
- Opinion and sentiment lexicons and other resources by **Bing Liu**, author of books and scientific papers on sentiment analysis (Liu 2015).[16]
- **General Inquirer Home Page**.[17] The Harvard General Inquirer is a lexicon attaching syntactic, semantic, and pragmatic information to part-of-speech tagged words.
- **Sentiment in finance and accounting**. Words appearing in documents from 1994 to 2014. The dictionary reports count statistics, proportion statistics, and nine sentiment category identifiers (e.g., negative, positive, uncertainty, litigious, modal, constraining) for each word.[18]
- **Creating your own sentiment lexicon**. Sometimes, especially when working in a very specific application domain, it is necessary to build a specific sentiment lexicon. This is a very demanding task, but it is possible to expand existing resources to facilitate this process. In Feldman (2013), there is a description of how to do this from WordNet by using a vocabulary of seed adjectives and introducing synonyms with "sentiment consistency."

15.3.3 Emotion Recognition from Facial Expressions and Gestures

Some authors have identified facial expressions as the most important clue for emotion detection, and in fact, emotion recognition from facial features is one of the research

[12] http://www.nltk.org/. Accessed February 27, 2016.
[13] https://opennlp.apache.org/. Accessed February 27, 2016.
[14] http://sentiwordnet.isti.cnr.it/. Accessed February 27, 2016.
[15] http://csea.phhp.ufl.edu/media.html#bottommedia. Accessed February 27, 2016.
[16] https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html. Accessed February 27, 2016.
[17] http://www.wjh.harvard.edu/~inquirer/. Accessed February 27, 2016.
[18] http://www3.nd.edu/~mcdonald/Word_Lists.html. Accessed February 27, 2016.

topics with a longer trajectory in the area (Ekman 1999). There are two approaches for facial expression analysis: message based and sign based (Calvo et al. 2014).

Message-based analysis is based on the assumption that the face "is the mirror of the soul" and that it displays a representation of a person's emotional state. Some authors have provided evidence of facial expressions that signal a reduced number of basic emotions that are recognizable across cultures. Darwin described facial expressions for more than 30 emotions, and the work by Ekman is quite paradigmatic on "universal" basic emotions (Ekman 2003; Ekman and Rosenberg 2005). There are even studies of homologous emotion recognition from facial expressions in primates.

However, many other authors are not comfortable with the assumption of message-based measurement and believe that interpreting the meaning of an expression depends on the context. For example, the same expression may indicate different emotional states depending on the context in which it was produced and the person who produced it. Also, facial expressions could be posed, and thus, there would be no correspondence between the real emotional state of the user and their facial expression.

Sign-based analysis is more similar to the speech signal analysis described earlier. The idea is to obtain relevant features from corpora of annotated facial expressions and by means of a machine learning approach to learn patterns that show the relation between the feature tuples and the annotated emotion.

There are different methods employed to discretize facial expressions into relevant inputs for classification. The most relevant is the Facial Action Coding System (FACS) (Ekman and Rosenberg 2005), which deconstructs facial expressions into action units (AUs) chosen from a repertoire of more than 40 indicating their presence or absence or their intensity (see some examples in Fig. 15.3).

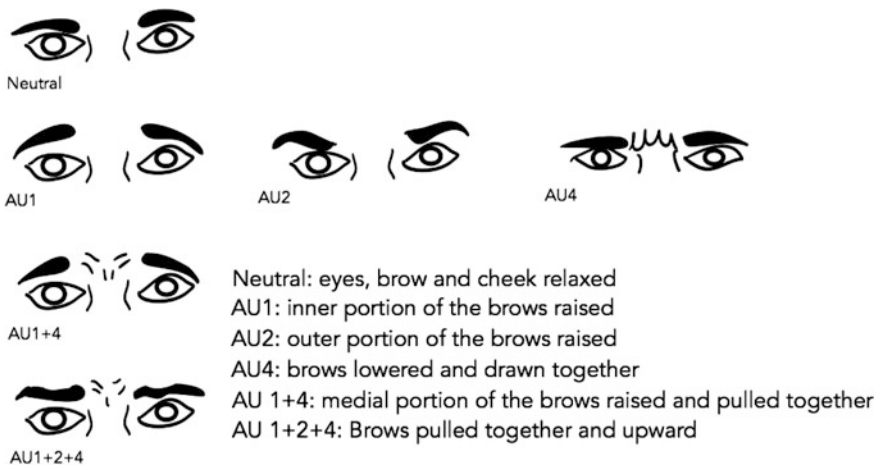

Fig. 15.3 Sample action units

Once the AUs have been detected, a classification process can be used to determine the emotion, although some authors have also detected mappings that allow the use of rule-based approaches.

