Chapter 1

Theories of Emotion

In this chapter we are going to analyse the most important theories of emotion proposed in the scientific literature. The aim of this chapter is to obtain a good understanding of these theories in order to be able to make a sensible choice about which theory of emotion will constitute the theoretical background of our work.

1.1 Definition of Theories of Emotion

Theories of emotion are theories, developed primarily in the field of psychology, which aims at explaining emotions and their nature. Rigorously, the first theories of emotion were developed around the half of the nineteenth century, when the term and the concept of emotion made their appearance in the discussions of psychologists and philosophers; before it, speaking of theories of emotion should be considered an anachronism, even if, undoubtedly, many philosophers and thinkers elaborated theories to explain human passions and affections [28].

A theory of emotion is a general set of principles and rules which aims at making sense and explaining emotions and their dynamics; these theories are usually developed by psychologists relying on observations, evidences and experiments. As every scientific theory, theories of emotion are supposed to have explicative power (i.e., being able to explain emotions and provide a comprehensive understanding of them) and predictive power (i.e., being able to make predictions about future events). It should be underlined that theories of emotions proposed in the psychological field are to be conceived, most of the time, as working hypothesis which have a sound theoretical and conceptual foundation, but which were rarely tested against extensive empirical evidence as it happens in the field of affective computing [20].

Any study of emotions, including the present one, has a deep interest in understanding and mastering these theories as they can offer a reasonable and coherent framework for research; these theories provide the concepts and the theoretical tools which will constitute the foundations of research.

1.2 Taxonomy of Theories of Emotion

Literature from the field of psychology abounds with several different theories of emotions. Theories of emotion may be significantly different from each other because they rely on different conceptions of what emotions are or because they focus on diverse aspects of emotions. Evaluating these theories and confronting them is a long and difficult task as some of them may be incomparable with others. We want to devise a systematic way to present these theories in order to be able to make a conscious and rational decision on which theory of emotions to adopt for our research. We will then draw some lines which will help us put these theories in place and help us orient among all the options.

Here we suggest a way in which we can order theories of emotion.

First of all, theories of emotion may be classified according to the different **aspects of emotions** they focus on. Theories of emotion may be concerned with:

- *Anthropological origin of emotions*: how in the course of the evolution human subjects came to experience emotions, which role did the emotions play in the development of the human race, which impact did emotions have on the survival;
- *Actual development of emotions*: how human subjects experience emotions in their everyday life, what elicits emotions, how emotions come to be;
- *Descriptive characterization and categorization of emotions*: how emotions can be described, what makes emotions different from each others, what are the common and peculiar traits of each emotion;
- *Behavioural and cognitive effects of emotions:* how emotions impact on the behaviour of human subjects, how emotions affect the cognitive processes of human subjects, what are the consequences of emotions.

Theories of emotion may be mainly focused on a single one of these areas of interest (e.g., *adaptive theories* which are mainly interested in the anthropological origin of emotions [22, 11]) or they may span many areas of interest (e.g., the *appraisal theories* which mainly consider the actual development and the descriptive characterization of emotions [105]).

Second, when dealing with concrete emotions, theories of emotion may be classified according to the different **components of emotions** they focus on. Theories of emotion may be concerned with [**?**]:

- *Physiological response*: which inner physiological responses, such as increased heartbeat or change in body temperature, are caused or correlated with specific emotions;
- *Motor response*: which outer motor responses, such as vocal expression or body posture, are caused or correlated with specific emotions;
- *Subjective feeling*: which personal subjective feeling, such as pleasure or arousal, are caused or correlated with specific emotions;

- *Behaviour preparation*: which behaviours or action tendencies, such as fight or flight, are caused or correlated with specific emotions;
- *Cognitive processes*: which cognitive processes, such as attention or judgement, are influenced by specific emotions.

Again, theories of emotion may be restricted mainly to a single components (e.g., the *dimensional theories* which limit themselves to subjective feelings [**?**, 78]) or they may try to embrace many components (e.g., the *appraisal theories* which try to take into consideration all the components listed above [**?**, 105]).

Third, when considering the emotion as a process in time, theories of emotion may be classified according to the different **phases of emotions** they focus on. Emotions theories may be implied in [**?**]:

- *Low-level evaluation processes*: automatic and basic evaluation processes taking place at a sub-cortical level;
- *High-level evaluation processes*: controlled and complex evaluation processes taking place at a cortical level;
- *Evaluation of goals and needs*: evaluation processes taking into account the hierarchy of goals and the set of constraints defined by needs;
- *Generation of alternatives*: process generating different possible alternatives of action and evaluating their likelihood of success;
- *Behaviour preparation*: process implementing the actions tendencies and making the body ready to act;
- *Behaviour execution*: process carrying out the planned actions;
- *Communication*: process relating an emotional experience to other human subjects in a social context through language.

Once again, theories of emotion may be restricted to a single phase (e.g.: the *constructivist theories* which are restricted mainly to the communication phase [**?**]) or extend to more phases (e.g.: the *dimensional theories* which stretch from the low-level activation to the high-level activation [**?**, 78]).

These dimensions (aspects of emotions, components of emotions and phases of emotions) constitute a framework which allow us to classify theories of emotion and evaluate if a given theory may be useful for a given task.

In the next section we list and briefly describe the main theories of emotion.

1.3 Main Theories of Emotion

The literature from the field of psychology and affective computing contains several theories of emotion [11, ?, 66, 60]. Here we list the most important theories:

Adaptive Theories Adaptive theories date back to Darwin [22]; they describe emotions as the result of evolutionary adaptive processes which maximized the biological fitness of an organism in its environment [22]. Adaptive theories study the anthropological origin of emotions, examine mainly the physiological and cognitive components of emotions and focus on the low-level and high-level evaluation phases of emotions.

Embodiment Theories Embodiment theories were first proposed by James [50]; turning upside-down the traditional conception that emotions cause physiological changes, embodiment theories consider emotions as the sum of the physiological changes inside the body [50]. Embodiment theories study the actual development of emotion and their characterization, examine mainly the physiological and motor components of emotions and focus on the low-level evaluation phase.

Constructivist Theories Constructivist theories assert that emotions are relative sociocultural artefacts shaped by social and cultural processes such as social coordination, self-regulation, power and status relationships; emotions exist only within a certain social context, within a given framework of values or within a language [62, 55, 61]. Constructivist theories study the anthropological origin of emotions and their actual development, examine mainly the motor, feeling and behavioural components of emotions and focus on the communication phase.

Appraisal Theories *or* Cognitivist Theories Appraisal theories maintain that emotions are the result of a process of cognitive evaluation (i.e., *appraisal*) of an event; an appraisal is usually conceived as an unconscious process during which an event is evaluated in relation to different variables, such as pleasantness, relevance for personal goals or novelty [34]. Appraisal theories study the actual development of emotions and their characterization, try to examine all the components of an emotion and focus on high-level evaluation phase.

Neuroscientific Theories Neuroscientific theories aim at explaining emotions by finding those brain circuits or patterns which are responsible for the emotions experienced by human subjects; by grounding the emotions in their neural substrates, neuroscientific theories explore how emotions are generated and how they interact [58]. Neuroscientific theories study the actual development of emotions, examine the physiological components of emotions and focus on the low-level evaluation phase.

Motivational Theories Motivational theories describe emotions in relation to the goals of a human subject; they explain the behaviour determined or induced by emotions in relation to the goals of a human subject [69]. Motivational theories study the characterization of emotions and their behavioural and cognitive effects, examine the physiological, motor, feeling and behavioural components of emotions and focus on the goal setting and

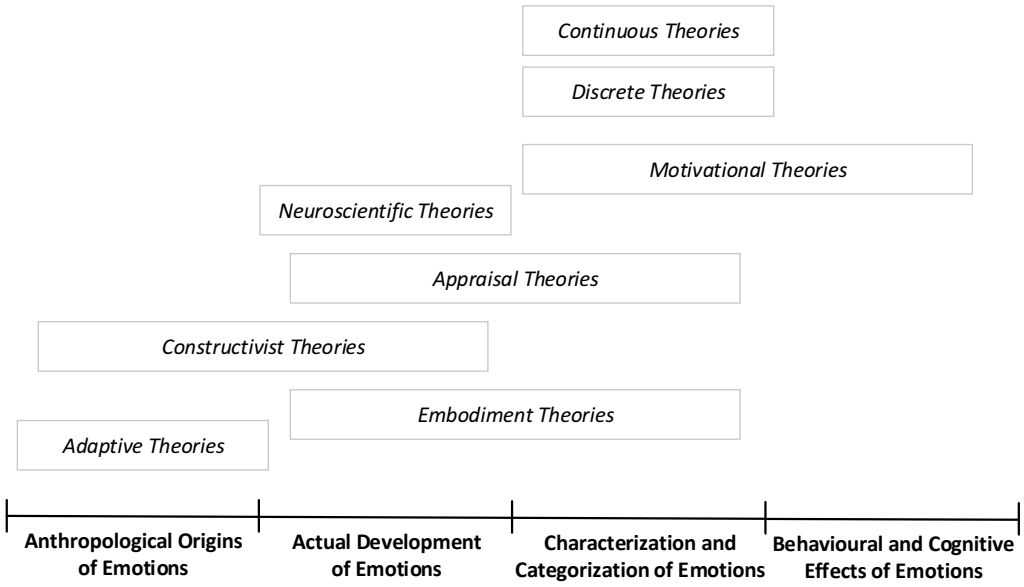

Figure 1.1: Theories of emotion classified according to the aspects of the emotions they consider

generation of alternatives.

Discrete Theories Discrete theories consider any emotion as the specific instance of few universal emotions; discrete theories usually define a finite and limited dictionary of basic emotions, such that every emotion can be reduced to one of these basic emotions or to a combination of them [31, 32]. Discrete theories study the characterization of emotions, examine physiological, motor, feeling and behavioural components of emotions and focus on the behaviour preparation and behaviour execution phase.

Continuous Theories *or* Dimensional Theories Continuous theories look for a set of continuous dimensions which allow to explain human emotions; emotions are then visualized as points in the $n-$dimensional space defined by the chosen dimensions [111, 76]. Continuous theories study the characterization of emotions, examine feeling components of emotions and focus on the low-level and high-level evaluation phase.

Note that each element of this list does not correspond to a single specific theory of emotion, but to a family of theories of emotion; each element denotes a set of theories of emotion with similar underlying assumptions and ideas; within the same family of theories of emotion, different researchers may propose specific theories differing from each other in their details. Moreover, sometimes these theories can be combined together in the attempt of defining more general and complete theories of emotion (e.g., Ekman's *neuro-cultural theory of emotion* [52] or Russell's *theory of core affect* [78])

Figure 1.1 visually summarizes how the different theories listed above can be classified according to the aspects of emotions they consider, while figure 1.2, modelled after [**?**], summarizes them according to the components and the phases of the emotions they consider.

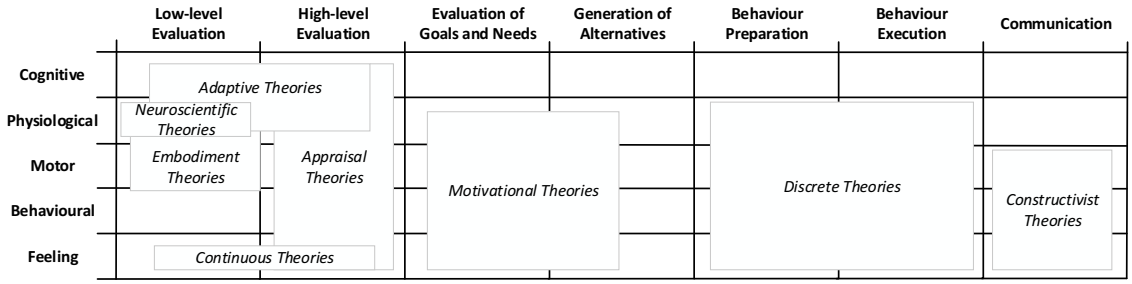

Figure 1.2: Theories of emotion classified according to the components and the phases of emotions they consider (adapted from [?])

1.4 Computational Theories of Emotion

Several theories of emotion were listed above; however, now, we would like to restrict our attention to *computational theories of emotion*. We define computational theories of emotion those which are rigorous enough or which can be made rigorous enough to be used to develop models which can be implemented on a computer. Computational theories of emotion are formalized and quantitative theories which can be executed or simulated on a computer to study or predict emotions. By restricting our attention only to computational theories of emotion we aim at narrowing down the range of alternatives we are given to choose a theory of emotion backing up our work.

Literature in the field of affective computing, artificial intelligence and human-computer interaction discuss possible computational theories of emotion [66, 60]. Three theories are particularly suitable to be implemented on a computer: discrete theories, continuous theories and appraisal theories.

In the following sections we give a detailed explanation of the most important computational theories of emotion.

1.4.1 Discrete Theories of Emotion

Discrete theories of emotion postulate the existence of a small finite number of basic emotions. Each basic emotion corresponds to an emotion category. Human emotions are expected to fall into one of these main basic categories or to be explained as a combination of other basic emotions, in analogy with colours (*palette theory* [18]). The idea of discrete basic emotions is easy and intuitive, rooted in the everyday use of terms denoting emotions. Human languages possess a wide array of nouns and terms to describe emotions; discrete theories of emotion posit that some of them have a very general and universal validity.

Many supporters of the discrete theories of emotion consider basic emotions universal emotion patterns; studies have been made to prove that basic emotions are patterns which are independent of human society and culture and can be virtually found at any latitude [32].

While these premises are shared by most of the proponents of these theories, much disagreement exist about which emotions should be deemed basic emotions. Clearly, human languages

provide a plethora of emotion-related terms; in [1], Averill collected 558 English word related to emotions; other researchers, aiming at creating comprehensive list of emotions, complied lists usually containing between one hundred and two hundred words [18]. Potentially, for each emotion-related term, we could instantiate a category of emotions. However such a high number of categories raises challenges from both theoretical and practicals point of view. From a theoretical point of view we want to reduce the complexity generated by the abundance of terms related to emotions; we want our theory to be simple and elegant, that is to be able to explain the high variability of emotions starting from few elements; therefore we want to filter all these terms to identify the smallest set of basic emotions. From a practical point of view working with hundreds of categories is, at the time being, infeasible; many emotional terms are different from each other only by little nuances; discriminating between so many terms would be a very hard task for machines.

Reducing the set of emotion-related terms we find in natural language to discover a finite set of basic emotions is a controversial task. Many variables influence the outcome of this task, such as the number of primitive emotions that should be identified or the very definition of what emotions are (see, for example, the discussion whether love should be considered an emotion, a basic emotion or something different from an emotion, such as an emotional plot). One of the first list of basic emotions was proposed by Ekman in 1969 [31, 16]; he listed six basic emotions (*fear*, *anger*, *happiness*, *sadness*, *surprise* and *disgust*) which became known as "Big Six". This list was meant to define the minimal set of basic emotions identified through cross-cultural studies. However, many alternative lists of basic emotions have been put forward since 1969; some researchers, including Ekman himself, provided extended lists by inserting additional basic emotions to the original "Big Six"; other researchers carried on their own studies and generated new independent lists of basic emotions. Cowie in [18] provide a synoptic table of the main sets of basic emotion proposed in literature; we reproduce that table with a minor addition in Table 1.1.

Discrete theories of emotion have advantages and disadvantages.

From a practical point of view, discrete theories are very easy to understand and to interpret; they use words and concepts borrowed from everyday language; this makes discrete theories very suited for those applications which interact with users who do not know anything about theories of emotion. The output of a human-computer interface which uses the categories defined by a discrete theory of emotion can be easily and readily understood by any user.

Moreover, given their easy interpretability, the categories offered by discrete theories of emotion are often adopted to label data collected during experiments. Using intuitive categories, as the ones listed in Table 1.1, allows expert and non-expert alike to process and annotate data.

However, discrete theories of emotion have been the subject of strong criticism, too. The conceptual assumption of using discrete categories to classify emotions has been attacked by different researchers: emotion labels may be only coarse human artefacts with very tenuous relation with actual emotions [15]; they may be the non-scientific product of a folk psychology [66].

Ekman (1969) "Big Six"	**Ekman (1999)**	**Lazarus (1999)**	**Buck (1999)**	**Lewis and Havilland (1993)**	**Banse and Scherer (1996)**	**Cowie (1999)**
Anger	Anger	Anger	Anger	Anger / Hostility	Rage / Hot Anger	Angry
					Irritation / Cold Anger	
Fear	Fear	Fright	Fear	Fear	Fear / Terror	Afraid
Sadness	Sadness / Distress	Sadness	Sadness	Sadness	Sadness / Dejection	Sad
					Grief / Desperation	
		Anxiety	Anxiety	Anxiety	Worry / Anxiety	Worried
Happiness	Sensory pleasure	Happiness	Happiness	Happiness	Happiness	Happy
					Elation / Joy	
	Amusement			Humour		Amused
	Satisfaction					Pleased
	Contentment					Content
			Interested			Interested
			Curious			
Surprise			Surprised			
	Excitement					Excited
			Bored		Boredom / Indifference	Bored
						Relaxed
			Burn out			
Disgust	Disgust	Disgust	Disgust	Disgust	Disgust	
	Contempt		Scorn			
	Pride	Pride	Pride	Pride		
			Arrogance			
		Jealousy	Jealousy			
		Envy	Envy			
	Shame	Shame	Shame	Shame	Shame / Guilt	
	Guilt	Guilt	Guilt	Guilt		
	Embarassment			Embarassment		
						Disappointed
	Relief	Relief				
		Hope				
						Confident
		Gratitude				
		Love		Love		Loving
						Affectionate
		Compassion	Pity			
			Moral rapture			
			Moral indignation			
		Aesthetic				

Table 1.1: Set of basic emotions (adapted from [18])

The idea of the universality of basic emotions has also been the topic of intense discussions. Psychologists supporting constructivist positions argued that emotions are social constructs and they brought forward examples of emotions changing their meaning and value according to space (e.g., the emotional concepts of *metagu* or *fago* among the Ifaluk studied by Lutz [61, 75]) or time (e.g., the emotional concept of *acedia* undergoing a semantic shift since the Middle Ages [28]). Hence, they doubt that a true universal set of basic emotions valid for all human beings could ever be defined.

Discrete theories of emotion were often accused of offering a too simplistic categorization of emotions and of overlooking important information related to emotions; for example, the information regarding the intensity of an emotion, which is considered by many an essential part of an emotion, is usually overlooked; analogous emotions, with very different intensity, are put in the same category with a loss of relevant information[15].

The widespread disagreement on a definite set of basic emotions is often seen as a weakness of discrete theories. The impossibility of finding a criterion to rigorously define which emotions can be considered basic emotions highlights a practical limit of these theories. There is no agreed principle to determine the number of basic emotions as two different needs collide: on one side, we want this number to be small enough to make the theories simple and to make their computation easy; on the other side we want this number to be big enough to take into consideration the vast amount of different human emotions. At the end the choice is left to the sensibility of the psychologists or the computer scientists. This has led to a proliferation of alternative lists of basic emotions, generated both at a theoretical level or at a practical level; at the theoretical level, several psychological studies were made, each one generating a new ideal sets of basic emotions as shown in Table 1.1; at the practical level, the lists produced by the theoretical studies were often edited and customized by other psychologists or computer scientists to tailor the need of specific applications.

1.4.2 Continuous Theories of Emotion

Continuous theories of emotion postulate that human emotions can be studied along continuous dimensions representing fundamental properties and qualities of all human emotions. Human emotions can be visualized as points in the $n-$dimensional emotional space defined by the chosen dimensions.

Psychologists supporting continuous theories assert that emotional experience is a continuous-time process; the emotional state of a human subject can be represented as a point moving in time in the $n-$dimensional emotional space; therefore, at every instant, a human subject is experiencing an emotion and can be located in the $n-$dimensional emotional space. Continuous theories embrace a very wide family of emotions, not only full-blown emotions, but also those states which are often labelled as moods or feelings. Continuous $n-$dimensional emotional space generally includes an area for a neutral state (which can be imagined as the state in which a human subject is not experiencing any specific emotion or feeling) which corresponds, at least, to the geometric origin of the $n-$dimensional emotional space.

Several *dimensions* (or *primitives* [42, 44]) can be taken into consideration to describe emotions. Two dimensions which are virtually ubiquitous [71] are:

- *Valence:* describing if the human subject experiencing the emotion feels it as positive and pleasurable or as negative and unpleasant. According to the particular theory and the individual scholar, this dimension can be named as *hedonic tone*, *pleasure/displeasure* [77], *appraisal* [71], *evaluation*, *liking*, *positive/negative*, *approach/avoidance*, *utility* [78];
- *Intensity:* describing if the human subject experiencing the emotion feels it as strong and intense or as weak and mild. According to the particular theory and the individual scholar, this dimension can be named as *arousal, arousal/sleep* [77], *excitement/calm, excitement/quiet* [18] *action readiness* [38], *disposition to action* [7].

These dimensions are backed by a broad consent in the psychological community and are adopted in many theories and models [78]; even if these two dimensions can not fully describe an emotion, but approximate it, they are considered established tools in affective computing [19].

However, side by side with these two main dimensions, other dimensions were suggested:

- *Dominance*: describing if the human subject experiencing the emotion feels in control of what is happening or if he feels no control on what is happening. This dimension is sometimes named *power/control*.
- Expectation: describing if the human subject experiencing the emotion is anticipating the emotion or if he is taken by surprise [44].
- *Positive affect*: describing if the human subject experiencing the emotion feels active and full of energies [102, 103, 97].
- *Negative affect*: describing if the human subject experiencing the emotion feels in an aversive and unpleasurable state [102, 103, 97].
- *Self-control*: describing if the human subject experiencing the emotion is able to maintain his self-control or if he loses it and starts panicking [27].
- *Engagement*: describing if the human subject experiencing the emotion feels involved in it or if he does not feel any involvement [61, 55].
- *Approach/Withdrawal*: describing if the human subject experiencing the emotion feels attracted by what is happening or if he feels repulsed by what is happening [23].
- *Stance*: describing if the human subject experiencing the emotion has an open and outgoing attitude or a close and inward attitude towards what is happening [101].
- *Difficulty of overcoming*: describing if the human subject experiencing the emotion feels it as something easy to overcome or as something difficult to overcome [75].
- *Transience*: describing if the human subject experiencing the emotion feels it as something transient or as something enduring [18].
- *Time direction*: describing if the human subject experiencing the emotion feels it as something directed towards the past or towards the future [56]. This dimension is sometimes named *retrospective/prospective* [28].

