Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing

Pages: 345
Publisher: Wiley; 1 edition (September 17, 2013)
Format: pdf, epub
Language: English

Contents

This edition first published 2014 © 2014 John Wiley & Sons, Ltd *Registered office*John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

Library of Congress Cataloging-in-Publication Data Schuller, Björn.   Computational paralinguistics : emotion, affect and personality in speech and languageprocessing / Björn Schuller, Anton Batliner. – First Edition.   pages cm   Includes bibliographical references and index.   ISBN 978-1-119-97136-8 (hardback) 1. Psycholinguistics–Data processing. 2. Linguistic models–Data processing. 3. Paralinguistics.4. Language and emotions. 5. Speech processing systems. 6. Human-computer interaction.7. Emotive (Linguistics) 8. Computational linguistics. I. Batliner, Anton. II. Title.   P37.5.D37S38 2014   401′.90285–dc23 2013019325 A catalogue record for this book is available from the British Library.

To Dagmar and Gisela

Preface It might be safe to claim that 20 years ago, neither the term ‘computational paralinguistics’ nor the field it denotes existed. Some 10 years ago, the term did not yet exist either. However, in hindsight, the field had begun to exist if we think of the first steps towards the automatic processing of emotions in speech in the mid-1990s. For example, Picard’s book on *Affective Computing* published in 1997, and the International Speech Communication Association (ISCA) workshop on emotion and speech in 2000, just to mention some of the many topics and events related to and belonging to computational paralinguistics. The term ‘paralinguistics’ had already been coined in the 1950s – with different broad or narrow denotations; we will try and sketch this history in Part I of this book. Yet, in the realm of ‘hard core’ automatic (i.e., computational) processing of speech, the topic was still not fully acknowledged; as one of our colleagues said: ‘Emotion recognition, that’s esoterics with HMMs.’ Today, it might be safe to claim that computational paralinguistics has been established as a discipline in its own right – although surprisingly, not the term itself. It is only natural that, as a new and still somewhat exotic field, it has to cope with prejudices on the one hand, and unrealistic promises on the other hand. This book represents the first attempt towards a unified overview of the field, its extremely ramified and diverse ‘genealogy’, its methodology, and the state of the art. ‘Computational paralinguistics’ is not an established subject that can be studied, and this fact is mirrored in the ‘scientific CVs’ of both authors. B.S. studied electrical engineering and information technology. However, his doctoral thesis dealt with one aspect of computational paralinguistics: the automatic recognition of human emotion in speech. During his habilitation period he broadened the scope of his work to ‘intelligent audio analysis’ – dealing with quite a number of further paralinguistic aspects, including those found in sung language and many other audio processing problems such as emotion in music and general sound. At the time of finalisation of the manuscript, he was a professor in computer science. A.B. started within philology, came from diachronic phonology to phonetics in general to prosody in particular, and via prosody in the interface for and within syntax and semantics to the automatic processing of acted and very soon naturalistic emotions – realising, moreover, that he had been dealing with different topics for a long time that all can be located within (computational) paralinguistics. Originally, the intended focus of this book was on the computational processing of emotions and affect in speech and language, taking into account personality as well. In the early conceptual stages, however, we realised that this would be sub-optimal and thus decided to deal with everything ‘besides’ linguistics – namely, the computational processing of ‘para’-linguistics in a broad sense. However, we confine ourselves to the acoustic/phonetic/linguistic aspects, that is, we only deal with one modality, namely speech/language including non-verbal components, and disregard other modalities such as facial gestures or body posture. Moreover, we do not aim at a complete description of human–human or human–machine communication which would include the generation and production (synthesis) of speech, the interaction with other components within a multimodal system, the role within application systems, or real-life applications and their evaluation. Apart from the fact that most of these aspects would not be part of our core competence, we feel that it makes sense to try to establish computational paralinguistics as one building block amongst several others. Besides, there are already good overview and introductory books available on these other topics. And last but not least, it would be rather too complex for one book. We wish to provide the reader with a sort of map presenting an overview of the field, and useful for finding one’s own way through. The scale of this map is medium-size, and we can only display a few of the houses in this virtual paralinguistic ‘city’ with their interiors, on an exemplary basis. In so doing, we hope to provide guidelines for the novice and to present at least a few new insights and perspectives to the expert. Many studies are referred to and core results are summarised. For all of them, the caveat holds that basically all such studies are restricted – confined to a specific choice of subjects, research questions, operationalisations, and features employed, just to mention a few of the decisive factors. There are errors such as the famous erroneous decimal point that made spinach more healthy than anything else – note that reports of this error might be erroneous themselves. And of course, there is much more that can go wrong – and hopefully, we will find out: scrutiny of results and replications of studies will eventually converge to more stable claims. We decided not to describe basic phonetic/linguistic knowledge such as vowel or consonant charts, details on pitch versus F0, loudness, morphological and grammatical systems, basics on production and perception of speech, and the like. Such information can easily be obtained in introductory and overview books from the respective fields, as well as from online sources. In a similar way, selections had to be made for the computation aspects. For example, many approaches to linguistic modelling exist, and the fields of machine learning and signal processing each deserve at least one book in its own right. Thus, we limited our choices to the methods most established and common in the field – serving as a solid basis and inspiration for the interested reader to look further. A connected resource of information is the book’s homepage found at http://www.cp.openaudio.eu which includes features such as links to the openSMILE toolkit and (part of) the data described. Björn W. Schuller and Anton M. Batliner *Munich, February 2013*