15.3.3.1 The Facial Expression Recognition Process

The process of facial expression recognition is shown in Fig. 15.4:

Face detection. Face detectors are reviewed in Zhang and Zhang (2010). The main challenge here is to overcome events that make image analysis difficult, as the user's face is not captured in its totality or in the correct position for an optimal feature extraction. For instance, head motion, partial coverage of the face (e.g., if the user puts their hand in front of their face or there is another obstacle between the camera or the face), and non-frontal poses.

Face normalization. There are individual differences in head shape, skin color, facial proportion as well as effects of the spatial face position that can be reduced by converting the face detected to a canonical size and orientation.

Facial feature detection and tracking. The features typically used for emotion recognition from facial expressions are based on the local spatial position or displacement of specific points and regions of the face:

- Position and shape features (also called geometric features) that account for shapes (e.g., eyebrows) and positions (e.g., edges of the mouth).
- Motion features that account for the movement of facial muscles (e.g., optical flow or dynamic models of specific regions).
- Appearance features that represent changes in skin texture (e.g., wrinkles).

Geometric features refer to facial landmarks such as the eyes or brows. They can be represented as points, a connected face mesh, active shape model, or face component shape parameterization.

As described in Calvo et al. (2014), we can further divide geometric features into sparse (e.g., eyes or eye corners) or dense (e.g., the contours of the eyes and other permanent facial features). An advantage of the latter is that they provide information from which to infer a 3D pose. To track a dense set of facial features, active

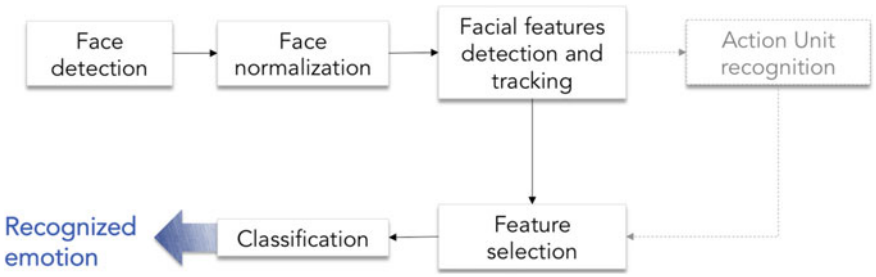

Fig. 15.4 The process of facial expression recognition

Fig. 15.5 Schema of an AAM mesh

appearance models (AAMs) are often used that describe shape by a 2D triangulated mesh. In particular, the coordinates of the mesh vertices define the shape and the vertex locations correspond to a source appearance image from which the shape is aligned (see Fig. 15.5).

Motion features include optical flow and dynamic textures or motion history images (MHI). These methods all encode motion in a video sequence.

Appearance features represent changes in skin texture such as wrinkling and deepening of facial furrows and pouching of the skin. Many techniques for describing local image texture have been proposed. A major challenge is that lightning conditions affect texture. Biologically inspired appearance features, such as Gabor wavelets or magnitudes, are more robust.

AU recognition. The features described can be directly used to recognize an emotional state or can be used to recognize action units that are then used as a basis for emotion recognition (i.e., *Features* → *Emotion* or *Features* → *AUs* → *Emotion*).

Feature selection. As happened with speech and physiological signals, there can be many features derived from facial expressions that could be used for classification, but it is important to reduce dimensionality. The same methods as in the previous cases apply (e.g., component analysis, bootstrapping), but there are also some that are specific to visual features, including eigenmaps and locality preserving projections.

15.3.3.2 Classification of Emotions from Facial Expressions

Most approaches use supervised learning with previously annotated data. For a review, see Ryan et al. (2009). Calvo et al. (2014) discuss two approaches to supervised learning:

1. Static modeling—typically posed as a discriminative classification problem in which each video frame is evaluated independently.
2. Temporal modeling—frames are segmented into sequences and typically modeled with a variant of dynamic Bayesian networks (e.g., Hidden Markov models, conditional random fields).