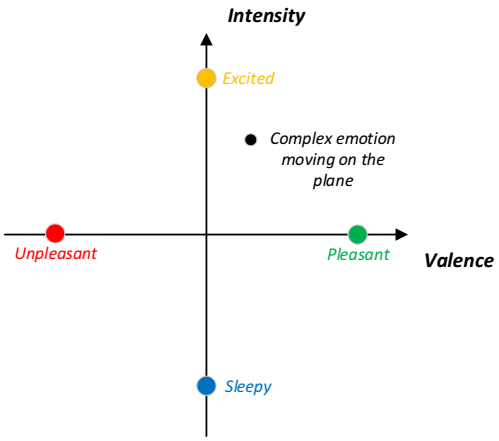

Figure 1.3: Valence×Intensity Emotional Space

These dimensions can be combined to give rise to emotional spaces which can be used to represent and to track the emotional state of a user. The emotional space which are most used are:

Valence×Intensity Space it is the emotional space made up by the two orthogonal dimensions of *valence* and *intensity*. As already mentioned, this space is widely used by psychologists to the point that Cowie stated that it "can be thought of as a minimal representation of themes whose centrality to emotion is not in doubt" [18]. This space is often referred to simply as *two-dimensional space*. See figure 1.3. Other names suggested for this space are *cold-hot* or *meaning-intensity* or *rational-emotive,* where the judgement on the pleasantness of an emotion is considered as a cold, rational, meaning-related process, while the judgement on the arousal of an emotion is considered as a hot, emotive, intensity-related process [65].

PAD (Pleasure×Arousal×Dominance) it is the emotional space given by the previous space (valence and intensity are simply renamed as *pleasure* and *arousal*) with the addition of a third orthogonal dimension, *dominance*. The dimension of dominance is usually added in order to increase the discriminative power of the emotional space; in particular dominance has been introduced to make it easier to distinguish between fear and anger; fear and anger have very similar values for pleasure and arousal, but the first one is considered having a strongly negative value of dominance, while the second one is considered having a strongly positive value of dominance [80]. Analogously to the previous one, this space is often referred to simply as *three-dimensional space*. See figure 1.4.

PA-NA (Positive Affect×Negative Affect) it is the emotional space made up by the two orthogonal dimensions of positive affect and negative affect [103, 97, 114]. The PA-NA emotional space can be identified with the valence-intensity emotional space through a 45° degree rotation of the axes. See figure 1.5.

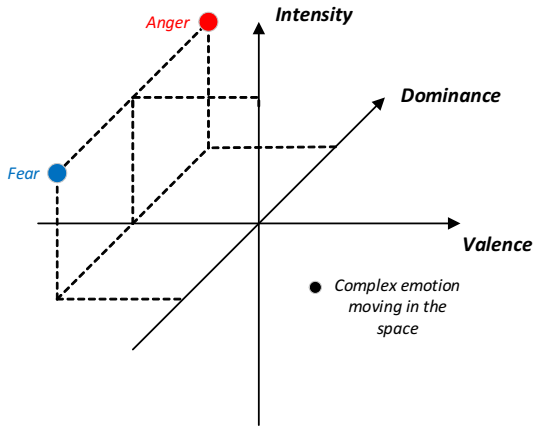

Figure 1.4: PAD Emotional Space

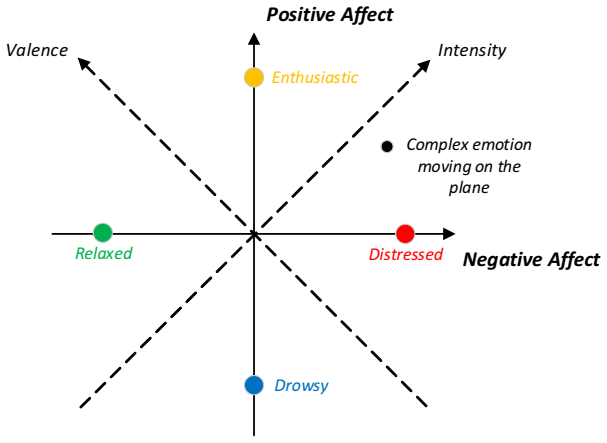

Figure 1.5: PA-NA Emotional Space

Continuous theories of emotion have advantages and disadvantages.

Continuous theories of emotion have a high degree of versatility as they allow the description of an infinite number of emotions; theoretically, every emotion experienced by a human subject can be given a position in the emotional space defined by the chosen dimensions; as long as it is possible to characterize an emotion defining the quantitative values it has along the chosen dimensions, then it is possible to locate it in the emotional space.

Continuous theories of emotion allow the researchers to consider not only full-blown emotions, as it is usually the case in discrete theories of emotion, but also states, which, because of their moderate value of intensity or valence, fail to meet the standards of full-blown emotions and are often labelled as moods or affections. This states, which may be overlooked in other theories, can be coherently dealt with in continuous theories of emotion.

Continuous theories of emotion allow a dynamic analysis of emotions, too. Instead of categorizing the emotional state of a user only at fixed time intervals or only when a specific full-blown emotion is displayed, continuous theories of emotion can track the emotional shifts of a user in time and track his emotional evolution.

This theoretical advantages are supported by the broad implementation and use of continuous theories to study emotions [78]. These theories proved to be very successful in social psychological and neurophysiological studies of emotions [**?**].

However, continuous theories of emotion have undeniable shortcomings, too. One of the most frequent criticism, is that continuous theories of emotion lack an implicit linguistic description [13]. This means that continuous theories of emotion provide by default a tuple of coordinates in a $n-$dimensional emotional space, but not a readable label; differently from the discrete theories of emotion, they do not implicitly suggest a categorization of emotions in classes which can be easily understood by a human user. Some critics suggested that this may account as a failure of continuous theories of emotion to model or distinguish human emotions [78]. A solution to this flaw is to define a mapping from subspaces of the emotional space to emotional labels. Russell mapped 28 emotional terms into the intensity-valence emotional space [76]. He showed how the 28 emotional terms, corresponding to full-blown emotions, tend to arrange themselves in a circular shape far removed from the origin; this arrangement is indeed coherent with the intuition that full-blown emotions corresponds to those emotional states having high value of valence or intensity; this pattern was called *two-dimensional circumplex* (see figure 1.6). Whissel extended the work done by Russell computing the values of valence and intensity for longer lists of emotional states [106]. Plutchik defined emotions by their angular coordinate in the two-dimensional circumplex and named this pattern *emotion wheel* [73]. See [17] for a comprehensive list of emotions and their values of valence, intensity and angular coordinate.

Some psychologists are critical about the lack explicative power, too. While continuous theories of emotion can be useful to map emotions in a $n-$dimensional space, they do not offer any explanation on how the different dimensions of the space interact to determine the final emotion [71]. Continuous theories provide a tuple of values specifying the position of an emotion in space, but they do not explain how these values contribute to shape an emotion.

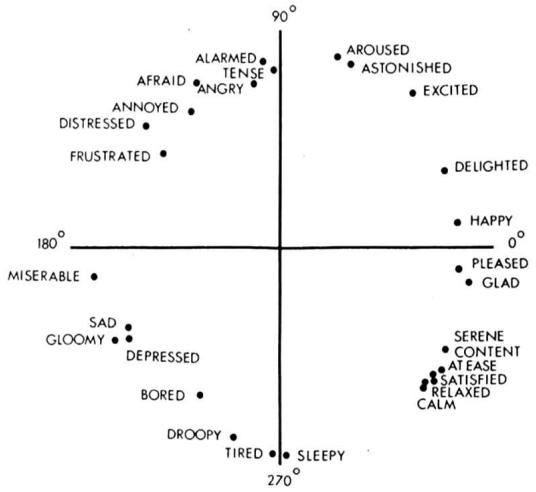

Figure 1.6: Two-dimensional Circumplex (taken from [76])

Other psychologists see the wide array of emotional states taken into consideration by continuous theories of emotion as a weakness. Continuous theories of emotion may be accused of dealing more with moods and affects than with emotions; this can therefore be considered a dangerous loss of focus. This criticism is clearly connected with the problem of the definition of what emotions are and whether moods and affects should be part of the study of emotions.

Practically, one of the main problem of adopting and implementing continuous theories of emotion is that it heavily relies on self-report [78]. This mean that often, during psychological experiments, the only way to assess quantitatively the value of valence or intensity of an emotion is by relying on the subjective judgement of the person experiencing the emotion.

Connected to this issue, is the fact that assessment of an emotion by external observers using abstract dimensions like the ones we described in this chapter is more difficult than tagging an emotion using discrete classes; this means that labelling emotional data requires annotators with at least a minimal training and understanding of the theory. To address this problem, Cowie et al. developed a software tool called *Feeltrace* to help researchers tracking emotions in the intensity-valence space; *Feeltrace* aims at making as intuitive as possible the navigation in the emotional space by returning visual cues which can guide the user around the valence-intensity space [21]. Another solution frequently employed is to discretise the values of intensity and valence that an emotional state at a given instant can have; few values having a clear and agreed meaning are selected; every annotator can then assign one of these fixed values; at the end, the final continuous values of intensity and valence are given by the average of all the values assigned by the annotators [80].

Finally, continuous theories of emotion can be subject to a criticism about the choice of dimensions. As in the case of discrete theories of emotion it is hard to determine which emotions are basic emotions, so in the case of continuous theories of emotion it is hard to choose which dimensions should be used to study emotions. As we wrote, there is a general agreement on the use of the first two dimensions, valence and intensity. But, beside them, other dimensions can be added. We described, for example, how the dimension of dominance was introduced

to allow the discrimination between fear and anger; however, from a theoretical point of view, the problem with this practice is that, as soon as we start introducing dimensions just to discriminate two specific emotions, then we can keep adding more and more dimensions to solve specific problems of discrimination [18]. Another problem with defining new dimensions is that they can be ambiguous [78]. As the varied nomenclature suggests, it is hard to define rigorously new dimensions which are orthogonal to the pre-existing dimensions. Just in the case of valence, we listed 8 possible synonyms (*hedonic tone*, *pleasure/displeasure*, *appraisal*, *evaluation*, *liking*, *positive/negative*, *approach/avoidance*, *utility*); these concepts may not be identical with valence, and there may be subtle nuances between these terms; however they do not define new dimensions and thus can be reduced to the best known dimension of valence.

1.4.3 Appraisal Theories of Emotion

Appraisal theories of emotion postulate that cognition and emotion are strictly interrelated; an emotion is, therefore, the result of a rational, though unconscious and often unexpressed, process evaluating a given situation. An appraisal is a prompt and intuitive judgement of an event, which allow a human subject to quickly grasp what is happening and, in case of need, to take a fast decision.

Appraisal is the process that determines an emotional experience. According to the specific theory considered, appraisal may precede and cause an emotion or it can be itself part of what we call an emotion. The complex process of evaluating events and circumstances can be subdivided in a collection of simpler judgements, each one considering only a particular feature of the event. These features are usually called *appraisal variables* or situational meaning structures [66]. The most important appraisal variables, common to several appraisal theories of emotion, are [34, 60]:

- *Novelty*: evaluating if the event taking place is expected or unexpected; the concept of novelty is also related with the concepts of *familiarity*, *expectedness*, *predictability* and *probability* of an event;
- *Pleasantness*: evaluating if the event taking place is pleasant or unpleasant; the concept of pleasantness is also related with the concept of *valence*, *complexity, desirability, liking* and *aesthetic* of an event;
- *Goals*: evaluating how the event taking place relates with personal goals, and whether it furthers or it hinders personal objectives; the concept of goals is also related with the concepts of *needs*, *values*, *motive consistency*, *importance*, *perceived obstacles*, *concern relevance*, *goal conduciveness*, *goal hindrance*, *motivational relevance*, *pertinence*, *urgency*, *immorality*, *self-consistency*, *probability* and *certainty*;
- *Agency*: evaluating who or what is the cause of the event taking place, and whether the event is under control or not; the concept of agency is also related with the concepts of *responsibility*, *causation*, *controllability*, *coping ability* and *direction of the event.*
- *Norms*: evaluating how the event taking place relates with social norms, and whether it complies or it violates cultural standards; the concept of norms is also related with the

	Frijda (1986)	**Roseman (1984)**	**Scherer (1984)**	**Smith and Ellsworth (1985)**
	Change		Novelty	Attentional Activity
			Suddenness	
Novelty	Familiarity		Familiarity	
Valence	Valence		Intrinsic Pleasantness	Pleasantness
	Focality	Appetitive/ Aversative Motives	Goal Significance	Importance
			Concern Relevance	
Goals	Certainty	Certainty	Outcome Probability	Certainty
Agency	Intent/ Self-Other	Agency	Cause: Agent	Human Agency
			Cause: Motive	
Norms	Value Relevance		Compatibility with Standard	Legitimacy
			External Compatibility	
			Internal Compatibility	

Table 1.2: Major appraisal variables (adapted from [34])

concepts *legitimacy*, *value relevance*, *compatibility with external standards*, *deservedness*, *praiseworthiness* and *justice*.

As in the case of dimensions in continuous theories of emotion, further appraisal variables can be conceived and combined with the existing ones; in table 1.2 we reproduce the comparative overview of major appraisal variables proposed by Ellsworth and Scherer in [34].

An event is then evaluated in a set of appraisal variables; practically, the appraisal process can be thought as a process in which a quantitative, continuous value is assigned to each appraisal variable. The final combination of the values assigned to each appraisal variable determines the emotion experienced by a human subject.

Appraisal theories of emotion are plainly different from discrete theories of emotion. Discrete theories of emotion works only with a limited set of emotion; appraisal theories of emotion defines a finite set of appraisal variables, but, since each appraisal variable can assume a continuous value, these theories deal with a potentially infinite number of emotions.

On the other hand, at first sight, appraisal theories of emotion and continuous theories of emotion seems to have relevant similarities. A set of n appraisal variables can be conceived as a

set of dimensions defining a $n-$dimensional space; then, an appraisal process identifies a point in this $n-$dimensional space. Moreover several appraisal variables and continuous dimensions can be easily mapped to each other, the case of the valence dimension and the pleasantness variable being probably the most evident. So, both theories deal with objects in a continuous space defined by dimensions having strong resemblances. However, despite these analogies, there are also deep conceptual differences between the two theories. First of all, an appraisal variable refers to the evaluation of an event, while continuous dimensions refer to the description of an emotion; in other words, the pleasantness variable refers to the pleasantness of an event, while the valence dimension refers to the feeling of pleasantness of an emotion. Second, an appraisal and the related emotion take place only when triggered by an event, while in continuous theories of emotion a human subject is always in an emotional state. Third, an appraisal is a always a relational construct, which means that it is the result of the compared evaluations of the event happening and the values and expectations of a human subjects; on the other hand, emotional states in continuous theories of emotion may be non-relational construct, that is description of the general affective state of a human subject independently of external events or inner values. Fourth, many appraisal may take place inside a human subject at the same time, potentially determining mixed or contrasting emotions experienced at the same time; differently, the emotional state of a human subject in continuous theories of emotion is always unique, determined by a single point in space. Last, appraisal theories of emotion put a strong emphasis on the dynamics of emotion generation, studying how the subject-environment relationship and how cognitive judgements contribute to the development of an emotion; continuous theories of emotion, while recognizing the influence of the environment and of cognition, focus mainly on the description of the affective state of a human subject [66].

Appraisal theories of emotion have advantages and disadvantages.

Like continuous theories of emotion, appraisal theories allow the dynamical analysis of a potentially infinite number of emotional states; they allow the researchers to study emotions in time and to focus not only on full-blown emotions but on an unlimited range of moods and affects.

Appraisal theories of emotion too enjoy a broad support from the community of psychology and affective computing. This theories are often implemented in artificial intelligence and within autonomous agents which are expected to show cognitive and emotive skills. Appraisal theories of emotion provide indeed an excellent framework for combining emotion and cognition; the very idea of appraisal links emotions to rational evaluation processes which can be readily implemented and simulated; moreover, as appraisals are started from external events, it is natural to model appraisals as cognitive processes triggered by events taking place in the environment [66, 60].

Finally, differently from continuous theories of emotion, appraisal theories have a strong explicative power. Appraisal theories were developed in order to explain how emotions arise in human subjects; the decomposition of an emotive process in a sequence of cognitive appraisal processes is meant to explain how an emotion arises in front of a given event.

Concerning the limits of these theories, appraisal theories are often the target of criticisms similar to the one moved to continuous theories of emotion. In particular, appraisal theories were accused of losing the focus on emotion by considering too a wide array of moods and affects; of relying heavily on self-reports; of being based on the subjective choice of appraisal variable, though appraisal theories of emotion show a good level of agreement on the choice of appraisal variables to be considered (for a more detailed explanation of these criticisms, please refer to section 1.4.2 on continuous theories of emotion).

Once again, one of the most relevant shortcomings of these theories is the lack of an implicit linguistic description. Emotions are defined by a set of variables which specifies how the evaluation of a given event gives rise to that emotion; however, the emotion itself lack any descriptive label which could help us categorizing the emotion itself and which could be returned and easily understood by a human user. In this case too, solution were proposed to relate abstract numerical appraisal variables to well-understood emotional terms. This mapping from appraisal variables to emotions is named *affect derivation*; beside connecting appraisal variables to emotions, this process often includes the computation of the intensity of the derived emotion; in other words, affect derivation maps a set of real values to an emotional label and to a value defining the intensity of this emotion [60].

1.5 Computational Theories of Emotion for Emotion Representation

After restricting our attention from all the theories of emotion to three computational theories of emotion, we would like now to further narrow our options in order to converge to the choice of an optimal theory of emotion for our research. In order to reduce our alternatives, we are going to analyse the three computational theories of emotion evaluating them in relation to the parameters for classifying theories of emotion which we described in section 1.2 and in relation to our objectives.

First of all, let's consider which *aspects of emotions* we are interested in (refer to section 1.2 and figure 1.1). As our research deals with measuring and clustering emotions, we are interested in theories of emotion which focus on the categorization and characterization of emotion; in our study, we are not concerned with the anthropological origins, with the actual development nor with behavioural and cognitive effects of emotions. This makes the discrete and continuous theories of emotion perfectly suited for our aim; appraisal theories of emotion satisfy our requirement, too, even if part of these theories is devoted to study and explain the actual development of emotions. Indeed, if we adopt appraisal theories for our work and if we neglect the part regarding the actual development of emotions, then appraisal theories and continuous theories end up being equivalent theories for our objectives [44]; despite the undeniable and significant conceptual differences between the two theories which we pointed out in section 1.4.3, both continuous theories and appraisal theories characterize emotions as continuous quantities; continuous theories use the concept of dimensions, while appraisal theories use the concept of

appraisal variables; for this research (and it should be underlined that this statement is not a general statement, but it is relative to the present work) continuous theories and appraisal theories can be considered equivalent.

Let's move on to question which *components of emotions* we are interested in (refer to section 1.2 and figure 1.2). In our research, our attention is devoted primarily to the study of the feeling components of emotion; in other word, our aim is to try to understand the feeling that a human subject is experiencing and which are conveyed by his speech. Motor components are partly and instrumentally considered in our research: we do not aim at studying them as we do with feeling components, but we aim at recording just a particular motor response (voice) in order to assess and predict the feeling component of an emotion. Physiological, behavioural and cognitive components of emotions are outside the scope of this research; while they are undeniably part of an emotion, we are not going to study them. At the end, then, we require a theory of emotion which is able to deal with the feeling components of emotions. Referring to figure 1.2, we can easily see that any of the three theories of emotion we are considering satisfy this requirement.

Finally, let's wonder which *phases of emotions* we are interested in (refer to section 1.2 and figure 1.2). This question, however, is bound to remain unanswered. Indeed, our study of emotions do not go deeply into details to the point of analysing an emotion phase by phase. For the time being, we are not interested in analysing how an emotion develops; we approach emotions as monolithic episodes to be studied in their unity. Therefore, considering phases of emotions unfortunately does not help us making a decision between the three theories of emotion we are considering.

This brief analysis confirmed two important facts. The first fact is that it showed that all the three theories of emotion we considered are well-suited for our objectives; this means that potentially any of them could be used as a theoretical background for our work. The second fact is that it pointed out that, for our objectives, continuous theories of emotion and appraisal theories of emotion are, practically equivalent; this allow us to reduce appraisal theories to continuous theories.

Now, coherently with the aim we stated at the beginning of this section, we can restrict our range of choices only to two alternative families of computational theories of emotion: continuous theories and discrete theories.

1.6 A Unified Theory of Emotion

At this point we are left with two families of theories of emotion from which to choose the theory which will define the theoretical foundation of our work. In the previous sections, we have analysed in details these two families of theories of emotion; in particular, we have underlined all their advantages and their disadvantages (see sections 1.4.1 and 1.4.2). Both

discrete theories and continuous theories have interesting strong points; discrete theories of emotion are very intuitive and can be easily interpreted by any user; continuous theories are more general and can allow us to track the emotional state of a user in a continuous fashion in space and time. In a way, discrete theories can compensate some of the shortcomings of continuous theories (e.g., lack of linguistic description) and, vice versa, continuous theories can counterbalance some of the shortcomings of discrete theories (e.g., inability to deal with all the emotions). The ideal solution would be to combine these theories . Russell, in [78], suggests a theory which merges continuous theories with discrete theories. In this section we summarize this theory which can be adopted as the theoretical foundation of this research.

In order to introduce Russell's theory of emotion, we need to define a couple of key terms:

Core Affect A core affect is a neurophysiological state experienced as a feeling described by hedonic and arousal values; a core affect is non-relational or object-less construct which define the general feeling of a human subject at each instant of time; core affects include moods, general affect and emotional states.

Emotional Episode An emotional episode is a complex process in which, by interacting with an object in the environment, a core affect can be modified. Generally speaking an emotional episode is triggered by an event taking place in the environment (*antecedent event*); the antecedent event may be caused by a person, a thing or a mental state (*object*); the object is required to have the property of influencing a core affect (*affective quality*); affective qualities of the objects dramatically alter the core affect of a human subject (*core affect shift*); after the core affect shift, a human subject researches the object which caused the change in his core affect and attributes to this object the causality of the change in core affect (*affect attribution*); once identified, the object is appraised (*appraisal*); finally, a set of actions and changes take place in the subject (*effects*). This whole process corresponds to an emotion. In particular cases, if the emotional episode meets a certain set of culturally-determined requirements, then we are in front of a case of a full-blown emotion or a prototypical emotion; for example, if a human subject finds himself in front of a lion (object), which is physically threatening (affective quality), which causes the arousal value to increase dramatically and the hedonic value to decrease dramatically (core affect shift), which is recognized as a danger (affect attribution), which is appraised as an obstacle for the goal of survival (appraisal) and which causes the human subject to run (effects), then we are in front of a ideal or prototypical case of fear. However, in general, an emotion manifests itself in many different ways and it is not tied to a precise set of requirements; for example, (non-prototypical) fear can determine the (non-prototypical) decision to fight instead of fleeing.

