Part I
Foundations

1

Introduction

1.1 What is Computational Paralinguistics? A First Approximation

So difficult it is to show the various meanings and imperfections of words when we have nothing else but words to do it with.

(John Locke)

The term *computational paralinguistics* is not yet a well-established term, in contrast to *computational linguistics* or even *computational phonetics*; the reader might like to try comparing the hits for each of these terms – or for any other combination of 'computational' with the name of a scientific field such as psychology or sociology – in a web search. This terminological gap is a little puzzling given the fact that there is a plethora of studies on, for example, affective computing (Picard 1997) and speech – which can partly be conceived as a sub-field of computational paralinguistics (as far as speech and language are concerned). But let us first take a look at the coarse meanings of the two words this term consists of: 'computational' and 'paralinguistics'.

Here, 'computational' means roughly that something is done by a computer and not by a human being; this can mean analysing the phenomenon in question, or generating humanlike behaviour. Note that nowadays computers are used for practically all systematic and scientific work, even if it is only for listing data, detailed information on subjects, or annotations in an ASCII (American Standard Code for Information Exchange) file. In traditional phonetic or psychological approaches, this can go along with the use of highly sophisticated signal extraction and statistical programs. A borderline between the 'simple' use of computers for tedious work and the use of computers for actually modelling and performing human behaviour is of course difficult to define. Here, we simply mean both: doing the work with the help of computers, and letting computers do the work of analysing and processing.

'Paralinguistics' means 'alongside linguistics' (from the Greek preposition $\pi\alpha\rho\alpha$); thus the phenomena in question are not typical linguistic phenomena such as the structure of a language, its phonetics, its grammar (syntax, morphology), or its semantics. It is concerned with *how* you say something rather than *what* you say.

Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing, First Edition.
Björn W. Schuller and Anton M. Batliner. © 2014 John Wiley & Sons, Ltd. Published 2014 by John Wiley & Sons, Ltd.

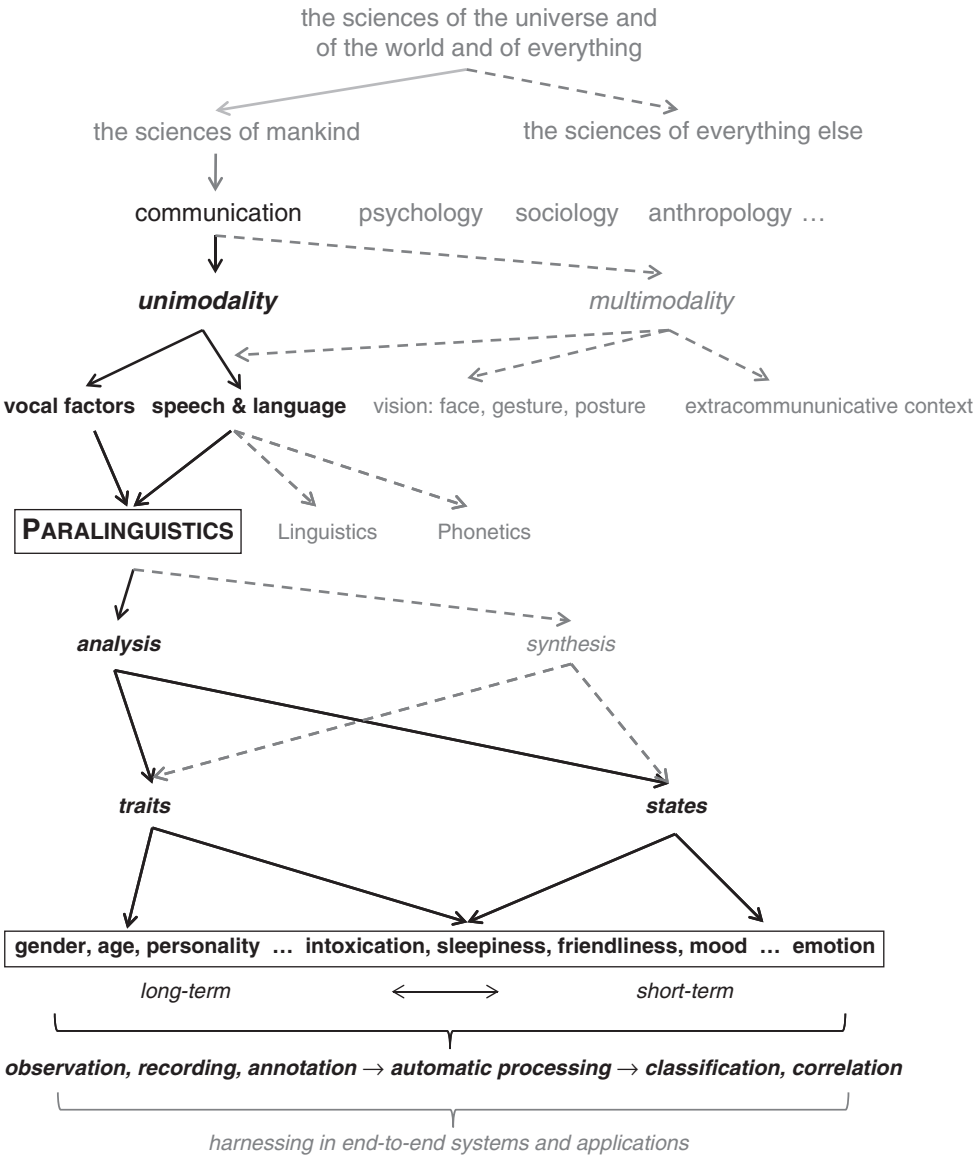

Figure 1.1 The realm of computational paralinguistics

In Figure 1.1 we try to narrow down the realm of paralinguistics in a reasonable way, as we conceive it and as we will deal with it in this book. Of course, there are other conceptualisations of paralinguistics, some broader, some narrower in scope. Figure 1.1 is a sort of flowchart that we will follow from top to bottom. A grey font indicates fields and topics that are not part of paralinguistics, for instance, the global science of mankind or of everything else that can be found in this world. Dashed lines lead to fields that are more or less disregarded in this book.

The first word shown in black is 'communication', denoting that interactions between human beings are focal. Paralinguistics deals with speech and language which both are primarily means of communication; even a soliloquy has to be overheard and eventually recorded and processed by the computer in order to be an object of investigation. The same holds for a private diary in its written form: it might not be intended as communication with others, but as soon as it is read by someone else, it is. Of course, human communication is an important part of related fields such as psychology, sociology, or anthropology. Thus, we have to follow the flowchart further down to point out what distinguishes paralinguistics from all these related fields.

In traditional linguistics, the term 'language' refers to the (innate and/or acquired) mental competence, and the term 'speech' to the performance, that is, to the ability to convert this competence into motor signals, acoustic waves, and percepts. In this book we adhere to a shallower definition of these two terms, based on their use in speech and language technology. Language is more or less synonymous with 'natural language' which is modelled and processed within computational linguistics; speech is the object of investigation within automatic speech processing, that is, 'spoken language', as opposed to written language.