Temporal dynamics also help the study of transitions between emotions. In conversational interfaces, temporal models are also interesting for coping with the effect of the movement of the facial muscles while talking and how they may interfere with one another.

15.3.3.3 Emotion Recognition from Gestures

Although gestures convey important affective information, they have not been exploited much yet. The process followed is the same as in Fig. 15.5, but with other features focusing on the body instead of on the face.

Usually, to avoid interpersonal differences of body shape and other aspects, work focuses on different abstract representations:

- Skeleton. A representation of the skeleton is used. In order to do this, either professional equipment or generally available commercial applications such as Microsoft Kinect have been used.
- Silhouette and blobs. For example, hand blobs from which motion features are extracted (acceleration, fluidity, symmetry, duration). See, for example, the work by Castellano et al. (2010).

15.3.3.4 Tools for Recognizing Facial Expressions and Gestures

ANVIL.[19] ANVIL is a free video annotation tool. It offers multilayered annotation based on a user-defined coding scheme. During coding, the user can see color-coded elements on multiple tracks in time alignment. Some special features are cross-level links, non-temporal objects, coding agreement analysis, 3D viewing of motion capture data, and a project tool for managing whole corpora of annotation files (Kipp 2012).

MUMIN annotation model. Implemented in various annotation tools, this model deals with communicative nonverbal behaviors such as facial expressions, head movements, hand gestures, body postures, and gaze. The MUMIN coding scheme, developed in the Nordic Network on Multimodal Interfaces, is intended as a general instrument for the study of gestures (in particular, hand gestures and facial displays) in interpersonal communication, focusing on the role played by multimodal expressions for feedback, turn management, and sequencing (Allwood et al. 2008).

[19] http://www.anvil-software.org/. Accessed February 27, 2016.

Databases

- **Danish first encounter NOMCO corpus**[20] (Paggio and Navarretta 2011).
- **Cohn–Kanade database**.[21] The Cohn–Kanade AU-coded facial expression database is for research in automatic facial image analysis and synthesis and for perceptual studies. Cohn–Kanade is available in two versions. The first version comprises 486 sequences from 97 posers, and the second includes both posed and non-posed (spontaneous) sequences.
- The **MMI facial expression database**[22] (Pantic et al. 2005). This database consists of over 2900 videos and high-resolution still images of 75 subjects annotated for the presence of AUs in videos (event coding) and partially coded on the frame level, indicating for each frame whether an AU is in either the neutral, onset, apex, or offset phase.
- A complete overview of publicly available data sets that can be used in research on automatic facial expression analysis is provided in Pantic and Bartlett (2007).

15.4 Emotion Synthesis

Emotion synthesis is based to a large extent on the same features described for emotion recognition, so we will focus mainly on the tools and resources available.

There is an extensive body of work that shows that humans assign human characteristics to artificial interlocutors. Especially relevant are the experimental results achieved by Nass who shows that we often assign emotional content to synthetic voices and treat conversational systems as social counterparts (Nass and Lee 2000; Nass and Yen 2012).

15.4.1 Expressive Speech Synthesis

In the case of speech synthesis, the same parameters described in Sect. 15.3.2 can be applied to color a message conveyed with emotion. This area of study is known as expressive speech synthesis (ESS).

There are several approaches to ESS. On the one hand, it is possible to modify naturally synthesized speech based on the prosodic rules that generated the desired expression. That is, once we have a series of rules that determine which parameters to change to synthesize emotional speech, we tweak those parameters to convey the desired emotion (e.g., make it faster and louder in order to sound angry).

[20] http://metashare.cst.dk/repository/browse/danish-first-encounters-nomco-corpus/6f4ee056444211e2b2e00050569b00003505d6478d484ae2b75b737aab697e99/. Accessed February 27, 2016.

[21] http://www.pitt.edu/~emotion/ck-spread.htm. Accessed February 27, 2016.

[22] http://mmifacedb.eu/. Accessed February 27, 2016.

On the other hand, we could use recordings corresponding to the target emotion that are already available in a database. To gain flexibility, the recordings can consist of small units that can correspond to the target emotion or to other emotions that are blended to generate new styles.

The procedure followed in both cases is the same as described for TTS (see Chap. 5), but accepting the target emotion as another input. Additional details on work addressing each of these techniques can be found in the comprehensive reviews by Govind and Prasanna (2012), Schröder (2009), and van Santen et al. (2008).