On one side, given their nature, core affects can be naturally studied within a continuous two-dimensional space; indeed, since core affects are described by hedonic value (i.e., valence) and arousal value (i.e., intensity), they can be mapped inside the intensity-valence emotional space that we described for the continuous theories of emotion.

On the other hand, emotional episodes are discrete events during which the core affect of a

human subject undergoes a sensible change. By considering all the phases of an emotional episode, it is possible to classify and label emotional episodes using emotional terms corresponding to the basic emotions suggested by the discrete theories of emotion. With the exception of prototypical emotions, (non-prototypical) emotions show high variability in the phases of an emotional episode; a natural way to model this variability is to use fuzzy categories instead of rigid categories.

For our research, this unified theory provides us with an excellent paradigmatic framework for our task.

The concept of core affect allows us to track the underlying emotional state of a user speaking in front of a computer in real-time; by mapping his state on a two-dimensional space, we are virtually able to record any emotion that the user will express through his voice.

The concept of emotional episodes allows us to detect the display of significant emotions; by following the movement of the emotional state of a user in the two-dimensional emotional space, we can identify sudden and dramatic change in the core affect and we can label this shift with emotional labels. We expect sudden and dramatic shift in core affects to move the position of the user far from the origin of the two-dimensional intensity-valence space; and, coherently with the observation made by Russell on the emotional circumplex (see section 1.4.2), we expect to identify fuzzy clusters for emotional episodes in an area far removed from the origin of the intensity-valence space [78].

So, by working with the concept of core affects, we can ground our research in the widely used and successful continuous theory of emotion which uses the valence-intensity space to describe emotion.

By working with the concept of emotional episodes, we can use emotional labels and emotional categories borrowed from natural language; we can partition the continuous emotion space and label each subspace with terms from natural language; this guarantees that our results are easily understandable by any user, including subjects with no experience at all in psychology and affective computing.

The combination of a continuous theory of emotion with a discrete theory of emotion is possible only if we define a precise mapping from the continuous domain to the discrete domain. Determining this mapping is itself a big challenge for the research in affective computing [84].

Mapping from the continuous domain to the discrete domain means solving a classification problem: we need to find a function which maps points in the valence-intensity space to emotional classes; from a geometrical point of view, we want to find the boundaries of the regions corresponding to each particular emotional state. However, we are not guaranteed that emotional classes can be separable in the valence-intensity space. Fuzzy categories may be used to take into account the overlapping of emotional classes.

Mapping from the discrete domain to the continuous domain means solving a sampling problem: we need to find an algorithm which returns a point in the area of the valence-intensity space corresponding to a given emotional label. As mentioned in section 1.4.2, it is possible to

find in literature studies establishing the value of intensity and valence of many full-blown emotions; however these studies usually report the values for the average or prototypical emotion; to avoid mapping every discrete emotion to the same pair of values of intensity and valence, we can use a probabilistic algorithm to generate reasonable and different values of intensity and valence for discrete emotions.

The unified theory described above could be seen as an attempt to combine a low-level description of emotions (continuous theories of emotion) with a high-level description of emotion (discrete theories of emotion). Considering the wide acceptance and virtual universality of the continuous theory of emotion using the intensity-valence space, this theory could be adopted as the founding and shared ground above which different discrete theories of emotion can be adopted. In other words, the intensity-valence space may provide a common domain from which and to which different sets of basic emotions (corresponding to different discrete theories of emotion) may be mapped.

Alternatively, a discrete theory of emotion could be seen as a limit case of a continuous theory of emotion. By discretising time (the position of a user in the valence-intensity space is sampled at fixed intervals) and space (the position of a user can assume only those values corresponding to prototypical emotions) a continuous theory of emotion can be reduced to a discrete theory of emotion; instead of moving continuously in all the valence-intensity space, the emotional state of a user jumps at discrete time steps between fixed point on the plane.

An undeniable advantage of combining a continuous theory of emotion with a discrete theory of emotion is the possibility of having access to a wider set of resources and data. A large number of databases use labels drawn from the everyday language to annotate the collected emotional data. Developing an algorithm to map those discrete labels into the continuous space of the valence-intensity plane would allow us to exploit those data for our study.

Chapter 2

Emotional Datasets

In this chapter we are going to analyse existing emotional datasets. The aim of this chapter is to give an understanding of emotional datasets, specifically speech emotional datasets, and to explain the choice we made on which emotional datasets to use for our research.

2.1 Development of Emotional Datasets

Emotional datasets play a crucial role in the field of affective computing. As in every application of machine learning, data provide, at the same time, the resources to learn a computational model of emotion (*training samples*) and the resources to test our theories and models (*testing samples*); given a machine learning algorithm, data define what we learn; machine learning algorithms can not learn what is not contained in the data [19].

Moreover, data define not only what we learn, but also which conclusion we can draw from the learning process. As Cowie points out in [20], it is easy to be deceived into wrong conclusions when we have just a superficial awareness of the content and the structure of a database; for example, by training a machine learning algorithm using a database containing instances of happy episodes, we may be induced to state that our system learnt to distinguish the emotion "happiness"; however, such a conclusion is imprecise and deceiving; most likely our system learnt to distinguish events of high valence or of general activation from a neutral state; asking our system to discriminate happiness from other emotions with similar valence or activation, such as anger, will very likely result in poor performances.

Now, given the complexity and the vagueness of the definition of emotions, it follows that producing and collecting data for emotional datasets is a complex and exacting task. Experience proved that the coarse application of data collection techniques successfully applied in other fields generates poor results when applied to the field of affective computing [20].

The process of building a sound affective computing database is a long and costly effort which requires a strong psychological, statistical and computational understanding of emotions; not only the content, but also the aim and the way of collecting data must be carefully defined and validated.

Here we summarize some of the key challenges that must be tackled and solved in order to

build a consistent database of emotions [20, 27]. We list these questions because the answers are important not only to scholars who wants to build a coherent database, but are also relevant to those researchers who need to select datasets for their applications.

- *Content*: what will be the content of the database? what are the emotions, affections or moods which will be contained in the database?
  These questions are strictly connected to the theory of emotion which backs up the database; indeed, a theory of emotion defines what should be considered an emotion and what should not and, therefore, identifies what should be contained in the database and what should not. Basically, every database is an implicit or explicit expression of a given theory of emotion [19]. Moreover, the content is strictly related to the application for which the database is developed; if the final application focuses only on some emotions, then the database may contain only a subset of emotions. Ideally, an emotional database should contain samples of all the emotions considered by the theory of emotion underlying it (see chapter 1 for a detailed description of theories of emotion).

- *Type*: should emotions be natural, induced or acted?
  This question is connected with the definition of emotions and with the expected use of data. Emotions could be placed on a continuous spectrum representing their artificiality; at one end of this spectrum, there are natural emotions arising spontaneously in everyday life and everyday interactions; at the opposite end, there are acted emotions which are performed by actors; in the middle of the spectrum, we can find several types of induced emotions arising in artificial settings. The choice of which type of emotion to record is often given by the evaluation of the trade-off between the cost of recording natural emotions and the artificiality of acted emotions (see section 2.2 for a more detailed analysis of the types of emotional datasets).

- *Size*: what will be the size of the database? how many samples for each emotion should be recorded?
  These questions are strictly connected to the expected use of data; if the samples are to be used for learning, then statistical evaluations must be done to determine the minimum amount of samples necessary to learn and to draw statistically reliable conclusions from the data. Statistics can suggest the optimal size which guarantees the possibility of doing statistical analysis on the data and which minimize the cost for the collection of data. However, the size of the data is related also to the protocol used for the collection; the type of protocol adopted is connected to psychological considerations about the method to be used to collect data and the type of data to be collected.

- *Balance*: how many samples for each emotion should there be? are all the emotions equally important?
  These questions are strictly connected to the expected use of data; if we want to train an application which learn to distinguish evenly among emotions, then it is important to have approximately the same number of samples for each emotion; if we want to bias the learning, then it is a good idea to have an uneven number of samples per emotion. Sometimes having the same number of samples for each emotion could be a costly task,

especially if eliciting certain emotions happens to be more difficult than eliciting other emotions.

- *Modalities*: which signals should be acquired to represent an emotion?
  This question is connected with the definition of emotion and to the practical limits of recording. Different physiological signals (e.g., heartbeat or skin conductivity) and motor responses (e.g., voice or gestures) carry relevant information about the emotion experienced by a user. Recording can be uni-modal (i.e., taking into consideration a single signal) or multi-modal (i.e., taking into consideration multiple signals). Experiments suggest that different signals combine in a non-independent and non-redundant way. The main drawbacks of recording multiple signals are the cost of multiple recording devices and the hindrance constituted by all the devices; if the aim is to collect emotions in a natural environment, then the presence of cameras, microphones and cables to record physiological signals may invalidate the realism of the environment.

- *Cross-culturality*: should the dataset contain samples belonging to a single culture?
  This question is connected with the definition of emotions and with the expected use of data. Collecting samples belonging to different cultures (e.g., different languages or different gestures) allow to draw conclusions which can be extended beyond the boundary of a single culture or society. Notice that, even if each single dataset belongs to a single culture, more datasets belonging to different cultures could be combined to extend the validity of the conclusions.

- *Fragmentation*: what will constitute a unit of analysis of an emotion?
  This question is connected with the definition of emotions and with the expected use of data. Recordings of emotional episodes may be subdivided in smaller units which can be processed and analysed individually. The division may be coarse-grained or fine-grained; it may use rigid boundaries or fuzzy boundaries; it may define distinct, overlapping or nested units. Notice that the process of fragmentation may be carried out by the creators of the database, but it is often left to the users of the database.

- *Annotation*: how should the data be labelled and annotated?
  This question is strictly connected to the theory of emotion which backs up the database. Deciding how to annotate the collected data includes choosing how rigorous the annotation should be (e.g., free annotations or formalized annotations), which annotations to use (e.g., discrete labels or continuous values), who should do the annotation (e.g., the subjects of the emotions, external naive annotators or external skilled annotators) and how different annotations on the same sample should be combined together (e.g., averaging the annotations or using a majority vote procedure). Statistical validation of the annotation should also be taken into consideration.

- *Format*: which formats should be used to store the data? are there standards than can be followed?
  A fundamental decision to make the collected data usable is the choice of formats and quality standards. Good quality of recording is a prerequisite for processing the data;

noisy recordings may not be handled well by machine learning algorithms and may compromise the final results. The adoption of widely used formats make processing of the data straightforward for other researchers; by complying to standards, pre-existing tools and libraries are readily available to process standardized data.

- *Ethics*: does the data collection comply with ethical norms? is the privacy of the subjects guaranteed?
  As any data collection task, also the construction of an emotional datasets requires that all the norms set up by ethical and legal standards must be respected. Privacy and informed consent of all the subjects taking part in the data collection must be respected. Notice that in the development of emotional databases, compliance with ethical norms may pose significant and practical challenges to data collection; for example, consider the fake scenarios designed to elicit certain emotions: if the subject must be made aware from the beginning of the inauthenticity of what is going to happen, the naturality of his emotions may be compromised; or, even worse, if a subject can not be manipulated at all, then emotions could not be elicited in an artificial environment.

Different answers to the questions listed above lead to the creation of datasets which are deeply different from each other. Nowadays, there is a wide selection of different datasets, even if recent years saw a trend in data collection towards natural, multi-modal and cross-cultural datasets [85].

2.2 Types of Emotional Datasets

One of the most essential traits of an emotional dataset is the type of emotion recordings contained in it. As explained in section 2.1, by *type of emotion recordings* we refer to the degree of naturality or artificiality of the recordings. Even if the dimension of naturality or artificiality may be considered a continuous dimension, three main categorical types of emotion recordings may be easily identified [19, 27]:

Acted Emotion Recordings in this scenario emotions are played by professional or non-professional actors who are given a situation to rehearse. In the case of speech emotional datasets, actors may be given a text to be read while simulating a specific emotion, such as anger or boredom; text may be as long as full sentences or may be just a single word; text may be semantically correlated with the emotion the actor is supposed to play or it may be a random text which has no relation with the specific emotion.
Acted recordings have many practical advantages that made them very popular. First, the cost for recording in a studio is usually very contained. Second, emotions can be collected in controlled environments where noises can be reduced to a minimum. Third, since emotions are defined a priori in the script given to the actor, emotional recordings can be easily categorized and annotated.
However, these advantages are balanced by severe drawbacks. First, the quality of the acted emotions is dependent on the actor. Second, some psychologists argued that, being acted, these fake emotions fail at eliciting all the physiological and motor reactions that

an authentic emotion would elicit. Third, the extension to real-life of applications which learned from acted data often proved unsatisfactory [20].

To sum up, on one side, acted emotions can be considered as idealized instances of emotions, easy to process and to deal with; on the other side, they can be seen as decontextualized and too idealized instances of emotions, too removed from reality.

Natural Emotion Recording in this scenario emotions are recorded from everyday life; ideally, the behaviour of the subject should not be affected by the data collectors or by the knowledge that his emotions are being recorded for research purposes; the goal is to get natural recordings which are not biased by artificial conditions introduced by the data collectors. In the case of speech emotional datasets, TV talk-shows, call-center recordings, lectures and meetings, children playing and medical dialogues may be used [27].

Natural recordings are the focus of increasing attention. The first and main advantage of a natural recording is its realism: each sample represents an authentic emotion experienced and expressed in its context.

Unfortunately, working with natural recordings may be extremely challenging. First, according to the source of the recordings, it may be difficult to obtain samples of specific emotions which happen to be very rare in a given context. Second, the quality of the recordings may be very low; in the case of speech recordings, even small movements in front of a microphone or background events may add noise which makes it difficult to process the recordings; it is clear that in a natural context, in which the subjects are not constrained in any way, noise is very high. Third, emotions and their expression may be shadowed by other normal activities such as moving, speaking or engaging in other activities as everyone usually does in a natural setting. Fourth, recorded emotions may be not well-defined and it may be hard or controversial to categorize or to annotate them. Fifth, recording devices have to be small and unobtrusive in order not to make the subject realizing he is being recorded. Sixth, experimenters collecting natural emotion recordings must take into consideration ethical and privacy issues about data collected from subjects unaware of the recordings.

So, contrary to acted emotions, natural emotions can be considered real instances of emotions, the true object which we would like to study; unluckily, they may be very difficult to collect and to process, thus reducing their appeal.

Induced Emotion Recording in this scenario, lying in the middle between acted emotion recordings and natural emotion recordings, emotions are artificially elicited in human subjects; various techniques were studied to make a human subject experience a certain emotion, such as exposing him to emotionally-charged music and videos or making him play emotion-eliciting games. In the case of speech emotional datasets, a common technique is to use Wizard-of-Oz (WOZ) scenarios, that is scenarios in which a human subject is convinced of interacting with a computer, while, in reality, a human experimenter is directly operating the machine and returning answers to the inputs of the user; protocols have been developed in which a human subject is expected to interact with the machine through vocal commands, while the operator gives back answers aimed at eliciting certain

emotions.

A strong point of induced emotion recordings is that they are able to blend some of the advantages of acted emotions and natural emotions. Since data collection usually takes place in a controlled laboratory environment, the quality of the recordings is usually high, the cost for collection may be contained and the emotions can be categorized more easily than in the case of natural emotions. Moreover, being authentic, though not spontaneous, induced emotions are closer to reality than acted emotions.

Still, induced emotion recordings have some weak points. Some psychologists doubt that induced emotions can be considered truly authentic emotions; the very awareness of a human subject of being inside a laboratory affects his emotions in their display. Some emotions may be easier to induce than others. More importantly, the design of laboratory protocols for emotion induction may be extremely complex: it must be shown that a protocol actually elicits the emotions it is supposed to elicit; it must guarantee that emotions are actually elicited and not acted (e.g., in the case of WOZ scenarios it means that the illusion the subject has of interacting with a machine must never be broken); it must comply with ethical and legal standards (e.g., in the case of WOZ scenarios emotions must be prompted but not manipulated).

All in all, induced emotions represent a valid and convenient alternative to natural or acted emotions. However, the difficulty of designing proper protocols, which was defined by Cowie as an art [19], limits their adoption.

2.3 Speech Emotional Datasets

As the interest of this research focuses on speech datasets, in this section we analyse the features of those datasets which can be used for our research.

A speech emotional dataset is a collection of audio emotional samples, recorded from one or more human subjects and tagged by one or more annotators. Strictly speaking, a speech dataset is a uni-modal database containing only audio data; however, we will take into consideration multi-modal audio-visual datasets, too; in this case, our attention will be focused only on the audio content, while the video part will be ignored.

The first speech databases used in affective computing were often developed for other applications (e.g., quality measurement of synthesis), but since the late 1990s new datasets primarily meant for research in affective computing were developed [89]. Beginning with small and limited databases, the quality and the breadth of speech datasets improved with time, to the point that now audio datasets are among the most abundant resources for affective computing [20].

The high number of available datasets is associated with significant differences between them: speech databases vary deeply according to their content, their type, their size and their annotations.

Historical databases such as *Danish Emotional Speech (DES)* [35] and the *Berlin Emotional Corpus* [9] are prototypical uni-modal, mono-lingual datasets built by recording actors expressing basic emotions and have been widely used by the affective computing community as a benchmark and a reference; newer datasets collecting acted emotional recordings include

the *Serbian Emotional Speech Corpus (GEES)* [51], the *Mandarine Affective Speech Corpus* (*MASC*) [110], the *Surrey Audio-Visual Expressed Emotion Database* (*SAVEE*) [49], the *Emotional Prosody Speech and Transcript* (*EPST*) [46], the *Russian Language Affective Speech Database* (*RUSLANA*) [64], the *Montreal Affective Voices Database* (*MAV*) [4] and the *MindReading Database* [41].

More recently, efforts were made to develop more realistic datasets; *Vera-am-Mittag (VAM)* [43] corpus collects emotional recordings from a TV talk show, *EmoTV* [25] corpus collects emotional recordings from TV news programs, *Audio-Visual Interest Corpus (AVIC) [86]* records natural reactions of interest or boredom of human subjects in front of advertisements, *Speech Under Simulated and Actual Stress (SUSAS)* [45] collects the natural reactions of human subjects under stressful situations, *BabyEars* [94] collects emotional recording of parents interacting with their children, *IDIAP Wolf Corpus* [48] records emotional utterances of subject playing a role-playing game, *Doors* [95] records emotional sentences of users playing a computer-based gambling game, *Japan Science and Technology / Core Research for Evolutionary Science and Technology Database (JST/CREST)* records emotions in everyday locations, *British Telecom OASIS Database* (*BT-OASIS*) [30] and *CallHome American English Corpus* (*CHE*) [37] collect calls to a customer service.

Induced datasets too were the focus of wide interest as new and alternative protocols were developed to elicit emotions; *Airplane Behaviour Corpus (ABC)* [82] and *eNTERFACE* [67] uses storylines to induce emotion, *Smartkom* [96] and *NIMITEK* [40] relies on Wizard-of-Oz scenarios to elicit emotions, *Sensitive Artificial Listener (SAL)* [20] uses an induction technique to make the user navigate its emotions, *EmoTaboo* [20] relies on a game designed to elicit emotions, *AIBO* [2] uses a robot to make children express their emotions.

Longer catalogues of speech datasets available to researchers can be found in [100], [19] and [20]; alternatively, an updated list of available datasets is maintained by *Humaine* research group on their Wiki page[1]. In the next section we are going to give the details of the datasets which will be used in our research.

2.4 Selected Datasets

As the data are the primary source of information about emotions in our research, it is crucial to select datasets which are consistent with our objectives.

Nowadays the field of affective computing can rely on a good number of speech datasets. It is impossible to determine which databases are objectively the best ones. The choice of a good dataset is inextricably bound to the assumptions, the aims and the constraints of a specific study [40]; for example the assumption of a continuous theory of emotion, the aim of studying anger-related emotions and the constraint of using only free public datasets reduce the choices to a small subsets of datasets optimized for these requirements. In order to choose the datasets we are going to use in our research, we are going to address the questions we outlined in section 2.1 in relation to our assumptions, our aim and our constraints. The following answers determine the requisites that the datasets we are going to select must have.

[1] http://emotion-research.net/wiki/Databases

- *Content*: as we adopted a theory of emotion unifying continuous theories of emotion and discrete theories of emotion (see section 1.6), we can theoretically work with datasets built over a continuous theory of emotion or over a discrete theory of emotion. We can take into consideration a broad range of emotions as the continuous theories of emotion underlying the unified theory allow us to deal with affections and moods, while the discrete theories of emotion constituting the upper layer of the unified theory allow us to distinguish full-blown emotions. Upon the development of good strategies to map from the continuous domain to the discrete domain, the unified theory we assumed guarantees us a good degree of content-wise versatility in selecting datasets.

- *Type*: in section 2.2 we explained the advantages and disadvantages of different types of recorded emotions; in particular, we highlighted the trade-off between ease of processing, maximized by acted emotions, and realism, maximized by natural emotions. Ideally, our aim is to be able to work with any kind of recordings (natural, induced and acted); therefore we do not have an explicit preference for any of them; instead, we would like to use several datasets containing different types of recordings and try to draw conclusions which are independent of the type of recording.

- *Size*: as we are going to use the datasets for statistical processing, we are going to select databases which contain an amount of data which will be big enough to draw significant and reliable statistical conclusions.

- *Balance*: as our research is not biased to the recognition and discrimination of particular emotions, we are going to favour databases containing an even amount of samples for each emotion.

- *Modalities*: since the focus of our research is on emotions expressed in speech, we are going to use mainly uni-modal audio databases; as an exception, we will also consider some multi-modal audio-visual databases, but in these cases we will simply discard the video signal and rely exclusively on the audio signal.

- *Cross-culturality*: similarly to the case for the type of emotions, we would like our conclusions to be independent of a specific culture. In speech datasets, culture is expressed mainly by the language of the human subject speaking; to make our conclusions truly cross-cultural, we aim at using datasets containing recordings in different languages or at combining mono-cultural datasets each belonging to a different culture; we want to process speech belonging to different languages (e.g., English and German) and, ideally, belonging to different families of languages (e.g., Indo-European and Sino-Tibetan).

- *Fragmentation*: as fragmentation of speech is not one of the main focuses of our research, we will not take into consideration dataset-specific fragmentation of the content; we will work with speech samples which will be fragmented according to the standards of the literature in affective computing.

- *Annotation*: as we adopted a theory of emotion unifying continuous theories of emotion and discrete theories of emotion (see section 1.6), we have a discrete degree of flexibility on

the annotation we can use. Ideally, the best type of annotation we can ask for is continuous annotation coherently with the fact that a continuous theory of emotion makes up the foundational layer of our theory of emotion; however, by developing strategies to map categories belonging to discrete theories of emotion into the continuous space we will be able to exploit the vast number of datasets tagged with discrete labels.