We want to restrict paralinguistics to the unimodal processing of events primarily produced with the voice, or secondarily encoded in written language. 'Communication' is used in a broad sense: speech and language are primarily means of interaction between human beings; however, they can be decoupled from this function and analysed on their own, that is, when not used in a communicative setting. Note that this sort of secondary communication is natural for written language, because here the communication of sender and addressee is normally decoupled as for time and place.

We do not want to distinguish between *extralinguistics* and paralinguistics; Laver (1994, pp. 22f) attributes extralinguistics to *informative functions* denoting age, sex, and suchlike, and paralinguistics to *communicative functions*. Implicitly, extralinguistic functions are always communicated, as a sort of background; we will subsume these functions under *biological trait primitives*; see Section 5.1. Moreover, it would be simply too cumbersome to introduce 'computational extralinguistics' as an additional field.

There are alternative conceptualisations, the most important one arguably paralinguistics in the sense of *'multimodality'*; this holds practically for every aspect, whether it be emotion, or personality, or social signals (see Chapter 5). Undoubtedly, the most natural human–human interaction is face-to-face, with each partner employing all the means available to them: voice, linguistic message, face, gestures, and body. However, there are some common and natural (interaction) scenarios where only the acoustic channel is used, for instance, in telephone conversations or in radio plays. Moreover, it is simply natural that interaction/conversation partners sometimes speak, and sometimes only listen. In the latter case, there is no speech available that can be investigated; in the first case, when people are talking, analysing faces is more difficult because the movements of speech are superposed onto the facial gestures.

In this book, apart from vocal factors, speech, and language, everything else such as facial expressions, gestures, body posture, and any extracommunicative context is not part of paralinguistics. Needless to say, all these other aspects are very important within human–human and human–machine communication; we will return to such multimodal aspects throughout this book.

Moreover, paralinguistics is sort of defined *ex negativo*; it comprises everything which is *not* the object of investigation in phonetics or linguistics: It does not address the systematic aspects of speech and language which are dealt with in sub-fields such as phonology, morphology,

syntax, or semantics. Note, however, that the *use* of specific phonological or grammatical structures within a specific context may very well be object of investigation within paralinguistics. All this will be illustrated more extensively below.

In this book we will focus on analysis, basically excluding generation and synthesis. At first sight, this might appear exotic: after all, analysis and generation/synthesis can be considered as two sides of the same coin. Both are necessary for a complete account. However, methodologies differ considerably; in Batliner and Möbius (2005) the methodological differences between analysis on the one hand and generation/synthesis on the other hand have been detailed for the modelling and processing of prosody. From a methodological point of view, the analysis of gestures has perhaps more in common with the analysis of speech than its synthesis. Moreover, analysis and not synthesis is the core competence of the authors. We therefore decided to treat synthesis the same way as vision and extracommunicative context, as a fringe phenomenon in this book.

So far, we have addressed broad *fields of science*, either including or excluding them from our definition of paralinguistics. We will now briefly sketch the *phenomena* we are dealing with, as well as the *processing chain*. All this will be dealt with in more depth in the chapters to follow. In simple terms, paralinguistics deals with *traits* and *states*; traits are long-term events, whereas states are short-term. Examples are given in Figure 1.1. Typical traits are gender, age, and personality, and typical states are emotions. Then, there are phenomena which are somehow in between: People can be friendly towards everybody, or, towards a specific person, only for a very short time. You can get tipsy, that is, intoxicated, for a short time, or you can be a regular heavy drinker. The title of this book mentions three exemplary phenomena:

> **personality** denoting long-term character traits which are specific to individuals or groups. In a broader sense, this encompasses everything that characterises a specific individual, including traits such as age, gender, race, and suchlike. In a narrower sense, this encompasses psychological traits such as neuroticism.
>
> **emotion** denoting short-term states: prototypical ones such as anger, fear, joy, or less prototypical ones such as surprise.
>
> **affect** as a broader term, encompassing all kinds of manifestations of personality such as mood, interpersonal stances, or attitudes as displayed in Table 2.1 – a very common term since Picard (1997).

The last terms to be commented upon in the title are *speech and language processing*: In basic research, the two fields of phonetics and linguistics deal with different data: phonetics with (the production/acoustics/perception of) speech, and linguistics with (written) language. Accordingly, there are two different lines of research traditions in paralinguistics: one dealing mostly with the acoustic signal (called, for instance, 'emotion/affect processing'), and one dealing exclusively with written language (called, for instance, 'sentiment analysis'). In automatic speech processing, the approach is different: acoustic and linguistic information are combined in a hybrid fashion. Following this tradition, we will address both acoustic and linguistic phenomena in this book.

With observations, recordings, and annotations, we decide which phenomena we are dealing with, and how long the single event takes. In *computational* paralinguistics, we then try to process these phenomena automatically. Ultimately, this means producing some performance

measures which tell us how good we are at doing that. All this is the core topic of this book. Eventually, we of course want to evaluate our models not as single components but within end-to-end-systems and to harness them in applications; this will be touched upon and exemplified *passim*.

1.2 History and Subject Area

> *Language is not an abstract construction of the learned, or of dictionary makers, but is something arising out of the work, needs, ties, joys, affections, tastes, of long generations of humanity, and has its bases broad and low, close to the ground.*
>
> (Noah Webster)

So far, we have outlined the realm of computational paralinguistics. In this section, we want to sketch the history of paralinguistics and to narrow down its subject area.

Ever since the advent of structuralism (Saussure 1916), the study of (speech and) language has been more or less confined to the skeleton of language: phonetics/phonology, morphology, syntax, and grammar in general; there were only rather anecdotal remarks on functions of language which go beyond pure linguistics, for example, the following from Bloomfield (1933):

> *. . . pitch is the acoustic feature where gesture-like variations, non-distinctive but socially effective, border most closely upon genuine linguistic distinctions. The investigation of socially effective but non-distinctive patterns in speech, an investigation scarcely begun, concerns itself, accordingly, to a large extent with pitch.*

Pike (1945) was amongst the few who noticed these additional functions of intonation:

> *Other intonation characteristics may be affected or caused by the individual's physiological state – anger, happiness, excitement, age, sex, and so on. These help one to identify people and to ascertain how they are feeling . . .*

The basic neglect of paralinguistics holds for both European and American linguistics at that time – both displaying different varieties of structuralism. Thus, the central focus of linguistics in the last century was on structural, on genuine linguistic and, as far as speech is concerned, on formal aspects within phonetics and phonology. Language was conceived of as part of semiotics which deals with *denotation*, that is, with the core meaning of items.