15.4.1.1 Tools

The same tools described in Sect. 15.3.2 apply here. For the specific case of speech synthesis, **EmoFilt**[23] is a very interesting tool for readers who want to experiment with ESS, as it shows very clearly how to generate emotions from a neutral voice by configuring the synthesis parameters. EmoFilt is an open-source program based on the free-for-non-commercial-use MBROLA synthesis engine (Burkhardt 2005).

15.4.2 Generating Facial Expressions, Body Posture, and Gestures

Embodied conversational agents (ECAs) are able to display facial expressions and gestures (see Chap. 14). Based on the earlier discussion of emotion perception in facial expression and gestures (Sect. 15.3.3), it is possible to generate static snapshots of emotional expression (e.g., by generating the relevant action units). However, expressive behaviors for conversational interfaces involve not only choosing the appropriate features but also deciding how they are realized and especially what are the dynamic qualities of the signals generated.

With respect to facial expressions, different approaches have been used in the literature, from interpolating discrete emotions to fuzzy logic, or by superposing different facial areas corresponding to different emotions. For gestures and body movements, different dynamic models are used based on features that correspond to temporal, spatial, power, and fluidity aspects. A comprehensive description can be found in the following references (Pelachaud 2009; Niewiadomski et al. 2013).

In addition to the synthesis of these general qualities, given that the generation of multimodal expressive behavior encompasses many details, it is possible to find very complex research on narrower aspects, such as emotional eye movement and gaze or smiling.

[23] http://emofilt.syntheticspeech.de/. Accessed February 27, 2016.

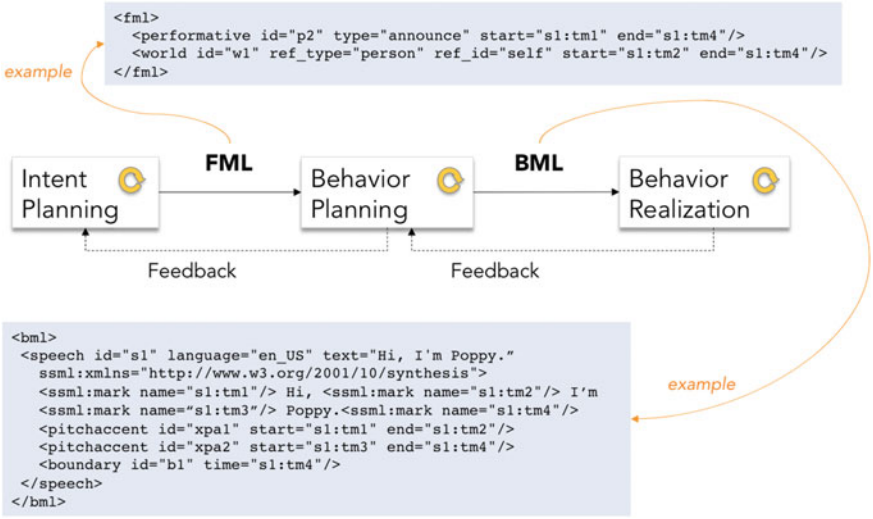

Fig. 15.6 The SAIBA model

15.4.2.1 Tools for Generating Facial Expressions, Body Posture, and Gestures

SAIBA, FML, and BML. SAIBA is a model for unifying multimodal behavior generation for ECAs. It consists of three stages: intent planning, behavior planning, and behavior realization. Figure 15.6 shows the general architecture and an example of the FML and BML files generated.[24]

The intent planner decides the communicative intention, the behavior planner schedules the communicative signals, and finally the behavior realizer realizes the behaviors scheduled to generate the corresponding animation. As shown in Fig. 15.6, the interface between stages (1) and (2)—intent planning and behavior planning—describes communicative and expressive intent without any reference to physical behavior. This information (e.g., the agent's current goals, emotional state, and beliefs) can be specified with the Functional Markup Language (FML),[25] which provides a semantic description that accounts for the aspects that are relevant and influential in the planning of verbal and nonverbal behaviors (Cafaro et al. 2014).

The interface between stages (2) and (3)—behavior planning and behavior realization—describes multimodal behaviors as they are to be realized by the final

[24]The example code shown is from the SEMAINE project: http://semaine.opendfki.de/wiki/FML. Accessed February 27, 2016.