Acknowledgements We would like to thank in particular Florian Eyben and Felix Weninger at Technische Universität München in Munich, Germany, for their great help in finishing Chapters 7–13; we also extend our thanks to Martina Römpp for her help with the illustrations in this part, and to Tobias Bocklet and Florian Hönig at the Friedrich Alexander Universität in Erlangen-Nuremberg, Germany, for commenting on Chapter 5. We stand on the proverbial shoulders of giants – and of many other ‘normal’ people who all contributed to our understanding. In the same way, we are grateful to colleagues who do not share our points of view: opinions are not only reinforced by encouragement but also by dissent which helps us to rethink or revise our own approaches and theoretical positions, or simply to stick with them. There are too many people in the field(s) that supported us, discussed with us, and exchanged ideas with us, too many to list in detail, so we wish to express our gratitude to all of them in a generic way. For the encouragement in the first place, excellent continued support and guidance as well as patience, we sincerely thank the editor and publisher – John Wiley & Sons. Finally, we would like to thank the HUMAINE Association (henceforth the Association for the Advancement of Affective Computing) for providing an excellent network not only for affective computing, but also for the broader field of computational paralinguistics. The authors also acknowledge funding from the European Commission under grant agreement no. 289021 from the ASC Inclusion Project committed to provide interactive serious emotion games for children with autism spectrum condition.

List of Abbreviations

ACF Autocorrelation Function
ACM Association for Computing Machinery
AD(H)D Attention Deficit (Hyperactivity) Disorder
AEC (FAU) Aibo Emotion Corpus
AL Active Learning
ALC Alcohol Language Corpus
AM Acoustic Model
ANN Artificial Neural Network
API Application Programming Interface
ARFF Attribute Relation File Format (WEKA)
ARMA Autoregressive Moving Average
ASC Autism Spectrum Condition
ASCII American Standard Code for Information Exchange
ASD Autism Spectrum Disorder
ASR Automatic Speech Recognition
AUC Area Under Curve
AVEC Audiovisual Emotion Challenge
AVIC (TUM) Audiovisual Interest Corpus
BAC Blood Alcohol Content
BES Berlin Emotional Speech Database
BFI Big Five Inventory
BLSTM Bidirectional Long Short-Term Memory
BoCNG Bag-of-Character-*N*-Grams
BoNG Bag-of-*N*-Grams
BoW Bag-of-words
BPTT Back-Propagation Through Time
BRAC Breath Alcohol Concentration Test
BRNN Bidirectional Recurrent Neural Network
C-AuDiT Computer-Assisted Pronunciation and Dialogue Training
CALL Computer-Aided Language Learning
CAPT Computer-Aided Pronunciation Training
CC (Pearson) Correlation Coefficient
CD Compact Disc
CEICES Combining Efforts for Improving Automatic Classification of Emotion in Speech
CFS Correlation based Feature Selection
CGN Spoken Dutch Corpus from Centre for Genetic Resources, the Netherlands
CR Compression Rate
CSV Comma Separated Value (format)
DAT Digital Audio Tape
DBN Dynamic Bayesian Networks
DCT Discrete Cosine Transformation
DES Danish Emotional Speech (Database)
DET Detection Error Trade-off (curve)
DFT Discrete Fourier Transform
DLL Dynamic Link Library