This conviction was clearly expressed by Sapir (1921):

> *If speech, in its acoustic and articulatory aspect, is indeed a rigid system, how comes it, one may plausibly object, that no two people speak alike? The answer is simple. All that part of speech which falls out of the rigid articulatory framework is not speech in idea, but is merely a superadded, more or less instinctively determined vocal complication inseparable from speech in practice. All the individual color of speech – personal emphasis, speed, personal cadence, personal pitch – is a non-linguistic fact, just as the incidental expression of desire and emotion are, for the most part, alien to linguistic expression. Speech, like all elements of culture, demands conceptual selection, inhibition of the randomness of instinctive behavior.*

On the other hand, in Sapir (1927) we can find an – albeit informal – conceptualisation of 'speech as a personality trait', giving a rough but fair enumeration of parameters which are relevant for characterising personality – and, by the way, emotion as well:

> *To summarize, we have the following materials to deal with in our attempt to get at the personality of an individual, in so far as it can be gathered from his speech. We have his voice. We have the dynamics of his voice, exemplified by such factors as intonation, rhythm, continuity, and speed. We have pronunciation, vocabulary, and style. Let us look at these materials as constituting so and so many levels on which expressive patterns are built.*

Such remarks were, however, normally anecdotal and somehow spurious. Generally, non-linguistic aspects were conceived as fringe phenomena, often taken care of by neighbouring disciplines such as anthropology, ethnology, or psychology. This attitude slowly changed in the middle of the last century; linguists and phoneticians began to be interested in all these phenomena mentioned by Bloomfield (1933) and Pike (1945), that is, in a broader conceptualisation of semiotics, dealing with *connotation* (e.g., affective/emotive aspects) as well.

According to Trager (1958), Laver (1994), and Rauch (2008), the term 'paralanguage' was first introduced by the American linguist Archibald Hill (1958).

Terms such as 'extralinguistic', 'paralanguage', and 'paralinguistics' were used by Trager (1958), and later elaborated on by Crystal (1963, 1966, 1971, 1974, 1975a,b). To start with, Crystal (1963) mentions the neglect of paralinguistics by linguistics:

> *The last decade has brought renewed study of this linguistic backwater, now called paralanguage; but there has been surprisingly little attempt to approach the subject in a sufficiently systematic and empirical way to satisfy the critical linguist.*

This critical attitude seems to have persisted during the decades to come (cf. Rauch 1999):

> *. . . paralinguistics is to linguistics, unfortunately, a neglected stepchild at most . . . (p. 165)*
>
> *. . . the seeds for obscuring the domain of paralanguage were inherent in its twentieth-century rebirth for linguists by linguists. (p. 166)*

One of the few who not only dealt with paralinguistic phenomena but also tried to really propagate this field was Fernando Poyatos (1991, 1993, 2002).

On the other hand, both within linguistics proper and especially with the advent of human–computer interaction, we can say that paralinguistics and neighbouring disciplines have been safely established. Yet, the subject areas are still defined differently. These are the definitions given in two renowned dictionaries:

> ***paralanguage*** (n.) *A term used in SUPRASEGMENTAL PHONOLOGY to refer to variations in TONE of voice which seem to be less systematic than PROSODIC features (especially INTONATION and STRESS). Examples of* ***paralinguistic features*** *would include the controlled use of BREATHY or CREAKY voice, spasmodic features (such as giggling while speaking), and the use of secondary ARTICULATION (such as lip-ROUNDING or NASALIZATION) to produce a tone of voice signalling attitude, social*

> *role, or some other language-specific meaning. Some analysts broaden the definition of paralanguage to include KINESIC features; some exclude paralinguistic features from LINGUISTIC analysis. (Crystal 2008)*

> ***paralanguage*** *. . . 1. Narrowly, non-segmental vocal features in speech, such as tone of voice, tempo, tut-tutting, sighing, grunts, and exclamations like* Whew! *2. Broadly, all of the above plus non-vocal signals such as gestures, postures and expressions – that is, all non-linguistic behaviour which is sufficiently coded to contribute to the overall communicative effect. . . . (Trask 1996)*

Thus, since it first came into use in the middle of the last century, 'paralinguistics' has been confined to the realm of human–human communication, but with a broad and a narrow meaning. We follow Crystal (1974) who excludes visual communication and the like from the subject area and restricts the scope of the term to 'vocal factors involved in paralanguage'; cf. Abercrombie (1968) for a definition along similar lines. 'Vocal factor', however, in itself is not well-defined. Again, there can be a narrow meaning excluding linguistic/verbal factors, or a broad meaning including them. We use the last one, defining *paralinguistics* as the discipline dealing with those phenomena that are *modulated onto* or *embedded into* the verbal message, be this in acoustics (vocal, non-verbal phenomena) or in linguistics (connotations of single units or of bunches of units). This scope is mirrored and, at the same time, instantiated by the possibility of late fusion in multimodal ('non-verbal') processing and by the (relative) independence of computational paralinguistic approaches from other fields. Many tools and procedures have been developed specifically for dealing with the speech signal or with (written) language; many sites and researchers, specialising in speech and language, have extended their focus onto computational paralinguistics.

To give examples for acoustic phenomena: everybody would agree that coughs are not linguistic events, but they are somehow embedded in the linguistic message. The same holds for laughter and filled pauses (such as *uhm*) which display some of the characteristics of language, for example, as far as grammatical position or phonotactics is concerned. All these phenomena are embedded in the word chain and are often modelled the same way as words in automatic speech processing; they can denote (health) state, emotion/mood, speaker idiosyncrasies, and the like. In contrast, high pitch as an indicator of anxiety and breathy voice indicating attractiveness, for example, are modulated onto the verbal message. As for the linguistic level, paralinguistics also deals with everything beyond pure phonology/morphology/syntax/semantics. Let us give an example from semantics. The 'normal' word for a being that can be denoted with these classic semantic features [+human, +female, +adult] is *woman*. In contrast, *slut* has the same denotation but a very different connotation, indicating a strong negative valence and, at the same time, the social class and/or the character of the speaker. Bunches of units, for instance the use of many and/or specific adjectives or particles, can indicate personality traits or emotional states.

Whereas the 'garden-fencing' within linguistics, that is, the concentration on structural aspects, was mainly caused by theoretical considerations, a similar development can be observed within automatic speech (and language) processing which, however, was mainly caused by practical constraints. It began with concentrating on single words; then very constrained, read/acted speech, representing only one variety, that is, one rather canonical speech

register, was addressed. Nowadays, different speech registers, dialects, and spontaneous speech in general are processed as well.

At least amongst linguists, language has always been seen as the principal mode of communication for human beings (Trager 1958) which is accompanied by other communication systems such as body posture, movement, facial expression, cf. (Crystal 1966) where the formal means of indicating communicative stances are listed: (1) vocalisations such as 'mhm', 'shhh', (2) hesitations, (3) 'non-segmental' prosodic features such as tension (slurred, lax, tense, precise), (4) voice qualifiers (whispery, breathy, . . .), (5) voice qualification (laugh, giggle, sob, cry), and (6) non-linguistic personal noises (coughs, sneezes, snores, heavy breathing, etc.).