- *Format*: we will favour datasets containing data which comply with the standards and containing recordings with a good level of quality.
- *Ethics*: we will select databases which conform to ethical and legal regulations.

Coherently with these requirements, we give details of the datasets we considered for our work:

Berlin Emotional Database The Berlin Emotional Database is an audio collection of emotional utterances developed between 1997 and 1999 by Burkhardt et al. at the Technical University Berlin [9]. It was originally built for studies of prosodic features, articulatory features and verification by resynthesis. Burkhardt et al. adopted a discrete theory of emotion with 7 basic emotions: four "big" emotions (*anger*, *fear*, *joy* and *sadness*), two additional emotions (*boredom* and *disgust*) and the *neutral* state. The database contains recordings of 10 non-professional actors (5 female and 5 male) uttering emotionally coloured sentences. Every actor was required to utter 10 sentences; each sentence was repeated for every single emotion considered in the study; in total, the database contains 700 sentences (10 actors * 10 sentences * 7 emotions) plus 100 additional sentences as a backup. In order to induce specific emotions, actors were suggested to use Stanislavki method (i.e., a self-induction strategy often used by actors based on recalling emotionally-charged events from the memory). The recordings were performed in an anechoic chamber; actors, standing in front of a microphone, were free to gesture, but they were asked to keep at a constant distance from the microphone in order to maximize the quality of the recordings.

Danish Emotional Database (DES) The Danish Emotional Database is an audio emotional database built within the framework of the VAESS-project at the Aalborg University [35]. The recordings were collected in order to develop synthetic voices expressing emotions. Engberg et al. adopted a discrete theory of emotion with 5 basic emotions: *anger*, *happiness*, *sadness*, *surprise* and the *neutral* state. The database contains recordings of 4 Danish radio theatre actors (2 female and 2 male). Every actor was required to utter single words, sentences and passages; in total, the database contains 260 utterances (4 actors * 13 utterances * 5 emotions) plus 81 additional recordings. The recordings were performed in a acoustically damped sound studio with the support of two operators; a high-quality microphone was used to record the utterances.

Serbian Emotional Speech Database (GEES) The Serbian Emotional Database is the first Serbian audio emotional database developed by Jovicic et al. at the Belgrade University [51]. The dataset was built in order to evaluate which emotional content in speech can be recognized by humans and if such emotional information can be used in human-computer interactions. Jovicic et al. adopted a discrete theory of emotion with 5 basic

emotions: *anger*, *fear*, *happiness*, *sadness* and the *neutral* state. The database contains the recordings of 6 students (3 female and 3 male) belonging to the Faculty of Dramatic Arts of the the Belgrade University. Every student was required to utter words, short sentences, long sentences and a passage; in total, the database contains 2790 recordings. The recordings were performed in an anechoic chamber under the supervision of three operators; students stood in front of a microphone and they were relatively free to move and gesture as long as they kept at a constant distance from the microphone.

Mandarine Affective Speech Corpus (MASC) The Mandarine Affective Speech Corpus is the first Mandarine audio emotional database developed by Wu et al. at the Zhejiang University in Hangzhou [110]. The aim of this dataset was to further the prosodic and linguistic study of emotions in the Mandarine language and to provide data for automatic emotion recognition systems. Wu et al. adopted a discrete theory of emotion with 5 basic emotions: *anger*, *elation*, *panic*, *sadness* and the *neutral* state. The database contains the recordings of 68 students (23 female and 45 male) belonging to the Advanced Computing and Systems Laboratory of the Zhejiang University. Every student was required to utter words, sentences and passages; in total, the database contains 25636 recordings. In order to induce specific emotions, students were suggested to read a short emotionally-charged story. The recordings were performed in a quiet office.

BabyEars BabyEars is an audio emotional database developed by Slaney et al. [94]. The aim of this dataset was to collect natural emotion recordings and to evaluate the level of discrimination in a classification task. Slaney et al. adopted a discrete theory of emotion with 3 basic emotions: *approval*, *attention* and *prohibition*. To record utterances that are natural and, at the same time, emotional, Slaney et al. collected recordings of 12 parents (6 female and 6 male) talking and playing with their children. Parents and children were left in a quiet room with different toys and they were asked to interact naturally. During each session, lasting one hour, several recordings were made; in total, the database contains 509 recordings. The recordings were performed using a light headset with a wireless microphone.

Vera Am Mittag Corpus (VAM) The Vera Am Mittag Corpus is an audio-visual emotional database developed by Grimm et al. [43]. The aim of this dataset was to collect natural emotion recordings in order to generate a high-quality audio-visual database for research in affective computing. The Vera Am Mittag Corpus is made up of three databases: VAM-Audio, VAM-Video and VAM-Faces, containing, respectively, audio samples, video samples and face still images; the only database we are considering is VAM-Audio. Grimm et al. adopted a continuous theory of emotion with 3 dimensions: *intensity*, *valence* and *dominance*. The emotions contained in the database are natural emotions; all the recordings are extracted from episodes of "Vera am Mittag" ("Vera at Midday"), a German talk show during which the guests discussed emotionally charged topics. Grimm et al. selected 47 guests (36 female and 11 male) which showed a wide enough range of emotions and whose recordings were of a good quality for audio processing. In total, the database contains 1018 recordings. Annotations were made using the SAM (Self-Assessment

Manikins) method [6]; SAM requires each annotator to rank intensity, valence and dominance of an emotion on a 5-values discrete scale; the annotations of all the annotators are then averaged to produce the final pseudo-continuous values of intensity, valence and dominance.

Sensitive Artificial Listener (SAL) The Sensitive Artificial Listener is an audio-visual emotional database developed by Schroder et al. [79] within the framework of the SEMAINE project. The aim of this dataset was to collect data in order to study emotion-related non-verbal aspects of a one-to-one conversation and to develop a conversational agent. Schroder et al. adopted a continuous theory of emotion with 3 dimensions: *intensity*, *valence* and *dominance*. The emotions contained in the database are induced emotions; 4 speakers (2 female and 2 male) interacted with a system which, by prompting specific questions, led the subjects to express different emotions. In total, the database contains 1692 recordings. Annotations were made using the FEELTrace tool [21] to assess the value of the emotions along three continuous dimension.

Humaine Database The Humaine Database is a mixed audio-visual emotional database developed by Douglas-Cowie et al. [29] within the framework of the HUMAINE project. The database was built in order to provide the researchers with a high-quality database containing a wide range of emotions in different contexts. The dataset is built by assembling high-quality, relevant emotional samples from pre-existing databases (e.g, Belfast Naturalistic Database, SAL, Gemep Corpus); in total the database contains 48 clips and includes mainly natural and induced emotions, even if acted emotions are present, too. Douglas-Cowie et al. adopted a continuous theory of emotion with 8 dimensions: the three main PAD dimensions (*intensity*, *valence* and *dominance*) plus five additional dimensions (*acting*, *masking*, *activation*, *expectation* and *word-relatedness*). Different techniques for annotation were experimented including set of prespecified terms, soft vectors and the FEELTrace tool [21].

NIMITEK Corpus The NIMITEK Corpus is an audio-visual emotional database developed by Gnjatovic et al. [40]. The aim of this dataset was to apply an innovative WOZ protocol for collecting realistic emotional samples. Gnjatovic et al. adopted different discrete theories of emotion; the original theory they used is based on 6 emotions: *anger*, *disgust*, *fear*, *joy*, *sadness* and the *neutral* state; this theory was then extended by allowing an unrestricted use of emotional terms, which raised the number of emotions to 22; finally, the number of emotions was reduced defining 3 cover classes which encompassed all the 22 emotional terms: *positive emotions*, *negative emotions* and *neutral emotions*. The emotions contained in the database are induced emotions; 10 subjects (7 female and 3 male) interacted with a system which proposed to the user games and challenges; inputs were given by voice and they were processed by a human operator (WOZ scenario) in order to return answers which would elicit specific emotions. In total, the database contains 1847 recordings. Annotations were made both by German-speaking and non-German-speaking annotators.

Speech Under Simulated and Actual Stress Corpus (SUSAS) The Speech Under Simulated and Actual Stress is an audio database developed by Hansen et al. [45]. The aim of this dataset was to collect recordings of speech in condition of high stress and noise in order to evaluate the quality of speech recognition. The SUSAS Corpus is made up of two databases: Simulated and Actual, containing, respectively, acted and natural speech recordings; the main database of interest is Actual. This database contains information on the talking style, the conditions in which the speech was produced and psychiatric analysis data related with the emotion experienced by the human subjects. Hansen et al. adopted a discrete theory of emotion with 4 emotions: *angry*, *anxiety*, *depression* and *fear*. Hansen et al. selected 36 subjects (13 female and 23 male), including 4 Apache helicopter pilots. The recordings were made in a variety of stressful conditions, such as on roller-coasters, in free fall or, in the case of pilots, while on their helicopters. In total, the database contains more than 16000 recordings.

AIBO Database The AIBO Database is an audio emotional database developed by Batliner et al. [2]. The aim of this dataset was to collect children speech and children emotional utterances in order to evaluate the quality of speech recognition. Batliner et al. adopted a discrete theory of emotion with 11 emotions: *anger*, *bored*, *empathic*, *helpless*, *joyful*, *motherese*, *reprimanding*, *rest*, *surprised*, *touchy* (i.e., irritated) and the *neutral* state. The emotions contained in the database are induced emotions; 51 German children (30 female and 21 male) played with a dog-like Sony AIBO robot and they were required to make their robotic pet perform certain tasks; inputs were given by voice and they were processed by a human operator (WOZ scenario) who made the AIBO robot react in a predefined way in order to elicit emotions in the children. In total, the database contains 51393 uttered words. Annotations were made by non-expert annotators and their evaluations were averaged to get the final annotations. The original AIBO database contains recordings from English children, too, but these data were not released because of privacy restrictions.

Airplane Behaviour Corpus (ABC) The Airplane Behaviour Corpus is an audio-visual emotional database developed by Schuller et al. [82]. The aim of this dataset was to collect emotional utterances which could be used for research in the field of surveillance. Schuller et al. adopted a discrete theory of emotion with 6 emotions: *aggressive*, *cheerful*, *intoxicated*, *nervous*, *tired* and the *neutral* state. The emotions contained in the database are induced emotions; 8 subjects (4 female and 4 male) were invited to take part in a simulated scenario; the setting was a return flight from holidays and different events were staged according to a guided storyline in order to induce emotions in the subjects. In total, the database contains 396 clips.

eNTERFACE The eNTERFACE database is an audio-visual emotional database developed by Martin et al. [67]. The aim of this database was to generate a collection of audio, video and audio-visual samples to test emotion recognition algorithms. Martin et al. adopted Ekman's discrete theory of emotion with 7 emotions: the "Big Six" (*anger*, *disgust, fear*, *happiness*, *sadness* and *surprise*) plus the *neutral* state. The emotions in the database

are induced emotions; 42 subjects (8 female and 34 male) were given emotionally-charged stories to read in order to immerse themselves in specific situations; the subjects were then required to utter emotionally-related sentences. All the subjects were required to read and to express themselves in English, even if many of them came from background as diverse as Belgium, Cuba or Russia. In total the database contains 1166 samples. The recordings were performed in a small room using high-quality microphones and cameras and asking all the subjects to keep at a constant distance from the microphone.

Audio-Visual Interest Corpus (AVIC) The Audio-Visual Interest Corpus is an audio-visual emotional database developed by Schuller et al. [86]. The aim of this database was to collect reliable and realistic data to study interest and curiosity detection. Schuller et al. studied interest manifested in speech and in facial expression and designed a discrete theory of emotion with 5 emotions or level of interest (LOI): *disinterest* ($LOI = -2$), *indifference* ($LOI = -1$), *neutrality* ($LOI = 0$), *interest* ($LOI = +1$), *curiosity* ($LOI = +2$). The emotions in the database are natural emotions; 21 subjects (10 female and 11 male) were invited to sit in front of an experimenter describing them different types of products or topics; the subjects were encouraged to behave naturally, not to show special politeness towards the experimenter and to freely ask questions according to their interest. In total the database contains 12839 turns. Speech recording was performed using a headset microphone and a far-field microphone. The samples were annotated by four non-expert annotators; the final discrete value of LOI was determined using a majority algorithm.

SmartKom SmartKom is a multi-modal (audio, visual and gesture) database developed at the University of Munchen [96, 98]. The aim of this database was to collect multi-modal realistic data which could be used to study human-machine interaction. The SmartKom dataset is made up by 4 databases: SmartKom Public, SmartKom Mobil, SmartKom Home and SmartKom Audio; the first three databases differ on the specific scenarios presented during the data collection, while the last database assembles all the audio samples from the other three multi-modal databases; the database we are interested in is SmartKom Audio. SmartKom is built over a discrete theory of emotion with 7 emotions: *anger/irritation*, *helplessness*, *joy/gratification*, *neutrality*, *pondering/reflecting*, *surprise* and *unidentifiable episodes*; the last class was defined in order to contain utterances devoid of emotional content, specific emotions which do not fall in any other category of the theory (e.g., *disgust*) and emotion on which the annotators could not agree. The emotions in the database are induced emotions; 224 subjects interacted with a machine which was remotely controlled by human operators in order to elicit specific emotions (WOZ scenario). In total the database contains 448 recording sessions. Speech recording was performed using a headset or clip microphone, a directional microphone and a microphone array with 4 channels. The samples were annotated by trained annotators.

EmoTV EmoTV is an audio-visual emotional database developed by Devillers et al. [25]. The aim of this database was to provide realistic and natural emotion samples for the analysis of emotional speech. EmoTV adopts two distinct theories of emotions, a continuous theory and a discrete theory. The continuous theory of emotion has 4 dimensions: the

standard PAD dimensions of *valence*, *intensity* and *dominance* plus an *activation* dimension, describing if the subject is active or passive. The discrete theory of emotion was originally implemented without the definition of basic emotions a priori; 176 fine-grained label were initially used, but then reduced to 14 broader categories: *anger*, *despair*, *disgust*, *doubt*, *exaltation*, *fear*, *irritation*, *joy*, *pain*, *sadness*, *serenity*, *surprise*, *worry* and the *neutral* state; coarser emotional cover classes, corresponding to Ekman's "Big Six" plus *neutral* and *other*, were also defined. The emotions in the database are natural emotions; audio-video samples featuring 48 subjects uttering emotionally-charged monologues were extracted from TV news programs. In total the database contains 51 clips. The samples were annotated by expert annotators who labelled the samples referring both to the discrete theory of emotion and the continuous theory of emotion of the database.

EmoTaboo EmoTaboo is an audio-visual emotional database developed by Devillers et al. [26, 115]. The aim of the database was to collect realistic and multi-modal emotional expressions in an interactive context. EmoTaboo adopts a discrete theory of emotions; 21 basic emotions were considered at the beginning: *amusement, annoyance, anxiety, boredom, cold anger, confidence, contentment, disappointment, effervescent happiness, embarrassment, excitement, frustration impatience, nervousness, pleasure, pride, sadness, satisfaction, stress, surprise* and *other*; all these categories were then organized in a hierarchical structure. The emotions in the database are induced emotions; 10 subjects (4 female and 6 male) took part into Taboo game sessions; Taboo is a guessing game and during each session the subjects played together with an experimenter who was instructed to behave in a predefined way in order to elicit emotions in the subjects. In total the database contains 10 clips, each featuring a game. The samples were annotated by expert annotators using the hierarchical structure defined by the adopted theory of emotion.

Russian Language Affective Speech Database (RUSLANA) The Russian Language Affective Speech Database is a Russian audio emotional database developed by Makarova et al. [64]. The dataset was built in order to collect emotional utterances to be used in affective computing. Makarova et al. adopted a discrete theory of emotion with 6 basic emotions: *anger*, *fear*, *happiness*, *sadness*, *surprise* and the *neutral* state. The database contains the recordings of 61 students (49 female and 12 male). Every student was required to utter 60 sentences; in total, the database contains 3660 utterances (61 subjects * 10 sentences * 6 emotions). Annotations were made by groups of listeners who labelled each emotion and ranked the quality of the each acted emotion.

IDIAP Wolf Corpus The IDIAP Wolf Corpus is an audio-visual emotional database developed at the Idiap Institute by Hung et al. [48]. The aim of the database was to collect multi-modal samples of speech in a competitive conversational setting in which the subjects were prone to adopt deceptive behaviours. The emotions in the database are natural emotions; 36 subjects took part into Werewolf game sessions; Werewolf is a role-playing game in which by acting, bluffing and seeing through each other's lies, players are required to discover the roles of other players. In total the database contains 15 clips, each featuring a game. Hung et al. annotated the data with 4 labels related to the roles

of the game: *Liar (Werewolf)*, *Non-Liar (Villager)*, *Special Role 1 (Seer)*, *Special Role 2 (Little Girl)*. The samples were annotated by expert annotators using the hierarchical structure defined by the adopted theory of emotion. The recordings were performed in a room using head-mounted omni-directional microphones and an array of microphones in the centre of the room.

Surrey Audio-Visual Expressed Emotion Database (SAVEE) The Surrey Audio-Visual Expressed Emotion Database is an audio-visual emotional database developed by Jackson et al. at the University of Surrey [49]. The aim of this dataset was to collect data for the development of an automatic emotion recognition system. Jackson et al. adopted Ekman's discrete theory of emotion with 7 emotions: the "Big Six" (*anger*, *disgust, fear*, *happiness*, *sadness* and *surprise*) plus the *neutral* state. The database contains the recordings of 4 students (all male) belonging to the University of Surrey. Every student was required to utter 120 sentences; in total, the database contains 480 recordings. In order to induce specific emotions, students were suggested to read a short emotionally-charged story. The recordings were made inside the 3D vision laboratory using a microphone located in front of the speaker.

Montreal Affective Voices Database The Montreal Affective Voices Database is an audio emotional database developed by Belin et al. at the University of Montreal [4]. The aim of this dataset was to provide to the community a set of validated emotional samples and to study the effects of gender differences in the production and the perception of emotional speech. Belin et al. adopted a discrete theory of emotion with 9 emotions: *anger*, *disgust, fear*, *happiness*, *pain*, *pleasure, sadness*, *surprise* and the *neutral* state; for annotations, Belin et al. also adopted a continuous theory of emotion with 3 dimensions: *arousal*, *valence* and *intensity*. The database contains the recordings of 10 amateur or professional actors (5 female and 5 male). Every actor was required to utter 9 non-verbal emotional bursts, one for each emotion identified in the discrete theory of emotion; in total, the database contains 90 emotional bursts. The recordings were performed in a sound-proof room in the Vocal Neurorecognition Laboratory of the University of Montreal using a microphone placed at 30 cm from the speaker.

MindReading Database The MindReading Database is an audio-visual emotional database developed by Baron-Cohen et al. at the University of Cambridge [41, 95]. The aim of this dataset was to collect a wide set of diverse and subtle emotions to be used to help individuals with Autism Spectrum Disorder recognize emotions. Baron-Cohen et al. adopted a fine-grained discrete theory of emotion with 412 emotions concepts, organized hierarchically and mapped to 24 high-level mutually excluding categories. The database contains the recordings of 17 adults (5 female and 12 male) recruited through an employment agency amateur or professional actors. Every participant was required to utter sentences, and the quality of these sentences was reviewed by a group of annotators; in total, the database contains 4411 emotional sentences.

Doors Database The Doors is a multi-modal emotional database developed at the Tel Aviv University by Sobol-Shikler et al. [95]. The aim of the database was to collect multi-modal

natural samples in a controlled environment to study the affective relevance of speech and facial expressions. The emotions in the database are natural emotions; 15 subjects took part into Iowa Gambling Test (IGT) sessions. IGT is a protocol defining a simple gambling game, in which rewards or losses are hidden behind closed doors and the user is required to choose which doors to open; during the recordings, the players were required to utter vocal commands to open or close the doors. In total the database contains 3000 sentences, including commands and other recordings. Four annotators reviewed and tagged the samples using annotations derived from the MindReading Database hierarchical set of emotional concepts. The recordings were made at the Tel Aviv University; together with the audio signal, video, electrocardiogram, Galvanic skin response and blood pressure signals were recorded.

JST - Core Research for Evolutionary Science and Technology (JST/CREST) The JST/CREST Database is an audio emotional database being developed by Campbell et al. within the framework of the Core Research for Evolutionary Science and Technology project [12]. The aim of this years-long project is to generate an extensive collection of natural emotions by recording the emotional expressions of volunteers in their everyday location and in their ordinary interactions. Together with natural recordings, this database will include acted and induced emotions too.

Emotional Prosody Speech and Transcript Database (EPST)

British Telecom OASIS Database (BT-OASIS)

CallHome American English Corpus (CHE)

Electromagnetographic Articulatory Study (EMA) [54, 42]

CINEMO [8]

JEMO [8]

Before making our decision about which datasets to use, we decided to study which of the aforementioned databases were used by other research groups. In order to analyse which datasets were used in which context, we decided to focus on the main publications we read in the field of affective. We took into consideration all the papers which use any of the aforementioned databases in their research; however, we ignored those articles written only to introduce a specific datasets; these papers (listed in in tables 2.1, 2.2 and 2.3 under the column *Source*), though interesting as they often contain insightful remarks and baseline performances for the databases they are introducing, were deemed not significant for the evaluation of the degree of adoption of such datasets. We also decided to group the papers we studied by the type of the specific research they are addressing; we identified the following areas of research:

Corpus	*Source*	*Year*	*Rec.*	*Lang*	*Speakers*	*Mod.*	*Emotions*	*#Sam*	*Rate (KHz)*	*Public*
Berlin Emotional Database	[9]	1997	acted (studio)	Ger	5F, 5M	A	Discrete theory with 7 basic emotions (anger, boredom, disgust, fear, joy, neutral, sadness)	700+100 sentences	48, downsampled at 16	Y
DES (Danish Emotional Speech)	[35]	1996	acted (studio)	Dan	2F, 2M	A	Discrete theory with 5 basic emotions (anger, happiness, neutral, sadness, surprise)	260+81 utterances	20	Y
GEES (Serbian Emotional Speech Corpus)	[51]	2004	acted (studio)	Serb	3F, 3M	A	Discrete theory with 5 basic emotions (anger, fear, happiness, neutral, sadness)	2790 recordings	44.1, downsampled at 22.05	Y
MASC (Mandarine Affective Speech Corpus)	[110]	2006	acted (room)	Mand	23F, 45M	A	Discrete theory with 5 basic emotions (anger, elation, neutral, panic, sadness)	25636 utterances	22.05 downsampled at 8	
RUSLANA (Russian Language Affective Speech)	[64]	2002	acted (studio)	Rus	49F, 12M	A	Discrete theory with 6 basic emotions (anger, fear, happiness, neutral, sadness, surprise)	3660 utterances	22.05 downsampled at 8	
SAVEE (Surrey Audio-Visual Expressed Emotion)	[49]	2007-2010	acted (lab)	Eng	4M	AV	Discrete theory with 7 basic emotions (anger, disgust, fear, happiness, neutral, sadness, surprise)	480 utterances	44.1	
MAV (Montreal Affective Voices)	[4]	2008	acted (studio)	Fre	15F, 15M	A	Discrete theory with 9 basic emotions (anger, disgust, fear, pain, happiness, neutral, pleasure, sadness, surprise) and continuous theory with 3 dimensions (valence, arousal, intensity)	90 bursts	96 downsampled at 44.1	Y
MindReading	[41]	2006	acted (lab)	Eng	5F,12M	AV	Discrete theory with 412 emotional concepts distributed hierarchically in 24 groups	4411 sentences		

Table 2.1: Acted datasets.
For each dataset, the table reports: the name and the acronym of the dataset (*Corpus*); the main source of information for the dataset (*Source*); the year the dataset was built, released to public or the year the main source of information was published (*Year*); the type of recordings (*Rec*); the language(s) of the utterances in the database (*Lang*); the number of speakers, subdivided into the number of female (F) and male (M) speakers (*Speakers*); the modality, audio (A) or audio-visual (AV), of the recordings (*Type*); the number of discrete emotions or the dimensions for continuous emotions (*Emotions*); the total number of samples (*#Sam*); the rate of recording in KHz (*Rate*); whether the dataset is publicly available (*Public*). For a more complete description of the datasets, refer to the text.

Corpus	*Source*	*Year*	*Rec.*	*Lang*	*Speakers*	*Mod.*	*Emotions*	*#Sam*	*Rate (KHz)*	*Public*
BabyEars	[94]	1998	natural	Eng	6F, 6M	A	Discrete theory with 3 basic emotions (approval, attention, prohibition)	509 utterances	22	
VAM (Vera Am Mittag Corpus)	[43]	2008	natural	Ger	36F, 11M	AV	Continuous theory with 3 dimensions (valence, intensity, dominance)	1018 utterances	44.1 downsampled at 16	Y
SUSAS (Speech Under Simulated and Actual Stress)	[45]	1988-1993	natural	Eng	13F, 23M	A	Continuous theory with 3 dimensions (valence, intensity, dominance)	16000 word utterances	8-16	Y
AVIC (Audio-Visual Interest Corpus)	[86]	2007	natural	Eng	10F, 11M	AV	Discrete theory with 5 basic emotions (curiosity, disinterest, indifference, interest, neutrality)	12839 turns	44.1	Y
EmoTV	[25]	2005	natural	Fre	48	AV	Continuous theory with 4 dimensions (valence, intensity, self-control, activation) and discrete theory with 14 emotions (anger, despair, disgust, doubt, exaltation, fear, irritation, joy, neutral, pain, sadness, serenity, surprise, worry)	51 clips		
IDIAP Wolf	[48]	2010	natural	Eng	36	AV	Discrete game-related classes	15 clips	48	
Doors	[95]	2009	induced	Heb	15	AV	Discrete theory with 412 emotional concepts distributed hierarchically in 24 groups	3000 sentences	32	
JST/CREST (Japan Science and Technology Agency - Core Research for Evolutionary Science and Technology)	[12]	2002+	natural	Jap, Eng, Chi		A			48	

Table 2.2: Natural datasets.
For each dataset, the table reports: the name and the acronym of the dataset (*Corpus*); the main source of information for the dataset (*Source*); the year the dataset was built, released to public or the year the main source of information was published (*Year*); the type of recordings (*Rec*); the language(s) of the utterances in the database (*Lang*); the number of speakers, subdivided into the number of female (F) and male (M) speakers (*Speakers*); the modality, audio (A) or audio-visual (AV), of the recordings (*Type*); the number of discrete emotions or the dimensions for continuous emotions (*Emotions*); the total number of samples (*#Sam*); the rate of recording in KHz (*Rate*); whether the dataset is publicly available (*Public*). For a more complete description of the datasets, refer to the text.

Corpus	*Source*	*Year*	*Rec.*	*Lang*	*Speakers*	*Mod.*	*Emotions*	*#Sam*	*Rate (KHz)*	*Public*
SAL (Sensitive Artificial Listeners)	[79]	2009	induced	Eng	2F, 2M	AV	Continuous theory with 3 dimensions (valence, intensity, dominance)	1692 turns		Y
NIMITEK Corpus	[40]	2010	induced	Ger	7F, 3M	AV	Discrete theory with 6 basic emotions (anger, disgust, fear, joy, neutral, sadness)	1847 turns		
AIBO Database	[2]	2004	induced	Ger	30F, 21M	A	Discrete theory with 11 basic emotions (anger, bored, empathic, helpless, joyful, motherese, neutral, reprimanding, rest, surprised, touchy)	51393 words		Y
ABC (Airplane Behaviour Corpus)	[82]	2007	induced	Ger	4F, 4M	AV	Discrete theory with 6 basic emotions (aggressive, cheerful, intoxicated, nervous, neutral, tired)	396 clips		Y
eNTERFACE	[67]	2004	induced	Eng	8F, 34M	AV	Discrete theory with 7 basic emotions (anger, disgust, fear, happiness, neutral, sadness, surprise)	1166 sequences	48	Y
SmartKom	[98]	1999-2003	induced	Ger	224	AV	Discrete theory with 7 basic emotions (anger, helplessness, joy, neutrality, pondering, surprise, unidentifiable episodes)	448 sessions	48 downsampled at 16	Y
EmoTaboo	[115]	2007	induced	Fre	4F, 6M	AV	Discrete theory with 21 basic emotions hierarchically organized.	10 clips		
Humaine Database	[29]	2008	mixed (natur., induced, acted)	Eng, Gre, Heb, Fre		AV	Continuous theory with 8 dimensions (valence, intensity, dominance, acting, masking, activation, expectation, word-relatedness)	48 clips	var	Y

Table 2.3: Induced and mixed datasets.
For each dataset, the table reports: the name and the acronym of the dataset (*Corpus*); the main source of information for the dataset (*Source*); the year the dataset was built, released to public or the year the main source of information was published (*Year*); the type of recordings (*Rec*); the language(s) of the utterances in the database (*Lang*); the number of speakers, subdivided into the number of female (F) and male (M) speakers (*Speakers*); the modality, audio (A) or audio-visual (AV), of the recordings (*Type*); the number of discrete emotions or the dimensions for continuous emotions (*Emotions*); the total number of samples (*#Sam*); the rate of recording in KHz (*Rate*); whether the dataset is publicly available (*Public*). For a more complete description of the datasets, refer to the text.

- *Classification Study*: this category embraces all the studies in which the database were used instrumentally in a study of classification (if a discrete theory of emotion was adopted) or regression (if a continuous theory of emotion was adopted) or clustering; these studies include the development of classification system, the analysis and the comparison of different algorithms for classification, the development of regression systems or the study of clustering techniques.

- *Feature Study*: this category embraces all the studies in which the database were used to extract and to generate features that could be used for emotion representation and recognition; these studies include the evaluation of techniques for extracting features, the evaluation of algorithms for selecting features, the computation of the impact of families or groups of feature on the final performance of an emotion-aware system.

- *Benchmark Study*: this category embraces all the studies in which the database were used to provide data for studies aimed at establishing the performances of emotion-aware systems; these studies include multi-datasets benchmarks, cross-corpora studies and the reports of international challenges in affective computing.

Notice that the division of the publications we considered into this three groups is not rigorous, but simply suggestive; it was undertaken in order to get a better understanding of which datasets were used in which situations. Table 2.4 reports for each database which studies were done using that database; moreover it also reports the total number of studies done on each database (in the last column *Total Studies*) and the total number of papers considered for each type of study (in the last row *Total*). Even if the sample of publications we considered is very limited and not exhaustive, it gives us some ideas on which databases are generally used by the affective computing community; in particular, we can see how the most used databases are mainly historical acted databases, such as the *Berlin Corpus* and *DES*, which constitute a sort of standard reference for many studies, and recent natural databases such as VAM and AIBO, which contain more realistic and challenging recordings.

Datasets	***Classification Studies***	***Feature Studies***	***Benchmark Studies***	*Total Studies*
Berlin	[112, 114, 70]	[92, 91, 10]	[90, 89]	*8*
DES	[93, 114, 112]	[92]	[90, 89]	*6*
AVIC			[88, 90, 81, 89]	*4*
eNTERFACE			[90, 89]	*2*
SUSAS	[114]	[57]	[90, 89]	*4*
SmartKom			[90, 89]	*2*
GEES	[93]	[92]		*2*
BabyEars		[92]		*1*
AIBO	[59]	[83, 57]	[81, 87]	*6*
SAL	[107, 108, 36]		[89]	*4*
VAM	[80, 108, 114, 42]	[109]	[89]	*6*
ABC		[91]	[89]	*2*
MindReading	[95]			*1*
Doors	[95]			*1*
CHE	[37]			*1*
BT-O	[37]			*1*
Total (24)	*13*	6	*5*	

Table 2.4: Datasets used in different studies.

Chapter 3

Emotional Speech Processing

In this chapter we are going to deal with emotional speech signals and their processing. The aim of this chapter is to give a general explanation of the nature of speech and to illustrate how emotional speech signals can be represented.

3.1 Definition of Speech

Speech is the intentional modulation of air pressure in order to transmit a message; it is a way to convey a message through compressions and rarefactions of air molecules [47]. Speech is indeed a special instance of a *sound*, that is a mechanical disturbance from a state of equilibrium propagating through an elastic medium [104].

Speech is articulated by humans through the use of the vocal apparatus and it is detected by auditive organs.

The variation in air pressure determined by speech can be studied as a waveform describing the change of pressure in time. Speech can therefore be seen as a signal having a source (i.e., vocal apparatus), being transmitted over a medium (i.e., air), encoding information (i.e., intentional meaning) and having a receiver (i.e., auditive apparatus).

As stated above, the main informational content of a speech signal, is the semantic meaning of the uttered words. However, beside the communicative explicit semantic content, other informative implicit contents are carried by a speech signal. According to the information conveyed, a speech signal can be decomposed in three layers:

- *Linguistic layer*: the linguistic layer carries the semantic content; the linguistic layer is made up by sounds constituting words; the composition of words according to the rules of language shared by the speaker and the listener, allow the speaker to communicate a message.
- *Extralinguistic layer*: the extralinguistic layer conveys information about the speaker and its background; the extralinguistic layer is made up by acoustic cues, such as tone and pitch; these cues can provide general information to the listener about the speaker, such as his gender, his age, his cultural background and other speaker-specific characteristics.

- *Paralinguistic layer*: the paralinguistic layer carries information about the intentions and the feelings of the speaker; the paralinguistic layer is made up by acoustic cues, such as stresses and pauses; these cues allow the listener to infer information about the inner state of the speaker.

It is worth noticing that this layerwise arrangement of speech is not a rigid decomposition. Often the boundaries between the layers are fuzzy. Moreover, acoustic cues can belong to more than one layer; for example, a stress can carry semantic information allowing the listener to distinguish between two homograph words (linguistic layer), it may suggest that the speaker belongs to a particular dialect group (extralinguistic layer) and it may underline the fact the speaker is concerned (paralinguistic layer).

Although these layers are approximate conceptualizations, this decomposition is very useful as it highlights the different contents of speech and it allow us to focus only that layer which is relevant for our study. Indeed, we could say that at different layers correspond different areas of research; the linguistic layer is the main field of research in *speech recognition* (i.e., reconstructing the single words uttered by a speaker starting from an acoustic signal); the extralinguistic layer is the main subject of research in *speaker recognition* (i.e., determining the identity of a speaker through the analysis of an acoustic signal); the paralinguistic layer is the main focus of research in *affective speech recognition*. However, it is worth to underline once again that the boundaries between the research areas we named are fuzzy; often each of these areas of research has to tackle complex problems which extend on layers which, strictly speaking, do not belong its focus of research (e.g.: speaker recognition techniques may be studied by affective speech recognition researchers in order to determine the baseline speech of a user and determine, therefore, his emotionally neutral state [88]). This strict interrelation between the research into speech recognition, speaker recognition and affective speech recognition makes the research very challenging and motivates the need for co-operation between the researchers; indeed, the exchange of information and results between these fields is frequent and productive.

Considering this layered structure of speech, our interest will be focused mainly on the paralinguistic layer of speech; by processing speech signals and extracting paralinguistic acoustic cues we aim at obtaining information about the emotional state of the speaker.

The acoustic cues which make up the paralinguistic layer (and the extralinguistic layer) include stress, intonation and rhythm. In general, all these non-semantic characteristics of speech which are independent of language and grammar and which still carry information to the listener are called *prosody* or *prosodic phenomena*; as prosodic phenomena are not limited to single phonetic segments of speech (e.g., syllables) but extend potentially to entire sentences, they are often called *supra-segmental phenomena* [53].

3.1.1 Prosody

Prosody is at the core of emotional speech. Every time a human listener perceives speech, he receives information about the speaker from prosody. Prosody encompasses several informative acoustic cues which shape the paralinguistic and extralinguistic layers of speech, but usually four phenomena are identified to describe the prosodic properties of speech, music and sounds in general [68, 47, 53]:

- *Intonation*: intonation describes the variation of pitch of the speaker and it is related to the vibration of vocal chords. Intonational patterns can carry communicative information (e.g., a falling or a rising intonational pattern allow us to distinguish between questions and assertions) and emotional information (e.g., falling intonational pattern may express doubt or uncertainty).

- *Stress* or *Accentuation*: stress describes which part of a word or of a sentence is emphasized by being uttered more strongly or loudly. In many languages, stress is a fixed property of words, meaning that each word has a stress falling on a given syllable; phrasal stresses usually agree with word stresses but they can be emphasized in different ways to convey diverse meaning and feelings.

- *Rhythm* or *Speech Rate*: rhythm describes the change in speed and the overall duration of speech. Varying the speed and grouping together words, a speaker can express his priorities and his feelings.

- *Voice Quality*: voice quality describes the specific qualitative characteristics of a sound produced by a particular person. Small differences in the vocal apparatus of different persons make sounds produced by different persons qualitatively different, even if all the other prosodic phenomena are identical.

All together, these phenomena describe how utterances which are identical at the linguistic layer (e.g., asking where a specific object is) may be very different on the paralinguistic layer (e.g., asking where a specific object is while being irritated or while being sad because unable to find it).

The study of prosodic phenomena is usually very complex [53]. First, even if differentiated, prosodic phenomena are not independent and a variation of one of them may affect all the others (e.g., a variation in rhythm, may cause a correlated variation in stress). Second, prosodic phenomena do not add up linearly; because they are tightly bound to each other, varying two prosodic phenomena at the same time generates an utterance which is not the simple superposition of the two prosodic phenomena. Third, prosodic phenomena may be affected by local segments, such as syllables; microprosody studies how prosodic phenomena may be changed because of particular combinations of sounds. Fourth, prosodic phenomena are tied to languages, too; even if some prosodic phenomena appear to be really cross-cultural, probably because of physiological or neurological evolution, other prosodic phenomena are strongly shaped by the culture and by the language in which they are expressed. Fifth, prosodic phenomena show high inter-subject variability; as they are generated by vocal apparatus which is different from individual to individual, prosodic phenomena conveying the same information may be different from one speaker to the other. Sixth, prosodic phenomena show also a certain degree of intra-subject variability; as vocal apparatus slightly changes in time and is affected by the environment, the same prosodic phenomena may change over time.

To sum up, acoustic signals representing the same phoneme expressed with equivalent emotions can show a high degree of variability in their prosodic phenomena.

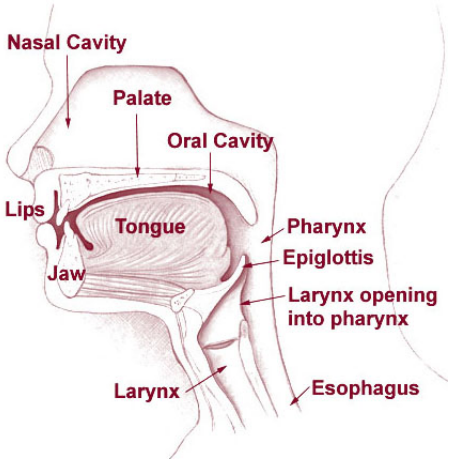

Figure 3.1: Vocal Apparatus

3.2 Speech Chain

In order to work with speech signals, it is necessary to focus on the *speech chain*, that is on that set of linked events that allow vocal communication between humans. The speech chain starts in the brain of the speaker and ends in the brain of the listener; in the following subsections we are going to analyse the two main physiological steps of the speech chain, that is production of speech and perception of speech.

3.2.1 Production of Speech

Speech is produced by the vocal apparatus, a set of organs which modulate air in order to produce sounds (see figure 3.1).

The process of speech production starts in the *lungs*. A compression of the lungs cause pressurized air to be released and flow across the *trachea*. The unperturbed mass of air put in motion by the lungs crosses the trachea until it reaches the *vocal tract*.

The vocal tract is the main part of the vocal apparatus and it embraces all the structures between the *glottis* and the *lips* where air is perturbed in order to produce sounds. In adults, the vocal tract reaches the length of 17 cm and, according to the sound to be pronounced, the cross-sectional area can vary from 0 to 20 cm^2 [74].

Traversing the glottis, the air enters the *larynx* (or *voice box*) the organ containing the *vocal folds* (or *vocal cords*); vocal cords are a set of close twin horizontal mucous membranes; when set into vibration, the vocal cords perturb the air generating a sound.

Transiting through the pharynx, the air then reaches the *oral cavity* within the mouth. In the oral cavity, we can identify four *ports* (or *valves*) that control different sections of the oral cavity; these sections work as resonating chambers which can further modulate the amplitude and the frequency of the generated sounds. The first port, is the *velum* (or *velar port*) on the soft palate, which control the access to the *nasal cavity*; when this port is open, air can flow through the velum till the nostrils and produce nasal sounds (e.g., the letters m and n). The

second and the third port are the *linguo-palatal port*, close to the palate, and the *linguo-alveolar port*, in the front side of the mouth; these two ports are mainly controlled by the tongue, which, acting like a flexible articulator, can close or open these chambers. The fourth port is the *labial port*, controlled mainly by the lips and the teeth; this port allow the production of labial sounds (e.g., the letters p and b) [53].

Through the coordinated use of all these organs, humans are able to produce sounds whose fundamental frequency lies in the range of approximately 80-200 Hz, for males, and 180-400 Hz, for females [68] and to generate a variation of pressure of approximately 0.01-1 Pa at 1 meter from their lips [104].

The generation of the basic unit of speech, phonemes (e.g., vowels or consonants), can be studied in relation to the physiological apparatus we described; vibration of vocal cords and control of ports can explain how the single phonemes are produced. However, the production of speech is not simply the linear concatenation of phonemes; the real and concrete process of uttering a word deeply affect the way phonemes are uttered to the point that the final sound can not be considered any more the simple juxtaposition of two or more elementary phonemes.

In order to study speech production more precisely, mathematical models of the vocal apparatus with different degree of abstraction can be developed. Rabiner and Schafer in [74] discuss a mathematical model in which they modelled the vocal and the nasal tract as a cylindrical tube with a non-uniform cross-sectional area and in which they took into consideration several acoustic phenomena such as variation in time of the vocal tract in shape, losses due to heat conduction and viscous friction at the vocal tract walls, softness of the vocal tract walls, radiation of sound at the lips, nasal coupling and excitation of sound in the vocal tract.

3.2.2 Perception of Speech

Speech is perceived by the auditory system, a set of organs which transduces sounds and relay a signal to the brain (see figure 3.2).

The human auditory system, usually called *ear*, can be divided into three parts: the outer ear, the middle ear and the inner ear.

The outer ear constitutes the visible part of the auditory system. The main element of this section is the *pinna* (or, in humans, *auricle*), a cartilage folded sheet which reflects and attenuates sounds. Sounds are then funnelled into the *auditory canal*, a short tube connecting the outer ear to the middle ear. The auditory canal measures about 2.5 cm in length [47] and it amplifies sound waves between 3 KHz and 12 KHz.

The end of the auditory canal is closed by the *eardrum* (or *tympanic membrane*); the eardrum is a membrane sealing the auditory canal which, upon being hit by sound pressure waves, oscillates at the same frequency of the sounds. Beyond the eardrum an air-filled cavity of about 6 cm^3 constitutes the middle ear. Within the middle ear, information about the sound wave is transferred from the eardrum to a system of three small ossicles: the *malleus* (or *hammer*), the *incus* (or *anvil*) and the *stapes* (or *stirrup*). These ossicles transfer mechanically the information

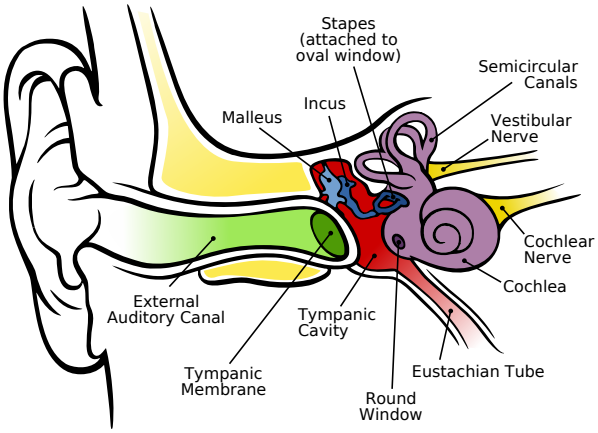

Figure 3.2: Auditory System

about the soundwave from the eardrum into the inner ear.

The ossicles of the middle ear acts on the bony *oval window* generating amplified waves. The oval window is itself the interface a more complex structure, the *cochlea*. The cochlea is a small tube of about 3.5 cm of length recoiling on itself 2.6 time and forming a small spiral [47]. Inside, the cochlea is subdivided into three sections, each filled with a fluid in which sound waves can propagate. The cochlea can be seen as a location-based (high frequencies are filtered by filters close to the apex of the cochlea, low frequencies are filtered by filters close to the base of the cochlea) filter bank that transduces mechanical vibrations into electrical impulses; the result of this conversion is then fed from the cochlea directly into the *auditory nerve* which carries the signal to brain for processing.

Thanks to the auditory system the human ear can perceive sounds between 20 Hz and 20 KHz. At different frequencies there are different minimal required intensities for a sound to be heard; the threshold of hearing (TOH) is the function determining this minimal intensity. TOH is non-linear function with its minimum between 1 KHz and 10 KHz; at 1 KHz the threshold of hearing is 0.0002 μbar.

3.3 Preprocessing of Speech Signals

Speech, and in general sound, can be modelled as an *acoustic* signal. Acoustic signals measure the amplitude of a sound as a function of time; the amplitude of a sound is a measure of the displacement of air molecules from their resting state [47]. The intensity of the displacement can be measured evaluating the air pressure (unit of measure: $[Pa]$) or evaluating the applied power per area (unit of measure: $\left[\frac{W}{m^2}\right]$). As the range of values that sound intensity can assume is very wide, usually a logarithmic scale in decibel is used; the amplitude of a sound signal is then computed as:

$$Amplitude\,(dB) = 20\log_{10}\frac{X}{X_0}$$

where X is the absolute measure of amplitude (in $[Pa]$ or $\left[\frac{W}{m^2}\right]$) and X_0 is a fixed reference value, usually assumed to be the threshold of hearing for a tone of $1\,KHz$ (i.e., $0.0002\,\mu bar$ or $10^{-12}\,\frac{W}{m^2}$) [47].

An acoustic signal can be captured using recording devices (e.g., microphones) and stored on a digital computer. Starting from an analogue signal we can rely on signal processing techniques in order to generate an accurate digital copy of the speech signal we want to process (*analogue-to-digital conversion, ADC*); the new digitized signal can then be efficiently and quickly processed using numerical software (*digital signal processing, DSP*); finally, if needed, the digital signal can be re-converted to a continuous analog signal (*digital-to-analogue conversion, DAC*).

The set of operations that leads from an analogue raw signal to a refined digital version of the same signal which can be easily processed is referred to as *pre-processing* of a signal. Pre-processing is a broad term which encompasses a series of operations which can be undertaken to modify a signal, not with the explicit aim of extracting specific information, but in order to make the signal easier to process. Some pre-processing steps which can be applied to a signal includes:

Amplification (or Enhancement) Amplification is the process in which the amplitude of a signal is increased in order to make its digitization or its analysis easier; signals of small intensity can be amplified by factors of several orders of magnitude.

Filtering Filtering is the process in which frequencies which carry no information of interest are removed from the signal; removing useless frequencies allows to filter out noise or misleading information, to reduce the size of data to be managed and to avoid aliasing effects during digitization. It is possible to implement and apply different types of filters, such as *low-pass filters* (filtering out all the frequencies over a given threshold), *high-pass filters* (filtering out all the frequencies below a given threshold), *band-pass filters* (filtering out all the frequencies outside a given interval) or *notch filters* (filtering out all the frequencies inside a given interval).

Digitization (or Sampling or Quantization) Digitization is the process in which a digital signal is generated; digitization is the core process of analogue-to-digital conversion. In order to generate a digital copy of an analogue signal we need to sample the signal in time and in amplitude. Sampling in time means selecting a frequency or a time step at which we sample the continuous speech signal; once we have determined the higher frequency we are interested to, Nyquist theorem tells us the minimal sampling frequency that we should use in order to be able to process and reconstruct the speech signal without distorting it. Sampling in amplitude means choosing a number of bits to represent the intensity of our signal; the amount of bits determines the number of intensity level we can use to encode the amplitude of the signal. Once digitized, a digital signal can undergo further rounds

of sampling (*sub-sampling*); additional steps of sampling reduce the definition and the amount of information contained in the signal, but they make the signal easier to store and faster to process.

Denoising Denoising is the process in which noise is removed or reduced in the signal. Different types of noise can affect a signal, from white random electronic noise to background human-generated noise; noise can influence the entire signal (e.g., the noise of traffic in the background) or take the form of an artefact limited in time and space (e.g., coughing). Several solutions exist to tackle this problem, from algorithms which simply remove sections of the signal hardly affected by noise to algorithms which try to untangle the noise from the signal and then try to reconstruct the original version of the signal.

Dereverberation Dereverberation is the process in which reverberation effects are removed from a signal. Reverberation is the physical phenomenon of soundwaves being reflected by walls and solid objects back on their original acoustic path with decreased amplitude [47]; because of this phenomenon, if the speaker is not in anechoic room or in free space, a microphone will record not only the original soundwave, but also the reverberated copies of it. While it is relatively easy for humans to ignore this effect, reverberation can pose serious challenges to the automatic processing of a speech signal.

Normalization Normalization is the process in which the recorded values of a signal are adjusted and aligned on a pre-determined scale. Normalization is a general-purpose statistics technique which can be applied to the amplitude of the recorded signal as well to other derived dimensions of the signal; normalization allows us to regularize our signal by mapping all the values on a fixed scale with a pre-defined mean.

These operations represent the main steps that can be implemented in order to pre-process a signal, in general, and an emotional speech signal, in particular. Even if extremely useful, no one of the above steps, with the exception of digitization, is compulsory; moreover, the order of these operations is not strict: they form the building blocks of a pipeline-like process in which the order of the single blocks can be rearranged.

3.4 Segmentation of Emotional Speech Signal

Once given a pre-processed emotional speech signal, a key decision before focusing on the problem of the representation of the emotional speech signal is the decision about the segmentation of the speech signal. Segmenting the emotional speech signal means determining which will be the atomic unit of analysis on which the following operations will be carried on. These atomic unit should be long enough to contain information related to emotions, but not too long to the point of embracing expressions of more than one emotion. The choice of an atomic unit of analysis is a compulsory choice both in *online processing* (i.e., when the emotional speech signal is recorded in real-time and segments of data must be identified and forwarded for analysis before the whole emotional speech signal has been recorded) and in *offline processing* (i.e., when the whole emotional speech signal is available and it must be segmented in order to be handled and processed).

Unfortunately, differently from other fields in which signal processing is employed, there are no fixed agreed standards on what should be the ideal dimension for emotional speech signal segments [24]. This is emphasized by the fact that during data collection there is no standard for the segmentation of the emotional speech signals; reviewing section 2.4 on the available emotional datasets, we can see that different research groups use very different temporal scales and very diverse degrees of segmentation for their recordings: from very short segments like *words* or *bursts*, through *utterances*, *sentences*, *turns,* and *recordings*, up to very long segments like *clips* or *sessions*. Given samples with so diverse levels of segmentation, it is necessary to find a common unit of segmentation which allows us to process every recording in a uniform way.

However, the problem of determining the dimension of atomic units of analysis is not a simple one; it implies selecting units which are representative of the emotions we want to study; in other words, we must find the atomic unit which carries emotional information and it is emotionally stable. At the end, choosing the right length of these atomic units entails evaluating a trade-off: we need units which are long enough to contain emotional information and to reliably apply statistical functions and, on the other hand, which are short enough to be emotionally stable. Indeed, if the units are too short, then it will be hard to extract emotional information and the application of statistical functions may return values which are not significant because of the low number of samples which can be obtained from a short segment. On the other hand, if units are too long, then multiple emotions can be expressed in the selected time frame and the application of statistical functions may return values which are not meaningful because of the non-stationariety of the emotional signal [101]. Moreover, the ideal length of the atomic unit of analysis may vary from application to application. Different emotions may show themselves on different time-scales; therefore if we are concerned with a set of emotions having a specific time-scale, we may tune the level of segmentation for our application.

In the affective computing literature, two main families of segmentation strategies have been developed:

- *Linguistically-aware Segmentation*: linguistically-aware segmentation relies on a speech recognition module and it segments emotional speech in meaningful linguistic units, such as words or sentences. Linguistically-aware segmentation proved to give good results [88, 3]; however the drawback of these techniques is that they require the support of an external module and that they may be computationally expensive; in an online scenario, waiting to have a speech recording long enough to be linguistically segmented and processing it through a speech recognition module may add a significant delay to the application [101].

- *Linguistically-agnostic Segmentation*: linguistically-agnostic segmentation relies only on temporal information to segment emotional speech recordings and it is totally unaware of the linguistic structure of speech. It is based on the idea that emotional information may be extracted reliably from the paralinguistic layer of speech. Research showed that the results obtained using linguistically-agnostic segmentation are lower than linguistically-aware segmentation; however, linguistically-agnostic segmentation still provides reliable results and it proved to be a very efficient solution from the computational point of view since it does not require additional processing of speech at the linguistic level [101].

In this research, we will focus only on the analysis of the paralinguistic layer of speech and, therefore, we will rely on linguistically-agnostic algorithms for segmenting our emotional speech data.

Linguistically-agnostic segmentation may take place in multiple steps.

When dealing with long sentences and long utterances, a first high-level segmentation is generated by the removal of pauses and silences. In this way *recordings*, *clips* and *sessions* may be reduced to a collection of *utterances* or *turns*. This first segmentation may be done manually or automatically by identifying silences in the emotional speech signal. Any time a pause lasts longer than a given threshold, the speech signal is segmented; common values for this pause threshold are 1000 ms [3] or 1500 ms [83]. More advanced methods for this high-level segmentation include techniques like Voice Activity Detection (VAD) and energy thresholding [36, 108].

Once utterances or turns have been generated, the actual linguistically-agnostic segmentation usually happens using one of the following techniques:

- *Absolute Time Intervals (ATI) Segmentation*: absolute time intervals segmentation is based on the definition of a fixed length unit of analysis; online and offline speech signals are then divided into frames of equal length [54]. This segmentation is fast and it leads to the creation of frames that, because of their uniform length, can be easily processed [91]. The minimal length and the optimal length for ATI segments are open problems. The minimal lenght, that is the shortest length that guarantees that the segment still contains some emotional information, is usually assumed to be in the order of milliseconds or tens of milliseconds [36, 89]. The optimal length, that is the ideal length that guarantess that the segment contains the maximum of emotional information, is often chosen empirically and it may vary from as little as 20 ms [36] or 25 ms [89], to 500 ms [83] and up to 5 s [54], depending on the type of analysis which will be carried out.

- *Relative Time Intervals (RTI) Segmentation*: relative time intervals segmentation is based on the definition of a fixed number of frames to be extracted from an utterance. In order to carry out this type of segmentation it is necessary for a whole utterance or a whole turn to be available; then the speech signal can be subdivided in a given number of frames. Differently from ATI segmentation, RTI segmentation generates a constant number of frames with variable length; this means that further processing must take into account this variability [91].

Notice that this two techniques can be combined together. The performances of ATI segmentation and RTI segmentation are strictly related to the length of the emotional speech signal to be processed and to the amount of information to be extracted. In the task of detection, as the length of utterances becomes longer, ATI segmentation tends to outperform RTI segmentation as it is able to generate more units of analysis; as the amount of information to be extracted from single units becomes higher, ATI segmentation and RTI segmentation tend to reach similar performances as there is a saturation of information [91].

Notice moreover, that the same segmentation technique may be applied multiple times using different parameters for segmentation. For example, ATI segmentation may be applied with

low and high parameters in order to generate very short and long segments; or, alternatively, short segments may be combined together into longer units. Applying multiple segmentation to the same data may be useful to extract different types of information from the emotional speech signal: short segments may provide local information while long segments may provide global information [92].

3.5 Representation of Emotional Speech Signals

Once we are given a pre-processed digital signal, we would like to extract from it information which is useful for the study of emotions. To do this, a pre-processed digital signal is not, for many reasons, the best description of information we can use. Therefore we look for better alternative *representations* of the emotional speech signal we want to analyse.

A *representation* is a formal, quantitative description of a phenomenon or of an aspect of a phenomenon (e.g., we can represent the movement of an object by recording its position in time); however, the same phenomenon can be described by different types of representations (e.g., we can represent the movement of an object tracking its position or computing its velocity); and, while the phenomenon described may be identical, different representations can give us different types of information (e.g., tracking the position of an object gives a scalar static information, computing the velocity of an object gives a vectorial dynamic information); mathematical transformations allow us to move from one representation to another (e.g., basic kinematic formulas teach us how to relate time, position and velocity); mathematical transformations can be composed in order to yield more complex or more abstract representations (e.g., starting from the representation of movement as position and composing two derivative transformations we obtain the representation of movement as acceleration); different types of transformations can be blindly or brute-forcedly composed to return new representations, but the biggest challenge is to define compositions of transformations which return representations which are meaningful or useful (e.g., starting from the representation of movement as position and composing more than two derivative transformations generates representations which, normally, will be meaningless or useless); however, the meaning or the usefulness of a representation is not an absolute value, but depends on the application (e.g., in some cases, starting from the representation of movement as position and composing three or four derivative transformations we obtain representations of movement as jerk or as jounce which can be useful in some scenarios).

Different representations allow us to extract or to highlight *different information* from the *same data* related with a *given phenomenon*.

In the case of emotional speech signals we consider as the *phenomenon* we want to study the emotion expressed by a user, as the *data* the raw analogical signal and as the starting *representation* the digital signal generated after pre-processing.

Figure 3.3 illustrates different types of representations which are commonly used to deal with emotional speech signals. We emphasize two main dimensions along which representations

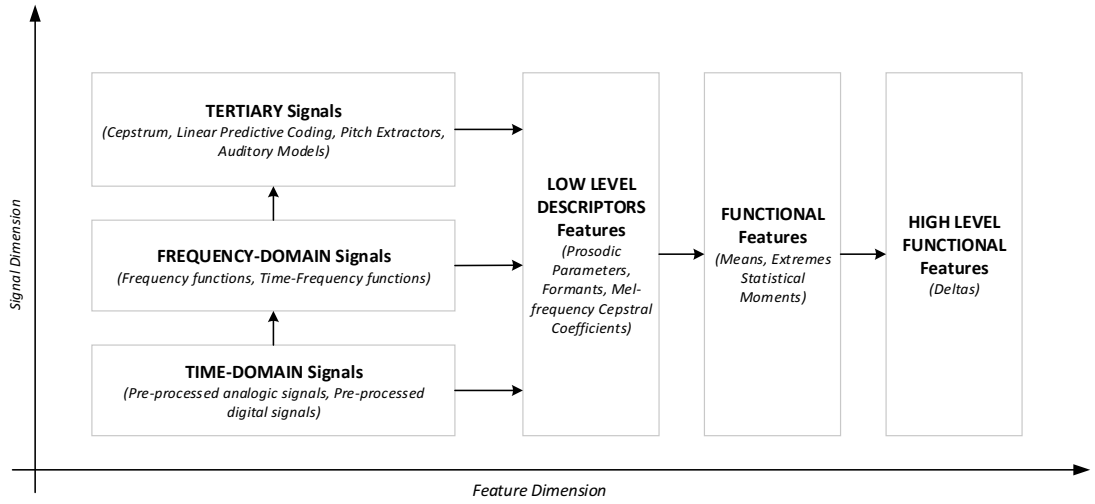

Figure 3.3: Representations of emotional speech signals

can be ordered. The first dimension is the dimension of *signals* and it is the domain of signal theory; starting from a representation as a time-domain signal and applying signal theoretic transformations (e.g., Fourier transform, short-time Fourier transform, spectral analysis) we can obtain representations as frequency-domain signals or as tertiary signals. The second dimension is the dimension of *features* and it is the domain of machine learning and statistics; starting from any signal theoretic representation and applying statistical transformations (e.g., sampling, evaluation of mean, computation of statistical moments) we can obtain representations as low-level descriptors, functionals and higher level functionals.

Notice that when working with representations in the domain of signal theory we tend to speak of representations as *signals*, while working with representations in the domain of machine learning and statistics we tend to speak of *features*. Even if at the implementation level there is no difference between signals and features, as they are both vectors or matrices of discrete values, the terminological difference highlights some conceptual differences. Signals are usually considered as general-purpose high-level uniform representations of a phenomenon which can be visualized in a low-dimensional space; features are usually sets of heterogeneous values selected in order to represent specific aspects of a phenomenon usually lying in a high-dimensional space.

In the next sections we are going to analyse the different representations we mentioned above and which are represented in figure 3.3.

3.5.1 Time-domain Signals

The first, most intuitive representation of an emotional speech signal is in the time-domain. In the time-domain a signal is represented as a variation of amplitude in time. Emotional speech signals at different stages, from raw analogical signals to refined pre-processed digital signals, can be represented in the time-domain.

The general form of an analogical continuous signal is:

$$x(t) = f(t)$$

where $f(\cdot)$ is the function of the amplitude of the emotional speech signal. Since speech signals propagate through an elastic medium, most commonly air, the variation in amplitude of the signal shows a periodic nature; in the case of air, this cyclic pattern is due to air molecules that bounding and rebounding against each other determine a cyclic variation of pressure [104]. Because of this behaviour, it is convenient to describe acoustic signals using periodic functions; the building block of any periodic signal is the sine (or cosine) wave:

$$\begin{aligned} x(t) &= A\,sin(\omega t + \phi) = A\,sin(2\pi f t + \phi) \\ x(t) &= A\,cos(\omega t + \phi) = A\,cos(2\pi f t + \phi) \end{aligned}$$

where A (in $[dB]$, in the case of speech) is the amplitude of the wave, ω (in $\left[\frac{rad}{s}\right]$) or f (in $[Hz]$) its frequency and ϕ (in $\left[\frac{rad}{s}\right]$) its phase. Pure sounds can be represented as simple sine waves. Speech is actually too complex to be represented by a single simple wave like the ones above. However, thanks to Fourier analysis, we know that we can decompose any periodic signal in a (infinite) sum of simple sine waves:

$$x(t) = \sum_{k=-\infty}^{\infty} a_k cos(2\pi k f_0 t) + b_k sin(2\pi k f_0 t) = \sum_{k=-\infty}^{\infty} c_k sin(2\pi k f_0 t + \phi_k)$$

where f_0 is the fundamental frequency and a_k, b_k and c_k are Fourier coefficients.

The general form of discrete digital signals is:

$$x[n] = x(nT) \quad n \in \mathbb{N}$$

where T is the timestep used to sample the continuous signal. Like analogical signals, discrete speech signals are cyclic and can be conceived as periodic functions made up by the discrete versions of the sine (or cosine) waves:

$$\begin{aligned} x[n] &= A\,sin[\omega nT + \phi] = A\,sin[2\pi f nT + \phi] \\ x[n] &= A\,cos[\omega nT + \phi] = A\,cos[2\pi f nT + \phi] \end{aligned}$$

where A (in $[dB]$, in the case of speech) is the amplitude of the wave, ω (in $\left[\frac{rad}{s}\right]$) or f (in $[Hz]$) its frequency and ϕ (in $\left[\frac{rad}{s}\right]$) its phase. Moreover, discrete signals too can also be decomposed in a (infinite) sum of simple sine waves:

$$x[nT] = \sum_{k=-\infty}^{\infty} a_k cos[2\pi k f_0 nT] + b_k sin[2\pi k f_0 nT] = \sum_{k=-\infty}^{\infty} c_k sin[2\pi k f_0 nT + \phi_k]$$

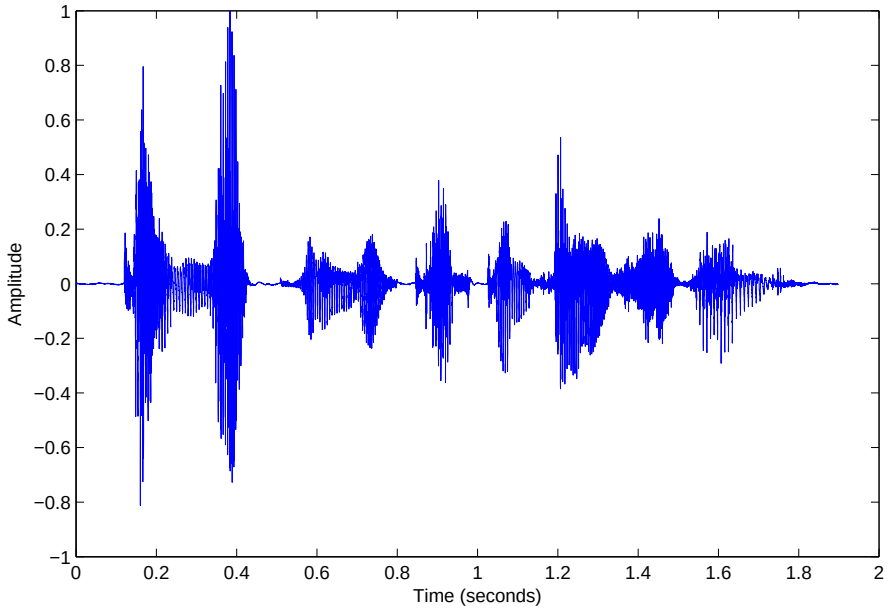

Figure 3.4: Speech signal in time domain

where a_k, b_k and c_k are Fourier coefficients.

Emotional speech signals represented in the time-domain can be easily visualized using two-dimensional plots in which the x-axis is time and the y-axis is amplitude (see figure 3.4). Time-domain representation is useful to extract information about the intensity of an emotional speech signal, its variation and its timing.

3.5.2 Frequency-domain Signals

Significant information about emotional speech is contained in the spectrum of a signal, that is, in the intensities of different frequencies. As explained in the previous section, an acoustic signal can be decomposed in a infinite sum of sine waves with different frequencies; the aim of the representation of a signal in the frequency domain is to highlight the contribution of each one of these sine waves.

The transformation which allows us to convert a time-domain representation in a frequency-domain representation is the *Fourier transform* (FT):

$$FT\{x(t)\} = X(f) = \int_{-\infty}^{+\infty} x(t)e^{-2\pi jft}dt$$

where $X(f)$ is the new representation of the signal as a function of frequency instead of time.

The inverse transformation, *Fourier inverse transform* (FT^{-1}), allows us to recover the

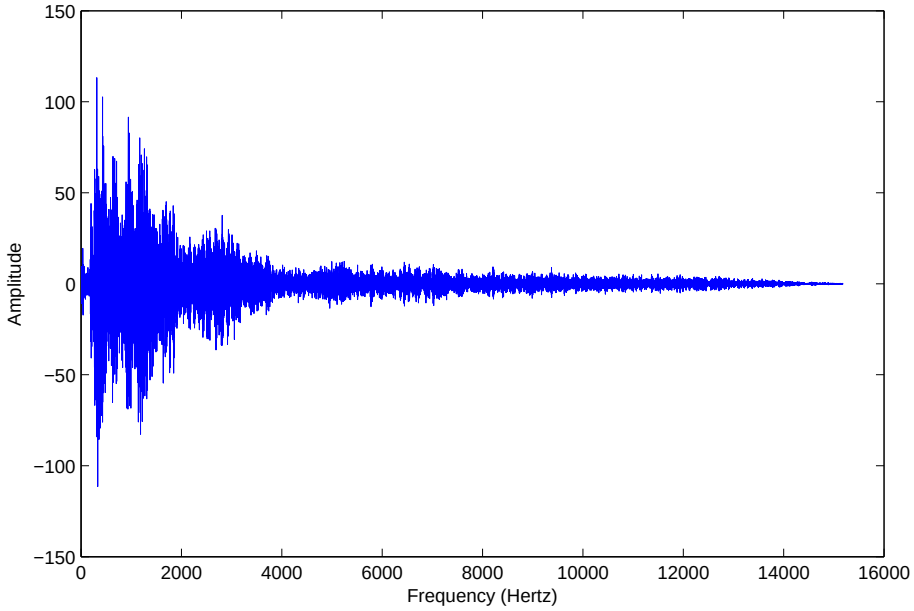

Figure 3.5: Speech signal in frequency domain

time-domain representation from the frequency-domain representation:

$$FT^{-1}\{X(f)\} = x(t) = \int_{-\infty}^{+\infty} X(f)e^{2\pi jft}df$$

where $x(t)$ is the original representation of the signal.

Fourier transform can be applied to continuous signals (*Fourier transform*) and to discrete signals (*discrete Fourier transform*, *DFT*); there exist different software implementations of the discrete Fourier transform, the most famous and used being the *fast Fourier transform* (*FFT*):

$$FFT\{x[n]\} = X[k] = \sum_{n=0}^{N-1} x[n]2^{2\pi jk\frac{n}{N}}$$

Emotional speech signals represented in the frequency-domain can be easily visualized using two-dimensional plots in which the x-axis is frequency and the y-axis is amplitude (see figure 3.5). Frequency-domain representation is useful to extract information about the spectral shape, the fundamental frequency and the formants of the emotional speech signal.

Beside focusing on the spectrum of an emotional speech signal as a whole, we could be interested in analysing how the spectrum of the signal varies in time. In this case, we need a transformation which allow us to convert our original time-domain representation in a frequency-time-domain representation. A Fourier-related transform, such as *short-time Fourier transform* (*STFT*), allows us to compute the spectral content of temporal segments of the emotional

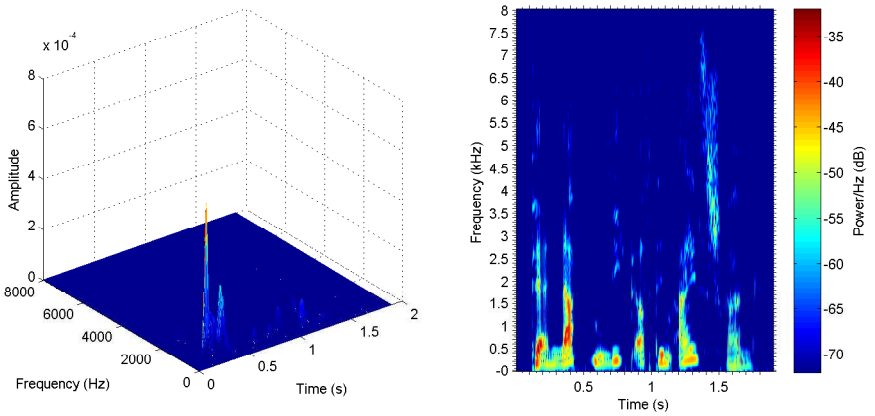

Figure 3.6: Speech signal in time-frequency domain (spectrograms). The figure on the left is a three-dimensional plot of a speech signal in the time-frequency domain; the figure on the right is a two-dimensional plot of a speech signal using colours to characterize the intensity of a signal.

speech signal we are analysing; the short-time Fourier transform relies on a windowing function to select sub-parts of the original signal to be processed:

$$STFT\{x(t)\} = X(f,\tau) = \int_{-\infty}^{+\infty} x(t)w(t-\tau)e^{-2\pi jft}dt$$

where $w(\cdot)$ is the windowing function and $X(f,\tau)$ is the new representation of the signal as a function of frequency and time.

As the original Fourier transform, the short-time Fourier transform admits an *inverse short-time Fourier transform* ($STFT^{-1}$) and a *discrete short-time Fourier transform* ($DSTFT$).

Emotional speech signals represented in the frequency-time-domain can be visualized using a spectogram. A spectogram can be drawn as a three-dimensional plot in which the x-axis is time, the y-axis is frequency and the z-axis is amplitude or it can be drawn as a two-dimensional plot in which the x-axis is time, the y-axis is frequency and colour of different intensities characterize the amplitude (see figure 3.6). Frequency-time-domain representation is useful to extract information about the variation of spectral shape in time.

Figure 3.7 summarizes the representations we discussed in this section and the transformations leading from one representation to another.

3.5.3 Tertiary Signals

From frequency-domain signals it is possible to generate more abstract representations, aimed at filtering out the specific characteristics of a speaker from the spoken content; these transformations can be very useful in the study of emotional speech signals as they allow us to

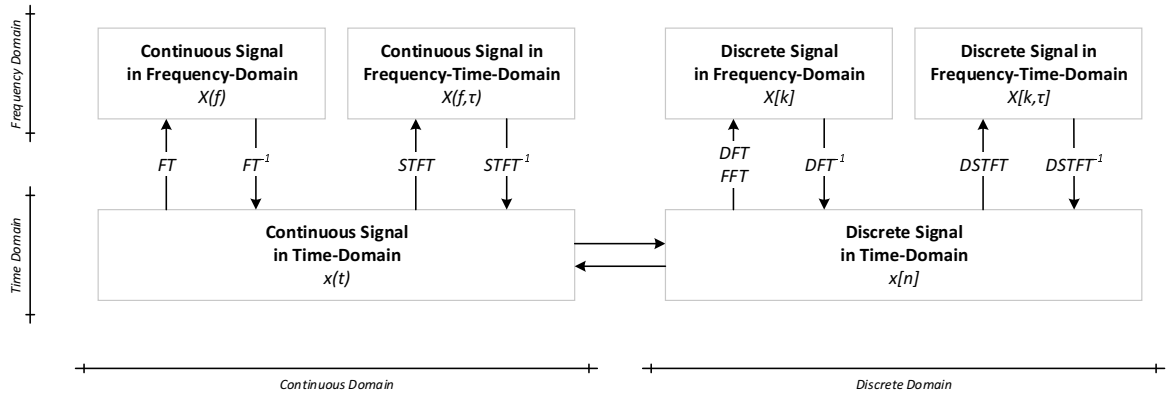

Figure 3.7: Representations in the frequency domain of emotional speech signals

extract the true emotional content of an utterance. These representations are often referred collectively as *tertiary signals*, as they are generated as a third step in the analysis of a signal [53].

In the next sections we are going to analyze two main tertiary representations which give a representation of the spectral envelope, *linear predictive coding* and *cepstrum, .* Among other tertiary representations which we will not consider in detail, we mention *pitch extractors*, aimed at highlighting and extracting pitch information from the emotional speech signal [39], *auditory models*, based on the development of models simulating the acoustic processing taking place in the human ear [63], *gammatone frequency cepstral coefficient (GFCC)* and *power normalized coefficient (PNCC)* [99].

3.5.3.1 Linear Predictive Coding

One of the most widely adopted representation in speech processing is *linear predictive coding* (*LPC*). LPC, also known as *autoregressive modelling* or *all-pole modelling*, is a powerful technique which allow the computation of a spectral envelope and the estimation of the main parameters of a speech signal [33].

LPC considers a speech signal $x[n]$ as a signal which is generated by a source $e[n]$ (i.e., the air going through the vocal cords) and which excites a linear filter $h[n]$ (i.e., the vocal upper tract); the speech signal $x[n]$ we record is:

$$x[n] = e[n] * h[n]$$

that is, the convolution of the excitation signal $e[n]$ and the linear filter $h[n]$. Because of the properties of the convolution operator, in the frequency domain, the signal $X[k]$ is:

$$X[k] = E[k]H[k]$$

that is, the product of the excitation signal $E[k]$ and the linear filter $H[k]$.

LPC works by estimating the values $\hat{E}[k]$ and $\hat{H}[k]$. First, it makes the assumption, reasonable in the case of speech, that $\hat{E}[k]$ has a flat spectral envelope, so that all the spectral information will be contained in $\hat{H}[k]$; then, it makes the assumption, reasonable for short time intervals, that $x[n]$ is stationary [72]. $\hat{H}[k]$ is then usually designed as filter modelled using resonances only, that is, as an *autoregressive* (AR) model having p poles:

$$\hat{H}[k] = \frac{X[k]}{\hat{E}[k]} = \frac{1}{1 - \sum_{i=1}^{p} a_i k^{-i}} = \frac{1}{A[k]}$$

where a_i are the AR model parameters and $A[k]$ is the *inverse filter*. The parameters of the AR model can be computed using different methods, such as *least-squares methods*. Notice that AR models are the most commonly used model in LPC, but other models such as *moving average* (MA) models or *autoregressive moving average* ($ARMA$) models can be used too.

Now, if the original signal $x[n]$ is filtered by the inverse filter $A[k]$ we obtain the excitation signal:

$$\hat{E}[k] = X[k]A[k]$$

$$\hat{e}[n] = x[n] - \sum_{i=1}^{p} a_i x[n-i]$$

The estimated excitation signal $\hat{e}[n]$ depends on the p previous values assumed by the signal $x[n]$. In this way, we recovered the excitation signal which was originally produced by the speaker before being filtered by his vocal tract [72].

Even with all the assumptions made, LPC is an efficient and effective technique to represent speech [47]; it is well-suited for slowly-varying linear filtering processes in which the filter is excited by few, short pulses [72]. However, LPC is based on a parametric model: in order to obtain a good accuracy in the representation of a signal, it is necessary to determine and set the parameters of LPC (e.g., p in the case of a AR model) to match the signal we want to describe.

3.5.3.2 Cepstrum

Another powerful and widely adopted representation for emotional speech signal is the *cepstrum* representation. Cepstrum[1] representations were developed with the aim of solving the problem of the deconvolution of two or more signals.

The (power) cepstrum representation is computed by taking the square of the inverse Fourier transform of the logarithm of the magnitude of the original spectrum:

$$\Xi[q] = \left(FFT^{-1}\{|\log(X[k])|^2\} \right)^2$$

So, computing the power spectrum amounts to applying a Fourier transform to the signal

[1] The terminology used in cepstral analysis was introduced by Bogert et al. in [5]; many of these terms are anagrams of terms used in signal theory, e.g., *cepstrum* being the anagram of *spectrum*, *quefrency* being the anagram of *frequency* and *liftering* being the anagram of *filtering*.

$x[n]$ a second time:

$$\Xi[q] = \left(FFT^{-1}\{|\log(FFT\{x[n]\})|^2\}\right)^2$$

Notice that if the inverse Fourier transform $FFT^{-1}\{\cdot\}$ were to be applied directly to the frequency-domain signal $X[k]$ it would recover the original time-domain signal $x[n]$; however, applying the module operator removes phase information from $X[k]$ and applying the logarithm operator changes the ratio between strong and weak components, so the final resulting signal $\Xi[q]$ is deeply different from $x[n]$ or $X[k]$. Notice, moreover, that the new signal highlights frequency components, even if, because of the inverse Fourier transform, it lies in the time-domain; to avoid confusion, this domain is called *quefrency* and its unit of measure is second [s].

So, suppose we are given a signal $x[n]$ which is the convolution of two signals $e[n]$ and $h[n]$:

$$\begin{aligned} x[n] &= e[n] * h[n] \\ X[k] &= E[k]H[k] \end{aligned}$$

By applying the module and the logarithm we can transform the product of the two spectra $E[k]$ and $H[k]$ in the frequency-domain in a sum:

$$\log|X[k]|^2 = \log|E[k]|^2 + \log|H[k]|^2$$

Now, by applying the inverse Fourier transform, because of the linearity of the Fourier transform, we obtain:

$$\Xi[q] = \Theta_E[q] + \Theta_H[q] + \xi$$

where ξ is a cross-product term; however, if the power cepstrum of $e[n]$ and $h[n]$ occupy different quefrencies, then $\xi = 0$ and the contribution of the two signals can be separated [14].

Applied to speech signal, cepstral processing allow us to recover an excitation signal $e[n]$ convoluted with a linear filter $h[n]$ without modelling the filter $h[n]$ as we did in LPC. Other more complete variants of the power cepstrum has also been developed, such as *phase cepstrum* and *complex cepstrum* [14].

3.5.4 Low-Level Descriptor Features

Working with signals usually turns out to be computationally expensive and inefficient. To solve this problem more aim-related representations can be devised. A way to generate new representations is to apply transformations to the signal in order to reduce it to collection of features. A *feature* is a quantitative value extracted from the signal describing a particular property or characteristic of the original signal. By computing set of features, a complex signal can be reduced to a collection of values which are easier to handle, which are sufficiently descriptive

for a specific objectives and which do not contain any redundant or misleading information.

Low-level descriptor (*LLD*) features denote those features that can be extracted from signal-theoretic representations in order to describe emotional speech signals. LLD features can be extracted by any representation of the signals we described above: it is possible to compute features from signals in the time domain, in the frequency domain or in tertiary representations. Starting from these signal-theoretic representations *feature extraction algorithms* work as transformation which can lead us to a new representation based on a set of features. It is possible to extract LLD features using expert-driven feature selection (*manual selection*) or computer-led brute-force selection (*automatic selection*). From a single signal, using automatic brute-force selection algorithms, thousands of LLD features can be extracted [91, 3]. Unfortunately, only a small subset of this huge number of features are practically useful; the vast majority of them are redundant, non-informative or, in some cases, even counterproductive. The definition and the choice of a good subset of LLD features is therefore crucial, to the point that researchers recognized that in the task of emotion classification good feature extraction is more critical than the selection of a good classifier [57]. For this reason, several studies tried to evaluate and define which LLD features are more informative and relevant in the context of analysis of emotional speech signals [83, 3, 92, 109, 113].

Given the high number of acoustic features that can be extracted from emotional speech signals, it is convenient to divide LLD features into families and groups. Table 3.1 lists the main families and the main groups of LLD features studied in literature; for each family and group notable examples of individual features are reported. Notice that this categorization of features is shared by many researchers, but the details on the specific group or family to which a certain feature belongs may vary from author to author; for example, features describing the spectrum of a signal may be attributed to the family of spectral features and to the group of spectral shape features ($|spectral|spectral\ shape$) or they may be attributed to the family of prosodic features and to the group of voice quality features ($|prosodic|voice\ quality$); or, features describing the variation of energy of a signal may be attributed to the family of prosodic features and to the group of intensity ($|prosodic|intensity$) or they may be attributed to the family of prosodic features and to the group of time features ($|prosodic|time$).

In the following sub-sections we are going to analyse more carefully the families of LLD features.

3.5.4.1 Prosodic Features

Prosodic features are extracted from prosody (see section 3.1.1). In order to describe the prosodic phenomena (*intonation, stress, rhythm and voice quality*) it is necessary to define a set of parameters; we will define two sets of parameters:

- *Physical parameters*, describing the physical properties of prosodic phenomena;

Family of Features	***Types of Features***	***Examples of Features***
Prosodic	Fundamental Frequency	F_0, characterising points, contours
	Intensity	Energy, characterising points, root mean energy
	Time	Duration, voice and unvoiced segments ratio, zero-crossing rate
	Voice Quality	Band-energies
Spectral	Formants	Formants
	Spectral Shape	Band-energies, roll-off, centroid, flux, spectral balance
Tertiary	Cepstral	Cepstral Coefficients, MFCC
	LPC	LPC Coefficients, PLPC
	Other Tertiary	Gammatone Frequency Cepstral Coefficient (GFCC) and Power Normalized Coefficient (PNCC)
Voice Source	Voice Source	Jitter, shimmer, microprosody, NHR, HRN
Wavelets	Wavelets	Band-energies, Teager energy, modulation spectograms, RASTA, Gabor features, cortical features
Harmonic	Harmonic	Filtered sub-bands amplitude, correlogram
Zipf	Zipf	Entropy of inverse Zipf of frequency coding

Table 3.1: Taxonomy of low-level descriptor features

- *Perceptual parameters* (or *psycho-acoustical parameters*), describing the perceived properties of prosodic phenomena after they have been processed by the human auditory system.

Notice that, while physical parameters and perceptual parameters are strictly related, they are not identical; for example, interacting sounds may be processed by the human auditory system in a way that the perceived properties of one sound do not correspond to its actual physical properties (*masking*) [47].

We can define the following physical and perceptual prosodic parameters:

Fundamental Frequency and Pitch The fundamental frequency (F_0) of a sound is its main frequency and it corresponds to the lowest frequency in the periodic waveform; it is the principal frequency over which all the harmonics are super-imposed. The fundamental frequency is measured in Hertz (Hz).

The perceptual correlate of the fundamental frequency is pitch. Pitch describes the perceived altitude of a sound; pitch is defined as *the auditory attribute of a sound according to which sounds can be ordered on a scale from low to high* [104].

Human can produce sounds with the fundamental frequency comprised between 80 Hz and 200 Hz, for males, or between 180 and 400 Hz, for females [68] and they can perceive sounds with a fundamental frequency between 20 Hz and 20 KHz. Pure tones with the same intensity always evoke the same sensation in human listeners [104]; however, perceived pitch may change by keeping the frequency of a sound constant and changing its intensity [47].

Intensity and Loudness The intensity of a sound is a measure of the variation of pressure generated by the soundwave and it corresponds to the amplitude of the periodic waveform. The intensity is usually measured in Pascal (Pa) using decibel units on a logarithmic scale.

The perceptual correlate of intensity is loudness. Loudness describes the perceived strength of a sound and it is usually conceived as the degree of audibility of a sound; loudness is defined as *the auditory attribute of a sound according to which sounds can be ordered on a scale from quiet to loud.*

Human can produce sounds with an intensity comprised between 0.01 Pa and 1 Pa at one meter from their lips [68] and they can perceive sounds with different intensity as a function of the frequency of a sound.

Timing and Duration The timing of a sound is a measure of the length in time of a sound or of a pause between sounds. Timing of ensemble of sounds is the base of the rhythm of longer utterances. The unit of measure of timing is second (s).

The perceptual correlate of timing is duration. Duration describes the perceived time length of single sounds and the overall rhythm of an utterance; duration can be defined as *the auditory attribute of a sound according to which sounds can be ordered on a scale from short to long.*

Humans are very sensitive to the variation of temporal patterns; it was shown that

Prosodic Phenomenon	***Perceptual Parameter***	***Physical Parameter [Unit of Measure]***	***Prosodic group of LLD features***
Intonation	Pitch	Fundamental Frequency $[Hz]$	Fundamental Frequency
Stress/Accentuation	Loudness/Energy	Intensity/Amplitude $[Pa]$	Energy
Rhythm	Timing/Duration	Onset-Offset Time $[s]$	Duration
Voice Quality	Timbre	Spectral Shape $[-]$	Voice Quality

Table 3.2: Relationships between prosodic phenomena, perceptual prosodic parameters, physical prosodic parameters and prosodic LLD features (adapted from [47])

infants are able to distinguish among rhythmic patterns even without any sort of linguistic knowledge of the uttered words [68].

Spectral Shape and Timbre The spectral shape of a sound is the general shape of a soundwave in the frequency domain; the spectral shape is determined by many variables such as the spectral power distribution, the temporal envelope, rate and depth of amplitude of frequency modulation and degree of inharmonicity of the harmonics [47]. No single unit of measure is defined to quantify the spectral shape.
The perceptual correlate of spectral shape is timbre. Timbre describes the qualitative nature of a sound. Timbre is also called tone quality, tone colour or tone register. Timbre is defined as *the auditory attribute in terms of which two sounds having the same loudness and the same pitch can be judged dissimilar.*
In the case of music, timbre is related to the quality of a sound and it allows us to distinguish sounds with same pitch and same loudness coming from two different instruments, such as a wind instrument and a string instrument. In the case of speech, timbre is related to the quality of voice and it allows us to distinguish between different voices with the same pitch and same loudness.

Now, the family of *Prosodic LLD* features is usually extracted by taking into consideration physical prosodic parameters. In this family of features we can identify four different groups of LLD features: *Fundamental frequency LLD, Energy LLD, Duration LLD* and *Voice Quality LLD.* Table 3.2, modelled after [47], summarizes how prosodic phenomena, physical prosodic parameters, perceptual prosodic parameters and prosodic LLD features are related to each other.

For each group of LLD features several different features can be computed [83]:

Fundamental frequency LLD Fundamental frequency LLD features describe the fundamental frequency of a signal; examples of these features include the value of F_0, intervals, characterising points and contours.

Energy LLD Energy LLD features describe the energy of a signal; examples of these features include amplitude at different frequencies, energy, root mean energy and characterising

points.

Duration LLD Duration LLD features describe the evolution of a signal in time; examples of these features include duration of the signal, duration of voiced segments, ratio between voiced and unvoiced segments, zero crossing rate, position of prominent events on the time axis.

Voice quality LLD Voice quality LLD features describe the specific acoustic properties of a voice signal; examples of these features include values of the spectrum in different bands.

All these parameters are widely used to represent the content of emotional speech signals; different studies, such as [83] or [109], proved the high representative power of these features, especially of fundamental frequency LLD features and energy LLD features.

3.5.4.2 Spectral Features

The family of *Spectral LLD* features is extracted from the frequency-domain representation of a signal. In this family of features we can identify two different groups of LLD features: *Formants LLD* and *Spectral Shape LLD*. For each group of LLD features several different features can be computed:

Formants LLD Formants LLD features describe the spectral maxima of a speech signal; examples of these features include the value of specific formants.

Spectral Shape LLD Spectral Shape LLD features describe the spectrum of a signal; examples of these features include values of the spectrum in different bands, spectral roll-off (measure of the steepness of a transition in the frequency domain), spectral centroid (measure of the "centre of mass" of the spectrum) and spectral flux (measure of the speed of variation of the power spectrum).

3.5.4.3 Tertiary Features

The family of *Tertiary LLD* features is extracted from the tertiary representations of a signal. In this family of features we can identify two different groups of LLD features: *Cepstral LLD* and *Linear Predictive Coding LLD*. For each group of LLD features several different features can be computed:

Cepstral LLD Cepstral LLD features describe the cepstrum of a speech signal; examples of these features include the value of cepstral coefficients, Mel-frequency spectrum coefficient MFCC (obtained by applying a cepstral transformation to a spectrum mapped on the Mel-scale, that is a logarithmic perceptual frequency scale modelled after the human ear system) and Mel FilterBank MFB.

Linear Predictive Coding LLD Linear Predictive Coding LLD features describe the LPC of a signal; examples of these features include the Linear Predictive Coding (LPC) coefficients and the Perceptual Linear Predictive (PLP) Coding coefficients (obtained by mapping a spectrum on the Bark-scale, that is a non-linear perceptual frequency scale modelled

after the human ear system)

Other Tertiary LLD Other Tertiary LLD features describe the spectral envelope of a signal; examples of these features include Gammatone Frequency Cepstral Coefficient (GFCC) and Power Normalized Coefficient (PNCC) [99].

3.5.4.4 Voice Source Features

The family of *Voice Source LLD* features is extracted by the analysis of the recorded voices. In this family of features we identify a single group of LLD features, in which we can list several different features:

Voice Source LLD Voice Source LLD features describe the characteristics of the recorded voices; examples of these features include jitter (measured as the undesired deviation from a periodic signal), shimmer (measured as the variability in peak-to-peak amplitude), noise-to-harmonic ratio (NHR), harmonic-to-noise ratio (HNR), microprosody parameters and glottal-to-noise excitation.

3.5.4.5 Wavelet Features

The family of *Wavelet LLD* features is extracted computing the wavelet decomposition of the speech signal. In this family of features we identify a single group of LLD features, in which we can list several different features:

Wavelet LLD Wavelet LLD features describe the contribution of different frequency bands to the final signal; examples of these features include amplitude in different bands, Teager energy, modulation spectograms, RASTA, Gabor features, cortical features [99].

3.5.4.6 Harmonic Features

The family of *Harmonic LLD* features is extracted computing the variation of energy as a function of the frequency [112]. In this family of features we identify a single group of LLD features, in which we can list several different features:

Harmonic LLD Harmonic LLD features describe the contribution of timbre and energy to the signal; examples of these features include the filtered sub-band amplitudes and correlograms [99].

3.5.4.7 Zipf Features

The family of *Zipf LLD* features is extracted coding the audio signal and evaluating the Zipf law and the inverse Zipf law [112]. In this family of features we identify a single group of LLD features, in which we can list several different features:

Zipf LLD Zipf LLD features describe a coding of the signal according to the Zipf law and the inverse Zipf law; examples of these features include the entropy of inverse Zipf of frequency coding and the resampled polynomial estimation Zipf of UFD coding.

3.5.5 Functional Features

The representation of information as a collection of features can be further refined; indeed, sometimes useful information is not contained in plain collections of features, but in specific or aggregate values of a set of features. One step above LLD features representation, there are functional features representation. Functional features representation are computed by applying *functionals or statistical operators* to set or subset of features. These transformations synthesize the information contained in a bunch of features in a single value.

As in the case of LLD features, it is possible to enumerate a high number of operators which can be used to generate functional features. For this reason it is practical to divide functionals into families [83]:

Extremes Extreme functionals are used to extract values related to the minima and the maxima of a signal; they include the value of minima and maxima, their position, their duration and their slope.

Percentiles Percentile functionals are used to extract values in a given percentile; they include the values of upper and lower quartiles and the values of other percentiles.

Means Mean functionals are used to extract averaged values of a signal; they include the arithmetic mean and the centroid.

Higher statistical moments Higher statistical moment functionals are used to extract statistical values of a signal; they include standard deviation, variance, skewness and kurtosis.

Specific functions Specific function functionals is a broad term for functionals which do not belong to any of the aforementioned classes; they include complex statistical operators, ratio, error measures, linear or quadratic regression coefficients, discrete cosine transform (DCT) coefficients.

3.5.6 High-Level Functional Features

More abstract representations of the information we are processing can be obtained by generating high-level functional features. *High-level functional features* are computed by applying transformations to a set or a subset of functional features. In this way, it is possible to generate features which contain global information about groups of functional features.

Widely used high-level functional features are *delta features* (Δ). Delta features are computed by evaluating the difference between values of a same feature sampled at different times; delta features are a kind of dynamic features which can be used to estimate the variation of a feature in time.

Bibliography

[1] James R. Averill. *A Semantic Atlas of Emotional Concepts.* American Psychological Association, 1975.

[2] A. Batliner, C. Hacker, S. Steidl, E. Noth, M. Russell, and M. Wong. "you stupid tin box" - children interacting with the aibo robot: a cross-linguistic emotional speech corpus. In *Proceedings of International Conference on Language Resources and Evaluation*, 2004.

[3] Anton Batliner, Stefan Steidl, Bjorn Schuller, Dino Seppi, Thurid Vogt, Johannes Wagner, Laurence Devillers, Laurence Vidrascu, Vered Aharonson, Loic Kessous, and Noam Amir. Whodunnit - searching for the most important feature types signalling emotion-related user states in speech. *Computer Speech & Language*, 25:4–28, 2011.

[4] Pascal Belin, Sarah Fillion-Bilodeau, and Frederic Gosselin. The montreal affective voices: a validated set of nonverbal affect bursts for research on auditory affective processing. *Behavior Research Methods*, 40:531–539, 2008.

[5] B. P. Bogert, M. J. Healy, and J. W. Tukey. The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. In *Proceedings of Symposyum on Time Series Analysis*, 1963.

[6] Margaret M. Bradley. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25:49–59, 1994.

[7] Margaret M. Bradley and Peter J. Lang. *Cognitive Neuroscience od Emotion*, chapter Measuring Emotion: Behaviour, Feeling, and Physiology, pages 242–276. Oxford University Press, 2000.

[8] Matyas Brendel, Riccardo Zaccarelli, Bjorn Schuller, and Laurence Devillers. Towards measuring similarity between emotional corpora. 2010.

[9] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F. Sendlmeier, and Benjamin Weiss. A database of German emotional speech. In *Interspeech*, volume 5, pages 1517–1520, 2005.

[10] Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 17:582–596, 2009.

[11] Rafael A. Calvo and Sidney D'Mello. Affect detection: An interdisciplinary review of models, methods, and their application. *IEEE Transactions on Affective Computing*, 1:18–37, 2010.

[12] Nick Campbell. The recording of emotional speech - jst/crest database research. 2002.

[13] Georgios Charopoulos. Speech emotion analysis, detection and recognition. Technical report, University of Manchester, 2011.

[14] Donald G. Childers, David P. Skinner, and Robert C. Kemerait. The cepstrum: a guide to processing. *Proceedings of the IEEE*, 65(10):1428–1443, 1977.

[15] Jason Clark and Amadeus Magrabi. Discrete vs dimensional theories of emotions, 2010.

[16] Randolph R. Cornelius. *The Science of Emotion: Research and Tradition in the Psychology of Emotion.* Pearson, 1995.

[17] R. Cowie, E. Douglas-Cowie, N Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor. *Emotion Recognition in Human-Computer Interaction.* IEEE Signal Processing Magazine, 2001.

[18] Roddy Cowie and Randolph R. Cornelius. Describing the emotional states that are expressed in speech. *Speech Communication*, 40:5–32, 2003.

[19] Roddy Cowie, Ellen Douglas-Cowie, and Cate Cox. Beyond emotion archetypes: Databases for emotion modeling using neural networks. *Neural Networks*, 18:371–388, 2005.

[20] Roddy Cowie, Ellen Douglas-Cowie, Sneddon Ian, Anton Batliner, and Catherine Pelachaud. *Emotion-Oriented Systems - The HUMAINE Handbook*, chapter Principles and History, pages 167–196. Springer, 2011.

[21] Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou, Edelle McMahon, Martin Sawey, and Marc Schröder. Feeltrace: An instrument for recording perceived emotion in real time. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 19–24, 2000.

[22] Charles Darwin. *Expression of Emotions in Man and Animals.* John Murray, 1872.

[23] Richard J. Davidson. Anterior cerebral asymmetry and the nature of emotion. *Brain and Cognition*, 20:125–151, 1992.

[24] Jun Deng, Zixing Zhang, Erik Marchi, and Bjorn Schuller. Sparse autoencoder-based feature transfer learning for speech emotion recognition. In *Proceedings of Affective Computing and Intelligent Interaction*, 2013.

[25] Laurence Devillers, Sarkis Abrilian, and Jean-Claude Martin. Representing real-life emotions in audiovisual data with non basic emotional patterns and context features. In *Proceedings of Affective Computing and Intelligent Interaction*, 2005.

[26] Laurence Devillers and Jean-Claude Martin. Coding emotional events in audiovisual corpora. In *Proceedings of International Conference on Language Resources and Evaluation*, 2008.

[27] Laurence Devillers, Laurence Vidrascu, and Lori Lamel. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18:407–422, 2005.

[28] Thomas Dixon. *From Passions to Emotions The Creation of a Secular Psychological Category.* Cambridge University Press, 2003.

[29] Ellen Douglas-Cowie. Final report on wp5. Technical report, HUMAINE, 2008.

[30] Peter J. Durston, Mark Farrell, David Attwater, James Allen, Hong-Kwang Jeff Kuo, Mohamed Afify, Eric Fosler-Lussier, and Chin-Hui Lee. Oasis natural language call steering trial. In *Proceedings of Interspeech*, 2001.

[31] P. Ekman and W. V. Friesen. Pan-cultural elements in facial displays of emotions. *Science*, 164:86–88, 1969.

[32] Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6:169–200, 1992.

[33] Daniel P. W. Ellis. *The Handbook of Phonetic Science*, chapter An Introduction to Signal Processing for Speech. John Wiley & Sons, 2009.

[34] Phoebe C. Ellsworth and Klaus R. Scherer. *Handbook of Affective Sciences*, chapter Appraisal processes in emotion, pages 572–595. Oxford University Press, 2003.

[35] Inger Samsø Engberg and Anya Varnich Hansen. Documentation of the danish emotional speech database. Technical report, Aalborg University, 2007.

[36] Florian Eyben, Martin Wollmer, Alex Graves, Bjorn Schuller, Ellen Douglas-Cowie, and Roddy Cowie. On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues. *Journal on Multimodal User Interfaces*, 3:7–19, 2010.

[37] Raul Fernandez. *A Computational Model for Automatic Recognition of Affect in Speech.* PhD thesis, MIT, 2004.

[38] Nico H. Frijda. *The Emotions.* Cambridge University Press, 1986.

[39] David Gerhard. Pitch extraction and fundamental frequency: History and current techniques. Technical report, University of Regina, 2003.

[40] Milan Gnjatovic and Dietmar Rosner. Inducing genuine emotions in simulated speech-based human-machine interaction: the nimitek corpus. *IEEE Transactions on Affective Computing*, 1:132–144, 2010.

[41] Ofer Golan, Simon Baron-Cohen, and Jacqueline Hill. The cambridge mindreading (cam) face-voice battery: Testing complex emotion recognition in adults with and without asperger syndrome. *Journal of Autism and Developmental Disorders*, 36:169–183, 2006.

[42] Michael Grimm, Kristian Kroschel, Emily Mower, and Shrikanth Narayanan. Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, 49:787–800, 2007.

[43] Michael Grimm, Kristian Kroschel, and Shrikanth Narayanan. The vera am mittag german audio-visual emotional speech database. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2008.

[44] Hatice Gunes, Bjorn Schuller, Maja Pantic, and Roddy Cowie. Emotion representation, analysis and synthesis in continuous space: A survey. In *Proceedings of International Workshop on Emotion Synthesis, Representation, and Analysis in Continuous Space*, 2011.

[45] John H. L. Hansen and Sahar E. Bou-Ghazale. Getting started with susas: a speech under simulated and actual stress database. In *Proceedings of Eurospeech*, 1997.

[46] LDC Consortium http://www.ldc.upenn.edu.

[47] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon. *Spoken Language Processing*. Prentice Hall PTR, 2001.

[48] Hayley Hung and Gokul Chittaranjan. The idiap wolf corpus: Exploring group behaviour in a competitive role-playing game. In *Proceedings of International Conference on Multimedia*, 2010.

[49] Philip Jackson and Sanaul Haq, 2010.

[50] William James. What is an emotion? *Mind*, 9:188–205, 1884.

[51] Slobodan T. Jovicic, Zorka Kasic, Miodrag Dordevic, and Mirjana Rajkovic. Serbian emotional speech database: Design, processing and evaluation. In *Proceedings SPECOM*, 2004.

[52] Arvid Kappas. Smile when you read this, wheter you like it or not: Conceptual challenges to affect detection. *IEEE Transactions on Affective Computing*, 1:38–41, 2010.

[53] Eric Keller. *Fundamentals of Speech Synthesis and Speech Recognition*. John Wiley & Sons, 1994.

[54] Samuel Kim, Panayiotis G. Georgiou, Sungbok Lee, and Shrikanth Narayanan. Real-time emotion detection system using speech: Multi-modal fusion of different timescale features. In *Proceedings of Multimedia Signal Processing*, 2007.

[55] Shinobu Kitayama, Hazel Rose Markus, and Hisaya Matsumoto. *Self-Conscious Emotions: The Psychology of Shame, Guilt, Embarrassment, and Pride*, chapter Culture, Self, and Emotion: A Cultural Perspective on Self-Conscious Emotions, pages 439–464. Guilford Press, 1995.

[56] Simo Knuuttila. *Emotions in Ancient And Medieval Philosophy*. Oxford University Press, 2006.

[57] Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, and Te-Won Lee. Emotion recognition by speech signals. In *Proceedings of Eurospeech 2003*, 2003.

[58] Joseph Ledoux. *The Emotional Brain: The Mysterious Underpinnings of Emotional Life.* Simon & Schuster, 1998.

[59] Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. Emotion recognition using a hierarchical binary decision tree approach. In *Proceedings of Interspeech*, 2009.

[60] Jerry Lin, Marc Spraragen, and Michael Zyda. Computational models of emotion and cognition. *Advances in Cognitive Systems*, 2:59–76, 2012.

[61] Catherine Lutz. The domain of emotion words on ifaluk. *American Ethnologist*, 9:113–128, 1982.

[62] Catherine Lutz and Geoffrey M White. The anthropology of emotions. *Annual Review of Anthropology*, 15:405–436, 1986.

[63] Richard F. Lyon, Andreas G. Katsiamis, and Emmanuel M. Drakakis. History and future of auditory filter models. In *Proceedings of IEEE International Symposium on Circuits and Systems*, 2010.

[64] Veronika Makarova and Valery A. Petrushin. Ruslana: a database of russian emotional utterances. In *Proceedings of International Conference on Spoken Language Processing*, 2002.

[65] George Mandler. *Mind and body: Psychology of emotion and stress.* Norton, 1984.

[66] Stacy Marsella, Jonathan Gratch, and Paolo Petta. *Computational Models of Emotion*, chapter Computational Models of Emotion, pages 21–47. Oxford University Press, 2010.

[67] O. Martin, I. Kotsia, B. Macq, and I. Pitas. The enterface 05 audio-visual emotion database. In *Proceedings of IEEE Workshop on Multimedia Database Management*, 2006.

[68] Leena Mary. *Extraction and Representation of Prosody for Speaker, Speech and Language Recognition*, chapter Significance of Prosody for Speaker, Language and Speech Recognition, pages 1–18. Springer New York, 2012.

[69] Abraham Maslow. A theory of human motivation. *Psychological Review*, 50:370–396, 1943.

[70] Stavros Ntalampiras and Nikos Fakotakis. Modeling the temporal evolution of acoustic parameters for speech emotion recognition. *IEEE Transactions on Affective Computing*, 3:116–125, 2012.

[71] Andrew Ortony, Gerald L. Clore, and Allan Collins. *The Cognitive Structure of Emotions.* Cambridge University Press, 1994.

[72] Douglas O'Shaughnessy. Linear predictive coding. *IEEE Potentials*, 7:29–32, 1988.

[73] Robert Plutchik. *Emotion: A Psychoevolutionary Synthesis.* Harper and Row, 1980.

[74] Lawrence R. Rabiner and Ronald W. Schafer. *Digital Signal Processing of Speech Signals.* Prentice Hall, 1978.

[75] William M. Reddy. *The Navigation of Feeling: A Framework for the History of Emotions.* Cambridge University Press, 2001.

[76] James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 1980.

[77] James A. Russell. Pancultural aspects of the human conceptual organization of emotions. *Journal of Personality and Social Psychology*, 45:1281–1288, 1983.

[78] James A. Russell. Core affect and the psychological construction of emotion. *Psychological Review*, 110:145–172, 2003.

[79] Marc Schroder, Elisabetta Bevacqua, Roddy Cowie, Florian Eyben, Hatice Gunes, Dirk Heylen, Mark ter Maat, Gary McKeown, Sathish Pammi, Maja Pantic, Catherine Pelachaud, Bjorn Schuller, Etienne de Sevin, Michel Valstar, and Martin Wollmer. Building autonomous sensitive artificial listeners. *IEEE Transactions on Affective Computing*, 3:165–183, 2012.

[80] Bjorn Schuller. Recognizing affect from linguistic information in 3D continuous space. *IEEE Transactions on Affective Computing*, 2:192–205, 2011.

[81] Bjorn Schuller. The computational paralinguistics challenge. *IEEE Signal Processing Magazine*, pages 97–101, 2012.

[82] Bjorn Schuller, Dejan Arsic, Gerhard Rigoll, Matthias Wimmer, and Bernd Radig. Audiovisual behaviour modeling by combined feature spaces. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2007.

[83] Bjorn Schuller, Anton Batliner, Dino Seppi, Stefan Steidl, Thurid Vogt, Johannes Wagner, Laurence Devillers, Laurence Vidrascu, Noam Amir, Loic Kessous, and Vered Aharonson. The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals. In *Proceedings of Interspeech*, 2007.

[84] Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9):1062–1087, 2011.

[85] Bjorn Schuller, Ellen Douglas-Cowie, and Anton Batliner. Guest editorial: Special section on naturalistic affect resources for system building and evaluation. *IEEE Transactions on Affective Computing*, 3:3–4, 2012.

[86] Bjorn Schuller, Ronald Muller, Florian Eyben, Jurgen Gast, Benedikt Hornler, Martin Wollmer, Gerhard Rigoll, Anja Hothker, and Hitoshi Konosu. Being bored? recognising natural interest by extensive audiovisual integration for real-life application. *Image and Vision Computing*, 27:1760–1774, 2009.

[87] Bjorn Schuller, Stefan Steidl, and Anton Batliner. The interspeech 2009 emotion challenge. In *Proceedings of Interspeech*, 2009.

[88] Bjorn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Muller, and Shrikanth Narayanan. Paralinguistics in speech and language - state of the art and the challenge. *Computer Speech and Language*, 27:4–39, 2013.

[89] Bjorn Schuller, Bogdan Vlasenko, Florian Eyben, Gerhard Rigoll, and Andreas Wendemuth. Acoustic emotion recognition: A benchmark comparison of performances. In *Automatic Speech Recognition & Understanding*, pages 552–557, 2009.

[90] Bjorn Schuller, Bogdan Vlasenko, Florian Eyben, Martin Wollmer, Andre Stuhlsatz, Andreas Wendemuth, and Gerhard Rigoll. Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing*, 1:119–131, 2010.

[91] Bjorn Schuller, Matthias Wimmer, Lorenz Mosenlechner, Christian Kern, Dejan Arsic, and Gerhard Rigoll. Brute-forcing hierarchical functionals for paralinguistics: A waste of feature space? In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2008.

[92] Arslan Shaukat and Ke Chen. Exploring language-independent emotional acoustic features via feature selection.

[93] Arslan Shaukat and Ke Chen. Emotional state categorization from speech: Machine vs. human. 2009.

[94] Malcolm Slaney and Gerald McRoberts. Baby ears: A recognition system for affective vocalization, 1998.

[95] Tal Sobol-Shikler. Analysis of affective expression in speech. Technical report, University of Cambridge, 2009.

[96] Silke Steininger, Susen Rabold, Olga Dioubina, and Florian Schiel. Development of the user-state conventions for the multimodal corpus in smartkom. In *Proceedings of Workshop on Multimodal Resources and Multimodal Systems Evaluation*, 2002.

[97] Auke Tellegen, David Watson, and Lee Anna Clark. On the dimensional and hierarchical structure of affect. *Psychological Science*, 10:297–303, 1999.

[98] Ulrich Turk. The technical processing in smartkom data collection: a case study. In *Proceedings of Eurospeech*, 2001.

[99] Maarten Van Segbroeck, Andreas Tsiartas, and Shrikanth Narayanan. A robust frontend for vad: exploiting contextual, discriminative and spectral cues of human voice. In *Proceedings of Interspeech*, 2013.

[100] Dimitrios Ververidis and Constantine Kotropoulos. A state of the art review on emotional speech databases. In *Proceedings of Richmedia Conference*, 2003.

[101] Thurid Vogt, Elisabeth André, and Johannes Wagner. *Affect and Emotion in HCI*, chapter Automatic Recognition of Emotions from Speech: A Review of the Literature and Reccomendations for Practical Realisation, pages 75–91. Springer, 2008.

[102] David Watson, Anna Lee Clark, and Auke Tellegen. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of Personality and Social Psychology*, 54:1063–1070, 1988.

[103] David Watson and Auke Tellegen. Toward a consensul structure of mood. *Psychological Bulletin*, 98:219–235, 1985.

[104] David Weenink. *Speech Signal Processing with Praat*, 2013.

[105] T. Wehrle and K. R. Scherer. *Toward Computational Modeling of Appraisal Theories*, chapter 20, pages 350–365. Oxford University Press, 2001.

[106] Cynthia M. Whissel. *Emotion: Theory, Research and Experience: Vol. 4, The Measurement of Emotions*, chapter The dictionary of affect in language. Academic Press, 1989.

[107] Martin Wollmer, Florian Eyben, Stephan Reiter, Bjorn Schuller, Cate Cox, Ellen Douglas-Cowie, and Roddy Cowie. Abandoning emotion classes - toward continuous emotion recognition with modelling of long-range dependencies. In *Proceedings of Interspeech 2008*, 2008.

[108] Martin Wollmer, Florian Eyben, Bjorn Schuller, Ellen Douglas-Cowie, and Roddy Cowie. Data-driven clustering in emotional space for affect recognition using discriminatively trained lstm networks. In *Proceedings of Interspeech 2008*, 2009.

[109] Dongrui Wu, Thomas D. Parsons, and Shrikanth S. Narayanan. Acoustic feature analysis in speech emotion primitives estimation. In *INTERSPEECH*, pages 785–788, 2010.

[110] Tian Wu, Yingchun Yang, Zhaohui Wu, and Dongdong Li. Masc: A speech corpus in mandarin for emotion analysis and affective speaker recognition. In *Speaker and Language Recognition Workshop*, 2006.

[111] Wilhelm Wundt. *Fundamentals of Physiological Psychology*. Engelmann, 1905.

[112] Zhongzhe Xiao, Emmanuel Dellandrea, Weibei Dou, and Liming Chen. Automatic hierarchical classification of emotional speech. In *Proceedings of IEEE International Symposium on Multimedia*, 2007.

[113] Serdar Yildrim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Carlos Busso, Zhigang Deng, Sungbok Lee, and Shrikanth Narayanan. An acoustic study of emotions expressed in speech. In *Proceedings of Interspeech*, 2004.

[114] Sungrack Yun and Chang D. Yoo. Loss-scaled large-margin gaussian mixture models for speech emotion classification. *IEEE Transactions on Audio, Speech, and Language Processing*, 20:585–598, 2012.

[115] Aurelie Zara, Valerie Maffiolo, Jean-Claude Martin, and Laurence Devillers. Collection and annotation of a corpus of human-human multimodal interactions: Emotion and others anthropomorphic characteristics. In *Proceedings of Affective Computing and Intelligent Interaction*, 2007.