[25]http://secom.ru.is/fml/. Accessed February 27, 2016.

stage of the generation process (e.g., speech, facial expressions, and gestures). The Behavior Markup Language (BML) was proposed to provide a general description of the multimodal behavior that can be used to control the agent (Kopp et al. 2006).

Alma.[26] Alma is a computational model of real-time affect for virtual characters. It contains appraisal rules for emotion, mood, and personality to control the physical behavior of ECAs. It also provides a CharacterBuilder tool based on the AffectML language. It has been programmed in Java, and its code is available in GitHub.

EMA (Marsella and Gratch 2009). EMA is another computational model of emotion, based on the time dynamics of emotional reactions that can be used to generate naturalistic emotional behaviors for ECAs.

Fatima.[27] Fatima is an autonomous agent architecture based on BDI and OCC, initially developed to control the minds of the agents in FearNot! (in the European project VICTEC). The architecture focuses on using emotions and personality to influence the agent's behavior.

Other tools. The general-purpose tools cited in Chap. 14 to develop ECAs can also be employed to render emotional behaviors.

15.4.3 The Uncanny Valley

Many experiments have demonstrated that a higher degree of human likeness increases the appeal of agents and robots. However, when building very realistic agents, there is the danger of falling into the so-called *uncanny valley* (Kätsyri et al. 2015). As shown in Fig. 15.7, as the agent becomes more sophisticated, it is more familiar and better accepted by users until it reaches a point when it becomes disturbing because it is very real but still not as natural as would be expected (that is the valley), and then as they become more human-like, the acceptability increases again.

The generation of rich expressive behaviors may lead to expectations in users that, when not addressed, may negatively affect their perception of the system. Ben Mimoun et al. (2012) present an interesting discussion of the reasons for the failure of ECAs, including an exaggeration of expectations. They present some solutions, such as explaining clearly to users the limitations of the agent and its functionality.

As discussed in Mathur and Reichling (2016), humans seem to appear to infer trustworthiness from affective cues (e.g., subtle facial expressions) that are known to contribute to human–human social judgments, and thus, affective interaction is a key to developing believable and likeable conversational interfaces.

[26] http://alma.dfki.de. Accessed February 27, 2016.

[27] http://sourceforge.net/projects/fearnot/files/FAtiMA/FAtiMA/. Accessed February 27, 2016.

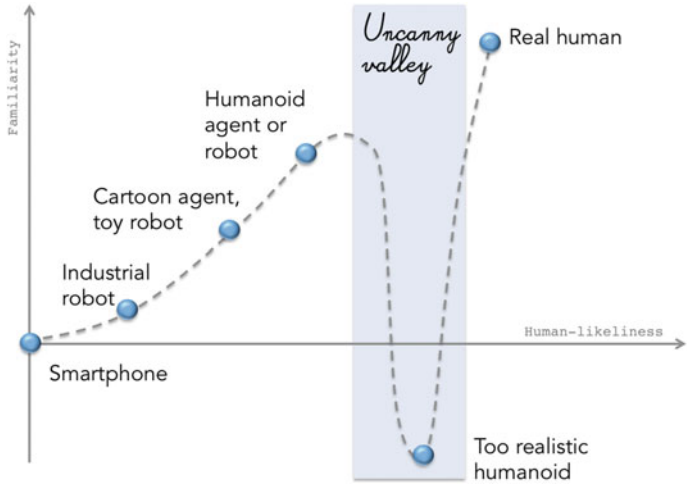

Fig. 15.7 Interpretation of the Uncanny Valley

15.5 Summary

Endowing conversational interfaces with the ability to display believable and expressive behaviors involves modeling and representing information from physiological signals, acoustic and paralinguistic features of speech and text, facial expressions, and gestures. Features of emotion can be marked up using the W3C Emotion Markup Language (EmotionML). The processes of implementing emotion recognition and synthesis include stages of data collection and annotation, learning, and optimization. There is a wide range of tools and databases available to developers who wish to incorporate emotional behaviors into conversational interfaces.