The extensional differentiation between terms such as verbal/non-verbal or vocal/non-vocal is sometimes not easy to maintain and different usages do exist; as often, it might be favourable to employ a prototype concept with typical and fringe phenomena (Rosch 1975). A fringe phenomenon, for example, is filled pauses which often are conceived of as non-verbal, vocal phenomena; however, they normally follow the native phonotactics, cannot be placed everywhere, can be exchanged by filler words such as *well*, and are modelled in automatic speech recognition the same way as words.

We can observe that different strands of research – having much in common – evolved more or less independently of each other; thus what sometimes has been subsumed under 'paralinguistics' by linguists has been called *non-verbal behaviour* research by psychologists (cf. Harrigan *et al.* 2008): facial actions, vocal behaviour, and body movement. Jones and LeBaron (2002) mention that '. . . the study of "nonverbal communication" emerged in the 1960s, largely in reaction to the overwhelming emphasis placed upon verbal behavior in the field of communication.' They argue in favour of integrating verbal and non-verbal approaches. Non-verbal communication from a multi-disciplinary perspective is dealt with in Burgoon *et al.* (2010).

Interestingly, the terms used are normally rather *ex negativo* such as 'para-/extra-linguistics' or 'non-verbal/non-vocal' parameters – again indicating that from its very beginning, the field had to be delimited from the more established discipline of linguistics.

1.3 Form versus Function

> *Form follows function – that has been misunderstood. Form and function should be one, joined in a spiritual union.*
>
> (Frank Lloyd Wright)

The distinction between *form* and *function* is arguably constitutive for modern phonetics and linguistics – form roughly meaning 'what does it look like, and how does it relate to other elements?', function meaning 'what is it used for?'. We can compare this basic distinction with the distinction between knowledge about fabrics (the substance for clothing) and fashion (the form, the code of clothing) on the one hand, and the function of clothing (used for an evening in the opera, or used for mountaineering) on the other hand. There are specialists in each of these aspects.

A phonetic form is constituted by some higher-level, structural shape or type which can be described holistically and analysed/computed using between 1 and *n low-level descriptors (LLDs)* such as pitch or intensity values and *functionals* such as mean or maximum values over

time. A simple example is a high rising final tone which very often denotes, that is, functions as indicating a question. This is a genuine linguistic function. In addition, there are paralinguistic functions encoded in speech or in other vocal activities. Examples include a slurred voice when the speaker is inebriated, or a loud and high-pitched voice when a person is angry. *Phonetics* deals with the acoustic, perceptual, and production aspects of spoken language (speech), and *linguistics* with all aspects of written language; this is the traditional point of view. From an engineering point of view, there is a slightly different partition: normally, the focus is on recognising and subsequent understanding of the content of spoken or written language; for speech, acoustic modelling is combined with linguistic modelling whereas, naturally enough, (written) language can only be modelled by linguistic means. Form is rather a means to handle the function of speech and language.

Laver (1994, p. 20) refers to the contrast between (phonological) *form* – how does an element relate to other elements? – and (phonetic) *substance* – for example, how does its acoustics look? Crystal (2008, pp. 194, 204) tells apart functions within and outside linguistics: linguistic and phonetic form and substance do have paralinguistic functions, for example, the word *somewhat* with its specific phonetic realisation in a specific syntactic position functioning as a hedge can characterise personality and/or communicative situations. In this book we will always contrast phonetic/linguistic form (consisting of form and substance) with paralinguistic function.

The distinction between form and function is also constitutive for discriminating paralinguistics as it typically is performed by linguists/phoneticians from paralinguistics as it typically is performed by engineers, psychologists, and other neighbouring disciplines. Linguists and phoneticians start with some formal element and try to find out which functions can be attributed to this specific form. Engineers and psychologists are primarily interested in modelling (manually or automatically) specific phenomena such as personality, emotion, or speech pathology, with the help of acoustic and/or linguistic parameter; that is, they are primarily interested in one specific (type of) function and want to find out which (form) features to use for modelling and classifying this function. A simple test whether the author of a study follows a formal or a functional approach is to estimate the number of pages dedicated to the one or the other aspect; of course, there are transitional forms in between.

Figure 1.2 illustrates the two different approaches. While the figure is straightforward, what is behind it can be extremely complicated. Conceptually, it is always a one-to-many mapping but the direction is reversed. To the left, there is the typical phonetic/linguistic approach. We start with one – more or less complex – formal element; this can be one word, one type of words *(part-of-speech)*, one syntactic construction; it can be one phoneme with its *allophonic* (free) variants, or one *supra-segmental* parameter, just to mention a few. Then we want to

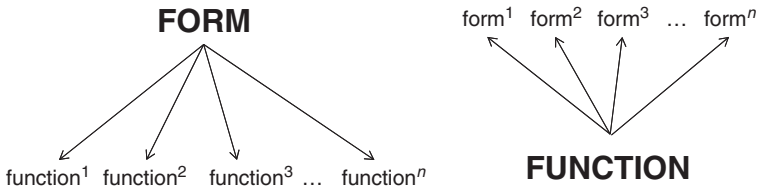

Figure 1.2 Form versus function: (left) linguistic/phonetic approach; (right) sociological, sociolinguistic, psychological, and psycholinguistic approach

find out which functions this formal element can be used for. This can be some intralinguistic function – for instance, a pronoun serves an *anaphoric* function if it refers to a noun that can be found earlier in the word chain; in this book, we are mostly interested in paralinguistic functions. To the right, the approach typical of psychology and other neighbouring fields is depicted. We start with one specific function – for instance, one emotion, one personality trait, or a specific non-native accent. Then we try to find out which formal elements denote this function and can be used for automatic modelling. In a few cases, this might be one form such as a high final pitch value, denoting questions or proneness towards questions. Nowadays, in brute-force approaches, we employ many formal elements, up to several thousand features.

In fact, it is mostly not a one-to-many but a many-to-many relationship because of the intrinsically multi-functional nature of acoustic-linguistic parameters. We will return to the distinction of form and function when presenting the different research strategies in Chapters 4 and 5.

1.4 Further Aspects

As pointed out above, we restrict the realm of paralinguistics to the analysis of vocal and verbal aspects. Of course, this is not the whole picture. There is *generation* and *synthesis* of paralinguistics as well, often embedded in a *multimodal* interaction. All this has to be modelled for human–computer interactions in prospective *application* scenarios. In order to be successful, *usability* has to be considered from the very beginning. Above all, and at a very early stage, *ethical* considerations have to be taken into account.

These aspects are not all relevant or pivotal for all subfields of paralinguistics: 'emotionally intelligent' virtual agents and robots might arguably be the main target group for generation and synthesis of adequate behaviour. In contrast, the synthesis of deviant speech (e.g., of a foreign accent or of some variety of pathological speech) most likely comes last, as far as meaningful applications are concerned. Of course, we can always imagine some application: there might be some place for a virtual agent in a computer game that impersonates a foreign language learner. Apart from being somehow exotic, such characters might be less attractive from a marketing point of view, and more difficult to implement.

In this section, we will first give a short account of synthesis, concentrating on emotion and personality. Then, both generation and analysis of multimodality are addressed. We will conclude with applications and usability, and ethical considerations.