DT Decision Tree (machine learning) / Determiner (linguistics)
EC Emotion Challenge (Interspeech 2009)
ECA Embodied Conversational Agent
EER Equal Error Rate
EM Expectation Maximisation
EMMA Extensible Multi-Modal Annotation (markup language, XML-style)
ERB Equivalent Rectangular Bandwidth
ETSI European Telecommunications Standards Institute
EU European Union
EWE Evaluator Weighted Estimator
F0 Fundamental Frequency
FAU Friedrich Alexander University
FFT Fast Fourier Transformation
FN False Negatives
FNN Feed-forward Neural Network
FNR False Negative Rate/Ratio
FP False Positives
FPR False Positive Rate/Ratio
GM Gaussian Mixture
GMM Gaussian Mixture Model
GPL GNU/General Public Licence
GUI Graphical User Interface
HMM Hidden Markov Model
HNR Harmonics-to-Noise Ratio
HTK Hidden Markov (model) Toolkit
ICA Independent Component Analysis
IDF Inverse Document Frequency
IG Information Gain
IGR Information Gain Ratio
IIR Infinite Impulse Response
IP Interruption Point
ISCA International Speech Communication Association
ISLE Italian and German Spoken Learners English (corpus)
ITU International Telecommunication Union
KL Kullback–Leibler (divergence/distance)
L1 first language
L2 second language
LDA Linear Discriminant Analysis
LDC Linguistic Data Consortium
LIWC Linguistic Inquiry and Word Count
LLD Low-Level Descriptor
LM Language Model
LOO Leave One Out
LP Linear Predictor / Linear Prediction
LPC Linear Predictive Coding
LPCC Linear Predictive Cepstral Coefficient
LSF Line Spectral (pair) Frequency
LSP Line Spectral Pair
LSTM Long Short-Term Memory
LSTM-RNN Long Short-Term Memory Recurrent Neural Network
LVCSR Large Vocabulary Continuous Speech Recognition
MAE Mean Absolute Error
MAPE Mean Absolute Percentage Error

MFB Mel-Frequency Band
MFCC Mel-Frequency Cepstral Coefficient
MIML Multimodal Interaction Markup Language
ML Maximum Likelihood
MLE Mean Linear Error
MLP Multilayer Perceptron
MSE Mean Square Error
NCSC NKI CCRT Speech Corpus
NHD Null Hypothesis Decision
NHR Noise-to-Harmonics Ratio
NHT Null Hypothesis Testing
NL Non-likeable
NMF Non-negative Matrix Factorisation
NN Noun (linguistic)
NP Noun Phrase (linguistic)
NPV Negative Predictive Value
OCEAN Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism
OOV Out Of Vocabulary (words)
openEAR open-source Emotion and Affect Recognition (toolkit)
openSMILE open-source Speech and Music Interpretation by Large Scale Extraction (toolkit)
PC Paralinguistic Challenge (Interspeech 2010)
PCA Principal Component Analysis
PCM Pulse Code Modulation
PDA Pitch Detection Algorithm
PDF Probability Density Function
PLP Perceptual Linear Prediction
PLP-CC Perceptual Linear Prediction Cepstral Coefficient
PLTT Post Laryngectomy Telephone Test
POS Part Of Speech
PP Prepositional Phrase (linguistic)
PR Precision
RASTA RelAtive SpecTrA
RASTA-PLP RelAtive SpecTrA Perceptual Linear Prediction (Coefficients)
RE Recall
RIFF Resource Interchange File Format
RMS Root Mean Square
RMSE Root Mean Squared Error
RNN Recurrent Neural Network
ROC Receiver Operating Characteristic
SAL Sensitive Artificial Listener
SAMPA Speech Assessment Methods Phonetic Alphabet
SEMAINE Sustained Emotionally coloured Machine-human Interaction using Non-verbal Expression (project)
SFFS Sequential Floating Forward Search
SFS Speech Filing System
SHS Sub-Harmonic Summation
SI International System of Units (French: Système international d’unités)
SIFT Simplified Inverse Filtering Technique
SIMIS Speech In Minimal Invasive Surgery (database)
SLC Sleepy Language Corpus
SLD Speaker Likability Database
SNR Signal-to-Noise Ratio
SP SPecificity