Further Reading

Schuller and Batliner (2013) give a complete survey of paralinguistics, including interesting aspects that we have not covered, such as the difference between acted versus spontaneous, felt versus perceived, intentional versus instinctual, universal versus culture-specific, and many details on emotion modeling such as type and segmentation units, features, balancing, partitioning, and laboratory versus life approaches. There is also a full chapter on corpus engineering, corpora and benchmarks, and a "hands-on" practical tutorial with openSMILE.

Petta et al.'s study (2011) is a collection of 41 chapters describing the HUMAINE project, funded by the European Commission. Calvo et al. (2014) present a comprehensive survey of affective computing, while Gratch and Marsella (2013) cover social aspects of emotion processing.

Exercises

1. **Not that easy**! Put yourself in the agent's shoes and take a test[28] of emotion recognition (if you take it in a language that you cannot speak, it will give you an even better perspective).
2. **A world of opportunities**. Go over the tools and databases that have been presented throughout the chapter and get familiarized with them, as they are open and provide instructions that will enable you to easily develop simple emotion recognizers and synthesizers.
3. **What a feeling**! Follow the demos created by Christopher Potts for the Sentiment Symposium Tutorial[29] to see how sentiment analysis works using natural language processing.

References

Allwood J, Cerrato L, Jokinen K, Naravetta C, Paggio P (2008) The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. Lang Resour Eval 41(3/4):273–287. doi:10.1007/s10579-007-9061-5

Baccianella S, Esuli A, Sebastiani F (2010) SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: International conference on language resources and evaluation (LREC2010). European Language Resources Association (ELRA), Malta, 17–23 May 2010, pp 2200–2204

Batliner A, Schuller B, Seppi D, Steidl S, Devilliers L, Vidrascu L, Vogt T, Aharonson V, Amir N (2011) The automatic recognition of emotions in speech. In: Cowie R, Pelachaud C, Petta P (eds) Emotion-oriented systems. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 71–99. doi:10.1007/978-3-642-15184-2_6

Ben Mimoun MS, Poncin I, Garnier M (2012) Case study—embodied virtual agents: an analysis on reasons for failure. J Retail Consum Serv 19(6):605–612. doi:10.1016/j.jretconser.2012.07.006

Boersma P, Weenink D (2016) Praat: doing phonetics by computer. http://www.fon.hum.uva.nl/praat/

Bos DO (2006) EEG-based emotion recognition; the influence of visual and auditory stimuli. http://hmi.ewi.utwente.nl/verslagen/capita-selecta/CS-Oude_Bos-Danny.pdf

Burkhardt F (2005) Emofilt: the simulation of emotional speech by prosody-transformation. In: Proceedings of the 9th European conference on speech communication and technology (Interspeech2005), Lisbon, Portugal, 4–8 Sept 2005, pp 509–512. http://www.isca-speech.org/archive/interspeech_2005/i05_0509.html

Cafaro A, Vilhjálmsson HH, Bickmore T, Heylen D, Pelachaud C(2014) Representing communicative functions in SAIBA with a unified function markup language. In: Bickmore T, Marsella S, Sidner C (eds) Intelligent virtual agents. Springer International Publishing, Switzerland, pp 81–94. doi:10.1007/978-3-319-09767-1_11

Callejas Z, Griol D, López-Cózar R (2011) Predicting user mental states in spoken dialogue systems. EURASIP J Adv Signal Process 1:6. doi:10.1186/1687-6180-2011-6

[28] http://www.affective-sciences.org/content/exploring-your-ec. Accessed February 27, 2016.

[29] http://sentiment.christopherpotts.net/. Accessed February 27, 2016.

Calvo RA, D'Mello S, Gratch J, Kappas A (eds) (2014) The Oxford handbook of affective computing, 1st edn. Oxford University Press, Oxford. doi:10.1093/oxfordhb/9780199942237.001.0001

Castellano G, Leite I, Pereira A, Martinho C, Paiva A, McOwan PW (2010) Affect recognition for interactive companions: challenges and design in real world scenarios. J Multimodal User Interfaces 3(1–2):89–98. doi:10.1007/s12193-009-0033-5

Clavel C, Callejas Z (2016) Sentiment analysis: from opinion mining to human-agent interaction. IEEE Trans Affect Comput 7(1):74–93. doi:10.1109/TAFFC.2015.2444846

Cowie R, Cornelius R (2003) Describing the emotional states that are expressed in speech. Speech Commun 40(1–2):5–32. doi:10.1016/S0167-6393(02)00071-7

Ekman P (1999) Basic emotions. In: Dalgleish T, Power MJ (eds) Handbook of cognition and emotion. Wiley, Chichester, pp 45–60. doi:10.1002/0470013494.ch3

Ekman P (2003) Emotions revealed: recognizing faces and feelings to improve communication and emotional life, 1st edn. Times Books, New York

Ekman P, Rosenberg EL (eds) (2005) What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system (FACS), 2nd edn. Oxford University Press, Oxford. doi:10.1093/acprof:oso/9780195179644.001.0001

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'13), Barcelona, Spain, 21–25 Oct 2013, pp 835–838. doi:10.1145/2502081.2502224

Feldman R (2013) Techniques and applications for sentiment analysis. Commun ACM 56(4):82. doi:10.1145/2436256.2436274

Govind D, Prasanna SRM (2012) Expressive speech synthesis: a review. IJST 16(2):237–260. doi:10.1007/s10772-012-9180-2

Gratch J, Marsella S (eds) (2013) Social emotions in nature and artifact. Oxford University Press, Oxford 10.1093/acprof:oso/9780195387643.001.0001

Haag A, Goronzy S, Schaich P, Williams J (2004) Emotion recognition using bio-sensors: first steps towards an automatic system. In: André E, Dybkjær L, Minker W, Heisterkamp P (eds) Affective dialogue systems. Springer Berlin Heidelberg, New York, pp 36–48. doi:10.1007/978-3-540-24842-2_4

Jang E-H, Park B-J, Kim S-H, Chung M-A, Park M-S, Sohn J-H (2014) Emotion classification based on bio-signals emotion recognition using machine learning algorithms. In: Proceedings of 2014 international conference on information science, Electronics and Electrical Engineering (ISEEE), Sapporo, Japan, 26–28 April 2014, pp 104–109. doi:10.1109/InfoSEEE.2014.6946144

Jerritta S, Murugappan M, Nagarajan R, Wan K (2011) Physiological signals based human emotion recognition: a review. In: 2011 IEEE 7th international colloquium on signal processing and its applications (CSPA), Penang, Malaysia, 4–6 March 2011, pp 410–415. doi:10.1109/CSPA.2011.5759912

Kätsyri J, Förger K, Mäkäräinen M, Takala T (2015) A review of empirical evidence on different uncanny valley hypotheses: support for perceptual mismatch as one road to the valley of eeriness. Front Psychol 6:390. doi:10.3389/fpsyg.2015.00390

Kim J, André E (2008) Emotion recognition based on physiological changes in music listening. IEEE Trans Pattern Anal 30(12):2067–2083. doi:10.1109/TPAMI.2008.26

Kim KH, Bang SW, Kim SR (2004) Emotion recognition system using short-term monitoring of physiological signals. Med Biol Eng Comput 42(3):419–427. doi:10.1007/BF02344719

Kipp M (2012) ANVIL: a universal video research tool. In: Durand J, Gut U, Kristofferson G (eds) Handbook of corpus phonology. Oxford University Press, Oxford. doi:10.1093/oxfordhb/9780199571932.013.024

Kopp S, Krenn B, Marsella S, Marshall AN, Pelachaud C, Pirker H, Thórisson KR, Vilhjálmsson H (2006) Towards a common framework for multimodal generation: the behavior markup language. In: Gratch J, Young M, Aylett R, Ballin D, Olivier P (eds) Intelligent virtual agents. Springer International Publishing, Switzerland, pp 205–217. doi:10.1007/11821830_17

Liu B (2015) Sentiment analysis: mining opinions, sentiments, and emotions. Cambridge University Press, New York. doi:10.1017/CBO9781139084789

Marsella SC, Gratch J (2009) EMA: a process model of appraisal dynamics. Cogn Syst Res 10 (1):70–90. doi:10.1016/j.cogsys.2008.03.005

Mathur MB, Reichling DB (2016) Navigating a social world with robot partners: a quantitative cartography of the Uncanny Valley. Cognition 146:22–32. doi:10.1016/j.cognition.2015.09.008

Nasoz F, Alvarez K, Lisetti CL, Finkelstein N (2003) Emotion recognition from physiological signals using wireless sensors for presence technologies. Cogn Technol Work 6(1):4–14. doi:10.1007/s10111-003-0143-x

Nass C, Lee KM (2000) Does computer-generated speech manifest personality? An experimental test of similarity-attraction. In: Proceedings of the SIGCHI conference on human factors in computing systems (CHI'00), The Hague, Netherlands, 1–6 April 2000, pp 329–336. doi:10.1145/332040.332452

Nass C, Yen C (2012) The man who lied to his laptop: what we can learn about ourselves from our machines. Penguin Group, New York

Niewiadomski R, Hyniewska SJ, Pelachaud C (2013) Computational models of expressive behaviors for a virtual agent. In: Gratch J, Marsella S (eds) Social emotions in nature and artifact. Oxford University Press, Oxford, pp 143–161. doi:10.1093/acprof:oso/9780195387643.003.0010

Paggio P, Navarretta C (2011) Head movements, facial expressions and feedback in danish first encounters interactions: a culture-specific analysis. In: Stephanidis C (ed) Universal access in human-computer interaction users diversity. Springer Berlin Heidelberg, New York, pp 583–590. doi:10.1007/978-3-642-21663-3_63

Pantic M, Bartlett MS (2007) Machine analysis of facial expressions. In: Delac K, Grgic M (eds) Face recognition. I-Tech Education and Publishing, Vienna, Austria, pp 377–416. doi:10.5772/4847

Pantic M, Valstar MF, Rademaker R, Maat L (2005) Web-based database for facial expression analysis. In: IEEE International conference on multimedia and expo (ICME), Amsterdam, The Netherlands, 6–8 July 2005, pp 317–321. doi:10.1109/ICME.2005.1521424

Pelachaud C (2009) Modelling multimodal expression of emotion in a virtual agent. Philos Trans R Soc B Biol Sci 364(1535):3539–3548. doi:10.1098/rstb.2009.0186

Petta P, Pelachaud C, Cowie R (eds) (2011) Emotion-oriented systems: the Humaine handbook. Springer, Berlin Heidelberg. doi:10.1007/978-3-642-15184-2

Picard RW, Vyzas E, Healey J (2001) Toward machine emotional intelligence: analysis of affective physiological state. IEEE Trans Pattern Anal 23(10):1175–1191. doi:10.1109/34.954607

Polzin TS, Waibel A (2000) Emotion-sensitive human-computer interfaces. In: International speech communication association (ISCA) tutorial and research workshop on speech and emotion. Newcastle, Northern Ireland, UK, pp 201–206

Ryan A, Cohn JF, Lucey S, Saragih J, Lucey P, De La Torre F, Rossi A (2009) Automated facial expression recognition system. In: 43rd annual international Carnahan conference on security technology, Zurich, Switzerland, 5–8 Oct 2009, pp 172–177. doi:10.1109/CCST.2009.5335546

Schröder M (2009) Expressive speech synthesis: past, present, and possible futures. In: Tao J, Tan T (eds) Affective information processing. Springer, London, pp 111–126. doi:10.1007/978-1-84800-306-4_7

Schuller B, Batliner A (2013) Computational paralinguistics: emotion, affect and personality in speech and language processing. Wiley, Chichester, UK. doi:10.1002/9781118706664

Van Santen J, Mishra T, Klabbers E (2008) Prosodic processing. In: Benesty J, Sondhi MM, Huang Y (eds) Springer handbook of speech processing. Springer, Berlin Heidelberg, pp 471–488. doi:10.1007/978-3-540-49127-9_23

Väyrynen E (2014) Emotion recognition from speech using prosodic features. Doctoral Dissertation, University of Oulu, Finland. http://urn.fi/urn:isbn:9789526204048

Ververidis D, Kotropoulos C (2006) Emotional speech recognition: resources, features and methods. Speech Commun 48(9):1162–1181. doi:10.1016/j.specom.2006.04.003

Wagner J, Lingenfelser F, Baur T, Damian I, Kistler F, André E (2013) The social signal interpretation (SSI) framework: multimodal signal processing and recognition in real-time. In: Proceedings of the 21st ACM international conference on Multimedia (MM'13), Barcelona, Spain, 21–25 Oct 2013, pp 831–834 doi:10.1145/2502081.2502223

Zhang C, Zhang Z (2010) A survey of recent advances in face detection. Microsoft TechReport MSR-TR-2010-66. http://research.microsoft.com/apps/pubs/default.aspx?id=132077