1.4.1 The Synthesis of Emotion and Personality

Basically, speech synthesis is either rule-based, with acoustic parameters generated following specific rules (*formant synthesis*), or based on speech samples that are concatenated, which can be as short as minimal sets of transitions between sounds (*diphone synthesis*) or whole phrases/utterances (*unit selection*). The speech samples are normally obtained from controlled recordings and a small sample of single speakers. *HMM (Hidden Markov model) synthesis* is a statistical parametric synthesis, based on hidden Markov models (see Chapter 11), and trained from speech databases.

Formant synthesis allows for systematic manipulation but does not sound fully natural. Unit selection sounds most natural if the transitions between units can be smoothed correctly but

necessitates too much pre-recorded information, especially if different paralinguistic phenomena – normally different emotional states – have to be modelled. Explicit modelling is flexible but not fully natural; in contrast, concatenative modelling sounds natural but is not flexible.

Of course, the building blocks are the same for synthesis and analysis, such as phones, words, phrases, and utterances. Methodologies, however, differ considerably. These differences did not show up very clearly in the early days of phonetic and emotion research, when only a few features were explicitly modelled, that is, manipulated or analysed. Nowadays, however, it seems difficult to bridge the gap between the thousands of features used for brute-force modelling of many speakers on the one hand, and the relatively few features or speakers modelled for rule-based or concatenative synthesis. The basically different procedural approaches towards analysis and synthesis of prosody – which is one of the main building blocks for emotional modelling, apart from voice quality – are elaborated on in (Batliner and Möbius 2005). *Embodied conversational agents (ECAs)* can be cartoon-like or very pronounced; they can be based on acted emotions produced by one single actor. It is conceivably not possible to manipulate and generate thousands of acoustic-prosodic features – which is no problem in a brute-force automatic classification. Thus, the perspective of paralinguistic synthesis differs considerably from that of paralinguistic analysis; it is more similar to that of traditional lab phonetics where specific hypotheses are proved with the help of carefully manipulated stimuli presented in *identification* or *discrimination* tests.

Let us give a short account of the synthesis of emotional speech. First attempts towards emotional, rule-based speech synthesis were reported in Murray and Arnott (1993, 1995). Schröder (2001) gave an overview of what had been done in the field; this was continued in Schröder (2004); see also Gobl and Ní Chasaide (2003) and Schröder *et al.* (2010). Black (2003) deals with unit selection and emotional speech. The synthesis of 'personality primitives' such as age and gender is straightforward. The synthesis of personality traits is not yet a fully established field. It is addressed in Trouvain *et al.* (2006); Schröder *et al.* (2012) describe a framework for generating and synthesising emotionally competent embodied conversational agents having four different personalities – aggressive, cheerful, gloomy, and pragmatic – within a prototype of a multimodal dialogue system, the Sensitive Artificial Listener (SAL) scenario. Schröder *et al.* (2011) present a conceptual view on the generic representation of emotions using an *Emotion Markup Language* (an agreed-upon computer-readable representation) and *ontologies* (formal specifications of shared conceptualisations such as paralinguistic states).

Of course, the divide between synthesis and analysis can be overcome, but this will take time. Indications are, on the one hand, the use of HMM synthesis based on multiple speakers, and on the other hand, the use of synthesised data for augmenting real-life databases used for training automatic classifications of paralinguistic phenomena.

1.4.2 Multimodality: Analysis and Generation

Evidently, it is not only speech and language that communicate personality, emotion, affect and the like. Darwin (1872) attributed a leading role to the face: 'Of all parts of the body, the face is most considered and regarded, as is natural from its being the chief seat of expression and the source of the voice.' In addition, there are gestures and body movements/posture (Kleinsmith and Bianchi-Berthouze 2012).

Although confined to a specific experimental condition, the results of Mehrabian and Wiener (1967) were often taken as proof that the verbal channel contributes only little (7%) to the communication of attitudes; this is called the 7%–38%–55% myth. However, already Ekman and Friesen (1980) state that the '. . . claims in the literature that the face is most important or that the nonverbal-visual channel is more important than the verbal-auditory channel have not been supported' in their experiments. O'Sullivan *et al.* (1985) elaborate further on the complex interrelationship between different types of messages and the relative importance of verbal compared to non-verbal factors. The answer is simply that 'no channel is always most important'. Further arguments can be found in Trimboli and Walker (1987), Lapakko (1997), and Krauss *et al.* (1981).

Generic statements on the relative importance of single modalities do not make any sense; we can only ask about the contribution of single modalities in specific communicative settings. Now, does it make sense to describe single modalities at all? Jorgensen (1998) claimed that researchers focusing only on one modality, for example, the verbal channel, 'are no longer studying valid communication processes, but rather disassociated parts of the whole'. A simple but important argument against this position can be found in Planalp and Knie (2002) where it is argued that even '. . . the simplest research on cue and channel combinations . . . produces incredibly complicated results'.

An overview on affect and emotion recognition methods in multimodal human–computer systems is given in Zeng *et al.* (2009) and Pantic *et al.* (2011). The complex task of coordinating multiple modalities in an affective agent – this holds for analysis as well – is nicely illustrated in the following list quoted from Martin *et al.* (2011):

- *Equivalence/substitution: one modality conveys a meaning not borne by the other modalities (while it could be conveyed by these other modalities)*
- *Redundancy/repetition: the same meaning is conveyed at the same time via several modalities*
- *Complementarity:*
  - *Amplification accentuation/moderation: one modality is used to amplify or attenuate the meaning provided by another modality*
  - *Additive: one modality adds congruent information to another modality*
  - *Illustration/clarification: one modality is used to illustrate/clarify the meaning conveyed by another modality*
- *Conflict/contradiction: the meaning transmitted on one modality is incompatible or contrasting with the one conveyed by the other modalities; this cooperation occurs when the meaning of the individual modalities seems conflicting but indeed the meaning of their combination is not and emerges from the conflicting combination of the meanings of the individual modalities.*
- *Independence: the meanings conveyed by different modalities are independent and should not be merged.*

The claim of Planalp and Knie (2002) might be slightly exaggerated; however, there is surely a trade-off between the basic complexity of multimodality and the possibilities for investigating complex phenomena within one single modality. Moreover, there are of course constellations where only one channel is used and available for analysis. We will come back to this in Section 2.12.

1.4.3 Applications, Usability and Ethics

The computational processing of paralinguistics could be conceived of as encapsulated – data in, measures out; thus neither applications nor usability need to be addressed. Potential applications are often mentioned in the introduction and/or final remarks of articles; yet, it is often not clear how the approach presented really can be harnessed in these applications. However, they are, together with ethical considerations, decisive for success or failure of approaches if we do not confine ourselves to pure research. Applications are, as it were, at the lowest, practical level; mostly they are presented in the form of single examples. However, we will try and present a tentative taxonomy. Usability is on a methodologically higher level; pertinent considerations are based on psychological and theoretical theories. Ethics is, of course, at the highest level, and not a genuine topic of this technologically oriented book. However, we definitely want to stress its importance.

Applications

Examples of applications for affective computing are given in Picard (1997), Picard (2003), and Batliner *et al.* (2006), and for paralinguistics in a broader sense in Burkhardt *et al.* (2007) and Schuller *et al.* (2013). In the following, examples and presentation are inspired by the last three references.

Basic types of application approaches are (1) speech recognition in itself, (2) analysis, screening and monitoring of paralinguistic events or phenomena, and (3) interaction, normally of humans with an ECA on a computer or with a robot; all these aspects can be employed alone or in combination. Speech recognition can hopefully be improved when the speaker's paralinguistic peculiarities are modelled, for instance, by a preceding attribution to speaker classes. For all the other aspects, it is the other way round: speech recognition should be as good as possible, especially for employing linguistic features. Human–human communication can be analysed and monitored. We are interested in the type of communication. That is, is the communication symmetric or asymmetric? Which roles are taken up to what extent by which communication partner? Is the communication 'normal', 'as it should be', 'not as it should be', or even 'pathological'? We can assess conversations of married couples having problems (Lee *et al.* 2010), and we can monitor and summarise meetings (Kennedy and Ellis 2003; Laskowski 2009) or call centre interactions. All types of deviant, especially 'pathological', speech (see Section 5.6) can be the object of analysis and monitoring, and of periodic screening. Human–machine interaction is a wide field, encompassing all kinds of gaming, tutoring, information, and assistive and communicative robotics. Media retrieval is a genuine object of investigation for written data. Even cross-modal control is possible by paralinguistic rather than linguistic means, for example, in helping artists with upper limb disabilities to use the volume of their voice to control cursor movements to create drawings on the screen.

A tentative taxonomy of basic properties for emotion-oriented systems is given in Batliner *et al.* (2006); this was extended to speaker classification systems in Burkhardt *et al.* (2007). Table 1.1, adapted from Batliner *et al.* (2006) and slightly edited, summarises the criteria that can distinguish different (types of) applications. 'Meta-assessment' means that we are looking at success or failure from the outside. For instance, telling a call-centre customer during the interaction that she is angry (single instance decision, system design: on-line, mirroring) can

Table 1.1 Some basic features of applications

features	description
	meta-assessment
critical	application's aims are impaired if the paralinguistic phenomenon is processed erroneously (single instance decision)
non-critical	erroneous processing does not impair application's aims (cumulative evidence)
	system design
on-line	system reacts (immediately/delayed) while interacting with user
off-line	no system reaction, or delayed reaction after actual interaction
mirroring	user gets feedback as for his/her (expressive) behaviour
non-mirroring	system does not give any explicit feedback
emotional	system reacts in an emotional way
non-emotional	system does not behave emotionally but 'neutrally'

be detrimental if she is not. Collecting cumulative evidence, for instance, checking whether customers are, in the long run, more content with one or another automatic system, is non-critical as long as there is some correlation with the ground truth. Generally, on-line, mirroring, and emotional system (re)actions will be more critical than off-line, non-mirroring, and non-emotional (re)actions. Note that 'non-critical' as used here refers to the immediate context of an application; later decisions based on the processing within such applications can of course still be wrong, that is, economically unwise or unethical.

Further examples of applications will be touched upon *passim* in the following chapters.

Usability

Trivially, usability is largely irrelevant in the case of pure recording and monitoring or screening, and subsequent analysis and evaluation – all this constitutes the largest part of paralinguistic research. Of course, usability becomes relevant if the user is somehow interacting with the system (see Table 1.1).

Kaye *et al.* (2011) present the historical background and contrast the traditional goals of software design, such as utility, effectiveness and learnability, with the goals of user-centred design (user experience goals: being motivating, fun, enjoyable) which is not a linear but an iterative process involving users at any step of design and evaluation. Historically, the concept of usability evolved from a narrow focus on the computational expert (engineers) in the 1950s, passing through a focus on psychological and cognitive experts in the 1980s – the users of the personal computer were no longer experts but rather lay people; the interface was not only a green or amber screen with a text console but a graphical interface, and nowadays, it can be ECAs and even robots as well. Most recent is a focus on experience-focused human–computer interaction. An up-to-date account of the whole field is given in Rogers *et al.* (2011). The specific requirements for multimodal interfaces are addressed in Oviatt (1999, 2008) and Oviatt and Cohen (2000).

Ethical Considerations

The first question often asked – not by engineers but by people directly concerned – about technology at large is whether we want it that way, and, in particular, whether we mind if it takes over human work. Secretaries fear unemployment when dictation systems are used, and speech therapists fear unemployment when screening and assessment of speech pathologies are taken over by machines. Is technology harmful to those whose expertise is substituted, or will they then be free to do more expert work?

The next question is whether technology does what it is supposed to do and what it promises to do. Failure to do that might not be detrimental in the case of dictation systems: it is straightforward to find out. It requires more effort in the case of automatic screening: a counter-check introduces exactly that kind of manual work that should be avoided. In the case of the lie detector, such a counter-check is not possible in real-life situations, for instance in court, thus we have to rely on transferability from scientific studies. A chapter in Kreiman and Sidtis (2011) describes nicely how this definitely should not be done because it simply would violate the principle *in dubio pro reo*. This would be the case of an erroneous single instance decision (Table 1.1), with monstrous unethical consequences.

Even if technology can do what it is supposed to do, we have to ask whether this is acceptable: is the monitoring of call-centre agents ethically acceptable – even if it might be reliable if done off-line and in an accumulative way?

Basic research on computational paralinguistics might not be much concerned with such questions but it definitely is necessary to know about them. Ethical concerns about privacy, however, are of utmost relevance. How can we ensure that ethical principles are observed during recruitment of participants in experiments, during recording, storing, and during dissemination/displaying of recordings and other types of results? Here, we should follow the principle of informed consent (Sneddon *et al.* 2011): amongst others things, participants should be informed about the goals of the study and the experiment, and they should be given the possibility to withdraw during and after the experiment. It should go without saying that strict anonymity must be guaranteed later on. All these provisions might be cumbersome to maintain but universities, research organisations and legal provisions will safeguard them.

Several further aspects of ethical concern are discussed in Ragin and Amoroso (2011), Cowie (2011), Döring *et al.* (2011), and Goldie *et al.* (2011).

1.5 Summary and Structure of the Book

In Part I, we will lay the foundations for the computational processing that is dealt with in Part II. In this first chapter, we began by defining 'computational paralinguistics', and sketched the history of the term and the subject area. The opposition between form and function is a guiding principle not only of phonetics and linguistics but also of paralinguistics, and was therefore described next. In the rest of this chapter, we sketched all those aspects – generation, synthesis, multimodality, usability, applications, and ethical considerations that we do not focus on in the following. Chapter 2 presents a taxonomy of oppositions that can – but need not in every case – be relevant for the different sub-fields, topics and phenomena within paralinguistics. Chapter 3 tries to point out important aspects of modelling, that is, theories and methodologies that heavily influence the way how we see, approach, deal with, and evaluate paralinguistic phenomena. Chapter 4 presents examples for formal elements – segmental and

supra-segmental, phonetic and linguistic, verbal and non-verbal – that constitute the building blocks for the marking of all those functions that are described in Chapter 5.

Corpus Engineering is dealt with in Chapter 6, especially annotations and exemplars of paralinguistics corpora; this constitutes the transition of Part I into Part II.

In Part II, we first give an overview of the chain of processing within computational paralinguistics in Chapter 7. Acoustic features and their extraction on a 'low' frame-by-frame level are described in Chapter 8. Linguistic features are then described in Chapter 9. Both types of features are used for feature generation on a supra-segmental level as described in Chapter 10. In Chapter 11 we deal with the field's most common approaches to modelling from a machine learning point of view. This also includes a statement on feature relevance analysis and testing protocols. In Chapter 12 insight is given into how best to embed computational paralinguistics in a running system's working context. The selected aspects cover distribution in a client–server architecture, weakly surpervised learning and confidence measure calculation. To provide the chance of experiencing what the book describes, Chapter 13 provides a 'hands-on' tutorial alongside a description of the 'usual suspects' when it comes to toolkits in the field.

Outside of these two main parts, Chapter 14 provides a short general epilogue.

Last but not least, the Appendix contains the description of standardised feature sets and a feature encoding scheme.

References

Abercrombie, D. (1968). Paralanguage. *International Journal of Language & Communication Disorders*, **3**, 55–59.

Batliner, A. and Möbius, B. (2005). Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground? In W. Barry and W. Dommelen (eds), *The Integration of Phonetic Knowledge in Speech Technology*, pp. 21–44. Springer, Dordrecht.

Batliner, A., Burkhardt, F., Ballegooy, M. v., and Nöth, E. (2006). A taxonomy of applications that utilize emotional awareness. In *Proc. of IS-LTC 2006*, pp. 246–250, Ljubljana.

Black, A. (2003). Unit selection and emotional speech. In *Proc. of Eurospeech*, pp. 1649–1652, Geneva.

Bloomfield, L. (1933). *Language*. Holt, Rinehart and Winston, New York.

Burgoon, J. K., Guerrero, L., and Floyd, K. (2010). *Nonverbal Communication*. Allyn & Bacon, Boston.

Burkhardt, F., Huber, R., and Batliner, A. (2007). Application of speaker classification in human machine dialog systems. In C. Müller (ed.), *Speaker classification I: Fundamentals, Features, and Methods*, pp. 174–179. Springer, Berlin.

Cowie, R. (2011). Editorial: 'Ethics and good practice' – Computers and forbidden places: Where machines may and may not go. In P. Petta, C. Pelachaud, and R. Cowie (eds), *Emotion-Oriented Systems: The Humaine Handbook*, Cognitive Technologies, pp. 707–712. Springer, Berlin.

Crystal, D. (1963). A perspective for paralanguage. *Le Maître Phonétique*, **120**, 25–29.

Crystal, D. (1966). The linguistic status of prosodic and paralinguistic features. *Proc. of the University of Newcastle-upon Tyne Philosophical Society*, **1**, 93–108.

Crystal, D. (1971). Prosodic and paralinguistic correlates of social categories. In E. Ardener (ed.), *Social Anthropology*, pp. 185–206. Tavistock, London.

Crystal, D. (1974). Paralinguistics. In T. Sebeok (ed.), *Current Trends in Linguistics 12*, pp. 265–295. Mouton de Gruyter, The Hague.

Crystal, D. (1975a). Paralinguistic behaviour as continuity between animal and human communication. In W. McCormack and S. Wurm (eds), *Language and Man: Anthropological Issues*, pp. 13–27. Mouton de Gruyter, The Hague.

Crystal, D. (1975b). Paralinguistics. In J. Benthall and T. Polhemus (eds), *The Body as a Medium of Expression*, pp. 162–174. Institute of Contemporary Arts, London.

Crystal, D. (2008). *A Dictionary of Linguistics and Phonetics*. Blackwell, Oxford, 6th edition.

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*. John Murray, London.

Döring, S., Goldie, P., and McGuinness, S. (2011). Principalism: A method for the ethics of emotion-oriented machines. In P. Petta, C. Pelachaud, and R. Cowie (eds), *Emotion-Oriented Systems: The Humaine Handbook*, Cognitive Technologies, pp. 713–724. Springer, Berlin.

Ekman, P. and Friesen, W. V. (1980). Relative importance of face, body, and speech in judgments of personality and affect. *Journal of Personality and Social Psychology*, **38**, 270–277.

Gobl, C. and Ní Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, **40**, 189–212.

Goldie, P., Döring, S., and Cowie, R. (2011). The ethical distinctiveness of emotion-oriented technology: Four long-term issues. In P. Petta, C. Pelachaud, and R. Cowie (eds), *Emotion-Oriented Systems: The Humaine Handbook*, Cognitive Technologies, pp. 725–734. Springer, Berlin.

Harrigan, J. A., Rosenthal, R., and Scherer, K. R. (eds) (2008). *New Handbook of Methods in Nonverbal Behavior Research*. Oxford University Press, Oxford.

Hill, A. (1958). *Introduction to Linguistic Structures: From Sound to Sentence in English*. Harcourt Brace, New York.

Jones, S. E. and LeBaron, C. D. (2002). Research on the relationship between verbal and nonverbal communication: Emerging integrations. *Journal of Communication*, **52**, 499–521.

Jorgensen, P. (1998). Affect, persuasion, and communication processes. In P. Andersen and L. Guerrero (eds), *Handbook of Communication and Emotion*, pp. 403–422. Academic Press, San Diego.

Kaye, J. J., Laaksolahti, J., Höök, K., and Isbister, K. (2011). The design and evaluation process. In P. Petta, C. Pelachaud, and R. Cowie (eds), *Emotion-Oriented Systems: The Humaine Handbook*, Cognitive Technologies, pp. 641–656. Springer, Berlin.

Kennedy, L. and Ellis, D. (2003). Pitch-based emphasis detection for characterization of meeting recordings. In *Proc. of ASRU*, pp. 243–248, Virgin Islands.

Kleinsmith, A. and Bianchi-Berthouze, N. (2012). Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing*, **99**.

Krauss, R. M., Apple, W., Morency, N., Wenzel, C., and Winton, W. (1981). Verbal, vocal, and visible factors in judgments of another's affect. *Journal of Personality and Social Psychology*, **40**, 312–320.

Kreiman, J. and Sidtis, D. (2011). *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception*. Wiley-Blackwell, Malden, MA.

Lapakko, D. (1997). Three cheers for language: A closer examination of a widely cited study of nonverbal communication. *Communication Education*, **46**, 63–67.

Laskowski, K. (2009). Contrasting emotion-bearing laughter types in multiparticipant vocal activity detection for meetings. In *Proc. of ICASSP*, pp. 4765–4768, Taipei, Taiwan.

Laver, J. (1994). *Principles of Phonetics*. Cambridge University Press, Cambridge.

Lee, C.-C., Black, M., Katsamanis, A., Lammert, A., Baucom, B., Christensen, A., Georgiou, P., and Narayanan, S. (2010). Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples. In *Proc. of Interspeech*, pp. 793–796, Makuhari, Japan.

Martin, J.-C., Devillers, L., Raouzaiou, A., Caridakis, G., Ruttkay, Z., Pelachaud, C., Mancini, M., Niewiadomski, R., Pirker, H., Krenn, B., Poggi, I., Caldognetto, E. M., Cavicchio, F., Merola, G., Rojas, A. G., Vexo, F., Thalmann, D., Egges, A., and Magnenat-Thalmann, N. (2011). Coordinating the generation of signs in multiple modalities in an affective agent. In P. Petta, C. Pelachaud, and R. Cowie (eds), *Emotion-Oriented Systems: The Humaine Handbook*, Cognitive Technologies, pp. 349–367. Springer, Berlin.

Mehrabian, A. and Wiener, M. (1967). Decoding of inconsistent communications. *Journal of Personality and Social Psychology*, **6**, 109–114.

Murray, I. and Arnott, J. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustical Society of America*, **93**, 1097–1108.

Murray, I. R. and Arnott, J. L. (1995). Implementation and testing of a system for producing emotion by rule in synthetic speech. *Speech Communication*, **16**, 369–390.

O'Sullivan, M., Ekman, P., Friesen, W. V., and Scherer, K. R. (1985). What you say and how you say it: The contribution of speech content and voice quality to judgment of others. *Journal of Personality and Social Psychology*, **48**, 54–62.

Oviatt, S. (1999). Ten myths of multimodal interaction. *Communications of the ACM*, **42**, 74–81.

Oviatt, S. (2008). Multimodal Interfaces. In A. Sears and J. J. A. (eds), *The Human-Computer Interaction Handbook. Fundamentals, Evolving Technologies, and Emerging Applications*, pp. 413–432. Lawrence Erlbaum, New York.

Oviatt, S. and Cohen, P. (2000). Multimodal interfaces that process what comes naturally. *Communications of the ACM*, **43**, 45–53.

Pantic, M., Caridakis, G., André, E., Kim, J., Karpouzis, K., and Kollias, S. (2011). Multimodal emotion recognition from low-level cues. In P. Petta, C. Pelachaud, and R. Cowie (eds), *Emotion-Oriented Systems: The Humaine Handbook*, Cognitive Technologies, pp. 115–132. Springer, Berlin.

Picard, R. (1997). *Affective Computing*. MIT Press, Cambridge, MA.

Picard, R. (2003). Affective Computing: Challenges. *Journal of Human-Computer Studies*, **59**, 55–64.

Pike, K. L. (1945). *The Intonation of American English*. University of Michigan Press, Ann Arbor.

Planalp, S. and Knie, K. (2002). Integrating verbal and nonverbal emotion(al) messages. In S. R. Fussell (ed.), *The Verbal Communication of Emotions: Interdisciplinary Perspectives*, pp. 55–77. Lawrence Erlbaum.

Poyatos, F. (1991). Paralinguistic qualifiers: Our many voices. *Language & Communication*, **11**, 181–195.

Poyatos, F. (1993). *Paralanguage: A Linguistic and Interdisciplinary Approach to Interactive Speech and Sounds*. John Benjamins, Amsterdam.

Poyatos, F. (2002). *Nonverbal Communication across Disciplines, Volume II*. John Benjamins, Amsterdam.

Ragin, C. C. and Amoroso, L. M. (2011). *Constructing Social Research: The Unity and Diversity of Methods*. Sage Publications, Thousand Oaks, CA.

Rauch, I. (1999). *Semiotic Insights: The Data Do the Talking*. University of Toronto Press, Toronto.

Rauch, I. (2008). *The Phonology/Paraphonology Interface and the Sounds of German across Time*. Peter Lang, New York.

Rogers, Y., Sharp, H., and Preece, J. (2011). *Interaction Design: Beyond Human - Computer Interaction*. Wiley, Chichester.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, **104**(3), 192–233.

Sapir, E. (1921). *Language: An Introduction to the Study of Speech*. Harcourt Brace, New York.

Sapir, E. (1927). Speech as a personality trait. *American Journal of Sociology*, **32**, 892–905.

Saussure, F. de (1916). *Cours de linguistique générale*. Payot, Paris.

Schröder, M. (2001). Emotional speech synthesis: a review. In *Proc. of Eurospeech*, pp. 561–564, Aalborg.

Schröder, M. (2004). *Speech and Emotion Research. An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis*, volume 7 of *Reports in Phonetics, University of the Saarland*. Institute for Phonetics, University of Saarbrücken.

Schröder, M., Burkhardt, F., and Krstulovic, S. (2010). Synthesis of emotional speech. In K. R. Scherer, T. Bänziger, and E. Roesch (eds), *Blueprint for Affective Computing*, pp. 222–231. Oxford University Press, Oxford.

Schröder, M., Pirker, H., Lamolle, M., Burkhardt, F., Peter, C., and Zovato, E. (2011). Representing emotions and related states in technological systems. In P. Petta, C. Pelachaud, and R. Cowie (eds), *Emotion-Oriented Systems: The Humaine Handbook*, Cognitive Technologies, pp. 369–388. Springer, Berlin.

Schröder, M., Bevacqua, E., Cowie, R., Eyben, F., Gunes, H., Heylen, D., Maat, M. t., McKeown, G., Pammi, S., Pantic, M., Pelachaud, C., Schuller, B., Sevin, E. d., Valstar, M., and Wöllmer, M. (2012). Building autonomous sensitive artificial listeners. *IEEE Transactions on Affective Computing*, **3**(2), 165–183.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., and Narayanan, S. (2013). Paralinguistics in speech and language – state-of-the-art and the challenge. *Computer Speech and Language, Special Issue on Paralinguistics in Naturalistic Speech and Language*, **27**(1), 4–39.

Sneddon, I., Goldie, P., and Petta, P. (2011). Ethics in emotion-oriented systems: The challenges for an ethics committee. In P. Petta, C. Pelachaud, and R. Cowie (eds), *Emotion-Oriented Systems: The Humaine Handbook*, Cognitive Technologies, pp. 753–768. Springer, Berlin.

Trager, G. L. (1958). Paralanguage: A First Approximation. *Studies in Linguistics*, **13**, 1–12.

Trask, R. (1996). *A Dictionary of Phonetics and Phonology*. Routledge, London.

Trimboli, A. and Walker, M. B. (1987). Nonverbal dominance in the communication of affect: A myth? *Journal of Nonverbal Behavior*, **11**, 180–190.

Trouvain, J., Schmidt, S., Schröder, M., Schmitz, M., and Barry, W. (2006). Modelling personality features by changing prosody in synthetic speech. In *Proc. of Speech Prosody*, Dresden.

Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**(1), 39–58.