SPC Speaker Personality Corpus
SSC Speaker State Challenge
STC Speaker Trait Challenge
SVM Support Vector Machine
SVQ Split Vector Quantisation
SVR Support Vector Regression
T-expression Ternary expression
TF Term Frequency
TFIDF Term Frequency and Inverse Document Frequency
TN True Negatives
TP True Positives
TPR True Positive Rate/Ratio
TUM Technische Universität München
UA Unweighted Accuracy
UAR Unweighted Average Recall
VAM Vera-Am-Mittag (German TV show, corpus)
VB Verb
VP Verb Phrase
VQ Vector Quantisation
WA Weighted Accuracy / Word Accuracy (ASR)
WAR Weighted Average Recall
WYALFIWYG What You Are Looking For Is What You Get
WYSIWYG What You See Is What You Get
XML eXtensible Mark-up Language
ZCR Zero Crossing Rate

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 *π α ρ α*); 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. 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.
Figure 1.1 The realm of computational paralinguistics 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 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. **Figure 1.2** Form versus function: (left) linguistic/phonetic approach; (right) sociological, sociolinguistic, psychological, and psycholinguistic approach 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*
.

This book presents the methods, tools and techniques that are currently being used to recognise (automatically) the affect, emotion, personality and everything else beyond linguistics (‘paralinguistics’) expressed by or embedded in human speech and language.

It is the first book to provide such a systematic survey of paralinguistics in speech and language processing. The technology described has evolved mainly from automatic speech and speaker recognition and processing, but also takes into account recent developments within speech signal processing, machine intelligence and data mining.

Moreover, the book offers a hands-on approach by integrating actual data sets, software, and open-source utilities which will make the book invaluable as a teaching tool and similarly useful for those professionals already in the field.

Key features: Provides an integrated presentation of basic research (in phonetics/linguistics and humanities) with state-of-the-art engineering approaches for speech signal processing and machine intelligence. Explains the history and state of the art of all of the sub-fields which contribute to the topic of computational paralinguistics. C overs the signal processing and machine learning aspects of the actual computational modelling of emotion and personality and explains the detection process from corpus collection to feature extraction and from model testing to system integration. Details aspects of real-world system integration including distribution, weakly supervised learning and confidence measures. Outlines machine learning approaches including static, dynamic and context‑sensitive algorithms for classification and regression. Includes a tutorial on freely available toolkits, such as the open-source ‘openEAR’ toolkit for emotion and affect recognition co-developed by one of the authors, and a listing of standard databases and feature sets used in the field to allow for immediate experimentation enabling the reader to build an emotion detection model on an existing corpus.

Computational Paralinguistics - Emotion, Affect - æ¥½å¤©ãƒ–ãƒƒã‚¯ã‚¹ - Emotion, Affect and Personality in Speech and Language Processing BjÃ¶rn Schuller, Anton

Wiley also publishes its books in a variety of electronic formats. Computational Paralinguistics von Bjorn Schuller - Computational Paralinguistics Emotion, Affect and Personality in Speech and Language Processingè±†ç“£è¯„å†ï¼š0.0 ç®€ä»‹ï¼š Computational Paralinguistics : Emotion, Affect and - eBay - [EPUB] Computational paralinguistics: emotion, affect and personality in speech and language processing by BjÃ¶rn Schuller, Anton Batliner. Book file PDF easily Computational Paralinguistics : Emotion, Affect and - Pandora - Most of the work on affective computing research is limited to the tion, how this affects the other speaker's emotional appraisal psycholinguistic features of the speech and linguistics sig- terizing the process of affective communication.. ioral and linguistic response... paralinguistics: emotion, affect and personality in. Computational Paralinguistics - Emotion, Affect and - Saraiva - SPEECH AND LANGUAGE PROCESSING Computational Paralinguistics Emotion Affect And Personality In Speech And ePUB structure of the book. Computational Paralinguistics: Emotion, Affect and Personality - You can download and read online Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing file PDF Book only if Emotion, Affect and Personality in Speech and Language - [EPUB] Computational paralinguistics: emotion, affect and personality in speech and language processing by BjÃ¶rn Schuller, Anton Batliner. Book file PDF easily Computational paralinguistics : emotion, affect and personality - Download file Free Book PDF Zeppelin Nights: London Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing. Zeppelin Nights: London in the First World War.pdf - Social Signal Processing - edited by Judee K. Burgoon May 2017. on the importance of emotional expression in speech communication and Computational Paralinguistics: Emotion, Affect - Amazon.ca - Personality in Speech and Language Processing file PDF Book only if you are Download file Book PDF Computational Paralinguistics: Emotion, Affect. Stefan Steidl (0000-0002-8937-3852) - ORCID - This book presents the methods, tools and techniques that are currently being used to recognise (aut... Computational Paralinguistics : Emotion, Affect and Personality in Speech and Language Processing. E-libro:
