T-Labs Series in Telecommunication Services

Tim Polzehl

Personality in Speech

Assessment and Automatic Classification

Springer

T-Labs Series in Telecommunication Services

More information about this series at http://www.springer.com/series/10013

Tim Polzehl

Personality in Speech

Assessment and Automatic Classification

Springer

Tim Polzehl
Quality and Usability Lab, Telekom
Innovation Laboratories
TU Berlin
Berlin
Germany

ISSN 2192-2810 ISSN 2192-2829 (electronic)
ISBN 978-3-319-09515-8 ISBN 978-3-319-09516-5 (eBook)
DOI 10.1007/978-3-319-09516-5

Library of Congress Control Number: 2014945959

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

If we want the vocal human–computer interaction to become more intuitive, it is inevitable to make the computer notice, interpret, and react to human ways of expression and patterns in communication beyond the recognition of the mere word strings. This is specifically important when it comes to subtle or hidden characteristics carrying connotations or status information on the speaker. One widely acknowledged model known from psychology is the all-encompassing empiric concept of the *Big 5* personality traits. Accordingly, personality is understood as defined and measurable set of habitual patterns of behavior, thoughts, and emotions. Throughout the entire presented work, vocal patterns are sought to be elicited, recorded, and examined in order to reveal, extract and model such defined patterns that correspond to personality expression in speech.

For statistical analyses and experimentation three databases comprising different speech conditions were recorded. These conditions comprise *acted* and *non-acted* recordings, *single*- and *multi-speaker* recordings, different *degrees of linguistic freedom* as well as differences in *microphone set up*. Extensive labeling sessions were conducted to annotate the speech data with personality assessments using the *NEO-FFI* personality assessment questionnaire. It provides estimates of the *Big 5* personality traits, i.e., *openness, conscientiousness, extroversion, agreeableness*, and *neuroticism*.

Analyses of correlations, consistencies, and latent factor structures show that the NEO-FFI can be applied to the new environment, namely speech input. Further, initial experiments focusing on time- and text-dependency show promising results, namely an overall insensitivity of personality attribution with respect to the time and textual domain the samples are drawn from. Findings are mostly congruent over the above-mentioned speech conditions.

On the basis of the recorded and examined databases, experiments on automatic modeling of personality from speech are conducted. This requires the relevant expressions to be identified, extracted, modeled, and retrieved—tasks eventually resulting in a comprehensive audio processing and machine learning problem. Therefore, a rather large-scale acoustic and prosodic feature extraction was developed, generating 1,359 features designed to capture intonation, dynamics, rhythmics, and spectral, i.e., voice-quality related behavior of human speech. Applying a ranking based on human personality annotations reveals prominent features and feature groups.

For modeling discriminative models are chosen, namely support vector models for both classification and regression tasks. Linear kernels were extended to nonlinear mapping. Ultimately, the joint performance of a subselection of the proposed features, different model configurations, and different speech conditions in the databases were systematically evaluated and analyzed.

In effect, results are very encouraging. Classification experiments aim to reveal the feasibility of telling apart different personality *types* or *classes* from speech automatically. Acted data could be classified with very high accuracies. Out of ten personality classes designed along the extremes of the *Big 5* trait scales, automatic signal-based classification results in up to 99 % accuracy for recordings where the linguistic content has been controlled, and up to 85.2 % accuracy for uncontrolled data. Another learning task sought to explore the feasibility of an actual prediction of trait scores along the *Big 5* independent traits. Reference for these experiments is the human personality perception as annotated during listening tests. The models reach correlations to human assessments of up to 0.91 for acted-data and 0.82 for non-acted data, depending on the actual trait to be predicted.

Eventually, these experiments provide systematic unprecedented indications for personality trait score modeling from speech, as well as systematic analyzes of personality classification from speech. In the authors hope and belief, the presented results and the explained approach will serve as both basis and reference for future works with regard to personality modeling from speech.

Acknowledgments

I would like to thank my supervisors Prof. Sebastian Möller from Technical University of Berlin, Germany, Prof. Florian Metze from Carnegie Mellon University, Pittsburgh, USA, and Prof. Tanja Schulz from Karlsruhe Institute of Technology, Germany, for support and review when concepting, experimenting and interpreting advancements and backlashes during my time at the Quality and Usability Lab of Technical University of Berlin. Further, I would like to thank all of my colleagues at the Labs, who have given me inspiring conversations along the way. Also, I would like to direct special thanks to the Telekom Innovation Laboratories, who build up the joint institute of Deutsche Telekom and Technical University of Berlin, short T-Labs. I am grateful for having received continuous ICT support and practical outreach into the company. Ultimately, my biggest thanks shall be directed to my beloved wife, who has given me trust and time for executing my studies throughout the whole time.

Contents

Introduction

Every day humans interact with computers. Computers help us organize extensive tables, drawing sketches, reminding us of open tasks and finally schedule much of our daily life. Moreover, computers help us to connect to and interact with other humans. In a growing number of interactions, the human interlocutor is replaced by computers. In this respect, intuitive human–computer interaction is dependent on the computer's abilities to communicate with us. But humans and computers interact in essentially different ways. In general, machine-to-machine communication accurately and explicitly follows protocols. These protocols have a fixed number of expected information. Every unexpected information may lead to errors or may simply be ignored. Human–human interaction also follows some guidelines but these guidelines give room for interpretation in case of unexpected or missing parts. In any case, speech will be processed and interpreted. As the level of human–computer interaction has passed the point of directed dialogs and simple command and control type interfaces, computers need to be equipped with the ability to understand and interpret more than just explicit protocol-based interaction. If we want the computer to better understand the human way of communication, it is our task to teach it. At the same time, this requires us to fully understand what, why, and how we communicate. But if we are to compile a list of what the computer would actually need to understand, we often face the challenge to understand what we actually express. Moreover, what we express might not necessarily be what we originally wanted to express or what we think we are aware of expressing. With respect to the abundance of information that can potentially be expressed, such as linguistic meaning, connotations, associations, emotions, speakers intentions and attitudes, or speaker characteristics, we finally commit that we are standing just at the very beginning of a comprehensive understanding of the human ways of expression.

Despite understanding, human's automatism to nevertheless interpret speech is inevitable and ubiquitous. From what we hear we instantaneously construct a kind of model. Again this model is not bound to the explicit content of what might have been uttered. On the contrary, this model also encompasses associations to expectations as learned from experiences as well as predictions of what might follow. If a speaker is unknown to us, we nevertheless assume certain characteristics, which then might be taken from persons we assume to be similar. Evaluating this model, we expect to encounter the predicted behavior. In many cases, we are even able to sense small deviations from this expected behavior, making us aware of even more meaning that might be uttered only between the lines. This relation is as interdependent as essential. Eventually, the richness of human communication is nourished by our appetite to interpret. We interpret much more than what has been addressed by the speaker. In psychology, an all-encompassing model exists, that tries to capture all those kinds of information that a person attributes to another person, namely the model of personality. Here each speaker is attributed a personality by other humans. Again, self-attribution is not necessarily congruent to attribution by others. Throughout the present work the author exploits a psychological definition of personality as *defined and measurable set of habitual patterns of behavior, thoughts, and emotions*, cf. Kassin (2003). The advantage of this definition is its empirical perspective, which allows to *measure observable* characteristics. Ultimately, the major aim of this work is to organize measurable personality-related speech characteristics and link them to signal-based speech characteristics, which are then modeled in order to serve for experiments on automatic classification and prediction of observable personality characteristics from speech.

One important aim along the way of the presented work is to explore whether the concept of psychological personality can actually be transfered to audible speech. While personality manifests itself in an abundance of cues, this work focuses on cues from speech only. The questions of how much personality is assessable from speech is the first out of two major research topics in this work. In order to be able to empirically analyze this research questions, three databases of different speech conditions were recorded. To ensure that different personality characteristics are contained in the data, a professional speaker was asked to perform speech from different personality perspectives. Two databases were recorded with this speaker, namely a text-dependent and a text-independent subset. Finally, a third set of recordings comprises non-acted realistic speech, which is recorded in order to be able to compare results between acted and non-acted data as well as to evaluate feasibility of personality assessment in realistic settings.

Extensive labeling sessions were conducted to annotate the speech data with personality assessments using the NEO-FFI questionnaire. This questionnaire is most widely acknowledged and applied in psychology. It provides an assessment scheme of the so called *Big 5* personality traits: *openness*, *conscientiousness*, *extroversion, agreeableness*, and *neuroticism*. Correlation and consistency-based analysis of labels as well as factor-analysis of the questionnaire responses leads to very good results. All results support the hypothesis that personality can be captured from speech by applying the NEO-FFI, with the resulting constructs resembling the same personality traits as known from psychology. Further, initial experiments also focus on time- and text-dependency. The question of whether a personality assessment from speech depends on the actual particular recording of the speaker is focused in order to come to indications of time-dependency. In terms of text-dependency the questions of whether the assessments from speech depend on what the speaker is actually speaking is focused. Also interdependencies are analyzed. Finally, knowing that the assessment is applicable as proposed, the speech assessments are analyzed for their personality characteristics as perceived by naive listeners. Results validated that the recorded data contains the desired personality variations.

Having this database ready for operations, the second major goal of the presented work is to explore automatic modeling of personality from speech. This requires the relevant expressions to be identified, extracted, prepared for modeling and retrieved by modeling experiments. Overall, these requirements result in a comprehensive audio processing and machine learning task. In order to be able to run automatic experiments, vocal cues need to be extracted automatically in the first place. In this work, a comprehensive acoustic and prosodic feature extraction unit generating 1359 features was implemented. Although still being just a subset of all proposed features from the speech and audio processing literature, these features can—to the best of the author's knowledge—be expected to resemble a good cross-section sample of frequently applied and robust features from the speech research field. Features are designed to capture melodic and dynamic speech behavior as well as rhythms and voice-quality related cues. In more detail, they exploit the extracted pitch, intensity, loudness, formants, spectral characteristics and some specific signal characteristics. This broad-band repertoire of acoustic and prosodic features provides the opportunity to apply a ranking based on human personality annotations in order to reveal prominent features and feature groups. At the same time, the interpretation of the ranking outcome depends on the subsequent modeling performance as well.

For modeling discriminative models were chosen. In more detail support vector models for classification and regression were trained and evaluated. Linear kernels were extended to non-linear mapping. As another major contribution of this work, the joint performance of a subselection of the proposed features, different models and different speech conditions in the databases were systematically evaluated and analyzed. Individual results range from very good to poor performance but overall results reveal very good and indicate overall success for future personality estimation from speech. In more detail, experiments are designed to account for two learning tasks. Classification experiments aim to reveal the feasibility of telling apart different personality *types* or *classes*. Acted data could be classified automatically with very high accuracies. Out of 10 personality classes in the databases, automatic signal-based classification results reached 99 % accuracy for the text-dependent subset, and 85.2 % accuracy for the text-independent data. Another learning task is to explore the feasibility of an actual prediction of trait scores along the five independent traits. Reference for these models is the human personality perception as annotated during listening tests. Comparing performance from acted data to realistic data, the actual trait scores were modeled by support vector regression models. Best models reached correlations to human assessments of up to 0.91 for acted-data and 0.82 for realistic data, depending of the actual trait to be predicted.

Ultimately, this work provides systematic unprecedented results for trait score modeling as well as systematic analyzes of personality classification from speech signals, which will hopefully serve as both, basis and reference for future works from the speech community members.

This work is organized as follows: Chap. 1 provides the basics of personality concepts from psychology including its assessment. Chapter 2 gives an overview of related works and terminology used with respect to personality assessment from speech. Chapter 3 explains the recording and labeling procedures for the three recorded databases. Chapter 4 focuses on the analysis of correlations and consistencies of human personality perception from speech. Also, the application of the assessment scheme to speech is analyzed by structural analysis of item responses using factor analysis. Likewise, results from statistical tests on the perception of the personality expressions in the databases are presented. Chapter 5 gives an overview of the audio-processing chain as well as experiments on automatic personality classification and trait score prediction. Starting with a brief description of audio extraction and feature definition, the applied ranking is introduced. Also modeling using support vector models is introduced along normalization and evaluation guidelines. Results from the modeling experiments are shown in Sects. 5.8–5.10 separately for the three recorded databases. Chapter 6 presents the discussion of results from human assessment and automatic modeling out of a personality-related perspective and presents a list of potentially influencing factors. Finally, Chap. 7 comprises a brief conclusion and the outlook including impulses for application.

Reference

Kassin S (2003) Psychology. Prentice Hall, Upper Saddle River

Chapter 1
Personality Assessment in Psychology

1.1 Definitions of Personality

In psychology, an abundance of attempts to define the concept of personality has been proposed. Following a comprehensive contemporary definition from Ryckman (2004) personality can be seen as

> a dynamic and organized set of characteristics possessed by a person that uniquely influences his or her cognitions, motivations, and behaviours in various situations.

Ryckman sees a concept of a certain *set* of characteristics, i.e. characteristics may be countable in number. He does not see this set to be of random order, consequently the impact on our behavior is determined, he calls it *organized*. Given a directed and orderly influence we, being observer, listener or interlocutor, experience the impact of these influences by perceiving observable manifestations in respective behavioral patterns or communication. This assumption is of essential importance. Accordingly, many researchers inquiring personality have worked on diverse methods to observe, isolate and assess such observable manifestations. A further important assumption is dynamics of personality. Strength, magnitudes and sets of characteristics are expected to differ.

Pervin et al. (2005) developed a definition that goes beyond Ryckman's theory, when he added that personality must essentially focus on *consistent patterns of feelings, thoughts and behaviors*. Here, personality is seen as a *consistent* concept, which does not contradict Ryckmans assumptions but adds another perspective that every person might have his or her unique set of consistent characteristics. Following Pervin, this also implies that personality profiles can be used to differentiate between people.

Congruent with this assumption Herrmann (1969) describes personality as an individual-dependent unique, outlasting and stable concept. Herrman adds an expected uniqueness. In theory, if we were able to assess all aspects of a person's personality, we would be able to differentiate this person from all other persons by his or her unique personality profile. Reality teaches us, that we can neither be sure

T. Polzehl, *Personality in Speech*, T-Labs Series in Telecommunication Services,
DOI 10.1007/978-3-319-09516-5_1

to capture all aspects of a person's personality, nor to collect enough data to verify the alleged uniqueness against all other persons.

The current focus in personality assessment is rather directed to the definition and collection of the most relevant characteristics. Researchers disagree on number and character of relevant aspects, and on how to assess these aspects. Still, when looking at different personality theories from a broader perspective, many of them seem to be at least partly complementary. Most theories do not challenge each other, they are rather seen as particularly helpful in particular situations or when looking from particular angles. For some applications or scientific questions researchers are interested in generation of a coherent personality profile of a single person, in other cases it might be helpful to investigate into how people can differ from each other considering numerous persons. Eventually, many researchers and psychologists do not explicitly identify themselves with a certain perspective but decide for the best possible explanation when examining particular issues in this respect.

This work focuses on personality as defined by Ryckmann and Pervin, and follows the hypothesis of Herrmann, assuming that people can be differentiated by their observable personality patterns. Looking from an empirical and conscious-driven perspective, a major approach is the so called *trait theory* of personality. According to Kassin (2003), personality can be seen as a *defined set of habitual patterns of behavior, thoughts, and emotions*, that manifests itself in terms of *measurable traits*. Traits are generally seen to be relatively stable over time and situation. By trait theory definition, a personality trait can be used to differentiate across individuals. As a result, there exist a multitude of possible traits that can be useful to describe personality. Following Goldberg (1993b), traits can be captured by continuous bipolar factors that can be distinguished from temporary states, e.g. emotions. After a brief overview of opposing theories, Sect. 1.2 provides an introduction to trait theories, their development and application.

In opposition to trait theory, another major stream of research stresses the generalizability and look for overarching characteristics in the psychology of human nature of all people's behavior. These researchers emphasize theory development, such as the theory of *psychodynamics*, as developed by Freud (1923). Accordingly, psychodynamics, also known as *dynamic psychology*, is the study of the interrelationship of various parts of the mind, personality, or psyche as they relate to mental, emotional, or motivational forces especially at the unconscious level. The core of any psychological process, according to Freud, is the *ego*. In his theory, the ubiquitous struggle the ego has to fight results in constant battle with three forces: the id, the super-ego, and the outside world. Ultimately, his theory focuses on the interactions between the id, ego, and superego and subsequently attempts to explain or interpret behavior or mental states in terms of these innate forces and processes. A comprehensive introduction and comparison of different assumptions along these perspectives can be found in Ahles (2004).

Yet other researchers see personality as determined by impacts external stimuli have on behavior, e.g. Skinner (1984), who has developed the so called *Behaviorist* theory. In the so called *social* or *cognitive* theory of personality, e.g. Bandura (1986), cognitive processes such as thinking and judging are emphasized. Here, cognitions are believed to guide behavior. Most important are cognitions such as expectations about the world, or expectations about other people. Proponents of the so called *humanistic* theory of psychology emphasize that people have a free will and they can actively determine how they behave. This theory focuses on analysis of experiences of individuals as exposed to definitive factors that determine behavior, cf. Snygg et al. (1998). Readers desiring a more detailed discussion about the presented and other personality perspectives can find a comprehensive collection of literature in the *Handbook of Personality* edited by John et al. (2008).

1.2 Trait Theory of Personality

1.2.1 Allport's Trait Organization

Gordon Allport was a pioneer in early trait research. He mainly focused on structure and membership of possible traits and created a structure of four groups, in which he arranged the trait candidates, which he also called dispositions (Allport 1937). In one of his early works with Henry S. Odbert (Allport and Odbert 1936) he extracted 17,935 trait candidate words from the most comprehensive dictionary of English language available at that time, i.e. Webster's *New International Dictionary*. He then reduced that huge list to a number of 4,504 adjectives, that he believed could be used to describe a person's traits. He arranged the terms due to the following semantic groups:

- Neutral terms designating personal traits
- Terms primarily descriptive of temporary moods or activities
- Weighted terms conveying social or character judgments of personal conduct, or designating influence on others
- Miscellaneous: designations of physique, capacities, and developmental conditions; metaphorical and doubtful terms

This work is widely seen as the foundation of the so called *lexical approach*. The basic assumption behind this approach is that important differences in characteristics to describe a personality have made their way into spoken language and the dictionary. Allport and Odbert declare:

> Those individual differences that are most salient and socially relevant in peoples lives will eventually become encoded into their language; the more important such a difference, the more likely is it to become expressed as a single word.

Characteristics that cannot be expressed by isolated terms were not considered. Allport argues that observing the behavior of people is a great clue indicating to personality traits. The observation of people who like to ski, hike, and ride bikes can be used to infer that they are athletic, which he then sees as a trait. Stressing the empirical nature of his approach he declares that observing others either in natural settings or through experiments can be used to infer some of their traits.

Organizing the remaining 4,504 terms designating personal traits he set up three membership groups:

1. Central traits, which are general characteristics inherent to every person. While degree or magnitude can vary, these are what he called *the basic building blocks that shape most of our behavior although they are not as overwhelming as cardinal traits*.
2. Secondary traits, which show only in specific situations and are thus more peripheral
3. Cardinal trait, which is a trait (or a set of very few traits) that dominates and shapes a person's behavior. Regarding number of traits he declares: *These are rare as most people lack a single theme that shape their lives*.

Allport added another limitation to his approach by defining the *common traits* which are those recognized within a certain cultural background, i.e. individual trait membership and inventories may vary from culture to culture.

1.2.2 Eysenck's P-E-N Theory of Super Factors

In 1947 Hans Eysenck worked on a development of his personality trait inventory. Influenced by improved techniques of statistical factor analyses at that time, he looked for correlations behind the observable traits. He referred to what he was looking for as *super factors*. At first, his experiments revealed two super factors, which he entitled *neuroticism* and *extraversion* (Eysenck and Lewis 1947). Initial tests were designed to capture neuroticism. As a result of his factor analyses he declared *that the factor of neuroticism is not a statistical artifact, but constitutes a biological unit which is inherited as a whole*. Eysenck was also driven by the question of heredity of personality. His early experiments were carried out with identical twins, i.e. developed from one zygote that splits and forms two embryos, as well as fraternal twins, i.e. developed from two separate eggs that are fertilized by two separate sperm, of 11 and 12 years of age. With this respect, his results lead to the conclusion that *neurotic predisposition is to a large extent hereditarily determined* (Eysenck and Prell 1951). Using the two super factors from his observations, he designed a personality attribution map, shown in Fig. 1.1.

Later he arranged personality types as known from ancient Greek philosophy (Hippocrates, ca. 400 B.C.) to match the factors and assigned unstable personalities as *choleric* when extraverted, and *melancholic* when introverted, while stable personalties were assigned *phlegmatic* when introverted, and *sangruic* when extraverted.

For successive works he extended his tests to a large number of test persons. In the late 1970s, Eysenck added a third super factor that resulted from analyses he conducted jointly with his wife (Eysenck 1967). The revealed factor was entitled *psychoticism* and referred to degrees of aggressiveness and interpersonal hostility. These three factors, i.e. neuroticism, extraversion, psychoticism, built the basis for

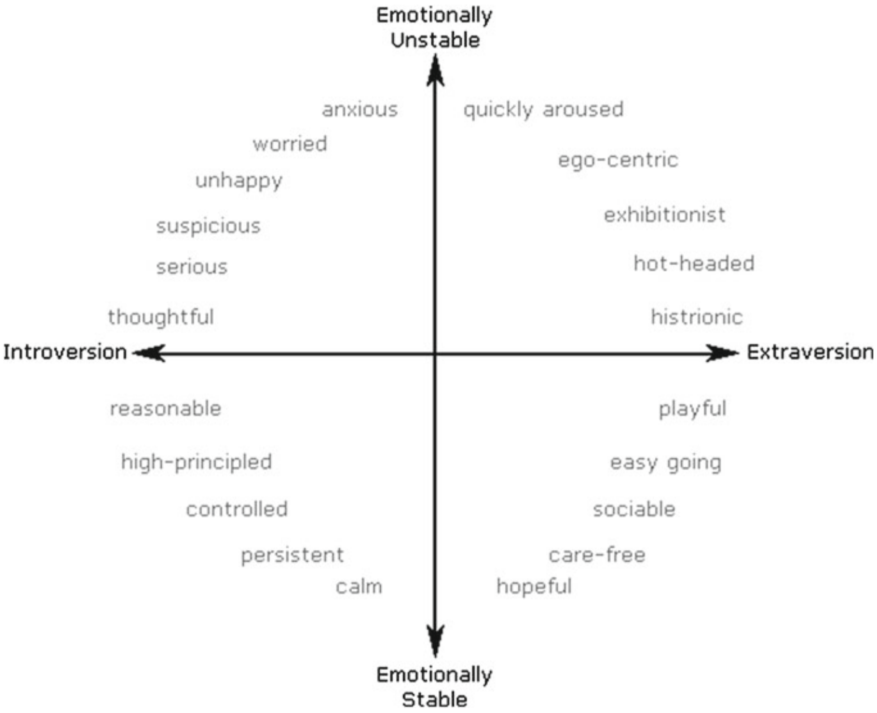

Fig. 1.1 Two dimensions of Eysenck's personality inventory including attributions

what he is mostly cited for today, i.e. the *PEN* inventory, a theory he has defended all his life time (Eysenck 1991).

1.2.3 Cattell's 16 Personality Source Traits

Also Raymond Cattell was an advocate of the empirical method of scientific psychology research and objective behavioral studies. Like Eysenck, he favored factor analytics instead of what he called *verbal theorizing*, which he saw as *subjective and poorly defined*. He wanted to align psychology with empirical principals, i.e. science, where a theory was to be tested in objective ways that could be understood and replicated by other researchers. To his mind, personality theorists often tended to provide little objective evidence. For example, in Cattell (1965) he analyzed over 400 contemporary works from literature of psychology colleagues for the topic "anxiety". As a result he stated

> The studies showed so many fundamentally different meanings used for anxiety and different ways of measuring it, that the studies could not even be integrated. […] Psychology appeared to be a jungle of confusing, conflicting, and arbitrary concepts. These pre-scientific

> theories doubtless contained insights which still surpass in refinement those depended upon by psychiatrists or psychologists today. But who knows, among the many brilliant ideas offered, which are the true ones? Some will claim that the statements of one theorist are correct, but others will favour the views of another. Then there is no objective way of sorting out the truth except through scientific research.

In his perspective, the means to that end were factor analyses as it could be used to comprehensibly reveal fundamental dimensions behind a certain number of measurable traits. He frequently used factor analyses, and successively developed improvements for this process, e.g. the so called *Screen Test*, which still today is widely used to determine the number of meaningful factors to be retained after factor analysis. Further, he developed factor analysis rotation for matching data at hand to a hypothesized factor structure, entitled *Procrustes* rotation. Although he is nowadays best known for his theory of dimensions of personality, he also used factor analyses to study basic dimensions of other domains, e.g. intelligence, motivation, and vocational interests. His main works concentrated on the exploration of the basic dimensions of personality, motivation, and cognitive abilities.

Cattell also pioneered in another sense, when he applied *multivariate* analyses. Most of his colleagues at that time examined psychology from an *univariate* perspective, i.e. the study of an observable effect that any single variable might have on any other variable. In Cattell (1966) he declares, that behavioral dimensions are highly complex and interactively interwoven. Trying to fully understand any one dimension in isolation was futile, he states. Cattell was highly aware of empirical work and its sensitivity to test design setup. He understood, that bringing test persons into artificial laboratory situations exercises influence on the test results as well as on the real-world validity. To his mind, the classical univariate approach represents such an artificial perspective. The multivariate approach on the contrary, allows psychologists to study the whole person and their unique combination of traits in a natural environment including real-world situations that could not be manipulated in a laboratory, he declares.

In the core of his work, Cattell applied multivariate research with respect to three domains: (1) the traits of personality or temperament; (2) the motivational or dynamic traits; and (3) the diverse dimensions of abilities. The constant assumption behind analyses in these domains was, that each of them extracts a finite number of *fundamental*, *unitary* elements hidden in the background, which can be identified. While believing that it was necessary to sample the widest possible range of variables in order to capture a full image of personality, he consciously distinguished between three conditions of influence in data acquisition:

- **L-data**: Life data, which measure an individual's natural, every day behavior and characteristic patterns in real world environment. This data encompasses, for example, the number of traffic accidents, the number of parties attended, or even the grade point average in school or number of illnesses or divorces.
- **T-data**: Experimental test data, emerging from reactions towards standardized empirical test setups created in a laboratory. This data can be generated by objective observation and measurement.

- **Q-data**: Questionnaire data, including introspective responses by the individual about behavior and feelings in self-analysis. Accordingly, this kind of questioning can capture more subtle internal states and viewpoints, that might not be captured from external observation.

Cattell required a factor extracted from factor analysis to be reproducibly present in all three data conditions in order to call it a *fundamental* and *unitary* dimension of personality. In the following years he constructed personality measures of a wide range of traits in each data condition, repeatedly performing factor analyses on the data.

In the 1940s he obtained the list of words from Allport–Odbert, cf. Sect. 1.2.1. Revising the list he added terms obtained from psychological research and deleted synonyms. He reduced the total list to 171 entries published in Cattell et al. (1957). He conducted user tests by asking subjects to rate people they knew based on his list. He then derived 35 major clusters of personality traits which he referred to as the *personality sphere*. In a next step he developed personality tests for these clusters. He also provided developing tests to measure these traits across different age ranges as well as distinguished between self-report and observer ratings.

Supported by works from contemporary colleagues, Cattell's factor-analytic studies were carried on over several decades, eventually producing the *16 fundamental personality factors*, abbreviated as *16PF*. Cattell labeled these 16 factors *primary factors* or later *source traits* because he believed they provide the underlying source for the surface behaviors we think of as personality (Hall and Lindzey 1978). Factor-analytic studies by many researchers in diverse cultures around the world have re-validated the number and meaning of these traits. An up-to-date collection of respective links to available literature can be found online in Cattell's Wikipedia listing (Wikipedia 2011).

Believing that observable source traits can be seen as manifestations of subtle personality tendencies Cattell later executed a factor analysis on the 16PF themselves. He derived a number of five factors, which he labeled *global traits*. These global traits were of broad, over-arching reference to domain and behavior, which he in turn used to assign meaning and structure to the source traits. For instance, the global trait he called *Extraversion* emerged from factor-analytic results based on the five primary traits that were of interpersonal focus. The combination of these two levels of personality, i.e. primary and global factors, provides a comprehensive picture of the whole personality. The global traits give overview of the persons broader personality setup, functioning in a generalized personality tendencies, and the primary trait capture more detailed information about the persons unique trait combinations.

1.2.4 Development of the "Big 5" Factor Inventories

Since the 1960s many researchers have worked independently in order to reveal underlying factor structures of personality. Many of them came to a five factor inventory or model, abbreviated *FFI* or *FFM*.

Already in 1961, two United States Air Force researchers, Ernest Tupes and Raymond Christal, conducted analyses of personality data including Cattell's trait measures using data from eight large samples. In a technical report they described five recurring factors, which they entitled *Surgency*, *Agreeableness*, *Dependability*, *Emotional Stability*, and *Culture* (Tupes and Christal 1961). Two years later this work was replicated by Warren Norman, whose analysis also revealed five major factors he entitled *Surgency*, *Agreeableness*, *Conscientiousness*, *Emotional Stability*, and *Culture* (Norman 1963). Norman also revised the list of attributes compiled by Allport and Odbert. He mainly cleaned the list from words referring to clinical, i.e. psychotic and anatomy related words.

The next two decades were characterized by dispute and cessation. Many researchers at that time aimed to use personality assessments to predict behaviour, most of whom failed. Walter Mischel, teaching Professor of Psychology in Stanford at that time, declared that personality tests could not predict behavior with a correlation of more than 0.3 (Mischel 1968). Sharing a sociological viewpoint on psychology with contemporary colleagues he demonstrated that attitudes and behavior were not stable, but depend on the situation. His publication triggered a paradigm crisis in personality psychology. Predicting behavior by personality tests became "impossible". It was not until the 1980s that this view was challenged, when the spirit of empiricism shifted from prediction of individual behavior towards prediction of patterns of behavior from large numbers of observations. Personality and social psychologists now generally agreed, that human behavior is determined by both personal and situational influences. Trait theories became respectable again.

By 1980, researchers had forgotten about the pioneering works by Tupes, Christal, and Norman for the most parts. In 1981 Lewis Goldberg began to conduct a lexical analysis reincorporating Normans' list of adjectives (Goldberg 1993a). Once again, results revealed an underlying five factors structure. Goldberg was the first to coin the nickname "Big 5" for these five factors observations, which he compare to Eysenck's super factors. Goldberg described the revealed factors and its title superscriptions as follows:

1. **Openness to Experience**: the tendency to be imaginative, independent, and interested in variety versus practical, conforming, and interested in routine.
2. **Conscientiousness**: the tendency to be organized, careful, and disciplined versus disorganized, careless, and impulsive.
3. **Extraversion**: the tendency to be sociable, fun-loving, and affectionate versus retiring, somber, and reserved.
4. **Agreeableness**: the tendency to be softhearted, trusting, and helpful versus ruthless, suspicious, and uncooperative.

5. **Neuroticism**: the tendency to be calm, secure, and self-satisfied versus anxious, insecure, and self-pitying

In the same year a symposium of prominent researchers, including Goldberg, was held. After selecting the most promising tests out of all available tests at that time, they coherently concluded that all these tests measured a subset of five common factors, similar to the ones proposed by Norman and Cattell (Goldberg 1980). This event triggered a general re-acceptance of the five factor inventory among personality researchers. Ultimately, the consortium agreed, that the found common factors can be calculated as continuous bipolar scales that are distinguishable from temporary states, e.g. emotions.

1.2.5 Costa and McCrae and the "NEO-" Inventories

In the 1970s, Paul T. Costa and Robert R. McCrae started to analyze the influence of age on personality. Starting to develop an own inventory, they focused on higher level representations such as Eysenck's or Goldberg's three or five trait inventories. They initially began to examine neuroticism and extraversion, two traits most widely agreed on by many researchers (Church and Katigbak 1976a). Conducting factor analyses they revealed another main factor, i.e. openness to experience. Costa and McCrae referred to the revealed inventory by their abbreviations, i.e. NEO or NEO-PI where "PI" abbreviates "personality inventory". Their results were initially published in the *Augmented Baltimore Longitudinal Study of Aging* (Shock et al. 1984) and one year later by the authors (Costa and McCrae 1985). The corresponding test comprised 48 items, i.e. questions in the questionnaire, for each of the three factors, which were also called *scales*. Further, each factor was built up by three sub-traits they entitled *facets*. Already four years later they extended their inventory by two more factors: agreeableness and conscientiousness (Costa and McCrae 1989) but it was not until 1992 that Costa and McCrae published a comprehensive five factor framework including facets for all factors, the revised NEO-PI-R (Costa and McCrae 1992b).

Figure 1.2 shows a list of the factors and facet names. As a memorable and mnemonic trick the five factors' abbreviations are sometimes arranged to build the acronym "OCEAN", or alternatively "CANOE". Eventually, the NEO-PI-R questionnaire that generates this inventory comprises 240 items, i.e. 48 items for each factor, each of which is built up by the displayed 6 facets. Each facet is captured by 8 items respectively. The NEO-PI-R takes about 45 minutes to fill out. Costa and McCrae (1992a) released a short version of NEO PI-R, the respective questionnaire is simply entitled *NEO-FFI*. It comprises 60 out of the 240 items, which take about 10 minutes to complete. Here, only 12 items build up the score for a factor. The most recent revision of the NEO-FFI was discussed in McCrae and Costa (2004). In this work, the authors discuss a variation in selection of the 60 items out of the NEO-PI-R items. The inventory shown in Fig. 1.2 however remains unchanged.

1. **Neuroticism**
   - Anxiety
   - Hostility
   - Depression
   - Self-Consciousness
   - Impulsiveness
   - Vulnerability to Stress
2. **Extraversion**
   - Warmth
   - Gregariousness
   - Assertiveness
   - Activity
   - Excitement Seeking
   - Positive Emotion
3. **Openness to experience**
   - Fantasy
   - Aesthetics
   - Feelings
   - Actions
   - Ideas
   - Values
4. **Agreeableness**
   - Trust
   - Straightforwardness
   - Altruism
   - Compliance
   - Modesty
   - Tendermindedness
5. **Conscientiousness**
   - Competence
   - Order
   - Dutifulness
   - Achievement Striving
   - Self-Discipline
   - Deliberation

Fig. 1.2 NEO-PI-R inventory consisting of 6 facets for each of 5 factors

For both forms, the long and the short one, a coding scheme transforms the item ratings into factor values, which range from 0 to 48. Low values designate a rather low strength or magnitude with respect to the factor at hand, while high values correspond to much of the characteristic. In the most recent releases, two forms of the questionnaire are offered, one for self report, and one for observer rating. In either way, raters answer the items on five point ordinal Likert scale, ranging from “strongly disagree” to “strongly agree”.

The NEO-FFI and the NEO-PI-R questionnaires for self or observer reports measure the same traits, which are defined as follows:

N **Neuroticism**: Low values indicate a calm and emotionally stable personality. People work well under pressure and are not easily agitated, while high values designate an emotionally unstable personality, i.e. people are easily shocked or ashamed, sometimes overwhelmed by feelings or nervousness, also generally not self-confident.

E **Extroversion**: Low values indicate a conservative personality, i.e. people are reserved and contemplating, while high values designate a rather sociable, energetic, independent personality.

O **Openness**: This scale estimates how people integrate new experiences or ideas in everyday life. Low values correspond to a conservativeness, preferring common-knowledge to avant-garde. High values indicate visionarity, curiosity, and open-minded behavior.

A **Agreeableness**: This scale corresponds to the ability of social reflection, commitment and trust. Low values indicate an egocentric, competitive and distrustful attitude. High values suggest that people are sympathetic, trustful, willing to be helpful.

C **Conscientiousness**: Low values indicate a careless, indifferent, reckless, even improperly acting personality, while high values designate accurate, careful, reliable and effectively planning behavior.

The NEO-FFI and the NEO-PI-R inventories have been validated with high consistency, including translations, cross-cultural experiments and retests. The internal factor consistency, measured as Cronbach's Alpha (cf. Sect. 4.2), based on a sample size of about 1.5 k test persons, results in high consistencies, i.e. $N = 0.92$, $E = 0.89$, $O = 0.87$, $A = 0.86$, $C = 0.90$, as reported in the NEO-PI-R manual. Internal consistencies for the facets ranged from 0.56 to 0.81. For the NEO-FFI, internal consistencies reported in the manual are: N = 0.79, E = 0.79, O = 0.80, A = 0.75, C = 0.83.

Eventually, the NEO-FFI seems to be used more frequently, and it also seems to be applied in more diverse situations. This may also be due to the shortened test time, which makes repeating measurements more feasible. Sherry et al. (2007) used the NEO-FFI to analyze perfectionism and reports internal consistencies of $N = 0.85$, $E = 0.80$, $O = 0.68$, $A = 0.75$, $C = 0.83$. Eggert et al. (2007) apply the NEO-PI-R to eating disorders and result in consistencies of 0.69–0.90 as well.

Also applications of NEO translations to other languages or cultures result in comparably high consistencies, e.g. Church and Katigbak (1976b) in the Philippines. Further details can also be obtained from McCrae and Allik, who published a selection of papers from researchers across the globe reporting on various issues in cross-cultural research using the NEO inventories in 2002 (McCrae and Allik 2002). The authors point out the robustness of the five factor inventory for, e.g., Chinese, Estonian, Finnish, German, Filipino, French, India, Portuguese, Russian, South Korean, Turkish, Vietnamese, and sub-Saharan cultures like Zimbabwe etc.

Rolland (2000) examined data from 16 different cultures and concludes that neuroticism, openness, and conscientiousness dimensions are robust in all cultures while extraversion and agreeableness are most sensitive to cultural context. The most recent publication includes 51 cultures (McCrae and Terracciano 2005). McCrae's results assert that the NEO-PI-R five factors and facets can be found across cultures. At this level, personality can be used to examine cultural differences, he declares.

Regarding dependencies between personality and age, much work has been conducted by McCrae and colleagues, e.g. McCrae et al. (1999). Most recent publications showed that the effect age exerts to personality is comparable in samples from Germany, Italy, Portugal, Croatia and South Korea. NEO-PI-R retest reliability is reported by the authors in the manual based on a study over 6 years. Results showed high reliabilities, $N = 0.83$, $E = 0.82$, $O = 0.83$, $A = 0.63$, $C = 0.79$. In Costa and McCrae (1992a) the authors declare, that the scores over 6 years are only marginally different from scores measured a few months apart.

As a limitation of age analyses McCrae points out, that personality development itself can be seen as relatively settled after adolescence. At the same time, the authors claim that, based on cross-cultural and longitudinal experiments, neuroticism and extraversion can decline with age, while agreeableness and conscientiousness can incline (Costa and McCrae 2006). A comprehensive analysis including 92 personality studies showed that social dominance, conscientiousness, and emotional stability increases with age, especially in the age of 20–40 (Roberts et al. 2006). In their work, the authors compared results from studies applying different personality inventories, among them NEO-PI-R.

1.2.6 The Development of the German NEO-FFI

While the English NEO-FFI and NEO-PI-R were published and explained in the same manual, the German versions were published individually. Borkenau and Ostendorf (1993) released the German version of the NEO-FFI, while the German version of the NEO-PI-R was releases in Ostendorf and Angleitner (2004). The initial sample size for the NEO-FFI analyses in German was 2,112 persons. Although the most recent revision of the English NEO-PI-R was discussed in McCrae and Costa (2004), Borkenau and Ostendorf (2008) declare, that the adjustment to the proposed changes in selection of NEO-FFI items out of NEO-PI-R items in English causes only marginal changes in the consistencies and factor parameters for German inventory. In contrast to this little improvement, the hitherto collected data, which grew to a sample size of 11,724 for the German version by that time, would thus be relegated for further use. The authors therefore decided not to incorporate the proposed changes. The most current sample sizes are monitored and updated by Costa and McCrae and can be obtained by downloading the following pdf file from the Internet.[1]

[1] http://www3.parinc.com/uploads/pdfs/NEO_bib.pdf.

Building up the German NEO-FFI inventory, Peter Borkenau and Fritz Ostendorf paid much attention to provide an exact translation of the English NEO-PI-R items. For reassurance, the German version was re-translated by a native English speaker and approved by Costa. In Borkenau and Ostendorf (1989), Ostendorf and Angleitner (1992) the authors conducted a joint analysis of the *Personality Research Form*, the *Freiburger Persönlichkeitsinventar*[2] (FPI), and the NEO-PI inventory. The *Personality Research Form* was published by Stumpf et al. (1985) and comprises a number of 234 items which aggregate to 14 traits. Retest reliability given by the authors range from 0.67 to 0.96. The test is available in German and English. The *Freiburger Persönlichkeitsinventar* was published by Fahrenberg et al. (1985) and comprises 138 items aggregating to 12 traits including Eysencks' neuroticism and extraversion. Internal consistencies were reported to range from 0.73 to 0.83. In the results of their analyses, Borkenau and Ostendorf revealed the NEO five factor structure. The most recent revision of the German NEO-FFI (Borkenau and Ostendorf 2008) is the second edition. The collection mostly consists of non-clinical data from approximately 50 individual studies. Also the German NEO-FFI has been validated using various stimuli in various situations, e.g. Angleitner and Ostendorf (2000) confirmed robustness of the Five Factor Model in German speaking countries like Austria, former East and West Germany and Switzerland.

1.2.7 Super Short Forms of Five Factor Inventories

Urged by the need to apply personality assessment also in situations where time is severely limited, Rammstedt and John (2007) introduced the so called *super short* "BFI-10". In this test, only 2 items constitute a scale value, the whole test consists of 10 items. For development of the super short inventory Rammstedt included samples from English and German language, choosing the Big Five Inventories from John et al. (1991), John and Srivastava (1999), known as "BFI-44", as a basis. The English sample consisted of 1552 students, half from a private university, half from a public university. The German sample consisted of 833 students. In accordance to the reference questionnaire, the test based on a 5-point Likert scales ranging from "strongly disagree" to "strongly agree". Also, it can be applied to both self-assessment and observer's-assessment. Reducing the inventory from 44 to 10 items, Rammstedt followed five criteria:

1. Inclusion of a high and a low pole item for each factor
2. Design the factors as broad as possible
3. Create identical versions for English and German
4. Prefer items with high correlation to BFI-44 scales
5. Prefer items showing low cross-loading effects

[2] translates to *personality inventory of Freiburg*; Freiburg is a city in the south of Germany.

In order to provide a comparison to NEO-FFI scales Rammstedt also includes NEO-FFI items in her test. As a result, she reports about correlations of 0.63 for O, 0.70 for C, 0.69 for E, 0.51 for A and 0.71 for N, given the English samples. For the German sample she reports correlations of 0.61 for O, 0.70 for C, 0.79 for E, 0.61 for A and 0.73 for N. While these results show moderate correlation for *C,E, and N*, weak correlation was found for *O, and A*. Rammstedt proposed to include another 11th item for adding information to A. Although figures increased, results stayed at a weak level. Mean consistency was reported with $\alpha = 0.75$. Looking at the somewhat low correlations, it is important to mention that lower correlations does not automatically mean that "less" personality information has been captured. This is because of the fact, that while the Big 5 are believed to be the most consistent and coherent method to capture personality, it cannot be proved that it is the only way to assess personality. On the other hand, what these numbers do show is that "other" information has indeed been captured, which somehow correlated to what the majority of researchers have agreed on as the Big Five Personality Traits. Hence the information captured clearly diverges from the overall agreement of the "Big Five" personality traits. Also other super-short inventories exist, e.g. Gosling (2003), which overall result in even lower consistencies and correlations. For a comprehensive overview see Grucza and Goldberg (2007), Amelang et al. (2006).

1.2.8 Trait Theories Comparison and Criticism

Most trait models, and even ancient Greek philosophy, incorporate concepts of extaversion and neuroticism, the latter sometimes entitled *emotional stability*. Eysenck's trait model consists of three factors: extroversion, neuroticism, and psychoticism (cf. Sect. 1.2.2). The Big Five model contains openness, extroversion, neuroticism, agreeableness, and conscientiousness (cf. Sects. 1.2.4 and 1.2.5). In common are extraversion and neuroticism, sharing associations with sociability and positive effect for high extraversion on the one side and emotional instability and negative effect for high neuroticism on the other side. Regarding Eysenck's psychoticism factor (Matthews et al. 2003) report, that it correlates with some of the facets of openness, agreeableness and conscientiousness, e.g. a high score on tough-mindedness in psychoticism correlates to a low score on tender-mindedness in agreeableness. Eventually, Eysenck's psychoticism factor has not become a factor in the Big Five model. A major reason for the denial was that psychoticism assessments do not fit a normal distribution curve. Here, scores are rarely high, causing a considerable overlap with psychiatric conditions such as antisocial and schizoid personality disorders. All other factors of the Big Five showed normal distribution.

Further, all Big Five factors are generated by methods of statistical factor analyses. Differences between Cattell and Eysenck Five Factor models are caused by preferences for different factor rotation techniques, i.e. Cattell used oblique rotation while Eysenck used orthogonal rotation. Despite the fact that there are often weak or even moderate correlations between the factors, they are widely interpreted as independent, i.e. uncorrelated and orthogonal. Congruent to this general assumption also Costa and McCrae used orthogonal rotation. As a consequence for application of personality analyses, traits are often analyzed separately.

As explained in Sect. 1.2.4 many psychologists currently believe that five factors are sufficient to represent a person's overall personality, while inventories incorporating more than five traits are widely interpreted to generate a more exhaustive and more detailed picture of the personality. Ultimately, these Big Five factors are seen as a conceptual framework which integrates all the major research findings and theories in personality psychology. On the other hand, due to this comprehensive integration, the Big Five have shown to be less powerful for prediction and explanation of isolated actual behavior when compared to the more numerous lower-level trait models, cf. Mershon and Gorsuch (1988), Paunonen and Ashton (2001).

Finally, there are only very few empirical studies comparing different personality inventories by means of correlation or congruence. One such work has been carried out by Grucza and Goldberg (2007), who compare 11 major inventories in a comprehensive article.

With regard to early analyzes and database generation, the Big Five have a substantial history in clinical application. Also, basic studies and user test have been conducted for clinical purposes or have even been executed in clinical environments when developing the Big Five factor structure. A comprehensive compilation of results can be found in Saulsman and Page (2004). In this work, Saulsman and Page analyze correlations between the Big Five and each of the 10 personality disorder categories as defined in the *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-TR)*, cf. American Psychiatric Association (2010), published by the American Psychiatric Association.[3] As a result, the authors were able to assign unique and predictable five factor profiles to each of the disorders.

Furthermore, the five factor methodology has been applied successfully in the area of assessment centers and job performance estimation. Barrick and Mount (1991), Mount and Barrick (1998) analyze 117 studies comprising 23,994 participants. According to their results conscientiousness showed consistent relations with all performance criteria for all occupational groups. The authors show, that extraversion is a valid predictor for occupations involving social interaction such as management and sales. Finally, extraversion and openness to experience were valid predictors of training proficiency criteria.

Much criticism of the Big Five is directed to the heterogeneity in choice of subjects when developing the factor structure. As the factors were acknowledged by a group of experts, the experts themselves relied on their own experiences incorporating both clinical and non-clinical studies.

Importantly, in some experiments ratings were generated by either the subjects at hand themselves or by observers or peers filling out the respective questionnaires. Although the underlying structures showed congruences with the Big Five, the actual ratings were not necessarily identical. Ball et al. (1997), for example, compared ratings of two German twins and revealed correlation between two peer-raters of 0.63

[3] http://www.psych.org/.

for NEO-FFI neuroticism. Correlation between self-rated neuroticism and peer-rated neuroticism showed only a magnitude of 0.55. In many studies researchers observed similar inconsistencies between self-ratings and ratings from observers. One possible explanation is offered by Edwards (1953). Accordingly, the mismatch between the two forms of assessment may be due to different perspectives, i.e. while objective observers are believed to use similar cues, self-reports may be more influenced by factors such as the desirability of a trait. Elaborating on the characteristics of personality assessments in a call center scenario Hogan (1982) declares:

> While the agents' perspective conceptually taps into a person's identity (or personality from the inside), the observers' perspective in contrast taps into a person's reputation (or personality from the outside). Both facets of personality have important psychological implications. A person's identity shapes the way the person experiences the world. A person's reputation, however, is psychologically not less important: it determines whether people get hired or fired (e.g., reputation of honesty), get married or divorced, get adored or stigmatised. [...] Given that in everyday life people act as observers of other peoples behaviours most of the time, the external perspective naturally has both high theoretical importance and social relevance.

1.3 Summary

This chapter gives an introduction into the understanding and assessment of personality. While many perspectives and models have been proposed in the literature in order to capture or estimate personality, the chosen approach underlines the empirical character of the phenomenon at hand. Following Kassin (2003) and the so-called *Trait Theory* of personality, personality itself is seen as a *defined set of habitual patterns of behavior, thoughts, and emotions*, that manifests itself in term of *measurable traits* throughout the present work. Moreover, these traits are generally seen to be relatively stable over time and situation. As a result, there exist a multitude of proposed traits and trait collections that have shown to be useful to describe personality in different works or scenarios. In this respect, major approaches have been outlined in this chapter. Details are given for those theories, that have contributed to build the one common sense model acknowledged by most researchers across the globe today, i.e. the *Big Five* personality traits. At the same time, underlying data and trait extraction methods were outlined. This work further follows the most frequently applied assessment scheme that is able to extract the Big Five out of empirical user data, i.e. the *NEO-FFI* questionnaire. It was defined by McCrae and Costa (2004) and translated into German language by Borkenau and Ostendorf (2008). The questionnaire is explained and analyzed with respect to strengths, weaknesses and factors of influence.

Eventually, when speaking about personality in the following chapters and experiments, the overall composition of personality is believed to be captured by the following five traits: *openness for experience, conscientiousness, extroversion, agreeableness*, and *neutoricism*. A short comparison of different models that have contributed to the NEO-FFI so far as well as common aspects between these models are given in the last section, cf. Sect. 1.2.8.

References

Ahles S (2004) Our inner world: a guide to psychodynamics and psychotherapy. Johns Hopkins University Press, Baltimore

Allport G, Odbert H (1936) Trait-names: a psycholexical study. Psychol Monogr 47(211):9–30

Allport GW (1937) Personality: a psychological interpretation. Holt, Rinehart & Winston, New York

Amelang M, Bartussek D, Stemmler G, Hagemann, D (2006) Differentielle Psychologie und Persönlichkeitsforschung. W. Kohlhammer

American Psychiatric Association (2010). Diagnostic and statistical manual of mental disorders : DSM-IV-TR. American Psychiatric Association, 4th edn

Angleitner A, Ostendorf F (2000) The FFM: A comparison of German speaking countries (Austria, Former East and West Germany, and Switzerland). Psychology Press Ltd, In: XXVIIth International Congress of Psychology, Stockholm, Sweden

Ball D, Hill L, Freeman B, Eley TC, Strelau J, Riemann R, Spinath FM, Angleitner A, Plomin R (1997) The serotonin transporter gene and peer-rated neuroticism. NeuroReport 8(5):1301–1304

Bandura A (1986) Social foundations of thought and action: a social cognitive theory. Prentice-Hall, Upper Saddle River

Barrick MR, Mount MK (1991) The big five personality dimensions and job performance: a meta-analysis. Pers Psychol 44(1–26):613–625

Borkenau P, Ostendorf F (1989) Descriptive consistency and social desirability in self- and peer reports. Eur J Pers 3:31–45

Borkenau P, Ostendorf F (1993) NEO-Fünf-Faktoren-Inventar nach Costa und McCrae (1.Aufl.). Hogrefe, Boston

Borkenau P, Ostendorf F (2008) NEO-Fünf-Faktoren Inventar nach Costa und McCrae (NEO-FFI). Manual (2., neu normierte und vollständig berarbeitete Auflage). Hogrefe, Boston

Cattell RB (1966) The meaning and strategic use of factor analysis. In Handbook of Multivariate Experimental Psychology. Rand McNally, Chicago, USA

Cattell RB, Marshall M, Georgiades S (1957) Personality and motivation: structure and measurement. J Pers Disord 19(1):53–67

Cattell RB (1965) The scientific analysis of personality. Penguin, London

Church AT, Katigbak MS (1976a) Age differences in personality structure: a cluster analytic approach. J Gerontol 31(5):564–570

Church AT, Katigbak MS (1976b) Indigenization of psychology in the philippines. Int J Psychol 37(3):129–148

Costa PT, McCrae RR (1992a) NEO PI-R professional manual. Psychological Assessment Resources

Costa PT, McCrae RR (1992b) Revised NEO personality inventory (NEO-PI-R) and NEO five-factor inventory (NEO-FFI) manual. Psychol Assess Resour 76(3):412–420

Costa PT, McCrae RR (2006) Age changes in personality and their origins: comment on roberts, Walton, and Viechtbauer. http://www.psych.utoronto.ca/~geneviev/age%20changes%20in20personality.pdf. Accessed 29 April 2011

Costa PT, McCrae RR (1985) The NEO personality inventory manual. Psychol Assess Resour 1:106–107

Costa PT, McCrae RR (1989) NEO PI/FFI manual supplement for use with the NEO Personality Inventory and the NEO five-factor inventory. Psychol Assess Resour 18:119–144

Edwards AL (1953) The relationship between the judged desirability or a trait and the probability that the trait will be endorsed. J Appl Psychol 37:90–93

Eggert J, Levendosky A, Klump K (2007) Relationships among attachment styles, personality characteristics, and disordered eating. Int J of Eat Disord 20(2):149–155

Eysenck H, Lewis A (1947) Dimensions of personality; with a foreword by Sir Aubrey Lewis. Routledge & Kegan Paul, London

Eysenck H, Prell D (1951) The inheritance of neuroticism: an experimental study. J Ment Health 97:441–465

Eysenck H (1967) The biological basis of personality. Thomas, New York

Eysenck H (1991) Dimensions of personality: 16: 5 or 3? criteria for a taxonomic paradigm. Pers Individ Differ 12:773–790

Fahrenberg J, Hampel R, Selg H (1985) Die revidierte form des freiburger persönlichkeitsinventars FPI-R. Diagnostica 31:1–21

Freud S (1923) The ego and the Id. Norton & Company, New York

Goldberg LR (1980) Some ruminations about the structure of individual differences: developing a common lexicon for the major characteristics of human personality. Technical report, symposium presentation at the meeting of the western psychological association, Honolulu, HI

Goldberg LR (1993a) Language and individual differences: the search for universals in personality lexicons. Rev Pers Soc Psychol 1:141–165

Goldberg LR (1993b) The structure of phenotypic personality traits. Am Psychol 48:26–34

Gosling S (2003) A very brief measure of the big-five personality domains. J Res Pers 37(6):504–528

Grucza RA, Goldberg LR (2007) The comparative validity of 11 modern personality inventories: predictions of behavioral acts, informant reports, and clinical indicators. J Pers Assess 89:167–187

Hall CS, Lindzey G (1978) Theories of personality, 3rd edn. Wiley, New York

Herrmann T (1969) Lehrbuch der empirischen persönlichkeitsforschung, 2nd edn. Verlag für Psychologie Hogrefe, Boston

Hogan R (1982) A socioanalytic theory of personality. Nebr Symp Motiv 1982:5589

John OP, Donahue EM, Kentle RL (1991) The big five inventory - versions 4a and 54. University of California, Berkeley, Institute of Personality and Social Research

John OP, Srivastava S (1999) The big five trait taxonomy: history, measurement, and theoretical perspectives. Eur J Pers 2(2):102–138

John O, Robins R, Pervin L (2008) Handbook of personality: theory and research, 3rd edn. The Guilford Press, New York

Kassin S (2003) Psychology. Prentice Hall, Upper Saddle River

Matthews G, Deary I, Whiteman M (2003) Personality traits, 2nd edn. Cambridge University Press, Cambridge

McCrae RR, Costa PTJ, Lima MP, Simões A, Ostendorf F, Angleitner A, Marušic I, Bratko D et al (1999) Age differences in personality across the adult life span: parallels in five cultures. Dev Psychol 35(2):466–477

McCrae RR, Allik J (2002) The five-factor model of personality across cultures. Kluwer Academic Publisher, Boston

McCrae RR, Costa PT (2004) A contemplated revision of the NEO five-factor inventory. Pers Individ Differ 36(3):587–596

McCrae RR, Terracciano A (2005) Personality profiles of cultures: aggregate personality traits. J Pers Soc Psychol 89(3):407–425

Mershon B, Gorsuch R (1988) Number of factors in the personality sphere: does increase in factors increase predictability of real-life criteria? J Pers Soc Psychol 55:675–680

Mischel W (1968) Personality and assessment. Wiley, London

Mount MK, Barrick MR (1998) Five reasons why the big five article has been frequently cited. Pers Psychol 51(849–857):613–625

Norman WT (1963) Toward an adequate taxonomy of personality attributes: replicated factor structure in peer nomination personality ratings. J Abnorm Soc Psychol 66:574–583

Ostendorf F, Angleitner A (1992) On the generality and comprehensiveness of the five factor model of personality: evidence for five robust factors in questionnaire data. In: Caprara GV, van Heck GL (eds) In modern personality psychology. Harvester Wheatsheaf, New York, pp 73–109

Ostendorf F, Angleitner A (2004) NEO-Persönlichkeitsinventar nach Costa und McCrae., NEO-PI-R. Revidierte FassungHogrefe, Boston

Paunonen S, Ashton M (2001) Big Five factors and facets and the prediction of behavior. J Pers Soc Psychol 81:524–539

Pervin L, Cervone D, John O (2005) Persönlichkeitstheorien, 5th edn. UTB, Stuttgart

Rammstedt B, John O (2007) Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. J Res Pers 41(1):203–212

Raymond Cattell – Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/wiki/Raymond_Cattell. Accessed 27 April 2011

Roberts BW, Walton KE, Viechtbauer W (2006) Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. http://www.psych.uiuc.edu/~broberts/Roberts,%20Walton,%20&%20Viechtbauer,%202005.pdf. Accessed 29 April 2011

Rolland JP (2000) Cross-cultural validity of the five factor model of personality. In: XXVIIth International Congress of Psychology, Stockholm, Sweden

Ryckman RM (2004) Theories of personality. Thomson/Wadsworth, Belmont

Saulsman LM, Page AC (2004) The five-factor model and personality disorder empirical literature: A meta-analytic review. Clin Psychol Rev 23:1055–1085

Sherry SB, Hewitt PL, Flett GL, Lee-Baggley DL, Hall PA (2007) Trait perfectionism and perfectionistic self-presentation in personality pathology. Pers Individ Differ 42(3):477–490

Shock NW, Greulich RC, Andres R, Arenberg D, Costa PT, Lakatta EG et al (1984) Normal human aging: the baltimore longitudinal study of aging. National Institutes of Health, Bethesda

Skinner BF (1984) Verbal behavior. Copley Publishing Group, Acton

Snygg D, Combs AW, Murphy G (1998) Individual behavior. A new frame of reference for psychology. Harper and Brothers. http://webspace.ship.edu/cgboer/snygg&combs.html. Accessed 7 June 2011

Stumpf H, Angleitner A, Wieck T, Jackson D, Beloch-Till H (1985) Deutsche personality research form (PRF). Hogrefe, Handanweisung

Tupes EC, Christal RE (1961) Recurrent personality factors based on trait ratings

Chapter 2
Speech-Based Personality Assessment

> *Voices are not merely a handy means to transmit information to the user. All voices—natural, recorded, or synthetic—activate automatic judgments about personality.*
>
> Nass et al. (1995)

As the previous chapter outlined approaches to personality assessment in psychology, this chapter summarizes works and insights of researchers from the speech community. Many of these researchers are linguists or computer scientists, hence the aim of approaching an individual's personality translates into the aim of modeling or experimenting with personality. Essentially, the assessment of perceivable manifestations of personality is the basis for any experimentation. When analyzing personality in terms of speech, the scope of interest is narrowed down from overall personality, i.e., maybe being able to judge about personality from previous knowledge about actions or incidents, towards focusing on perceivable characteristics, in this respect it means perceivable at the very point in time the conversation or the experiment occurs as well as comprehensible to any person including persons having no prior knowledge about the speaker. Resulting limitations and cleavages of this respective will be addressed in the present chapter.

One early remarkable study was carried out by Pear (1931). Conducting a recognition test, Pear played nine voices over the radio, all of whom spoke the same text passage taken from Dickens. He provided a feedback form in the newspaper and received about 5,000 responses. Only two voices were notably recognized, namely an actor and a judge, but his primary finding was the importance of vocal stereotypes he could deduce from 600 of the participants who also provided detailed comments on each of the speaker. Elaborating on these stereotypes, Allport and Cantril (1934), Cantril and Allport (1935) obtained more evidence. The authors conducted eight laboratory experiments using recordings of three male speakers. Listeners were asked to judge "outer" characteristics like age, hight and appearance, as well as "inner" characteristics like vocation, political preference, extroversion, dominant values, and

T. Polzehl, *Personality in Speech*, T-Labs Series in Telecommunication Services,
DOI 10.1007/978-3-319-09516-5_2

ascendance.[1] Further, the listeners also matched the speakers to summary sketches of personality. Subsuming results from this experiment and further subsequently conducted smaller radio studies the authors eventually observed, that although the obtained ratings showed consistency in stereotyping, the raters were oftentimes wrong in choice. The authors declared, that the uniformity of opinion regarding the personality of radio speakers was somewhat in excess of the accuracy of such opinion. Many authors in many studies were able to prove that a consistent and intuitive perception of personality from speech exists. Sanford (1942) declared:

> With respect to voice and personality, we can start with the evidence that they are related. The analytical approach, if judiciously employed, may clarify the relationship. If such an approach reveals no relationship, we would be forced to conclude that it may be the fault of the approach.

Like Sanford, many researchers at that time were of the opinion, that failure and inconsistencies in results must be due to the lack of sound analytical methodology rather than neglecting any association between personality and speech. Similar to results from Pear, Allport and Cantril, also Taylor (1934) obtained high consistency between ratings. In his study he asked 20 subjects to check a list of 136 items in a listening test including recordings from 20 male voices.

Taylors' primary focus was, however, set on the relation between self-ratings and observer's ratings on personality. When comparing these results he noticed that the ratings diverged considerably. More recent experiments also found this discrepancy, e.g., Hunt and Lin (1967), Ball et al. (1997).

But when looking closer at the alleged failure in self-impression and observer's impression congruence as well as when looking at alleged errors in the recognition tests from Pear, Allport and Cantril, one notices that the task of assessing a personality was processed in parallel and in addition to other tasks, transferring the discussion into a question of "truth" when asking for the "true" personality, assuming the person his- or herself would be able to tell the actual true personality. Similarly, asking for the "true" occupation of a speaker assumes that occupation was audible and related to perceived personality. Although these additional tasks failed, the ability to assess personality-related information from speech could be verified in its essence. Moreover, when taking a more general perspective on vocal communication, it is predominantly not reasonable to match personality impressions to any matter-of-fact inquiry, unless the aim of the conversation is to find out a matter-of-fact truth.

Rather, humans acquire a perception and reflection of a personality impression in terms of the very communicative behavior that the person exhibits. This behavior is not to be judged "true" or "false", but should be seen as manifestation of a set of characteristics possessed by a person that uniquely influences his or her cognitions, motivations, and behaviors in various speech situations, which is entirely congruent to the definition of personality given in the introductory quote above Chap. 1.

The introductory quote above the present chapter, resembling a contemporary theory from Nass and Brave (2005), describes a further basic observation of personality

[1] The terms "outer" and "inner" were chosen by the authors.

in vocal communication. In Nass et al. (1995) the authors report that people assign personality rapidly and automatically. Given this automatism, personality can also be seen as an instinct allowing humans to quickly construct a model of another person. This model encompasses all our expectations as learned by our experiences. If a person at hand is unknown, humans nevertheless assume certain characteristics, which may be taken from persons we assume to be similar. We automatically predict a wide range of attitudes, behavior, and other properties. After prediction, we expect to encounter the predicted behavior. In many cases, we are even able to sense small deviations from this expected behavior. In this respect, perceived personality may be seen as the entity of all possible behavior sensible in speech communication that manifest personality traits. Note that this definition is purposely set in broad terms, hence it also includes non-verbal and unconsciously exhibited information. In accordance, Nass unanimously concludes that personality is certainly the richest means for characterizing, and even classifying people.

Note, that Nass consciously includes synthetic voices in his statement. Accordingly, also voices in voice user interfaces leave a personality impression. Another major insight behind this statement is that *all* voices leave an impression. If designers of these interfaces are not fully aware of this automatism, they unconsciously leave the perception of their interfaces to fate, resulting in random or even unpleasant results. Also, while the concept of personality gives us cues on what to expect from others, it can also influence how we in turn behave ourselves in the communication. Consequently, the user's behavior will be influenced by his or her perception of any system's personality as well. Nass and Brave developed the so called *Computers as Social Actors (CASA)* interaction paradigm based on these observations. In Reeves and Nass (1996) the authors declare, that when people encounter someone who seems to have a personality like their own, they tend to have positive feelings toward that person. This paradigm became widely known as *similarity attraction*, cf. Nass and Brave (2005) for a detailed list of suggestions for further reading. Wrapping up their arguments, Brave and Nass proposed that designers of user interfaces should seek to manipulate the speech characteristics of their technology to consciously give it a personality.

Cassell et al. (2000) extend Nass' perspectives to include *Embodied Conversational Agents (ECAs)*. In Bickmore and Cassell (2004) the authors show, that perceived personality of the agent is a major factor in the overall perception of such user interfaces. When evaluating ECAs, Catrambone et al. (2002) argue, that personality is a factor to be included in evaluations of ECAs. Examining and understanding the mechanisms of observable personality from real interactions will eventually enable designers of ECAs to generate an appropriate system personality. More literature focusing on analysis on personality and ECAs can be found in Chen et al. (2010). The aspect of how to model personality and emotion in ECAs, once these user traits and states can reliably be estimated is discussed in Breese and Ball (1998). The authors propose Bayesian networks to manipulate parameters affecting word choice, syntactic framing, speech pace, rhythm, pitch contour, gesture, expression, and body language.

Finally, very few experiments have focused on interplay of rater and subject personality. Reeves and Nass (1996), Cassell and Bickmore (2003) have conducted experiments on attribution of personality of call center agents, also asking for the raters' personality. As a result, the authors indicated, that the attribution also depends on the users own personality.

2.1 Contemporary Terminology in Speech Analysis

When it comes to understand literature on personality estimation from speech, some explanations on terminology and information organization may be helpful to understand strengths and delineations of the approaches. When tackling perceivable manifestations of personality researchers have postulated a variety of terms and approaches regarding information organization, sub-grouping, and expansion of meaningful entities.

2.1.1 Prosodic Information and their Temporal Expansion

Analyzing vocal expressions researchers predominantly focus on the so called *prosodic* speech properties. The term *Prosody* originates from ancient Greek. It translates into a sound-related intrinsic "sung-along" to the actual words. Similar to the disagreement on definitions of personality, also various approaches to prosody coexist. As a common ground, prosodic analyzes often focus on perceptive phenomena like rhythm, stress and intonation patterns of speech, i.e., non-verbal speech properties. Theories mostly dissent in terms of the basic unit of a prosodic entity. Some researchers manifest prosodic phenomena on an expansion unit they call *supra-segmental*. Here, prosody is seen to be a characteristic influence affecting multiple segments, e.g., more than a word, a syllable or a phone. But when Paeschke (2003) analyzed phone onsets and constituents of quasi-periodic voiced speech sounds which were smaller than a phoneme, she found significant differences in terms of emotional expression on these levels, too.

At the same time, prosodic expression may overlay a multiplicity of words. Here, many researchers seek to determine meaningful phrases. In general, phrase structures occur due to different reasons:

1. The need to breath causing a perceivable speech pause
2. The need to insert pauses for linguistic organization and disambiguation
3. The expression of supra-segmental meaning, i.e., emotion, attitudes, interest, etc.

While the first need is simply a physiological imperative, speakers often initiate respiratory breaks at points where the resulting disjuncture will not negatively impact the information transmission flow. On the other hand, phrasing can be utilized to realize syntactic or organizational constrains. Here, intonational phrasing is frequently used to disambiguate and contribute to the intended meaning of an utterance (Pierrehumbert 1979). Accordingly, prosodic phrasing is described acoustically as the presence of *perceived disjuncture*. In the description four major acoustic indicators exist: the presence of silence, pitch and energy reset, pre-boundary lengthening and changes in speaking rate across the phrase boundary. All indicators contribute to the perception of increased disjuncture at a word boundary individually or jointly.

2.1.2 Extralinguistic and Paralinguistic Information

Organizing the information many researchers differentiate between *extralinguistic* and *paralinguistic* information. Paralinguistic information is mostly seen as any non-verbal element of communication used to modify meaning or convey emotions. This information can be expressed consciously or unconsciously. While in some situations we may be aware of our attitudes and emotions, in other we will not notice giving away these information. This is also true for extralinguistic information, which is mostly seen as long-term or even permanent non-verbal characteristic. Once more, when looking at definitions, the specific situation at hand influences the exact definition. As emotions are widely seen to be non-permanent, i.e., short-term spontaneous behavior, they are widely seen as paralinguistic information. A typical example for extralinguistic information is gender, as, in principle, the sex of a speaker does not change by itself.

Also age and personality are often characterized as extralinguistic information. But this is already where definitions need to be adapted to specific analyses objectives. Analyzing speech behavior, no absolute constants can be assumed since humans continuously age and physics will change while getting older. Many researchers therefore resort to the term "long-term" rather than permanent. If we take a closer look at personality, also personality can change within a lifetime. It develops while growing up and is believed to be relatively stable after adolescence, as has been motivated in Sect. 1.1. As for extralinguistic information, Winkler (2009) shows significant differences in intonation characteristics when analyzing aging voices. Winkler explains, that caused by physiological tissue aging, women's voices tend to become lower while men's voices tend to become higher. However, in short-term analyses, e.g., looking for emotions, age is most frequently and most certainly regarded as stable, i.e., extralinguistic.

2.2 Vocal Cues of Personality Perception

In this chapter related works are organized similar to Kreiman and Sidtis (2011). While the actual number and categorization differs, also Kreiman sorted findings from the literature according to the line of underlying research methodology.[2] The categories used in this work are:

1. Purely descriptive studies linking perceptual speech properties and personality mostly by means of rating-based studies
2. Studies incorporating correlations between perceptual ratings and acoustic measurements

[2] Kreiman's categories are basis for the current work, which extends the proposed categories.

3. Studies manipulating acoustic parameters by means of speech synthesis in order to cause changes in personality perception
4. Studies applying signal-based automated personality modeling
5. Other related studies.

2.2.1 Linking Perceptual Speech Properties and Personality

When looking at the category of purely descriptive, rating-based studies, most of the the works were carried out using observer's ratings with recorded voices unknown to the observers. One early study belonging to this category is the work on "outer" and "inner" characteristics by Allport and Cantril (1934), Cantril and Allport (1935), introduced in Sect. 2.2. The authors found the following associations:

- loud, boisterous voices were judged to be extroverted
- gentle, restrained voices were judged to be introverted
- forceful, aggressive voices were judged to be ascendant
- passive, meek voices were perceived as submissive

Stagner (1936) conducted a listening test asking students to rate on what he called "vocal characteristics" like aggressiveness, nervousness, "general impression", intensity, poise, flow of speech and clearness, given five male and five female voices. On the one hand, he observed high consistency on all scales except aggressiveness, but on the other hand no significant correlations to personality could be found. However, results showed moderate correlations between intensity and aggressiveness, as well as between flow of speech and clearness and nervousness. Moore (1939) reports on correlations between introversion and lower dominance and a breathy voice in a related study. Mallory and Miller (1958) found correlations between dominance and perceptive loudness, resonance and low pitch.

Addington (1968) trained two male and two female speakers to manipulate their speech delivery in terms of seven voice qualities: breathy, tense, thin, flat, throaty, nasal, "orotund",[3] while reading a text passage. Readings were also varied in terms of speaking rate and pitch variability. Conducting perception tests he asked listeners to rate the readings in terms of 40 personality scales and observed high agreement in general. However, agreement varies with respect to the scales, ranging from 0.94 for the scale capturing masculinity versus femininity to 0.64 for the scale capturing extroversion versus introversion. Table 2.1 shows the many gender dependencies found. Increasing throatiness lead to a more sophisticated impression regarding male voices, while the female speakers were perceived less intelligent. Also increased vocal tenseness lead to an impression of an older speaker regarding male voices, while it was perceived as younger and more emotional regarding female voices. Gender-consistent results were obtained by increasing flatness and varying speaking rates.

[3] The term translates to *sonorous, pompous*.

Table 2.1 Correlations between variation of speaking styles and personality ratings according to Addington (1968)

Variation	Gender	Correlation to ratings
Increased breathiness	m	Younger, more artistic
	f	Prettier, shallower, more feminine, petite, effervescent, highly strung
Increased flatness	m/f	Colder, more withdrawn, sluggish, masculine
Increased throatiness	m	Older, more realistic, sophisticated
	f	Lazier, uglier, more masculine, boorish, careless, neurotic, apathetic, less intelligent
Faster speaking rate	m/f	More animated, extroverted
More pitch variation	m	More dynamic, feminine, aesthetic
	f	More dynamic, extroverted
More tensed	m	Older, more unyielding
	f	Younger, more feminine, emotional, highly strung, less intelligent
More orotunded	m	Healthier, prouder, more energetic, sophisticated, artistic, interesting, enthusiastic
	f	Livelier, prouder, more gregarious, sensitive, humorless

Gender effects: *m* male, *f* female

Scherer (1974) obtained correlations between rated vocal effort and extroversion, sociability and emotional stability. He further found correlations between a lack of pause and/or irregularities in speaking rate and extroversion, competence and likability. For related work on vocal attractiveness see Zuckerman and Driver (1989). Stereotypes of babyish sounding voices and associations to weakness, dominance, vulnerability have been reported in Berry (1990).

Nass and Lee (2001) give more assumptions on personality-related speech characteristics. Accordingly, humans assume that extroverted persons talk more than they listen to other people. It is assumed that they use strong language. Introverted persons are assumed to listen more than they talk, and use qualifiers such as *maybe*, or *perhaps* more frequently. Nass further declares, that assessing personality from voice was an evolved skill, due to the fact that one could often hear people before seeing them, which lead to improve skills of anticipation of speaker characteristics from voice. Illustrating the combination of the personality indicators he hypothesizes

> When people meet someone who speaks loudly and rapidly, in a high pitch, and with a wide voice range they are feeling confident that they are dealing with the life of the party. Conversely, when people hear a soft, deep, monotone voice speaking slowly, they feel equally confident that this person is shy.

According to Kreiman and Sidtis (2011), many findings in the literature are bound to be speaker-dependent if the speaker sample size is small. Stagner (1936) conducted his tests on basis of five male and five female American voices. Addington (1968) included American 4 speakers only. Generalization of these findings therefore appears questionable. Scherer (1974) included 21 American and 22 German voices. At the same time, his raters were professional phoneticians, who might be more likely to detect personality expressions, so his results might change when laymen judge personality.

This point touches an ongoing controversy within the community of speech researchers. Very frequently, the ability of speakers or listeners to perform the required voice or to perceive the personality expression leads to discussions. Laymen may simply fail to generate the desired variations. Therefore, many studies relied on professionally acted manipulation, i.e., employing professional speakers and speech actors. Taking a closer look, the performances of actors are mostly perceived as instructed. At the same time they are often identified as professionally acted, translating into an over-enunciated or blunt impression. But this impression is predominantly additional to the targeted personality impression. In the first place, the actings may actually be perceived correctly by the raters, which then is a major strength. The drawback on the other side is the forfeited authenticity and realistic character when it comes to every-day speech communication.

2.2.2 Correlation between Perceptual Personality and Acoustic Measurements

In the second category studies seek to find correlations between ratings and acoustic measurements, frequently also examining or even arranging the speaker's variations in delivery of voice. For example, Aronovitch (1976) was able to establish links between acoustic characteristics and ratings as shown in Table 2.2. He obtained high consistencies within the ratings, although the revealed correlations differed considerably according to speakers' sex. Male voices showed high correlation to some personality ratings in terms of dynamic attributes like variances of pitch or intensity. Female voices showed more links to personality ratings, especially in terms of more static audio measures like means of intensity or pitch.

In recapitulation of literature that had been published by 1980, Scherer and Scherer (1981) declare:

> Three major parameters of the acoustic speech signal [...] are likely to be affected by personality variables: fundamental frequency (level and variability), vocal energy or intensity (level and variability), and energy distribution in the voice spectrum. These objectively measurable acoustic cues correspond to perceptual cues as experienced by observers, namely, pitch, loudness and voice quality.

Due to muscular affection, pitch seems to be a reliable indication of psychological arousal, Scherer declares. In a more extensive early study involving 372 female American students as speakers, Mallory and Miller (1958) found lower subjectively rated pitch for dominant and extroverted girls. On the other hand, male American speakers who rated themselves high on achievement, task ability, sociability, dominance and aggressiveness, and who were rated in the same way by observers,

Table 2.2 Correlations between acoustic measurements and personality ratings according to Aronovitch (1976)

Acoustic measure	Association
Male speakers	
Intensity variance	Self-doubting/self-confident Extraversion/introversion Boldness/caution Submissiveness/dominance
F0[a] variance	Submissiveness/dominance
Female speakers	
Mean intensity	Self-doubting/self-confident Extraversion/introversion Boldness/caution Laziness/energy
Speaking rate	Self-doubting/self-confident Boldness/caution
Mean F0[a]	Kindness/cruelty Maturity/immaturity Emotionality/unemotionality
Pausing	Self-doubting/self-confident Extraversion/introversion

[a] Acoustic measurement of perceived pitch, cf. Sect. 5.3.2

were observed with higher pitch in Scherer (1977). Eventually, Scherer presumes an inherent difference for male and female speakers with respect to aspects related to extroversion. In terms of intensity, which relates to loudness and power, results from Scherer were congruent with results from Mallory. Accordingly, extroversion and high intensity are positively related. Regarding speech fluency constituents, i.e., pauses, American extroverts were found to speak with fewer pauses, also including fewer filled pauses.

In a cross cultural study by Scherer (1974, 1979), however, German speakers are found to behave the opposite way. Extroverts are observed to show more silent pauses. Studies analyzing the effect or voice quality and speech disfluencies have shown rather scarce and inconclusive results, Scherer reports. Eventually consolidating results from analyses of pitch and intensity, he concludes that extroverted speakers speak louder, and with fewer hesitations. Furthermore, he declares that extroversion is the only factor that can be reliably estimated from speech.

2.2.3 *Work From Speech Synthesis*

With respect to the third category of experiments including sythesized speech, some works show a reverse perspective on experimentation. In category 1 and 2 researchers strove to observe and capture personality-relevant aspects in speech. Here the generation of parameters to synthesize maximal distinguishable speech in terms of personality expression is found to be the pronounced goal for a number of works in this category.

In synthesis, one of the major strengths is that designers are able to keep a maximum of characteristics constant while deliberately varying a specific parameter at hand. Brown et al. (1974) manipulates speech rate, mean pitch, and pitch variance of two male voices saying the sentence[4]: *We were away a year ago*. Rating *competence* and *benevolence* 37 listeners rated on each of 15 personality scales. As a major effect, lowering the speech rate resulted in a drastic loss of perceptual competence and a moderate loss of benevolence. Increasing the speech rate acted vice versa. As a less strong effect, increasing pitch variability raised the impression of increased benevolence while decreasing it also caused perceived competence to decrease. Finally, an increased pitch mean caused both low competence and low benevolence. In a subsequent study including a larger set of speakers, the relationship between speech rate and competence could be verified.

Apple et al. (1979) suggests four different aspects of vocal personality indication: *volume, pitch, pitch range, speech rate*. The term volume hereby refers to mean amplitude. Buller and Aune (1988), Pittam (1994) as well as Tusing (2000) later on verified that mean amplitude is positively associated with extroversion and dominance. Verifying his suggestions about speech rate and pitch, Apple used speech synthesis[5] for manipulation of speech from 27 male speakers. In fact, he built a three times three factorial design, each cell being populated by 3 speakers. Speech conditions within the cells were of lowered, original and raised pitch quality as well as compressed, original and expanded speech rate. While pitch was manipulated by 20 %, speech rate was altered by 30 %.

In listening test involving twenty undergraduate students Apple observed that speech rate influences the perception of a speaker's voice with regard to factors such as truthfulness, empathy, and potency. Slow speaking samples were judged less truthful, less persuasive, more passive but also more potent at the same time. These finding also verified findings from Miller et al. (1976), who found that more rapid speech was perceived to be more persuasive. He arranged speed alternations by selecting individuals due to their natural speech rate. Presumably, fast talkers are seen as more credible, Miller declared. Apple further found that high voices were judged less truthful, less potent and more nervous, which in turn was congruent to the findings from Brown et al. (1974) linking higher voices to less competent and less benevolent speakers. Claiming reasonable measure of confidence in the validity

[4] Note that the actual correct reference to pitch in term of speech synthesis is the acoustic correlate of the perceived pitch. i.e., F0. For simplification and comparability in the literature review, the term *pitch* is retained throughout this chapter. For details on how to obtain acoustic measurements for the perceived pitch and respective terminology please refer to Sect. 5.3.2.

[5] In his experiment Apple re-synthesized recordings after altering the speech using the LPC method of Atal and Hanauer (1971). LPC abbreviates *linear predictive coding* and is one out of many methods in speech synthesis.

of his three times three factorial results, Apple nevertheless refrains from drawing generalizations because all results were based on male speakers exclusively.

Also Smith et al. (1975) conclude that a strong linear relationship between speaking rate and perceptual competence exists, e.g., fast talkers are perceived significantly more competent than slow speakers. Further, very fast and very slow speech rates were associated with less benevolence, while increased rates also caused a more active impression. Normal, i.e., non-manipulated, speaking rates were, however, judged as most fluent, persuasive, emphatic and least nervous.

In an experiment using the German *Mary* synthesis system, Schröder and Trouvain (2003) and Trouvain et al. (2006) manipulate four prosodic parameters, namely pitch level (±30 %), pitch range (+2, +4 and +8 semitones), tempo (0, +15, +30 %) and loudness (soft, modal, loud). Using a 5 point Likert scale, 36 native German speakers were asked how much the modeled utterances fitted a set of intended variation along the Big 5 personality dimensions. Using just one male and one female synthetic voice, the authors wanted to explore if and how much it is possible to model the Big 5 dimensions with the same voice. As a result, many modifications to perceived sincerity, excitement, competence, and sophistication were rated as intended for the male voice. For the female voice, ratings showed only marginal affection by speech manipulations except from the excitement dimension. Finally, the authors conclude that excitement was best modeled for female voice, whereas competence and ruggedness is best modeled for male voice.

Experimenting with different personality-related perceptions like arousal, Schröder and Grice (2003) generated stimuli of high and low arousal for speech synthesis experiments. Higher arousal was realized by an increased pitch level and range, more prominent accents, steeper slopes for rising and falling pitch, faster speech rate, more but shorter pauses, a longer duration of obstruent consonants compared to vowel durations, and a voice quality expressing high vocal effort. After evaluating his stimuli in a user test, Schröder approved the proposed manipulation to be effective for high and low arousal, respectively.

Summing up the presented results from speech synthesis, speaking rate manipulation seems to cause a consistent change. Fast talkers are perceived as more active, dynamic, competent and extroverted, while slow talkers are perceived less truthful, less emphatic, less active and less serious. Experiments manipulating pitch show a more diverse outcome. However, higher pitch is mainly associated with greater extroversion and assertiveness and higher confidence and competence, but at the same time also with immaturity, emotionality and nervousness. Increasing pitch variance results in a more dynamic, extroverted, outgoing and benevolent impression.

2.2.4 Signal-Based Automated Personality Modeling

When it comes to signal-based modeling, only classification models trained by mostly few acoustic parameters can be found in the literature to date. Mairesse et al. (2007) found that prosodic and acoustic features are important for modeling extroversion, and that extroversion can be modeled best, followed by emotional stability (neuroticism) and openness to experience. In their work, the authors calculate various linguistic features regarding content, syntax and utterance type, as well as a relatively small number of 15 prosodic features. The prosodic features include intensity and pitch and capture mean, extrema and standard deviation measures as well as an estimation of speech rate. While results on two different corpora are produced, only one corpus contains spoken interaction, namely a subsample of conversation extracts recorded using the *Electronically Activated Recorder (EAR)*, cf. Mehl et al. (2001, 2006). Personality was assessed using a 44-item FFI questionnaire proposed by John and Srivastava (1999), cf. 1.2.7. Regarding self-assessment on a 5-point scale, reliabilities for Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness to Experience resulted in 0.90, 0.77, 0.83, 0.87, and 0.80, respectively. Using a 7-point scale and standardizing within all ratings from a single rater the average consistencies resulted in 0.93, 0.86, 0.92, 0.88, 0.87, respectively across the ratings of the 18 raters. The authors verify another major finding previously introduced, which is an observation of a general tendency on self- and observers' assessment: almost all results are much worse for self-ratings. This ranges from ratings consistencies calculated by (Mehl et al. 2006), towards correlation analysis, and classification evaluation by the authors. Eventually, they declare that they found many good results with models of observed personality, while models of self-assessed personality never outperform the baselines. Most prominent features are, however, word counts, pitch mean and intensity variation, eventually reaching binary classification, i.e., low versus high trait scores, accuracies of 74 and 73 % for neuroticism and extraversion respectively.

Mohammadi et al. (2010) and Mohammadi and Vinciarelli (2011) extracted pitch, first and second formants, energy and speaking rate. Calculating averages, minima, maxima, and relative entropy, 25 features are eventually used for automatic classification. Using a corpus of 640 clips extracted from news bulletins in French, nearly half the 10 s clips portray journalists, i.e., professional speakers, while the other portray non-journalists, i.e., naive speakers. Three raters annotated labels using a shortened form of the Big 5 containing only 10 items (Rammstedt and John 2007), cf. 1.2.7. To limit the effect of verbal interference, the raters did not understand French. In binary classification, discriminating the speakers of lower halves of each of the scales from the ones of upper halves, extroversion and conscientiousness reached best recognition rates of about 76.3 and 72 % respectively. Recognizing openness, however, revealed most difficult, remaining at no significant improvement over the baseline of randomly guessing. When using the combination of personality annotations and the prosodic features in order to discriminate journalists from non-journalists experiments showed accuracies of 75.5 % for annotation-based, and 87.2 % for prosodic-based classification. The combination of both resulted in 90 % accuracy. Interestingly, the recognition by a machine based on audio features outperformed human performance in the experiments. While the authors give no explanation for that observation, database design, training of the labelers, and the fact that the raters did not understand the bulletins could potentially have influenced the results.

Ivanov et al. (2011) used a boosting algorithm for classification (Schapire and Singer 2000) of 119 speech samples from 12 speakers engaged in a tourist call-center roll play experiment, cf. A Dix et al. (2004). The data was annotated with self-assessments by a short form of the Big 5 questionnaire (Gosling 2003), different to the one proposed from Rammstedt and John (2007). The authors applied a large-scale acoustic and prosodic feature set defined and implemented in Eyben et al. (2010). As a result, the authors were able to classify high versus low extroversion and high versus low conscientiousness with significant improvement over a random guessing baseline. Experiments on other scales failed to exceed the baselines.

2.2.5 Other Related Studies

Affecting all the presented literature from all three categories, further differences in listening test setup, underlying personality theory, scales at hand and raters training prohibits a more detailed comparison between the respective results. Eventually being another difference immanent in the literature, many experiments base on readings of a text, while other studies focused on speech in human-human conversations. Some studies transmit recordings via channels or play it using equipment in the laboratories while others involve a live-speaking situation. Also, cultural background and language differs. For a detailed comparison between Korean and American English personality impressions from speech see Peng et al. (1993). While loud voices conveyed impressions of power in both languages, Korean listeners rated personality regardless of any speech rate alternation. When van Bezooijen (1995) analyzed pitch hight perception in Dutch and Japanese language, he observed perceptual differences in attractiveness which he explained by sociocultural factors establishing different norm of highs for women in the languages. A cross cultural study comparing German and American English can be found in Scherer (1974, 1979), cf. results mentioned above. Few studies on perception of accents and dialects provide further insights into factors of relevance. Tsalikis et al. (1991) concludes, that speakers of General American or big-city accent were perceived as more intelligent, industrious, while regional accents were associated with greater integrity, sincerity, generosity, friendliness and warmth. With regard to the third category, i.e., experimentation including synthesized speech, the varying character of the overall speech synthesis quality could also be expected to have exerted considerable influence. Bad synthesis quality corresponds to unnatural listening experiences. The mere fact, that the speech may have sounded unnatural could potentially have changed the perception of the actual simuli. Note, that the addressed acoustic parameters are not completely independent from each other, as spectral characteristics and pitch change when humans speak louder, for example. Manipulation of only one factor out of a compound of naturally associated factors can also exert a distorting or unnatural influence.

Finally, many of the aforementioned authors observed interplay between the message (the text which was spoken) and the effect of a manipulation of the acoustic parameters hitherto mentioned. Gill and French (2007) investigate the relationship between the personality of an author of short Emails and blog texts, generated by self-assessment, and their language and conclude that personality also influences the text of a communication. Using co-occurrence techniques, the authors observe insufficient correlations. They assume, that personality will be represented in text using more complicated features. Oberlander and Gill (2004) examine the relation between part-of-speech (POS) distributions in Email texts and two distinct personality traits of the authors, i.e., neuroticism and extroversion. They conclude that POS can be characteristic. Walker et al. (1997), Mairesse et al. (2007) and Mairesse and Walker (2007, 2011) report on linguistic personality detection and generation. In Mairesse et al. (2007) and Mairesse and Walker (2011) the authors tackle the problem of modifying linguistic text style in statistical language generation according to a desired personality. Furthermore, the authors give a comprehensive overview of related works on linguistic personality generation as well as an introduction to the developed system *PERSONAGE*.

Zen et al. (2010) investigate the relationship between personality traits, visual attention, and spacial distances of speakers. In order to estimate the degree of extroversion and neuroticism the authors track head and body orientation from video. Tackling recognition of extraversion by means of audio and video information Pianesi et al. (2008) study multimodal assessment of personality in meeting scenarios. The authors declare, that results improve when combining the information.

Eventually, in Enos et al. (2006) the authors examine the ability to perceptually detect deception in speech given scores on openness, agreeableness and neuroticism of the raters. In order to find "good" raters for deception detection, their study resulted in a speculation that neurotic persons are more in need to presume truthfulness, since the neurotic individual suffers more than others when faced with upsetting thoughts or negative perceptions.

2.2.6 Own Prior Work

Polzehl et al. (2010c) presents early results on experiments for automatic personality classification using the Big 5 personality traits and the NEO-FFI. Introducing a new corpus of German speech containing personality actings, this work re-evaluated a selection of features and classifiers that showed high performance in earlier own related work on emotion recognition from speech (cf Polzehl 2006; Polzehl and Metze 2008; Burkhardt et al. 2009a, b; Metze et al. 2009; Polzehl et al. 2009a, b; Schuller et al. 2009; Metze et al. 2010; Polzehl et al. 2010a, d, e; Schmitt et al. 2010a, b, c, d; Polzehl et al. 2011b). Classifying for a high and a low target on each personality trait the 10-class task reveals approximately 60 % accuracy in automatic classification, which exceeds a random guessing baseline six times. Neuroticism and conscientiousness could be classified best, i.e., with an F-measure of more than 0.8 in magnitude.[6] Also the classification of extroversion revealed good results. The models

[6] More details on measurements are given in Sect. 5.7. As for now, the F-measure can be seen as accuracy-related measure accounting for a single class out of a multi-class classification task which is less biased by class distribution imbalance. The value of 0.8 corresponds to good classification success.

fail to capture relevant information for agreeableness while assessing openness seems to be problematic also for human annotators.

Comparing this work to the present work there exist a wide range of differences. The personality in these early experiments were analyzed with respect to the speaker's neutral characteristic. Fewer features which were less advanced and less targeted when compared to this work were extracted. Also the models were less advanced. No experimentation with continuous trait score prediction was executed in any form. Still, the recordings used for the early work are included in the presented work as the *text-dependent* subset. Using the same database, Polzehl et al. (2010b) execute factor analyses of the item responses from NEO-FFI personality assessment and observe overall high consistencies as well as distinct factor structures. In Metze et al. (2011) the authors present a review on personality in voice-based man machine interaction and elaborate on the the prospects of synthesis of speech with personality.

In Polzehl et al. (2011a) the authors introduce a new subset to their database, which includes spontaneous speech. Designing the database, special focus was directed to text-dependency and temporal effects when generating and assessing personality from speech. All personalities were performed by a single actor, who has been invited to recording sessions over several weeks. As a result, only very few text- and time-dependencies were observed. Applying a cluster analysis the authors give fist insights into similarities and oppositions inherent in the personality assessments of different traits. Again the introduced dataset from Polzehl et al. (2011a) are included into the present collection and will be referred to as *text-independent* recordings.

In Polzehl et al. (2012) another extension of the database comprising speech from over 60 naive speakers is introduced and compared to a database of French language introduced by Mohammadi et al. (2010). When comparing results for automatic personality predictions with human annotations for extroversion for both corpora, highest correlation results in 0.6 using various acoustic features. The recorded data is part of the presented *multi-speaker* collection.

More details regarding all three subsets and further extensions towards the presented work are given in Sect. 2.3. Major difference in between the isolated publication and the presented comprehensive work exist. These differences predominantly affect the label analyzes, the feature extraction and the modeling parts. Results from the current data exceed reported figures from the author's former publications. Also the broad scope of the systematic comparison of features, results, and consistencies between the datasets as well as the comprehensive analysis and modeling sections including an elaborated discussion are major and novel contributions of the present work.

2.3 Chapter Summary

After motivating the analysis of personality in speech, this chapter provides a short overview of contemporary terminology. This terminology is useful when trying to understand and compare existing literature as the literature oftentimes elaborates on different aspects of personality and speech. Therefore, this chapter also includes a definition of the so-called *prosodic* characteristics as well as a delineation of the underlying assumption that these characteristics must have a certain expansion space. Next, the frequently used terms *extralinguistic* and *paralinguistic* are explained and examples are given also for ambiguous cases.

Having provided this terminological background, Sect. 2.2 organizes the literature review into five categories:

1. Purely descriptive studies linking perceptual speech properties and personality mostly be means of rating-based studies
2. Studies incorporating correlations between perceptual ratings and acoustic measurements
3. Studies manipulating parameters my means of speech synthesis in order to cause changes in personality perception
4. Studies applying signal-based automated classification for personality estimation
5. Other related studies

Most literature can be found for extroversion and neuroticism. Descriptive studies predominantly look at extroverted or neurotic personalities and describe their speech characteristics along *loudness*, *speed* or *passiveness* impressions. Later studies add voice qualities like *breathy*, *tensed*, or *orotund*, while melodic perception is described along pitch *flatness* mostly. Also many other characteristics are included frequently.

The search for parameters in order to synthesize maximally distinguishable speech is found to be a dominant theme in experiments using synthesis. Parameters that can be manipulated are mostly volume, pitch range and speech rate, which show highest impact for perceived extroversion amongst other characteristics like truthfulness, persuasiveness, and competence. Alternations of speaking rate cause the most consistent change. Fast talkers are perceived as more active, dynamic, competent and extroverted, while slow talkers are perceived as the opposite and less serious. Further, results on pitch manipulation show a more diverse picture. Accordingly, higher pitch is oftentimes associated with extroversion, assertiveness and higher competence, but at the same time also with immaturity, emotionality and nervousness. Increasing pitch variation oftentimes leads to a more dynamic, extroverted, and benevolent impression.

However, overall bad synthesis quality and the fact that manipulating just one single parameter out of a multitude of related parameters can lead to an unnatural speech experience very easily. Nevertheless, results from synthesis experiments surely have indicative function, and they produce highly relevant insights into personality perceptions. Eventually, these insights can oftentimes not be implemented in extraction and modeling directly, which is mostly due to extraction inaccuracy or interdependence with overlaying factors.

Very recent studies have investigated more systematically the nature of personality perceptions and signal-based measurement. Only very few studies use more than a basic inventory of few acoustic and prosodic descriptors, even fewer report on automatic procedures and modeling. This is also due to the fact, that more systematic analyses require more systematically generated databases.

The lack of a versatile database with consistent annotations is one of the essential motivations why the author chose to record a new speech corpus as will be described in detail in Chap. 3. Hopefully, the database will inspire many follow-up analyses and modeling experiments. In the literature, only extroversion and neuroticism have been captured to a reasonable extent, yet resulting in moderate classification success. Results on the current data exceed reported figures form the literature in both, performance and opportunity for targeted analyses. With respect to modeling for personality trait score prediction, the results presented in Sects. 5.8.1–5.10 and discussed in Sect. 6.2 are even unprecedented. Ultimately, the preparation of the recorded and labeled database as well as the comprehensive analysis and modeling sections including an encompassing discussion are presumably the most important contributions of the present work to the research community.

References

Addington DW (1968) The relationship of selected vocal characteristics to personality perceptions. Speech Monogr 35(4):492–503

Allport GW, Cantril H (1934) Judging personality from voice. J Soc Psychol 5(1):37–55

Apple W, Streeter LA, Krauss RM (1979) Effects of pitch and speech rate on personal attributions. J Personality Soc Psychol 37(5):715–727

Aronovitch CD (1976) The voice of personality: stereotyped judgments and their relation to voice quality and sex of speaker. J Soc Psychol 99(2):207–220

Atal BS, Hanauer SL (1971) Speech analysis and synthesis by linear prediction of the speech wave. J Acoust Soc Am 50(2B):637–655

Ball D, Hill L, Freeman B, Eley TC, Strelau J, Riemann R, Spinath FM, Angleitner A, Plomin R (1997) The serotonin transporter gene and peer-rated neuroticism. NeuroReport 8(5):1301–1304

Berry DS (1990) Vocal attractiveness and vocal babyishness: effects on stranger, self, and friend impressions. J Nonverbal Behav 14(3):141–153

van Bezooijen R (1995) Sociocultural aspects of pitch differences between japanese and dutch women. Lang Speech 38:253–265

Bickmore T, Cassell J (2004) Natural intelligent and effective interaction with multimodal dialogue systems. Kluwer Academic, New York

Breese J, Ball G (1998) Modeling emotional state and personality for conversational agents. Technical Report MSR-TR-98-41, Microsoft Research

Brown B, Strong W, Rencher A (1974) Fifty-four voices from two: the effects of simultaneous manipulations of rate, mean fundamental frequency, and variance of fundamental frequency on ratings of personality from speech. J Acoust Soc Am 55:313–318

Buller DB, Aune RK (1988) The effects of vocalics and nonverbal sensitivity on compliance a speech accommodation theory explanation. Hum Commun Res 14:548–568

Burkhardt F, Ballegooy Mv, Engelbrecht K-P, Polzehl T, Stegmann J (2009a) Emotion detection in dialog systems: applications, strategies and challenges. In: Proceedings of international conference on affective computing and intelligent interaction (ACII (2009)) vol 1. IEEE Netherlands, Amsterdam

Burkhardt F, Polzehl T, Stegmann J, Metze F, Huber R (2009b) Detecting real life anger. In: Proceedings of international conference on acoustics, speech, and signal processing (ICASSP (2009)) vol 1. Taipei, Taiwan, IEEE, pp 4761–4764

Cantril H, Allport G (1935) The psychology of radio. Harper and Brothers, New York

Cassell J, Sullivan J, Prevost S, Churchill E (eds) (2000) Embodied conversational agents. The MIT Press, Cambridge

Cassell J, Bickmore T (2003) Negotiated collusion: modeling social language and its relationship effects in intelligent agents. User Model User Adapt Interact 13(1–2):89–132

Catrambone R, Stasko J, Xiao J (2002) Anthropomorphic agents as a user interface paradigm: experimental findings and a framework for research. In: 24th annual conference of the cognitive science society, pp 166–171

Chen Y, Naveed A, Porzel R (2010) Behavior and preference in minimal personality: a study on embodied conversational agents. In: International conference on multimodal interfaces and the workshop on machine learning for multimodal interaction, ICMI-MLMI' 10, 49:1–49:4, New York, NY, USA. ACM

Dix A, Finlay J, Abowd G, Beale R (2004) Human-computer interaction, 3rd edn. Prentice-Hall, Upper Saddle River

Enos F, Benus S, Cautin RL, Graciarena M, Hirschberg J, Shriberg E, (2006) Personality factors in human deception detection: comparing human to machine performance, ISCA, pp 813–816

Eyben F, Wöllmer M, Schuller B (2010) OpenSMILE—The Munich versatile and fast open-source audio feature extractor, 1459. ACM Press, New York

Gill AJ, French RM (2007) Level of representation and semantic distance: rating author personality from texts. In: Proceedings of the second european cognitive science conference (EuroCogsci07), Delphi, Greece

Gosling S (2003) A very brief measure of the Big-Five personality domains. J Res Pers 37(6):504–528

Hunt RG, Lin TK (1967) Accuracy of judgments of personal attributes from speech. J Personality Soc Psychol 6(4):450–453

Ivanov AV, Riccardi G, Sporka AJ, Franc J (2011) Recognition of personality traits from human spoken conversations. Most (August) pp 1549–1552

John OP, Srivastava S (1999) The Big Five trait taxonomy: history, measurement, and theoretical perspectives. J Personality 2(2):102–138

Kreiman J, Sidtis D (2011) Foundations of voice studies, an interdisciplinary approach to voice production and perception. Wiley-Blackwell, West Sussex

Mairesse F, Walker MA, Mehl MR, Moore RK (2007) Using linguistic cues for the automatic recognition of personality in conversation and text. J Artif Intell Res 30:457–500

Mairesse F, Walker M (2007) PERSONAGE: personality generation for dialogue, Association for computational linguistics

Mairesse F, Walker MA (2011) Controlling user perceptions of linguistic style: trainable generation of personality traits. Comput Linguistics 37(January 2009):1–34

Mallory EB, Miller VR (1958) A possible basis for the association of voice characteristics and personality traits. Speech Monogr 25:255–260

Mehl MR, Pennebaker JW, Crow DM, Dabbs J, Price JH (2001) The electronically activated recorder (EAR): a device for sampling naturalistic daily activities and conversations. Behavior Res Methods Instrum Comput J Psychon Soc Inc 33(4):517–523

Mehl MR, Gosling SD, Pennebaker JW (2006) Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. J Person Soc Psychol 90(5):862–877

Metze F, Batliner A, Eyben F, Polzehl T, Schuller B, Steidl S (2010) Emotion recognition using imperfect speech recognition. In: Proceedings of the annual conference of the international speech communication association (Interspeech 2009), Makuhari, Japan, IEEE, pp 1–6

Metze F, Black A, Polzehl T (2011) A review of personality in voice-based man machine interaction. In: Human-Computer Interaction. Interaction techniques and environments—14th international conference, HCI International 2011, Springer, pp 358–367

Metze F, Polzehl T, Wagner M (2009) Fusion of acoustic and linguistic speech features for emotion detection. In: Proceedings of international conference on semantic computing (ICSC 2009) vol 1. Berleley, USA, CA, IEEE

Miller N, Maruyama G, Beaber RJ, Valone K (1976) Speed of speech and persuasion. J Pers Soc Psychol 34(4):615624

Mohammadi G, Mortillaro M, Vinciarelli A (2010) The voice of personality: mapping nonverbal vocal behavior into trait attributions. In: Proceedings of the international workshop on social signal processing, pp 17–20

Mohammadi G, Vinciarelli A (2011) Humans as feature extractors: combining prosody and personality perception for improved speaking style recognition. In: Proceedings of IEEE international conference on systems, man and cybernetics, pp 363–366

Moore W (1939) Personality traits and voice quality deficiencies. J Speech Hear Disord 4:33–36

Nass C, Moon Y, Fogg B, Reeves B, Dryer DC (1995) Can computer personalities be human personalities? Int J Hum Comput Stud 43:223–239

Nass C, Brave S (2005) Wired for speech: how voice activates and advances the human-computer relationship. The MIT Press, Cambridge

Nass C, Lee KM (2001) Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. J Exp Psychol, pp 171–181

Oberlander J, Gill A (2004) Individual differences and implicit language: personality, parts-of-speech and pervasiveness. In: Cognitive Science Society, Chicago, IL, USA

Paeschke A (2003) Prosodische analyse emotionaler sprechweise—dissertation TU berlin, vol 1. Reihe Mündliche Kommunikation. Logos Verlag, Berlin

Pear T (1931) Voice and personality. Chapman & Hall, London

Peng Y, Zebrowitz L, Lee H (1993) The impact of cultural background and cross-cultural experience on impressions of american and korean male speakers. J Cross-Cultural Psychol 24(2):203–220

Pianesi F, Mana N, Cappelletti A, Lepri B, Zancanaro M (2008) Multimodal recognition of personality traits in social interactions. In: Proceedings of the 10th international conference on multimodal interfaces IMCI 08, 53

Pierrehumbert J (1979) The perception of fudamental frequency declination. J Acoust Soc Am 66:363369

Pittam J (1994) Voice in social interaction: an interdisciplinary approach. Sage Publications Inc., Baldwin City

Polzehl T (2006) Automatische klassifizierung emotionaler sprechweisen. In: Tagungsband 1.Kongress Multimediatechnik, Wismar, Germany

Polzehl T, Metze F (2008) Using prosodic features to prioritize voice messages. In: SIGIR, Singapore proceedings of speech search workshop at SIGIR

Polzehl T, Schmitt A, Metze F (2009a) Comparing features for acoustic anger classification in german and english IVR systems. In: Proceedings of international workshop of spoken dialogue systems (IWsDs 2009) vol 1. University of Ulm, Germany, Ulm

Polzehl T, Sundaram S, Ketabdar H, Wagner M, Metze F (2009b) Emotion classification in children's speech using fusion of acoustic and linguistic features. In: Proceedings of the annual conference of the international speech communication association (Interspeech 2009), Brighton, England. ISCA, pp 340–343

Polzehl T, Metze F, Schmitt A (2010a) Linguistic and prosodic emotion recognition. Deutsche Jahrestagung für Akustik (DAGA). DAGA, DAGA, pp 1–2

Polzehl T, Möller S, Metze F (2010b) Automatically assessing acoustic manifestations of personality in speech. In: Workshop on spoken language technology, Berkeley, USA IEEE

Polzehl T, Möller S, Metze F (2010c) Automatically assessing personality from speech. In: Proceedings of international conference on semantic computing (ICSC 2010), IEEE, pp 1–6

Polzehl T, Schmitt A, Metze F (2010d) Approaching multi-lingual emotion recognition from speech—on language dependency of acoustic/prosodic features for anger detection. In: Speech-Prosody, Chicago, IL, USA. University of Illionois, pp 1–6

Polzehl T, Schmitt A, Metze F (2010e) Salient Features for Anger Recognition in German and English IVR Portals. In: Spoken dialogue systems technology and design. Springer, Berlin, Germany, pp 81–110

Polzehl T, Möller S, Metze F (2011a) Modeling speaker personality using voice. In: Proceedings of the annual conference of the international speech communication association (Interspeech 2011), Florence, Italy. ISCA

Polzehl T, Schmitt A, Metze F, Wagner M (2011b) Anger recognition in speech using acoustic and linguistic cues. Speech communication, special issue: sensing emotion and affect—facing realism in speech processing

Polzehl T, Schoenenberg K, Möller S, Metze F, Mohammadi G, Vinciarelli A (2012) On speaker-independent personality perception and prediction from speech. In: Proceedings of INTERSPEECH 2012

Rammstedt B, John O (2007) Measuring personality in one minute or less: a 10-item short version of the big five inventory in english and german. J Res Pers 41(1):203–212

Reeves B, Nass C (1996) The media equation: how people treat computers, television, and new media like real people and places. Cambridge University Press, Cambridge

Sanford FH (1942) Speech and personality: a comparative case study. J Personality 10:169198

Schapire RE, Singer Y (2000) BoosTexter: a boosting-based system for text categorization. In: Machine Learning, 135–168

Scherer KR (1974) Voice quality analysis of american and german speakers. J Psycholinguistic Res 3:281–298. doi:10.1007/BF01069244

Scherer KR (1979) Personality markers in speech, Cambridge University Press, Cambridge, pp 147–209

Scherer KR, Scherer U (1981) Speech behavior and personality. Speech Evaluation Psychiatry, 115–135

Scherer KR (1977) Effect of stress on fundamental frequency of the voice. J Acoust Soc Am 62:25–26

Schmitt A, Pieraccini R, Polzehl T (2010a) For heavens sake, gimme a live person! designing emotion-detection customer care voice applications in automated call centers. Advances in speech recognition. Springer, US, Berlin, Germany, pp 81–110

Schmitt A, Polzehl T, Minker W (2010b) Facing reality: simulating deployment of anger—recognition in IVR systems. In: Spoken dialogue systems for ambient environments—lecture notes in computer science, vol V. 6392, Springer, Makuhari, Japan, pp 23–48

Schmitt A, Polzehl T, Minker W, Liscombe J (2010c) The influence of the utterance length on the recognition of aged voices. In Calzolari N, Choukri K, Maegaard B, Mariani JOJ, Piperidis S, Rosner M, Tapias D (eds) Proceedings of the seventh conference on international language resources and evaluation (LREC'10), Valletta, Malta. European Language Resources Association (ELRA), pp 1–6

Schmitt A, Polzehl T, Minker W (2010d) Modeling a-priori likelihoods for angry user turns with hidden markov models. In: SpeechProsody, Chicago, IL., USA. University of Illionoise, pp 1–6

Schröder M, Trouvain J (2003) The german text-to-speech synthesis system mary: a tool for research, development and teaching. Int J Speech Technol 6(4):365–377

Schröder M, Grice M (2003) Expressing vocal effort in concatenative synthesis, 25892592. Citeseer

Schuller B, Metze F, Steidl S, Batliner A, Eyben F, Polzehl T (2009) Late fusion of individual engines for imrpoved recognition of negative emotion in speech—learning vs. democratic vote. In: International conference on acoustics, speech and signal processing (ICASSP). IEEE

Smith R, Parker E, Noble E (1975) Alcohol's effect on some formal aspects of verbal social communication. Archives Gen Psychiatry 32(11):1394–1398

Stagner R (1936) Judgments of voice and personality. J Educational Psychol 27(4):272–277

Taylor HC (1934) Social agreement on personality traits as judged from speech. J Soc Psychol 5:244–248

Trouvain J, Schmidt S, Schröder M, Schmitz M, Barry WJ (2006) Modelling personality features by changing prosody in synthetic speech. Number Table 2. ISCA, pp 4–7

Tsalikis J, DeShields OJ, LaTour M (1991) The role of accent on the credibility and effectiveness of the salesman. J Pers Sell Sales Management 11:31–41

Tusing K (2000) The sounds of dominance. vocal precursors of perceived dominance during interpersonal influence. Hum Commun Res 26(1):148–171

Walker MA, Cahn JE, Whittaker SJ (1997) Improvising linguistic style: social and affective bases for agent personality. In: Proceedings of autonomous agents, p 10

Winkler R (2003) Merkmale junger und alter stimmen: analyse ausgewählter parameter im kontext von wahrnehmung und klassifikation, vol 6. Reihe Mndliche Kommunikation. Logos Verlag, Berlin

Zen G, Lepri B, Ricci E, Lanz O (2010) Space speaks: towards socially and personality aware visual surveillance. In: Proceedings of the 1st ACM international workshop on multimodal pervasive video analysis, p 3742

Zuckerman M, Driver RE (1989) What sounds beautiful is good: the vocal attractiveness stereotype. J Nonverbal Behav 13(2):67–82

Chapter 3
Database and Labeling

In analogy to the first studies in emotion recognition, initial database recordings were restricted in order to exclude a maximum number of variables that could potentially interfere. The first subset collection of our own database was therefore confined to a fixed set of sentences produced by a single professional speaker, who has a lot of experience in recording voice prompts for speech dialog systems and who is used to work with voice coaches. Ten discrete personality targets, i.e. speech conditions to be performed, were defined such that there is one low-score target and one high-score target for each of the Big 5 personality traits as presented in Sect. 1.2.4. The speaker was asked to prepare and perform this 10 voice personalities, i.e. acting as a person with either high or low values on each of the Big 5 scales individually and sequentially. The targets were explained to the speaker by presenting the definitions from the original, textual NEO-FFI personality trait descriptions by Costa and McCrae (1992), cf. Sect. 1.2.5. The speaker was given time in advance to cognitively prepare his performance. The order of the 10 personality classes to be performed was randomized for all recordings. After finishing the recordings and after basic data cleaning and cutting, listening tests produced the personality annotations for each recording. Each recording was assessed by a multitude of different raters, all of whom were asked to fill out a full NEO-FFI questionnaire after having listened to the recordings.

After recording the one professional speaker also a number of non-professional speakers were recorded. The way of expressing personality is expected to differ between recordings of professional and non-professional speakers in so far, as professional speakers are widely believed to express themselves more clearly and more enunciatedly, while native speakers are expected to exhibit perceivable characteristics to a diminished degree. Hence, better results in terms of increased separability and increased classification accuracy could be expected from professional speakers. On the other side, more realistic results are expected from non-professional speakers. This cleavage is well known from neighboring research fields such as speech recognition or emotion recognition.

T. Polzehl, *Personality in Speech*, T-Labs Series in Telecommunication Services,
DOI 10.1007/978-3-319-09516-5_3

Regarding the range of expected expressions, the totality of all recorded data is designed to account for a large effectual scope of personality variation. While speech actings are often criticized as unnatural, over-enunciated and staged, the resulting personality expression can be believed to extend the range of non-acted personality expressions. Acted personality gestures can be believed to exhibit more extreme personality characteristic, thus leading to an extended coverage in term of Big 5 scales coverage.

The whole data collection is stratified due to text- and speaker dependency in a way that the overall data collection comprises three mutually exclusive datasets:

1. Text-dependent dataset
2. Text-independent dataset
3. Multi-speaker dataset

When recording *text-dependent* data, one speaker was given a fixed paragraph of text to perform according to the 10 personality targets. When recording *text-independent* data the speaker was freely composing his own words and thoughts. Note that, although one speaker is believed to exhibit one personality profile only, the speaker at hand is an experienced professional speaker. When recording *multi-speaker* data, 64 non-professional speakers were invited. Here, all the speakers spoke spontaneously and using own words. No *reading* scenario was included as this would have caused a change in overall intonation patterns, cf. (Batliner et al. 2000). All recorded data in the *text-dependent* and *text-independent* scenarios build on the fact that the actor was able to act out personality expressions different from his own personality. This assumption will be validated by listening test explained in Sect. 3.4. Unlike these datasets, the *multi-speaker* dataset can be seen as independent from any fixed personality profile of a single speaker or any artificially elicited personality expression.

All speakers spoke German as their native language. All recordings are of high quality. Although the recording set-up was not completely identical for the different subsets, all the takes can be considered as noise-free, non-echoic recordings containing one speaker at a time and one recorded audio file only. Detailed subset descriptions are explained separately in the following sections. The text-dependent subset was introduced in (Polzehl et al. 2010a). The text-independent subset was introduced in (Polzehl et al. 2010b). Most recently, the multi-speaker subset was introduced in (Polzehl et al. 2011).

3.1 Text-Dependent Data Recordings

In this subset the speaker repeated a fixed paragraph of text all the times. The spoken text resembles a neutral, complete phrase as can be expected in a typical *Interactive Voice Response (IVR)* system or a call-center hot-line. It starts with a short welcome sentence followed by a sentence giving information on how to use a voucher redeemer service. The next sentence is a negative sentence, i.e. the user is told that he cannot redeem the voucher for the reason that he calls from within a wrong network. A solution is offered in the following sentence, where the user is advised to call again when connected to the corporate network. Eventually, the speaker says a short goodbye. On average, a complete recording of all 5 sentences took about 20 s.

The following text gives the exact wording in German and English translation.

> Willkommen beim Gutscheindienst der Deutschen Telekom! Um Ihren Gutschein einzulösen geben Sie bitte den Gutscheincode ein. Mit diesem Telefonanschluss können Sie Ihren Gutschein leider nicht einlösen. Rufen Sie bitte diese kostenlose Rufnummer noch einmal an, und zwar von dem Telekomanschluss aus, für den Sie den Gutschein einlösen möchten. Vielen Dank und auf Wiederhören.

> Hello, and welcome to the voucher redeemer service of Deutsche Telekom! To redeem your voucher please enter the voucher code. Unfortunately, you cannot redeem your credit points from the line you are using just now. Please call this toll-free service again, using the line you want to charge your credits points to. Thank you, and goodbye.

The speaker repeated this paragraph for all targets. Eventually, for all the 10 targets a minimal number of 15 paragraph repetitions were recorded. Because of operationalizational setup the actual usable number of recordings per class resulted in 16, 17, 15, 16, 17, 17, 15, 17, 17, 15 for classes O+, O−, C+, C−, E+, E−, A+, A−, N+, N− respectively. All usable data will be included in the experiments. Overall, the collection comprises 160 recordings, which sums up to more than one hour of speech recordings in total.

3.2 Text-Independent Data Recordings

In this subset the speaker spoke freely without any given text passage. The speaker was presented a series of images to trigger associations and talk about them. He was instructed to first give a very brief summary, i.e. one or two sentences describing what he sees. Then, he was asked to talk about any feelings or associations with respect to the image presented. To give some examples, the speaker reported about whether or not he feels familiar with what he interpreted from the pictures, and if so what his feelings towards the image are. Thus the speaker freely described or presumed emotions, feelings, harmony, joy, bravery, or distress he associated with the images or persons depicted in the sceneries. Each recording in this subpart of the recording collection took between 40 and 100 s. The overall average lengths results in 1 min roughly.

In order to include diverse descriptions and associations, a series of 20 images showing different motives was compiled. In the first recording session ($S1$), the speaker interpreted all 20 images. Analyzing the actings and the richness of potential interpretations, 12 images were selected for a series of three follow-up recordings referred to as $S2$, $S3$, and $S4$. In each session the speaker interpreted all 12 images. While it was only 2 weeks time between the first and the second recording session, the time lag between session 2 and session 3 was 4 weeks, between session 2 and session

4 even 6 weeks . Thus, the data comprises perspectives on the actor’s performance quality and the time-dependency of his performance at the same time.

In terms of the images presented, three black-and-white images were borrowed from the *Thematic Apperception Test (TAT)*, cf. (Murray 1938), images #1, #2, #4. The test is a projective psychological test designed in 1935 for clinical application. The TAT pictures are designed to trigger the subject’s unconscious so he reveals aspects of personality. Some psychologists nowadays see this theory and especially the test design of the TAT as controversial and outdated. However, the application of the TAT pictures in this work does not aim to assess the speakers’ personality on basis of reactions towards the pictures directly. On the contrary, an expedient variety of possible views and interpretation of these pictures has been observed in pre-tests. This interpretation space motivated the decision to include the selection of three TAT pictures into the experiments.

Image #1 shows a child in front of a violin. Due to TAT theory, this picture elicits aspects of attitude with respect to the relation to the interpreter’s parents and conscientiousness as well as self-fulfillment, craving for recognition, expectations for the future and daydreaminess. The second image borrowed from TAT, #2, is believed to elicit feelings and problems with respect to partnership and relations to human affection. The third picture, #4 is believed to reflect the interpreter’s attitude towards relations to groups, especially relations among men.

Irrespective of TAT theory, another image shows the character *Jason Voorhees*. It represents a rather violent situation depicting a masked person holding a long bloody knife. The scenery is taken from the slasher movie *Friday the 13th*. This picture is expected to act provocatively, pointing the interpreter to unpleasant feelings and anxiety.

Next, a drawing outlining showing a young women was presented. While the drawing is mostly realized as a rather simple, schematic sketch, the level of detail increases when one looks into the eyes of the young woman. Thus, the interpretation of the way she looks is expected to cause an emotive touch, which in turn can serve as rich trigger for associations and interpretations. Similar to the yoiung woman, another line art from Christoph Glarner[1] was included. The simple, rather schematic work delineating pieces of the human body is governed by obscure geometry and also emphasizes the eyes. The multitude of forms and formations are expected to trigger various interpretations and associations.

Sketching three faces the next work presented is a woodcarving work by Peter Padubrin-Thomys.[2] The interpreter sees three persons celebrating. The way the faces are sketched and the unusual color of the liquid in the glasses seem to allude to additional subliminal meaning open for interpretation.

The next work presented is a painting from Ignacio Trelis,[3] depicting a proudly looking Indian warrior. The strong powerful impression of this warrior becomes

[1] “Chaos”, Christoph Glarner, http://www.fofo-art.ch/.

[2] “Gesellig”, Peter Padubrin-Thomys, http://www.ppt-grafik.de/.

[3] http://commons.wikimedia.org/wiki/File:Chtrelis.jpg, Creative Commons Attribution 3.0 Unported license.

tragic when looking at the problems Indian natives have been facing, which in turn leaves much space to associations and feelings.

Another image was originally used to promote a coming-of-age film by the Danish female director Natasha Arthy entitled *Fighter* in 2008. The image shows a karate training scene of a young Turkish woman and her coach.[4] Because of the antagonism between Turkish traditional values and a Turkish woman fighting, this image is also expected to trigger associations and different interpretations.

Trying to include another perspective a photo which is likely to be seen as stereotypical for advertisement was presented. It shows two happily smiling seniors in a delightful and shiny setup. This image was selected because its artificial shiny scenery is set up to a level, where it occurs to be all to perfect, thus appearing bluntly exaggerated, potentially causing ironic reactions. This photo by Janik Fauteux is taken from Flickr.[5] Looking from yet another perspective a photo of an apparently elegant, cultivated, rather upper class woman is presented. The photo was published on the online blog *Netplosiv.com*.[6] While on the first glance the observer might notice the unostentatiously rich appearance and might thus be indulged to envy the woman, the second look reveals that she seems to be worried about something, yet trying to retain her composure. Being the beneficiary of a huge company in Germany, the woman, i.e. Madleine Schickedanz, was experiencing a total bankruptcy in 2009. While the case was talked about in many news on TV and radio, the speaker did not consciously recognize her by face, which was also the expected situation when selecting the picture. Nevertheless, the motto of *money can't buy you happiness* subsequent in this scenery can be expected to trigger diverse feelings and associations.

The last selected picture shows a post-surrealistic collage from Michael Maier entitled *The places where we go*.[7] Because of the surrealistic expression, the image can be thought of as alluding to reverie and fantasy.

While many other pictures could have potentially been selected, the selected images are not to be directly linked to personalities. This selection serves as pool offering impulses for diverse interpretation for the actor, when taking on a personality-driven perspective.

Eventually, the recorded speech data amounts to more than six hours of recordings. All text-dependent and text-independent recordings were done at the *Quality and Usability Lab* of the Technische Universität Berlin, which is equipped with an anechoic recording booth. The speaker was recorded using a high quality AKG 414 B-XLS microphone in hyper-cardioid polar pattern mode, and a digital Hammerfall RME Multiface device set to 24bit, 41.1 KHz. The speaker kept a distance between 20 and 30 cm to the microphone when speaking. In addition, he was given headphones in order to be able to listen to his own voice and the feedback from the chief

[4] http://www.delphisfilms.com/images/mail/images/movie_art/art-fighter3.jpg.

[5] http://www.flickr.com/photos/relocalisationentourage/5212014340.

[6] http://netplosiv.com/201037916/promis/quelle-pleite-madeleine-schickedanz-muss-luxus-villen-verkaufen.

[7] http://www.artoffer.com/r.asp?n=221585&i=627309128112011225183.

engineer in the control room. Acting as sound engineer and voice coach during the sessions, the chief engineer and the speaker kept visual contact all the time. Table 3.1 shows an overview of the recording conditions.

3.3 Multi-Speaker Data Recordings

The multi-speaker dataset, which comprises a multitude of different speakers, comprises of two distinct recording conditions. In both conditions recordings were taken from a conversational test scenario. The tests were designed in a two-fold way. On the one hand the conversational speech was recorded without further manipulation, on the other hand artificial transmission delay was introduced and the affection on the perceived quality was also determined in order to be used in a related study. The data included into the present collection comprises only the non-delayed part of the recordings. Pairs of two participants were asked to have conversations following scenarios known as *Short Conversation Tests* (*SCT*), cf. (Moeller 2007). All recordings are of full bandwidth using an RME Hammerfall Multiface interface set to set to 24 bit and 44.1 kHz bandwidth.

During the first conversation tests, participants were sitting in two separate soundproof cabins interacting via a narrow-band telephone connection using SNOM 870 as terminal devices. The conversations were actually recorded with high quality PMC 65 pressure zone microphones from Beyerdynamics, which had been placed on the table in front of each participant. The resulting distance between the microphone and the speakers was roughly 30–40 cm. The polar pattern of the microphone matched the direction of the incoming speech waves. Out of 29 recordings in the collection, 14 are done by male speakers and 15 are done by female speakers. The average age results in 29 years.

For the second conversation tests, recordings were done in a large anechoic chamber (area $= 120$ m^2, lower frequency limit $= 63$ Hz) at the Technische Universität Berlin. Speech was captured in a close-capturing manner, using Beyerdynamics DT 290 headsets. The room was big enough to conduct simultaneous test with three conversations at a time. Although close capturing and large distances in between the participants prevented cross-capturing from neighboring conversations, loudspeakers emitting smooth bubble-noise not perceivable in the recordings, and separating heavy curtains were set up to partition the room. Out of 35 recordings in the collection, 22 speakers are male and 13 female. The average age results in 27 years.

Technically, the two subsets offer the opportunity to examine the impact of the capturing type, i.e. close capturing using headsets or stand-alone microphone capturing resulting in 30–40 cm space between the mouth and the membrane.

In comparison to the text-dependent and text-independent subsets presented before, all non-professional stimuli are non-acted, meaning that no specific personality perspective was performed or induced. Consequently, this data does not necessarily exhibits a 10-class structure that could be directly compared to the 10 personality targets the professional speaker had been asked to perform. Thus, the

Table 3.1 Overview of recorded database and recording conditions

Subset	Text-dependent	Text-independent	Multi-speaker	
Domain	IVR-prompt	Image description	Conversations	
Text-dependency	Yes	No	No	
Speaker-dependency	Yes	Yes	No	
Acted/non-acted	Acted	Acted	Non-acted	
Linguistic diversity	1 paragraph “Gift Voucher”	12 chunks “Images Description”	64 extracts “Short Conversation Tests”	
Database size in total	1 h	6 h	1.2 h	
Number of speakers	1	1	64	
Microphone type	Stand alone	Stand alone	Stand alone	Headset
Capturing quality	24 bit, 44.1 kHz, mono			

range of personality diversity captured in the collection can only be analyzed after finishing the recording sessions and no criteria for a pre-selection of the test persons were available in advance. Table 3.1 shows a summary of all recorded data and conditions.

3.4 Annotating the Recordings with Personality Ratings

To generate human personality labels listening tests were conducted. In the tests raters filled out NEO-FFI personality questionnaires as proposed by Costa and McCrae (1992) and explained in Sect. 1.2.5. The raters were given time to listen to the stimuli at least once before starting to rate the 60 items of NEO-FFI the questionnaire. Also, the raters were free to listen to the stimuli as often as they wanted to, during rating time. Rating sessions took between 60 and 90 min including one break of 5–10 min and a number of smaller pauses the raters could decide for individually and autonomously. Raters were given high quality headsets, a pleasant level of playback velocity was determined with the raters manually before starting the listening tests. The stimuli were presented in random order.

In a pre-test using the text-dependent data, three out of the 15 recording repetitions of each of the 10 targets were selected for labeling. The reason for this selection is twofold: (1) in a substantial number of cases professional speech actings have shown to be perceived as not *realistic*, meaning not natural, over-acted, exaggerated or staged, thus imposing distortion to the intended personality impressions as well; and (2) the labeling of the whole database was not possible due to cost limits. In order to execute the pre-selection, 3 students rated all personality targets of all recordings according to their perceptional “artificiality”, using a Likert scale between 0 and 5. Eventually, three stimuli per personality target were selected that were rated low in artificiality. The recordings are of 20 s length on average. The selected recordings were further assessed by 20 different native raters using the NEO-FFI. Eventually,

Table 3.2 Image ID and number of labelers who have assessed all the 10 targets, grouped by recording session and image described

Image/ Session ID	Image 2	Image 4	Image 6	Image 9	Image 12	Image 16	Image 17	Sum
Session1	0	0	5	0	0	0	0	**5**
Session2	3	5	5	15	3	5	5	**41**
Session3	3	15	5	15	3	5	5	**51**
Session4	3	15	5	15	3	5	5	**51**
Sum	**9**	**35**	**20**	**45**	**9**	**15**	**15**	**148**

approximately 600 NEO-FFI questionnaires were filled out. The average fill-out time for one questionnaire including the time the raters listened to the speech recording results in 7 min roughly. For all stimuli, regardless of personality target, full NEO-FFI tests were done, so in total over 36,000 item responses have been generated. Since targets are set in pairs of two to denote low and high values along the Big 5 scales, 7,200 item ratings are available for each scale. Eighty-seven raters took part in the tests, most of whom were bachelor or master students. Raters were of 29 years of age on average, three out of five raters were male.

For the text-independent data, only 7 out of 12 images could be selected for human personalty labeling due to cost limitations. This selection was done manually and intuitively after the first recording session, cf. Sect. 3.2. Looking for maximal variation and maximal richness in the recordings, 5 images that turned out not to trigger much ideas, feelings and associations were discarded. Table 3.2 shows the image ID taken from the initially tested 20 images as well as the number of labelers who have assessed all the 10 targets for each recording session and for each image described.

In order to match the data conditions of the fixed-text recordings, the spontaneous-text recordings were cut to excerpts of roughly 20 s starting from the beginning of the utterances. This procedure was done manually, natural phrases were kept as entity. To generate human personality labels, the same procedure as described above for the fixed-text recordings was applied. For this subset of the database 102 raters participated in the tests, the mean age resulted in 29 years with equally distributed gender accounts.

Table 3.2 gives an overview about the number of labelers that have assessed all the 10 targets of an image description recorded at a respective recording session. For example, recordings of descriptions of image 9 have been taken from three different recording sessions and assessed by 15 raters each. Since every image description comprises the assessment of 10 personality targets, the number of conducted NEO-FFI test corresponds to 10 times the number given in the table. Raters were given stimuli in randomized order.

The varying number of available ratings per image and/or session was consciously designed. Image 4 and image 9 offer the possibility to analyze text- and time-dependency by providing a number of repetitions meeting the critical mass for

Table 3.3 Conditions and resulting data pools for the overall data collection

Subset	Text-dependent	Text-independent	Multi-speaker
Personality targets classes	10	10	n/a
Labeled takes per target class	3 (out of 15)	7 (out of 12)	n/a
Recordings with three or more labels	30	210	64
Number of labels per recording/pool	15	15 (for 5 ISP[a]) 5 (for 16 ISP[a]) 3 (for 22 ISP[a])	15
Average turn duration in seconds	20	17	17
Unique labelers	87	102	42

[a] Depending on the condition to pool on, which is a combination of recordings from either different recording session or different images presented to the speaker, or both
ISP = Image-Session Pairs

descriptive statistics. Image 6, image 16 and image 17 reach a critical mass when aggregating the recordings from different sessions. Image 2 and image 12 have been included to add more variety in textual description and increase the overall amount of data available for automatic classification as explained in Chap. 5. For statistical tests and correlation analyses exploring time- and text-dependency, subsets of the data can flexibly be pooled together in order to meet the critical mass required for descriptive statistics. Note that pooling would only be applicable when the data does not show significant inherent distinctions. Chapter 4 therefore analyzes the data and data structures in a subsequent order. Table 3.2 also gives the numbers of available labels when all recording sessions are pooled as well as the number of labels available when all images are pooled. Eventually, when taking all data and images into one group, 148 labels per personality target are available for analysis.

Also the multi-speaker recordings were cut to match a turn length of 20 s roughly. Again, complete phrases were kept and cuts were applied in between different phrases manually. Forty-two raters listened to the stimuli as described above and filled out personality questionnaires. Each stimulus was rated by minimum 15 raters, stimuli were in random order. Raters were of 29 years of age on average with 53 % of the raters being male. Eventually, 64 conversation extracts, each of which has been assessed by 15 raters, are available.

To sum up the availability of data and labels for all the recorded data, Table 3.3 gives an overview. This perspective predominantly provides the amount of data that could potentially be used for training in automatic processing, i.e. beyond descriptive statistics. Note, since no deliberate target classes were induced in the case of multi-speaker data, a number of discrete targets and respective labels are not available.

3.5 Summary

This chapter outlines the speech corpus characteristics. Starting with a reference from the related field of emotion recognition, the design of the corpus is explained. In more detail, the corpus is designed to match the spirit of early works on emotion recognition. Therefore, a number of three subsets comprising different constraints to the recordings has been included in the corpus.

A first set of recordings comprises speech that is confined in various respects. First, all stimuli were produced by a single professional actor. Second, the actor was given a fixed text passage to portray out of different personality perspectives. Third, the actor was given time in advance to prepare a clear targeted acting. These records are referred to as *text-dependent* records. They are *speaker-dependent* at the same time.

For a second series of records the actor was invited several times with recording sessions spread over several weeks. This time, the actor was presented images like photos, drawings or artwork. The images were selected to trigger emotions and feelings. The actor was asked to speak freely about any associations that arouse within his mind while acting out of different personality perspectives. Therefore, this set of recordings is referred to as *text-independent*, still being *speaker-dependent*. Because of database design this set provides the opportunity to analyze time- and text-dependency in personality expression at the same time.

For the two sets above, the speaker was asked to perform personality targets along the Big 5 personality traits. For each trait, one high target and one low target was explained to the speaker as described in the NEO-FFI, cf. (Costa and McCrae 1992) and Sect. 1.2.5. Eventually, he portrayed the fixed and spontaneous text speech out of 10 different personality targets in randomized order. All fixed text recordings were repeated 20 times, 4 recording sessions of spontaneous portrayals were scheduled.

For a third series of recordings 64 native speakers were recorded. The speakers were engaged in different conversational scenarios, no personality perspectives were induced. This set will be referred to as *multi-speaker* set, which is also of spontaneous speech but originating from conversational communication. In addition, the set comprises two technical recording conditions, as half of the speakers were recorded with stand alone microphones while the other half were recorded with a headset.

In order to generate personality labels listening tests were conducted. A multitude of raters filled out full NEO-FFI questionnaires after having heard 20 s excerpts from the recordings one or more times. Because of cost and time limitations, only a selection of recordings was annotated. For the fixed-text data 3 out of 20 stimuli for each of the 10 targets were selected according to an additional rating on "artificiality". Each selected stimulus was rated by 20 listeners, generating scores on all of the Big 5 traits. From the spontaneous-text data 7 out of 12 images were selected that triggered the most various associations with the speaker. Depending on the individual image, a number of 3–15 raters listened to each of the 10 targets of each of the 7 images. For the multi-speaker data each stimulus was rated by 15 listeners.

The overall aim of generating labels for the data is two-fold. On the one hand, sufficient numbers of labels for each of the personality targets must be generated in order to meet the critical mass required for descriptive statistics, as will be executed in Chap. 4. On the other hand, this leads to a situation where relatively few data points are annotated with a relatively high number of labels. For automatic modeling, as will be described in Chap. 5, a higher number of data points is desirable. Here automatic means must be presented sufficient numbers of samples to learn from, while the multitude of labels is reduced to only one prototypical label.

References

Batliner A, Fischer K, Huber R, Spilker J, Nöth E (2000) Desperately seeking emotions: actors, wizards, and human beings. In: ISCA workshop on speech and emotion

Costa PT, McCrae RR (1992) Revised NEO personality inventory (NEO-PI-R) and NEO five-factor inventory (NEO-FFI) manual. Psychological Assessment Resources, Odessa

Möller S (2007) Subjective evaluation of conversational quality (ITU-T Rec. P.805). Technical report, International Telecommunication Union, CH-Geneva

Murray HA (1938) Explorations in personality. Oxford University Press, New York

Polzehl T, Möller S, Metze F (2010a) Automatically assessing personality from speech. In: Proceedings of international conference on semantic computing (ICSC 2010), 1–6. IEEE

Polzehl T, Möller S, Metze F (2010b) Automatically assessing acoustic manifestations of personality in speech. In: Workshop on spoken language technology, Berkeley. IEEE

Polzehl T, Möller S, Metze F (2011) Modeling speaker personality using voice. In: Proceedings of the annual conference of the international speech communication association (Interspeech 2011), Florence. ISCa

Chapter 4
Analysis of Human Personality Perception

In essence, this chapter serves for exploratory insights into the personality-related expressions and interdependencies in the datasets. At the same time, the question of whether or not the chosen assessment scheme can be applied to speech input is tackled. For these reasons, different aspects of human personality perception, as captured by the use of the NEO-FFI questionnaire (Costa and McCrae 1992) annotations and explained in Sect. 1.2.5, are presented and analyzed by means of auditory and statistical description. Signal-based analyzes will be explained in the following Chap. 5. First, auditory impressions as attributed by the author of this work are given as hypotheses. Having dealt with voice analysis from various perspectives for many years, the author's individual impressions on intonation, dynamics, rhythm, and voice qualities with regard to the recorded speech are described in Sect. 4.1. This serves two purposes: (a) to help the reader to comprehend and understand the data and its properties on an intuitive level; and (b) to provide information for the comparison between results from statistical analyses, intuitive perception, and signal-based modeling as will be presented in Chap. 5. Analyzing the rating responses from the multitude of invited raters, correlation analyses are given in Sect. 4.2. Analyzing the applicability and portability of the underlying NEO-FFI test scheme, Sect. 4.3 presents latent factor analyses and compares the extracted structure with the pre-determined structure of traditional, i.e. non-speech application of the NEO-FFI assessment scheme. Looking for statistically significant differences in the perception of personality the given class structure is analyzed by means of variance analyses in Sect. 4.4. Finally, results from all analyses of human personality assessment from speech are presented in a short chapter summary. Results from the text-dependent subset were partly also published in Polzehl et al. (2010b). Results from the text-independent subset were partly published in Polzehl et al. (2010a). Most recently, a subset of the results on the multi-speaker data was published in Polzehl et al. (2011).

T. Polzehl, *Personality in Speech*, T-Labs Series in Telecommunication Services,
DOI 10.1007/978-3-319-09516-5_4

4.1 Auditory Impressions and Personality Perception Hypotheses

In addition to Sect. 1.2.5, where psychological definitions of the Big 5 personality traits are given, this section presents auditory impressions of personality from the recorded database. These impressions are based on the text-dependent and text-independent data sets. For the multi-speaker data no class structure can be assumed a priori. As motivated in Chap. 2, not all characteristics of psychological personality might be transmitted via speech. Also, present limitations like the ability of the speakers to act or express personality and the fact that the recordings are cut to 20 s in time on average can influence or even hinder personality expressions from being transmitted and/or received. Therefore, and also in order to help the reader to effectively envision the data on an intuitive level, individual personality perceptions are given in Table 4.1. These perceptual impressions are not validating any class structures or labeling of these structures. Still, for the presented analyses the classes will be referred to by the personality perceptions they are expected to comprise in order to facilitate reading and understanding of the following analyzes. For the same reason, also stereotypical examples are given where appropriate. Although these stereotypes are useful to have a quick pointer to sometimes very complex characteristics, the reader should be reminded that the very nature of stereotypes is to be incomplete and non-representative for the whole matter at hand.

Also, it is very likely that complex personality characteristics cannot exhaustively be characterized. Still, the main auditory impressions will be described by a small array of 8 characteristics, as presented in Table 4.1. The characteristics are chosen to represent the most commonly acknowledged characteristics found in the literature as presented in Sect. 2.2, i.e. *Relative Pitch, Pitch Variation, Intensity, Tempo, Rhythm, Pauses, Voice Quality,* and *Overall Impression*. Impressions or associations from the cells that are not explained below seem inconsistent or inconclusive. Finally, these perspectives are the expert's opinion of the author of this work, having dealt with voice analysis from various perspectives for many years. Also because of harmonic and rhythmic education that the author has undergone in the past, individual impressions on intonation, dynamics, rhythm, and voice qualities can be formulated with substantial background.

Characterizing intonation, harmony and melody perception, the first two columns of Table 4.1 show relative pitch and pitch variation estimation. Relative pitch has been drawn out of the context of all stimuli in the corpus, observed levels are *low*, *mid*, and *high*. Pitch variation denotes pitch movements and has been characterized as *high*, *mid*, and *low* in range. Overall, the auditory impression of **High Openness** and **High Extroversion** stimuli are of high relative pitch combined with high variation. This can reflect the stereotype, that there is more excitement, arousal, commitment and activation with these personalities. Also pitch target frequencies, i.e. longer perceptually prominent segments between actual jumps and slopes in the pitch movements, were perceived as melodic for these personality targets. In opposition, **Low Agreeableness** stimuli were perceived as rather disharmonic. Least variation is observed for **low openness** stimuli, while the perceptual impression of a damped, i.e. consciously flattened but still perceivable pitch movement, mostly corresponds to mid-level relative pitch movements. This might link to the stereotype of withdrawn, indifferent personalities, trying to minimize effort in physical movement of articulators. Possibly, this might as well be seen as an effort not to be perceived as activated or committed.

Similar impressions arises when judging on intensities. **High Openness** and **High Extroversion** stimuli are perceived as dynamic, i.e. actively changing behavior, with extroversion exceeding openness in range. Here, also low agreeableness stimuli leave a dynamic and energetic impression. In terms of a stereotypical explanation, this effort might be caused by the wish to enforce or push non-agreeableness, which is in opposition to the indifference with **Low Openness** stimuli mentioned before. **Low Neuroticism** stimuli on the other hand are perceived as flat in intensity, a characteristic they share with **Low Openness** stimuli. This might point to the stereotype of being withdrawn but stable and assured of one's attitude at the same time for these two personalities. More general, a kind of passiveness, calmness, non-activation seems to be perceptually inherent for **Low Openness**, **Low Conscientiousness**, **Low Extroversion**, and **Low Neuroticism** stimuli. Hence, with regard to the prosodic characteristics of pitch and intensity, these personalties build a cluster. The one low target missing in this cluster is **Low Agreeableness**. This might be again caused by the stereotype of actively pushing characteristics of non-agreeable behavior like egocentrism and competitiveness into vocal expression.

In terms of speaking rate, rhythm, and pauses, extroversion shows the fastest speech, combined with a straight, i.e. non-broken, non-changing rhythm. Here, only few pauses of short duration are observed. Slowest speech is perceived with **High Conscientiousness** stimuli. Another stereotypical example might be of help. Accordingly, very accurate personalities might be more likely to take their time in order to express things in more detail or more clearly. The impression of a staccato, i.e. a non-flowing, choppy rhythm pattern for **High Conscientiousness** stimuli might be caused by a high frequency of perceivable pauses. Because of the slow speaking rate these pauses are not necessarily short, as for extroversion, but perceptually prominent. On the opposite site, a rather legato or slurred rhythmic pattern is perceived from **Low Conscientiousness**, **Low Extroversion** and **Low Neuroticism** stimuli. Similar to finding from pitch and intensity, these perceptual non-emphatic impressions might be accounted to the low activity level.

The next column of Table 4.1 refers to perceptual voice quality. As a first observation, **High Openness**, **High Conscientiousness**, **High Extroversion**, **High Agreeableness** as well as **Low Neuroticism** are perceived as sonorous, i.e. resonant and orotund. Here, the voice is perceptually present and appears unconcealed and healthy. In opposition, neurotic stimuli are perceived as crumble. Similarly, introverted stimuli are perceived as damped and concealed. In terms of stereotypes, this could indicate to self-assured speakers using their voice overtly, clearly with the aim of being perceived and noticed, while withdrawn or instable personalities are rather unwilling to be in the center of any attention, thus avoiding orotund and attention catching voices. On another dimension, non-agreeable stimuli are perceived as tensed and

Table 4.1 Auditory descriptions of perceptual speech properties in the datasets

Target	Relative pitch	Pitch variation	Intensity	Tempo	Rhythm	Pauses	Voice quality	Overall impression
O+	High	High, melodic	Dynamic	Varying	Varying	Many, long	Straight, sonorous	Brisk
O−	Low	Low	Flat, low	Rather high	Straight	Many, long	Rather tensed	Monotonous
C+	Mid	Damped	Mid	Slow	Staccato	Many	Sonorous, rather tensed	Calm
C−	Low	Damped	Low	Rather high	Legato, slurred	Varying	Laxed	Calm
E+	High	Very high, melodic	Very dynamic	High	Straight, accented	Few, short	Sonorous, laxed	Lively
E−	Mid	Mid	Mid	Mid	Legato	Rather long	Damped	Calm
A+	Low	Mid	Dynamic	Mid	Straight	Mid	Sonorous	Calm
A−	Mid	Mid, rather disharmonic	Dynamic	Rather high	Straight	Many	Tensed, sharp	Monotonous
N+	Low	Damped	Damped	Rather slow	Legato, stumbled	Many, long	Fragile, crumbled	Monotonous
N−	Mid	Damped	Flat	Rather slow	Legato, slurred	Many, long	Sonorous	Calm, firm

sharp, which can refer to the desire of being perceived as forceful. This is in line with the stereotype that non-agreeable, e.g. egocentric, personalities seek challenge, competitions and stand up to other individuals, which requires forceful behavior.

The last column presents the author's overall impression given the hitherto mentioned characteristics. For the cluster comprising very accurate, very indifferent, very introverted, rather altruistic and non-neurotic personalities, a calm and unexcited impression results prominent. In clear opposition, brisk and lively speech is the most important perceptual characteristic of **High Openness** and **High Extroversion** stimuli.

As working hypothesis for further analyses and in order to be more comparable to findings from the literature as presented in Chap. 2, the most important characteristics for the present 10 personality targets can be summarized as follows:

High Openness …brisk, activated and sonorous vocal pattern showing much melodic and dynamic variation.

Low Openness …monotonous and flat in terms of melodic and dynamic aspects, combined with a high speaking rate and a rather tensed voice.

High Conscientiousness …calm, damped, and sonorous way of speaking, while the use of pauses and the slow tempo can add a rather staccato impression in terms of rhythm.

Low Conscientiousness …calm, damped and rather laxed way of speaking, while the rather high speaking rate can lead to an impression of slurred speech.

High Extroversion …melodic and lively impression accompanied by a very dynamic power pattern. Pauses are rather short, speaking rate is high.

Low Extroversion …mostly unobtrusive, calm way of speaking with damped voice quality and a mid-tempo.

High Agreeableness …sonorous and calm way of speaking with a dynamic power pattern and a straight rhythm.

Low Agreeableness …tensed and sharp voice enforced by a dynamic power pattern, a rather high speaking rate, and a straight rhythm. The overall impression is monotonous, rather disharmonic.

High Neuroticism …monotonous, rather slow and stumbled rhythmic pattern, a rather fragile voice and damped dynamic and melodic expressiveness.

Low Neuroticism …calm and firm impression with a sonorous voice, flat intensity and damped melodic variation.

4.2 Distributions and Correlation Analysis of Rating Responses

Using the NEO-FFI questionnaire for assessing personality from speech means to assess ratings on a five-point Likert scale, cf. Sect. 1.2.5. Since the response values on that scale are designed to be equidistant, and the summation of individual ordinal item responses spans a range for the eventual personality trait score to be of any integer value between zero and 48, the responses can be assumed to be of interval-level characteristic.

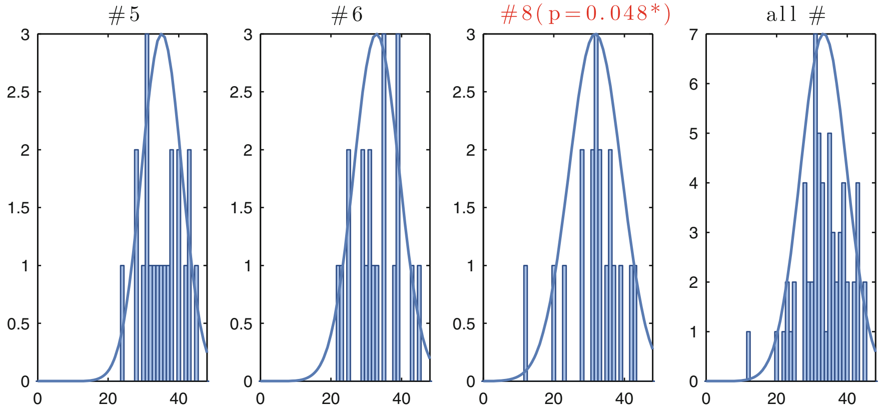

Fig. 4.1 Histograms (*bars*) and Gaussian normal fit (*lines*) plot for visualization of label distributions for the example of **High Conscientiousness** ($C+$). The three plots to the left account for 20 ratings for each of the individual recording repetitions #5, #6, and #8. The plot to the right shows the distribution of all labels, i.e. 60 labels, jointly. According to Lilliefors tests, only labels of take #8 do not belong to a normal distribution ($\alpha = 5\,\%$), but still show a high tendency towards a normal distribution

Regarding the text-dependent recordings, Fig. 4.1 shows the distribution of the 20 ratings available for the **High Conscientiousness** stimuli. The first three plots show histograms and a Gaussian normal fit for the 3 out of 20 selected recording repetitions that were selected for NEO-FFI annotation, i.e. take #5, #6 and #8. Lilliefors tests, which resemble Kolmogorov-Smirnov tests for normal distribution with mean and variance unknown, are applied. Accordingly, a test of the null hypothesis that the labels are drawn from a distribution in the normal family against the alternative, which is that they do not come from a normal distribution, is executed on a 5 % significance level. Only recording #8 significantly deviates from a normal distribution, although the visual impression and the high *p-value* suggest a strong tendency towards a normal distribution.[1] In the plot to the right the three recordings are analyzed jointly. When combining all ratings from the repetitions, 60 ratings are available for each personality target. In the present case, the joint distribution accounts for a normal distribution.

Looking at all available ratings for all available targets and recording repetitions, i.e. looking at three recording repetitions for each of 10 personality targets given by NEO-FFI annotation from 20 users each, only 7 % of all stimuli ratings reveal a non-Gaussian shape. Figure A.1 in Appendix A shows the distribution plots of all the data. In more detail, only two distributions reveal significant deviation from Gaussian shapes, i.e. the one mentioned above and repetition #8 from **Low Agreeableness** recordings. Because, after all, this number of non-Gaussian shapes appears very low, their occurrence is interpreted as irregularities and overall normal distribution of ratings will be assumed for all further analyses. Thus, the expected values

[1] The critical *p-value* for a 5 % significance level results in 0.05.

are assumed to follow the measurements of a distribution out of the Gaussian family, i.e. the arithmetic mean and the standard deviation account for the first and second order moments when describing the distributions. Note, that also due to the central limit theorem, stating that the sum of a large number of random variables is distributed approximately normally, normal distributions can coincidentally emerge when joining many repetitions or images.

To provide an intuitive feeling on the difference between all available ratings from low and high targets, Fig. 4.2 shows histograms along the Big 5 personality traits and a Gaussian fit overlay. The figure allows for a direct comparison between the ratings on low (brown) and high (light blue) targets. In all cases, low targets were perceived as expected, i.e. lower than the high targets. If the difference between the pairs of low and high targets was not perceivable, the Gaussians would predominantly overlap. While for some scales, e.g. neuroticism, the area of overlap appears to be small, for other scales, e.g. openness, the targets show much more overlap. This simply means that some high and low target pairs are more distinctly differentiable from this database.

Figure B.1 in Appendix B shows the distributions for the text-independent recordings by Gaussian fit. As explained in Sect. 3.4, 5 pictures taken from different recording sessions were annotated by 15 individuals. The plot shows the distributions for high and low targets for respective combinations of recording session and image display. Similar to results from text-dependent analysis, only very few rating distributions show a significant deviation from a Gaussian shape, i.e. 8 % (4 out of 50). We therefore assume overall normal distribution of ratings also for this subset of data. Again, all targets designed to be of high values were perceived as expected, i.e. higher as the low targets.

The multi-speaker data cannot be displayed in terms of high and low target classes for the simple reason that they do not comprise any induced structure. In order to still gain some insights into the data all item responses from all raters were tested for normal distribution. As a result, for 9 % of all stimuli the ratings on openness significantly diverted from a normal distribution. For C, E, A, N the respective share of non-normal distribution resulted in 8, 8, 11, and 12 % respectively.

When differentiating between the subsets comprising close-talk recordings and the subset comprising the stand-alone microphone recordings the same results hold true, in principle. Only two individual observations need to be emphasized at this point in time. First, the percentage of non-normal distribution is generally lower for the subset comprising the stand-alone microphone recordings. Second, ratings on agreeableness and neuroticism from the close-talk recordings show a clearly increased share of non-normal distributions, namely 20 % as well as 17 % respectively. Apart from these two exceptions, the observed figures match the figures seen from text-dependent and text-independent stimuli assessment. Hence, subsequent analyses will consider the ratings for one stimuli from different speakers to be of overall normal distribution, as well.

Regarding the joint trait scores generated out of all responses for each individual trait, Lilliefors tests show that non of these distributions belong to a normal distribution ($p < 0.001$). This also indicates, that the underlying population of respective personality perceptions does not come from a single characteristic or family, but

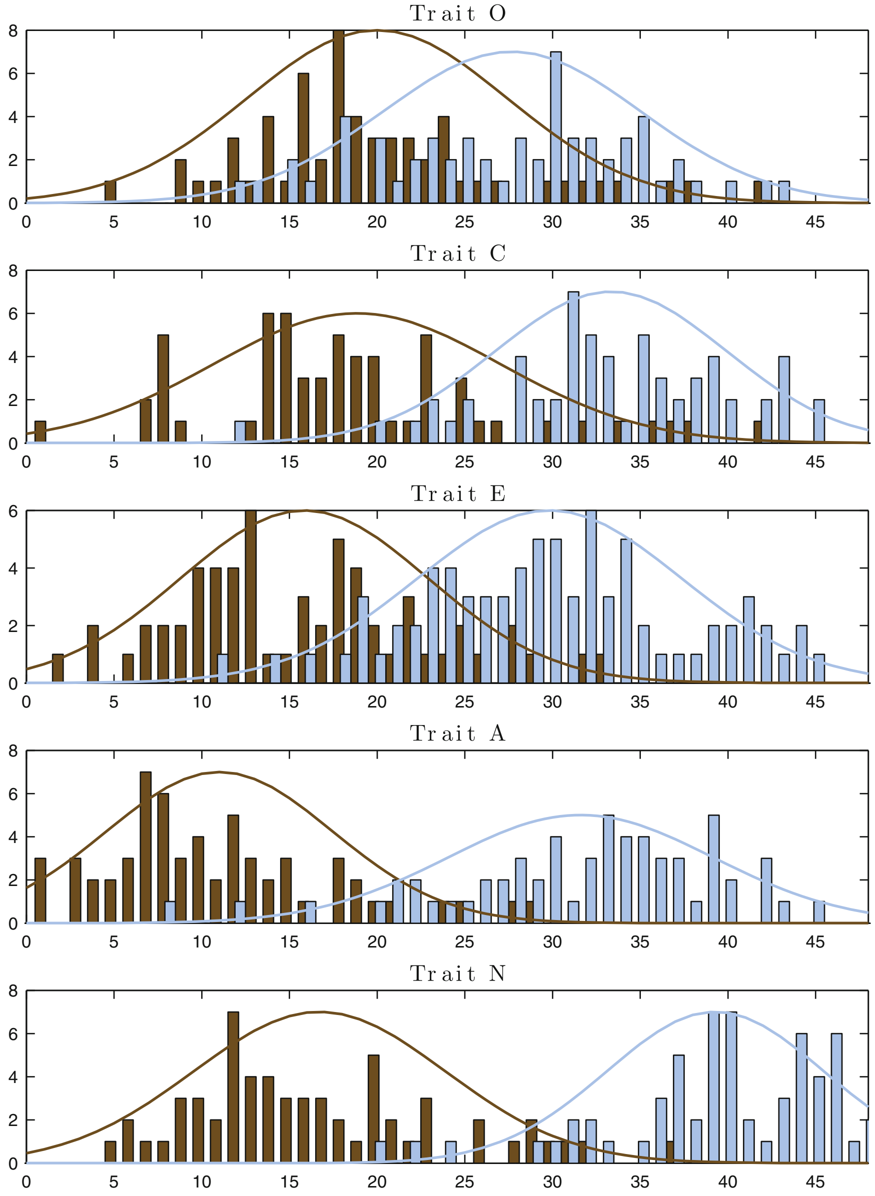

Fig. 4.2 Histograms (*bars*) and Gaussian normal fit (*lines)* plot for visualization of high (*light blue*) and low (*brown*) target ratings on the text-dependent recordings along the Big 5 personality traits

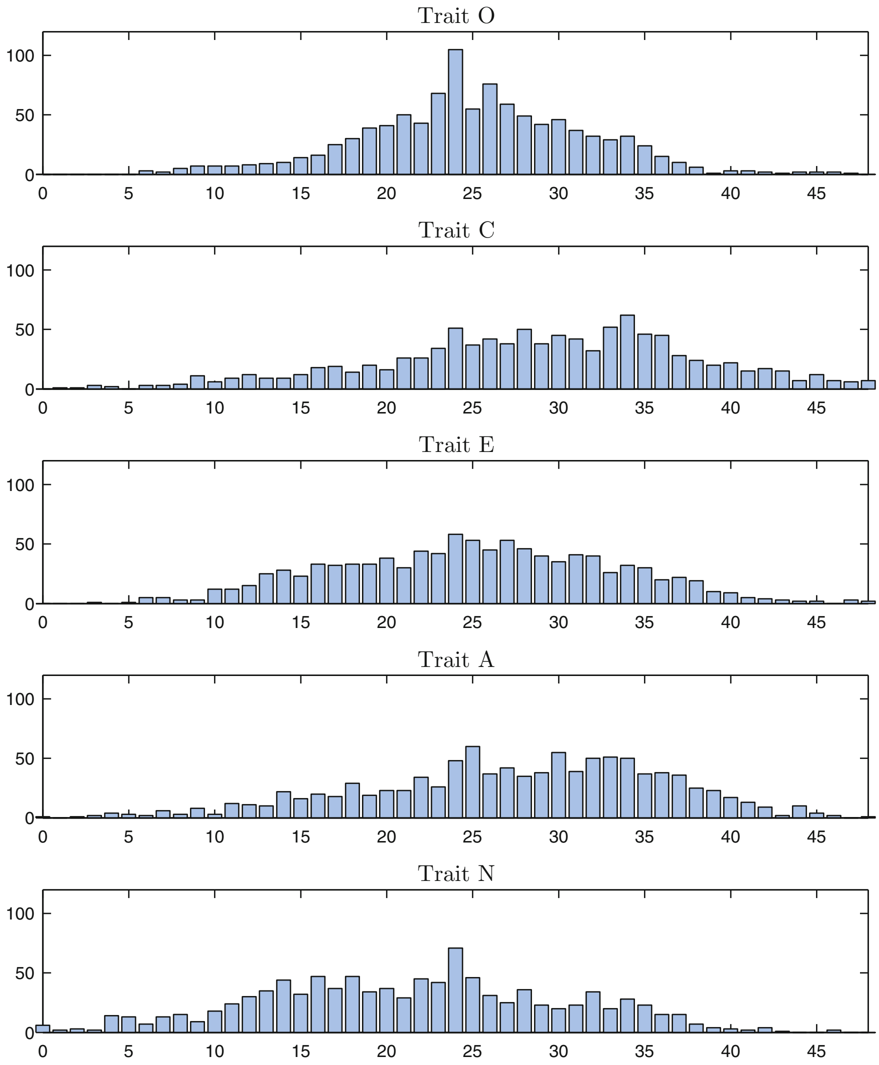

Fig. 4.3 Histograms showing the distributions of the joint ratings on all speakers in the multi-speaker subset for each individual trait

reflects multiple diverse personalities. The actual distributions of the joint ratings are shown in Fig. 4.3.

Given the observed overall normal distributions and an interval-leveled scale, correlations for speaker-dependent subsets, i.e. the text-dependent and the text-independent subsets, are calculated as Pearson *product-moment* correlation. This measure is a commonly used measurement for estimation of linear dependence between two variables of these characteristics. For the multi-speaker subset correlations are calculated as Spearman *rank correlation*. In general, both types of correlation r can range between $+1$ and 1, stating a perfect linear relationship for which both variables increase in the same direction at value $+1$. A value of -1 implies that all data points lie on a line for which one variable decreases as the other increases. At a value of 0 no linear correlation is stated. In terms of absolute figures, magnitudes below or equal to 0.2 are commonly seen as *very weak*, below or equal to 0.5 as *weak*, below or equal to 0.7 as *moderate*, below or equal to 0.9 as *good*, and magnitudes above as *very good*.

Table 4.2 shows the resulting correlation coefficients between the Big Five traits along the off-diagonals. All correlations presented are significant ($p < 0.01$). The three subparts correspond to: a) correlations calculated on basis of the text-dependent ratings; b) correlations calculated on basis of the ratings of the text-independent stimuli; and c) correlations calculated on basis of the ratings of the speaker independent setup. Correlations are given between the traits by the respective cells in the rows and columns except from the diagonal, which corresponds to the consistency as described below.

The question how to interpret these correlations among the traits can only be answered by looking at the correlations provided with the original German NEO-FFI as described in Sect. 1.2.6. As a first result one clearly notices a general increase of all correlations when applying the NEO-FFI to personality impressions based on speech data only. While the average absolute correlation results in 0.14, i.e. very weak, for the original NEO-FFI ratings, the correlation rises to 0.29, 0.35, and 0.52 for multi-speaker, text-independent, and text-dependent ratings respectively. This observation might be explained by the absence of background information, visual impression, context information, or other prior experience about the speaker. Here, the listeners needed to rely on a 20 s voice excerpt only, which potentially limits the very existence of cues to deduce personality from drastically. For a more elaborate comparison of common personality assessment practice in psychology and the expected differences to personality assessment from speech please revisit Chap. 2. Furthermore, only negligible changes in correlations were obtained when sub-dividing the multi-speaker data into the close-talk and stand-alone recording condition. A separated interpretation is therefore omitted.

Looking more closely at the NEO-FFI speech application some interesting results can be highlighted. Firstly, the reason why the correlations increase when moving from multi-speaker data towards text-independent data to text-dependent data could potentially be explained by the limited degree of freedom to which the actor was bound when performing the speech actings. Also with respect to the listener, judging personality from always the same linguistic content might affect the level of difficulty of the assessment task. While different speakers are very likely to produce speech more diversely and independently, a single speaker can be expected to have a fixed set of expressive cues. Likewise, while in the case of text-independent recordings the choice of words was free, the speaker repeated the identical text passage while doing the text-dependent recordings. That way, the perceptive distance in the resulting recordings is likely to be smaller than the difference in between conditions containing diverse wording, and the latter is consequently expected to show smaller differences than a set of multiple speakers with individual ways of expression.

Table 4.2 Consistencies (diagonal) and correlations (off-diagonals) between NEO-FFI scale ratings: left) on basis of the text-dependent ratings from speech; middle) on basis of text-independent ratings from speech; right) on basis of multi-speaker speech data

	Text-dependent					Text-independent					Multi-Speaker				
	O	C	E	A	N	O	C	E	A	N	O	C	E	A	N
O	(0.83)	0.30	−0.20	0.41	0.46	(0.83)	0.34	0.47	0.53	−0.34	(0.79)	0.16	0.43	0.50	−0.26
C		(0.88)	0.51	0.60	−0.67		(0.93)	0.19	0.22	−0.43		(0.93)	0.14	0.13	−0.41
E			(0.80)	−0.59	−0.55			(0.91)	0.28	−0.67			(0.88)	0.31	−0.55
A				(0.84)	0.78				(0.90)	−0.14				(0.89)	−0.05
N					(0.85)					(0.93)					(0.90)

Looking more closely at individual figures, the following list of findings can be given:

1. The consistent moderate negative correlation between N and E seems to be an inherent characteristic of speech for all our databases. Also results from traditional NEO-FFI application show the highest correlation between these two traits. At the same time, there is a relative high negative correlation between N and C in the data as well, which can also be found in the recorded data. Consequently, the more neurotic a speaker is being judged, the less extroverted and conscientious he is being perceived in parallel. This finding seems to be consistent for both traditional and speech application of the NEO-FFI. A consistent dependency between E and C could not be found. This result is in line with the auditory perception presented in Table 4.1.
2. Another mostly moderate and consistent correlation occurs between A and O. Hence, when being perceived as more open the speakers were also perceived as more agreeable. In other words, the more curiosity and open-mindedness the actor wanted to portray in the actings, the more commitment and social reflection he was perceived by the listeners.
3. Inconclusive indications result for the relation between A and E as well as between A and N. While the results from the multi-speaker and text-independent subsets suggest that the overall correlation would be weak or even very weak, results from text-dependent data show a somewhat different picture. For this subset, the more introverted actings were also perceived as more agreeable, the more extroverted as less agreeable, i.e. more egocentric. One the other hand, more agreeable actings were also perceived as more neurotic, or, in other words, more egocentric actings were perceived as less neurotic.
4. Common to all three databases is the oftentimes weak and sometimes even very weak overall correlation between all other traits, thus retaining statistical independence. After all, statistical independence is desired for the traits in order to assess different and unrelated aspects of personality.

The diagonals of the three sub-tables in Table 4.2 show intra-scale consistencies by means of Cronbach's Alpha (α), cf. Zinbarg et al. (2005). This measure, which is oftentimes also called internal consistency, is commonly used to estimate the reliability of a psychometric test score for a series of ratings. Originally introduced by Lee Cronbach (Cronbach 1951), Zinbarg's extension allows for calculation of consistencies among more than two raters. Magnitudes below or equal to 0.5 are commonly seen as *unacceptable*, below or equal to 0.6 as *poor*, below or equal to 0.7 as *questionable*, below or equal to 0.8 as *acceptable*, below or equal to 0.9 as *good*, and above as *excellent*. The total α range results between minus infinity and one. As a rule of thumb, results worse than $questionable$ should not be taken into account and the applicability or scale design should be pushed back to reconsideration. In these cases, one needs to verify if the scale at hand really measures a single construct or if the ratings could be influenced by multiple constructs inherent in the scale. Factor analyses can be of help to shed some light on latent scale structures. In this work, Sect. 4.3 will explore the underlying factor structures to verify the application of both

Cronbach's Alpha and the NEO-FFI item coding for NEO-FFI application to speech input only. At the same time, good consistencies do not exclude the possibility that the scale at hand might still measure more than one construct. As an example, a scale intermingling the constructs *depression* and *anxiety* can still show high consistency, if the constructs are set into a context where they act similar to each other. On the other hand, if dissimilar constructs are intermingled, the respective consistency decrease.

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K}\sigma_{Y_i}^2}{\sigma_X^2}\right) \tag{4.1}$$

Equation 4.1 shows the formula for calculation of α, with K being the number of items, σ_X^2 the variance of the observed scores in total X, and $\sigma_{Y_i}^2$ the variance of the item i for the current subsample of items or raters Y. Alternatively, a standardized version of Cronbach's Alpha is defined as shown in Eq. 4.2 with K defined as above, and $\bar{r}$ being the mean correlation between the ratings of the different items. Note, that the mean correlation corresponds to the mean of the $K(K-1)/2$ pairs of possible couplings from the upper or lower triangular correlation matrix. Also, because the reliability calculation bases on correlation, it is sensible to relative relations, not absolute. To give an example, let one construct be measured by three items on a five-point Likert scale. Rater A rates [1, 2, 3, 2, 1], rater B rates [3, 4, 5, 4, 3]. Although the two raters do not agree on the absolute ratings, their relative relation matches perfect, hence the consistency results perfectly high.

$$\alpha = \frac{K\bar{r}}{(1+(K-1)\bar{r})} \tag{4.2}$$

Looking at the results in Tables 4.2 and 4.3, all consistencies show good or excellent magnitudes, O and A being exceptions resulting in acceptable figures for the original NEO-FFI ratings only. Again, only negligible changes in consistencies were obtained when sub-dividing the multi-speaker data into the close-talk and stand-alone recording condition. A separated interpretation is therefore omitted.

The overall average level of consistency increases from the original NEO-FFI ratings to the text-dependent ratings, towards the multi-speaker ratings and is highest for the text-independent ratings. When taking into account that for the original NEO-FFI

Table 4.3 Consistencies (diagonal) and correlations (off-diagonals) between NEO-FFI scale ratings as given by the German version of the manual of the NEO-FFI

	O	C	E	A	N
O	(0.75)	−0.10	0.14	0.05	0.02
C		(0.84)	0.13	0.10	−0.26
E			(0.81)	0.16	−0.36
A				(0.72)	−0.10
N					(0.87)

and text-independent speech data multitudes of speakers have been assessed, the observation that consistencies for these data show slightly lower magnitudes seems to be comprehensible. The question why the raters show more consistency when rating the text-independent recordings than when rating the text-dependent recordings can again be answered by the hypothesis, that the speaker was more free to chose words including any type of connotation to the words when composing his own sentences. He was more severely limited when performing the text-dependent recordings. This degree of freedom could potentially have caused a more distinguishable and targeted expression with regard to personality in the text-independent actings.

When looking at individual consistencies, highest magnitudes occur with scales *N*, *C*, and *E* for the original NEO-FFI figures. Also for the ratings from speech data these consistencies show highest magnitudes, except from the slightly lower magnitude of *E* ratings for text-dependent data. The lower consistencies of ratings on *O* indicate that assessing openness in general seems to be more difficult. These consistencies result in the lowest magnitudes for all the datasets except from the text-dependent subset, where it is only undercut by the exceptionally low *E* value.

Interpreting these results, the overall good and excellent consistencies support the basic assumption of this work, namely the assumption that the NEO-FFI can be used to access personality from speech. Especially scales *N*, *C*, and *A* seems to be assessable with high consistency from speech data. While openness seems to cause most difficulties in assessment, the resulting consistencies still reach a good or acceptable level.

4.3 Factor Analyzing the NEO-FFI Item Responses

In this work, the NEO-FFI and its method to describe and assess the Big 5 personality traits are used in two ways. First, the verbal descriptions of the Big 5 are used to illustrate the targets that the actor has been asked to perform. Second, the NEO-FFI questionnaire with its respective 60 items and coding scheme to aggregate 5 trait scores out of these items are applied to assessment from speech input only. Consequently, the question of whether or not the questionnaire can actually be applied to speech input also includes the question whether or not the item coding scheme still functions with speech data as it does with personality assessment in psychology. In order to explore the underlying structure of the collected item responses, and compare the structures to the structure inherent in the NEO-FFI coding scheme, factor analyses hypothesizing the presence of 5 underlying factors are conducted. Hypothetically, these analyses are expected to reveal the same five factor structure with each factor being built up from the same individual items when compared with the item factor structure in the NEO-FFI. The finding of an identical or comparable structure would mean that the NEO-FFI can also directly be applied to speech input without loosing the validity of the NEO-FFI even when taken to the speech domain.

Conducting the factor analyses the maximum likelihood method for component extraction and orthogonal factor rotation *(varimax)* are applied. The aim is to reveal any latent variables that cause the assumed factors to correlate. During factor extraction the shared variance of variables is partitioned from its unique variance and error variance to reveal the underlying factor structure. Here, only shared variance appears in the solution, which is unlike another related and popular method of factor extraction named *principal components analysis (PCA)*, cf. Bortz (2005).

Observing the loadings of these revealed factors, loadings below 0.4 are disregarded. Table 4.4 shows the summary of the loading structures for the text-dependent, text-independent and multi-speaker datasets. Both columns and rows are sorted by the respective explanatory power according to covered variance share. The first three rows show the order of extracted factors, their explanatory power and the assigned trait interpretation. Rows below show the loadings sorted in descending order, i.e. on the multi-speaker dataset, the order of loadings on factor 1 was $C10$, $C2, \ldots, O1$ with the loading of $C10$ being highest in absolute magnitude, and $O1$ being lowest in absolute magnitude. The results given for the multi-speaker data resembles the results from the joint recording conditions, i.e. close-talk and stand-alone together, since the individual analyses resulted in almost identical structures and loadings.

Table 4.4 Latent factor structures in the speech databases

	Multi-speaker					Text-independent					Text-dependent				
Factor #	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
% Variance explained	25.0	23.5	23.5	16.8	11.3	22.2	21.8	21.5	18.6	15.9	23.1	22.1	21.9	20.8	12.2
Factor inter-pretation	C	N	A	E	O	C	N	A	E	O	E	N	C	A	O
Loadings by NEO-FFI	C10	N6	A8	E7	O3	C10	N6	A3	E7	O5	E7	N3	C10	A3	O3
label & item#	C2	N5	A3	E11	O9	C7	N3	A8	E8	O9	E1	N11	C1	A8	O9
	C8	N3	A12	E8	O5	C5	N5	A12	E2	O3	E8	N6	C2	A12	O2
	C1	N11	A5	E1	O10	C8	N2	A10	E1	O12	E11	N5	C4	A5	O5
	C5	N9	A10	E4	O2	C2	N11	A5	E4	O10	E5	N9	C8	A10	O10
	C7	N2	A9	E2	O12	C4	N9	A2	E5	O2	E2	N12	C7	A2	O12
	C3	N12	A7	E5		C6	N7	A7	E9	O11	E4	N2	C5	A1	O11
	C4	N7	A4	E9		C1	N12	A11	E11	O7	E9	N8	C12	A4	O7
	C11	N10	A1	O6		C9	N10	A9	E3	O6	E3	N7	C9	A9	
	C9	N8	A11	E3		C12	N8	A4	O6	O4	O6	N10	C6	A7	
	C6	N4	A2	E10		C3	N4	A1	E10		E10	N4	C11	O7	
	C12	N1	O7			C11	N1	O7	E6		A1	N1	C3	A11	
	O1	A6	E6			O11	C11	E6	E12		E6	C11	O11		
		E12	E2				A6	A6	N10		N10		O1		
		E9									N4				
											A7				

Loadings are printed by reference to their meaning in the original NEO-FFI coding scheme, where 12 items build up a trait scale

Starting with the least powerful factor $F5$, the amount of explained variance results between 11.3 % for the multi-speaker set and 15.9 % for the text-independent set. The interpretation seems to be clear since only O items are observed. In general, out of the 12 O items available from the NEO-FFI questionnaire, relatively few are observed to load on $F5$ and relatively many are spread all over the other factors as cross-loadings. In more detail, only 6 items are loading on the O-factor on all datasets (#2, #3, #5, #9, #10, #12). Four items (#1, #4, #6, #8) seem to be not applicable for capturing openness from speech at all, as they do not load above the lower threshold of 0.4 or predominantly result in cross-loadings. The factors interpreted as *C*, *N*, and A show a consistent picture for all the three datasets. The percentage of variance explained results in between 25 % for C on multi-speaker data and 20.8 % for A on text-dependent data. All 12 C-items load on one factor and almost no cross-loadings of C-items occur in the datasets. Likewise, only very few cross-loadings can be found on the C-factors as well. The factors interpreted as N and A show very similar results. Also the loading structure of the factors interpreted as E show a very coherent picture, apart from the individual items #6 and #12, which show some cross-loading impact. However, the share of variance captured in the E-factors results to be higher for the text-dependent dataset than for the other datasets. Overall, a difference of up to 6.3 % absolute can be observed. Hence, using the NEO-FFI, extroversion was better assessable when a fixed-sequence of words was given which was only colored by the speaker with prosodical means, than when additionally allowing the speaker(s) to choose the wording freely. At the same time, the text-dependent condition can be thought of being more staged, acted, and thus invariant, as the speaker repeated the same text passage many times and could thus more deeply immerse into the play.

The overall goodness-of-fit of the 5 factor models cannot cover all of the variance seen in the datasets, i.e. 53.7 % for the multi-speaker set, 57.4 % for the text-independent set, and 57.0 % for the text-dependent set. The models clearly fail to capture all the relevant variances in the datasets significantly. But when imposing the constrain of 5 factors to be revealed this model is not intended to capture all the variance in the first place. It is built in order to allow for insights on latent structures. The revealed latent structures, indeed, clearly support the application of the chosen inventories for personality assessment from speech. The lack of overall fit could also be explained by the fact, that the NEO-FFI items were not designed to capture all aspects relevant for perceived personality impressions from speech. Thus, some items are expected to capture irrelevant information while others would try to capture information that might just not be present in speech data at all. On the other hand the share of variance explained by the models outperforms the share of variance explained as presented with the original NEO-FFI. Here, 37 % of variance could be explained by the first 5 factors, which is seen as typical magnitude of variance explanation as reported on a sample size of 11,724 samples in the NEO-FFI Manual, cf. Costa and McCrae (1992). Departing from the hypothesis that speech cannot convey the whole range of personality impressions, the reverse conclusion would be that one also expects less variance to be present in speech. Obviously, the NEO-FFI items were well able to capture this amount of information, so the share of variance explained in the speech data results relatively high. According to the results shown here, a sub-questionnaire comprising only relevant items for speech application could be extracted and proposed from the conducted factor analyses. However, more data comprising a higher number of speakers and recordings from more diverse scenarios would be desirable before trying to provide a general revised NEO-FFI for speech application.

In summary, the presented factor analysis confirms the hypothesis that the NEO-FFI can be applied to capture personality perceptions from speech while retaining the original NEO-FFI coding scheme. Some items are observed to fail to capture relevant information, few are observed as cross-loadings only. But apart from these "loose" items, the revealed structure in the recorded data proves very similar to the hypothesized NEO-FFI coding structure for all recorded data.

4.4 Analyses of Variance

Starting with the analysis of the text-dependent recordings the first experiment analyzes whether the mean perceived personality attributions differ significantly for three groups. For each of the Big Five factors, there exist:

1. A group of high targets, symbolized with "**+**" ,
2. A group of low targets, symbolized with "**–**", and
3. A group of stimuli that were not manipulated with respect to the factor at hand, symbolized with "**0**" and referred to as *0-target* group

As introduced in Sect. 3, when recording the text-dependent data the performance of each of the 10 target groups was repeated 20 times before selecting the three least artificial examples out of it. Therefore, the high and low target groups are populated by 20 ratings for each of three example recordings, i.e. 60 ratings in total. At the same time, these 60 stimuli were annotated with 60 full NEO-FFI questionnaires producing all of the Big Five factors for each stimuli—not just the manipulated target.

On trait level, this means that for any trait as shown in Fig. 4.4 there exist 4 other traits, i.e. 8 other targets, which were not intentionally manipulated. Ratings on these stimuli were subsumed to the 0-group. Each of the non-manipulated targets are also populated by 20 ratings for each of three repetitions, so finally the 0-group size results in 480 NEO-FFI questionnaires.

Due to the observed normal distribution in the data, cf. Sect. 4.2, and in order to see whether the high, low and 0-groups differ in terms of their mean, Tukey's honestly significant difference criterion for post-hoc tests of one-way analyses of variance ($p < 0.05$) are applied. Figure 4.4 shows the outcome of the post-hoc tests. The 0-group is plotted in blue, low and high targets are colored in red. Circles show the population means, lines correspond to confidence intervals. Any non-overlapping lines state a significant difference in between the groups belonging to these lines. Since no overlap on any trait is observed, all the low targets are perceived significantly

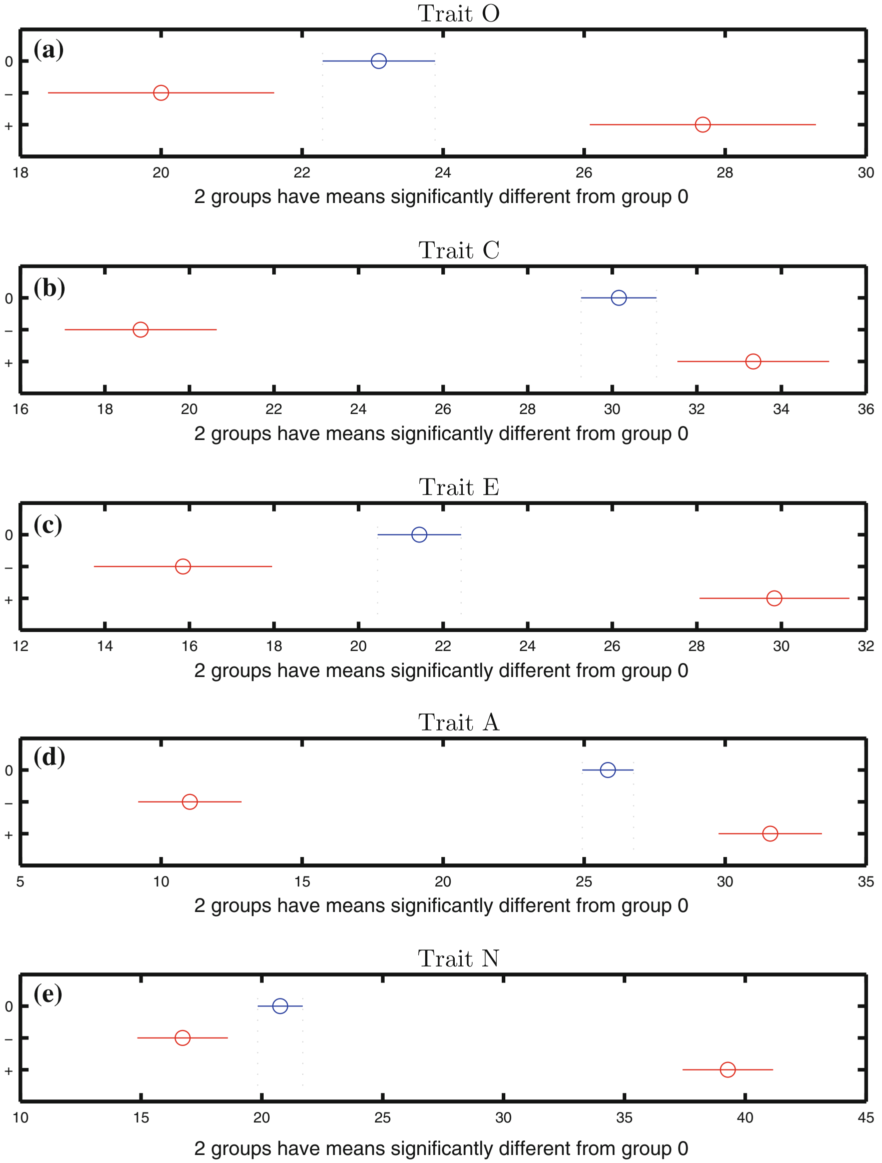

Fig. 4.4 Post Hoc tests, mean and confidence intervals for the text-dependent recordings. **a** Trait openness **b** Trait conscientiousness **c** Trait extroversion **d** Trait agreeableness **e** Trait neuroticism

lower than the groups of 0-targets, which in turn show significant lower perception than the high targets. In other words, these results prove, that all the intended manipulations, induced into the speaker and performed by the speaker, were also perceived as intended by the raters. The targets can therefore be expected to convey distinct acoustic or prosodic cues, that trigger the desired personality perceptions.

Because of the database design including temporal and textual diversity for the text-independent subset, the data offers the opportunity to analyze time dependency and text dependency of personality perceptions for the given speaker and stimuli.

In order to be able to analyze time-dependency, i.e. reproducibility of speech with the same personality character by the speaker, this subset was recorded in distinct sessions, several weeks apart. For analyses, only image descriptions which were assessed by at least 15 raters are selected and shown in Fig. 4.5. To see whether the mean perceived personality attribution depends on the recording date, Tukey's honestly significant difference criterion for post-hoc tests of one-way analyses of variance ($p < 0.05$) was applied. Combinations of recording sessions ($s1, \dots, s4$), and image shown to trigger associations ($I4$, $I9$) are shown on the vertical axis for low (m) and high (p) targets separately. As in the plots above, circles show the population means, lines correspond to confidence intervals. If the perceived personality in the speech recordings from different recording session triggers statistically different perceptions within the listeners, the respective confidence intervals do not overlap. The first group is plotted in blue, groups showing means significantly different from this group are colored in red. Any overlapping lines state a non-significant difference, i.e. a similarity, in between the groups belonging to the lines.

Figure 4.5a for example shows much overlap between ratings on image 4 recorded in session 3 and ratings on the same image from recordings of session 4, as can be seen in the first two lines of the plot, i.e. on the low target. For the high targets in line 3 and 4 confidence intervals seem almost identical. For ratings on speech triggered by image 9 the recordings of high targets from the three sessions show much overlap, as shown on the last three lines. For the low targets on image 9, the recording from session 4 shows a clear trend to be perceptively distinct from the recordings from session 2 and 3. However, this trend does not become significant. Figure 4.5b–e show the comparisons in the same order of targets, sessions and images (labels have been omitted). As a general observation from the first four lines, there is no difference in between any high or any low targets for the traits C, E, A and N triggered by image 4 in between recordings from session 3 and 4. Results on speech triggered by image 9, however, show more variation. Confidence intervals of all low targets generally match within the respective groups. The only exception is the perception of **Low Agreeableness**, as the ratings on recordings from session 3 were even significantly lower than the ratings of recordings from session 4. Looking at the high targets significant differences can be observed for traits E and N between ratings from recording session 4 and 3.

When looking closer at these differences, all variations in between the groups of high targets result in higher means than any variation mean of low targets. In other words, although there exists a small number of significant differences within the groups of high targets, it seems that the direction of these differences are towards even higher values, and not intermingling with the groups of low targets. In addition to these findings, one can generalize that from all sessions, images, and traits, all mean values of high targets result in higher values than any mean of any low target.

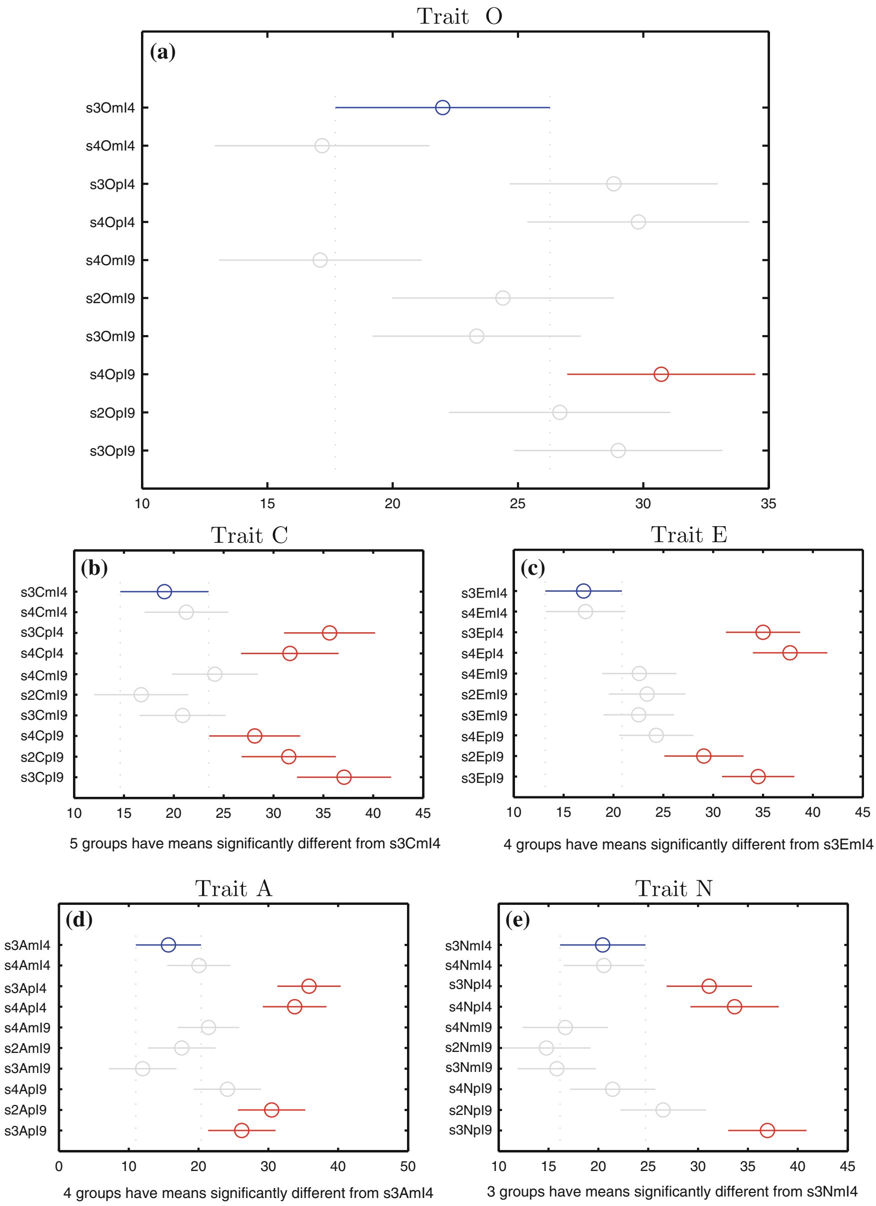

Fig. 4.5 Post Hoc tests, mean and confidence intervals for analysis of text-dependency and time-dependency on text-independent dataset. **a** Trait openness **b** Trait conscientiousness **c** Trait extroversion **d** Trait agreeableness **e** Trait neuroticism

There are a few hypothesis on explanations for the differences within the groups. The first possible explanation is that the actor was acting with different level of concentration leading to different quality of the actings. At one day he might have been able to portray a personality to a certain extend, at another day he might even be able to add to his expression and portray it in an even more distinguished way. A second hypothesis could reflect the fact that there exist more or more various triggers to draw personality from when looking at image 9 than there is for image 4.

Summing up on time-dependency it can be stated that, with few exceptions only, personality perceptions from the recorded actings are consistent with respect to temporal effects on speech expressions. The only exceptions to this general finding appear within the group of high targets in a way that pushes the intention of a high perception towards an even higher perception.

The next analysis aims to show differences that the spoken content could have caused. The recordings contain diverse speech, i.e. not even two recordings contain identical wording or sequences. On the contrary, because the speaker performed independently and freely, the wording and sounding is very dissimilar. Nevertheless, the very nature of the images could be expected to exert an unconscious impact. In analogy to the question of time-dependency, this question can be answered by executing a series of analyses of variance. Here, the grouping of the database is done according to the presented images. Revisiting Fig. 4.5 the differences between the images presented are now in the focus.

In case of openness we see that there is no significant difference in between the low targets on image 4 and image 9. Neither is there a significant difference in between high targets. Overall, more diversity within the low target group can be observed when compared to the high target group. In addition, for openness low and high targets seem to be very close to each other. A similar but more clear picture can be observed when looking at conscientiousness. Here, the groups of high and low targets contain all respective images. Hence, the difference with respect to image presentation does not cause a significant change in personality perception. For agreeableness a very similar picture can be observed, although the high target from image 4 and session 3 was perceived significantly higher than the high target of image 9. Once again, being perceived even higher, the high targets stay even more clear from being confused with the low targets.

Like observed when looking at the time-dependency, also for text-dependency the perception of the low targets on extroversion and neuroticism show consistency. Again, the high target groups show much variation, especially within the ratings of image 9. Because this within-group variation results large, this confidence interval of this group overlaps with the confidence interval of image 4 ratings. However, the inconsistencies within this group is directed to even higher values, and stays thus even more clear from being mixed up with the low target groups.

In summary of the above presented analyses on time- and text-dependency, one can say that in general the effect of different recording times and visual stimuli does not prove significant differences. The unexpected differences within the low and high target groups observed in the data are directed towards even higher or even lower values and result uncritical for the interpretation.

Table 4.5 Impact of deliberate acting of individual personality traits on secondarily affected traits

		Deliberate actings									
		Increase					Decrease				
		O	C	E	A	N	O	C	E	A	N
Impact	O	·		↑	↑		·	↓		↓	
	C		·			↓		·			
	E	↑		·	↑	↓	↓	↓	·		
	A				·	↑		↓		·	
	N	↓		↓		·	↑				·

Therefore, and in order to provide an overall understanding of the recorded data, the next experiments pool data from different recording sessions and images. In accordance to the analysis presented hitherto the pooled data will be split in three groups, i.e. low targets, high targets and 0-targets. The resulting sample sizes are 75 ratings in the high and low targets respectively, and 675 ratings in the 0-target group. Accordingly, the following analyses are executed with an α-level of 1 %. ($p < 0.01$). The results are shown in Fig. 4.6. They prove very similar to the results from the text-dependent data. Again, all targets are perceived as intended, and all target groups are significantly different from the 0-targets. The only exception is observed for **Low Neuroticism**. Here these ratings fall together with the group of 0-targets, i.e. the targeted **Low Neuroticism** performance is not perceived significantly lower the non-manipulated neuroticism stimuli. Although for text-dependent data this difference is significant, the absolute difference is small. As a first hypothesis, this could be attributed to the speakers unconscious character of naturally **Low Neuroticism** when acting on any other trait. Alternatively on hypothesis 2, the speaker's ability to consciously lower neuroticism in the voice might be limited. Hypothesis 3 is not directed to the speakers abilities and refers to the observed correlations between neuroticism and extroversion. Accordingly, the extroversion acting might have executed a noticeable impact on the neuroticism 0-target actings.

In order to indicate this kind of mutual impact Table 4.5 shows the secondary effect of primary manipulation in the data. Arrows represent significant influences ($p < 0.01$). Analysis of variance is executed as above. For instance, the perception of *C* is influenced only by a deliberately increased *N*, for which *C* decreases. When deliberately increasing *C* no secondary effect is observed. But when deliberately lowering *C*, a decrease of perceived *E*, *O*, and *A* is observed. Overall, *C* seems to be robust and relatively independent from other scales, while the *O* and *E* traits are most affected.

With respect to secondary effects exerted on neuroticism, the table shows significant change when increasing *E* or when manipulating *O*. For the former, the perceived score of neuroticism decreases. For the latter the relationship shows a reciprocal character, i.e. decreasing *O* increases *N* and increasing *O* decreases *N*. Interestingly, this finding is not true vice versa, i.e. when acting on *N* there is no

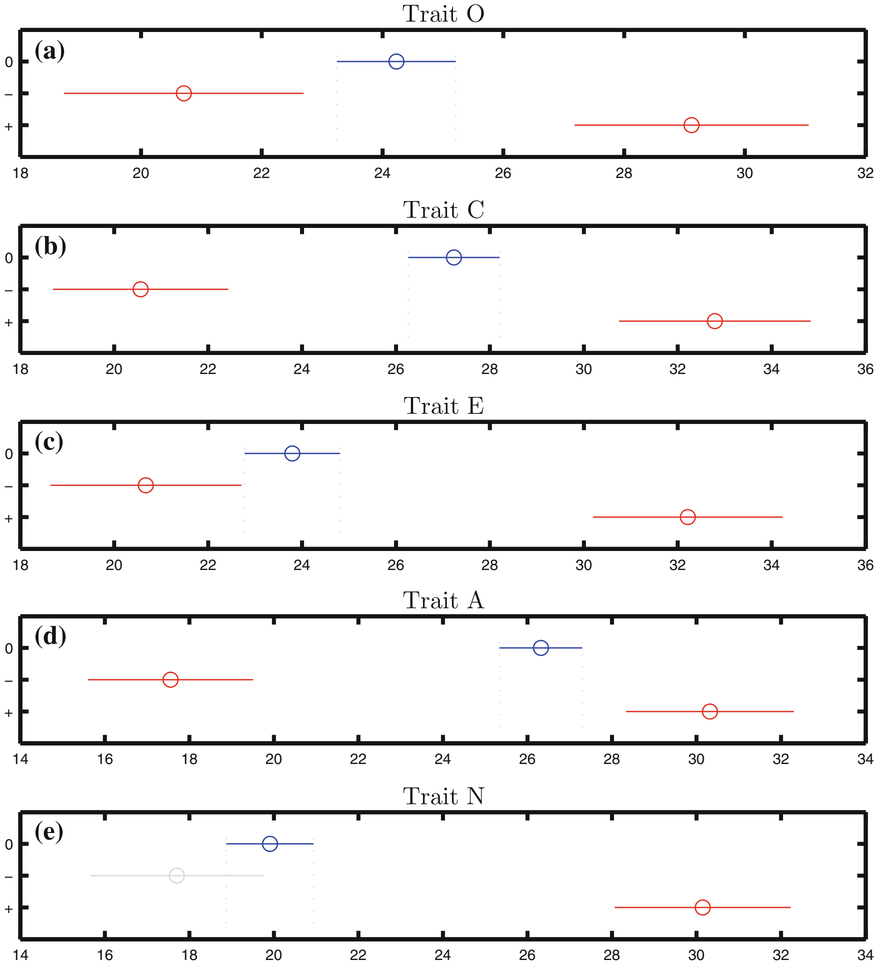

Fig. 4.6 Post Hoc tests, mean and confidence intervals for the text-independent recordings. **a** Trait openness. **b** Trait conscientiousness. **c** Trait extroversion. **d** Trait agreeableness. **e** Trait neuroticism

significant change in perception of *O*. Also, when deceasing *E* there is no significant effect on *N*, so hypothesis 3 cannot be confirmed from this perspective.

With regard to the aforementioned inverse correlation between *N* and *E* a more detailed image of dependencies can be drawn from Table 4.5. While deliberately increasing the perception of *N* and *E* shows the expected mutually reciprocal effect, actings deliberately decreasing the perceptions did not cause a significant effect. Similarly, the found correlation between *A* and *O* turns out to be effective when increasing or decreasing *A* primarily. When acting on *O* no significant effects could be found.

As there is no high and low target structure in the multi-speaker dataset, a comparison between any classes or groups cannot be drawn. Also when looking at Fig. 4.3, no class structure could be obtained from visual characteristics. Clustering experiments can be seen as a mean to partition the ratings along the scales. The effort to find an optimal cluster structure demands much effort and results in a research question different from the questions focused on in this work. Initial experiments separating the *low* form the *high* scores are executed in Sect. 5.4, but the remaining splits are not believed to be optimal for analyses of variance. Rather, they will be used to indicate a rank of individual features, as will be explained in the respective section. While clustering experiments remain future work, a deeper look into the expected problems and characteristics can be found in Chap. 7 when discussing the outlook of this work.

4.5 Summary

This chapter on human personality assessment starts with a presentation of an auditory analysis done by the single expert, i.e. the author. These information can be useful for the reader in order to understand and visualize the data. At the same time, it helps to estimate the scope of reasonable or comprehensible ratings as given by the author's expert opinion. Analyzing a multitude of labelers the following sections present detailed analyses of correlations, consistencies and significant differences in the data in terms of descriptive statistics. In Sect. 4.2 the distributions of the three recorded subsets are analyzed. Accordingly, the ratings of more than 92 % of all rated stimuli are of normal distributions. Hence, overall normal distribution is assumed for the text-dependent and text-independent data. Trait-wise analyses on the multi-speaker data shows highest non-normality shares with neuroticism and agreeableness ratings, both of which results in approximately 12 %.

Next, correlation analyses prove the found correlation in the recorded data to be comparable to correlation found with the NEO-FFI based on personality assessment in psychology. Also, the analysis reveals a consistent inversely directed link between neuroticism and extroversion as well as between neuroticism and conscientiousness in all the datasets. This correlation is of moderate magnitude. The more neurotic the speakers are perceived the less extroverted and the less diligent they are perceived as well. The analysis of consistencies in the ratings shows overall good or excellent results. With this respect, raters show least consistent results when assessing openness.

In Sect. 4.3 the latent structures of the recorded data are compared with the structure inherent to NEO-FFI application. A factor analysis hypothesizing 5 latent factors in the data revels very congruent structures. Note, that the NEO-FFI has been developed by factor analyses itself. Further it was generated to match the assessment of personality as applicable in psychology. With this respect the underlying question of whether or not the NEO-FFI can actually be applied to infer the same personality estimations when intending to assess personality perceptions of unknown persons from their voices only arises. As a result, only few items out of the NEO-FFI items fail to capture relevant information. At the same time, only few cross-loadings occur. Most importantly, the revealed factor structure results very similar to the NEO-FFI structure, which verifies the applicability of the NEO-FFI test scheme to the recorded data. As a secondary result, openness items seem to capture least information, which can be observed for all three databases.

The last section describes results from analyses of variance departing from different perspectives. First, the overall perception of the stimuli as desired when instructing the speakers could be verified. High targets are consistently perceived significantly higher than low targets on all traits. In between high and low targets on any trait a non-manipulated characteristic is assumed, which occurs when acting on other traits. Also these targets, which are referred to as *0-targets*, show significant differences to the high and low targets for the text-dependent and text-independent datasets. The structural design of the latter allows for insights into time- and text-dependency of the actings. Despite a very small number of outliers the analysis verifies overall time-independence as well as overall text-independence in the data. In other words, the assessments do not differ due to the spoken works or content, nor do they differ due to the time of the recordings, which were spread weeks apart.

Next, a more detailed analysis examines the impact of one deliberate trait acting on any other traits. While the results in general confirm the findings from the correlation analyses, a deeper insight shows that for the recorded speaker-dependent data the correlations between neuroticism and extroversion as well as the correlations between agreeableness and openness are verified by significant differences in analyses of variance for high targets mainly. When increasing these traits, the effect causes significant changes in the correlated trait as well, which is not consistently true for the case vice versa.

Eventually, the overall high consistencies, the observation of normal distributions in the ratings, the comparable correlation patterns in between the traits, the very congruent latent factor structure as well as the significant differences in between the target groups show that personality impressions can be generated and assessed by the NEO-FFI reproducibly when using speech input exclusively. The difference in recording condition in the multi-speaker data, i.e. when differentiating between close-talk and stand-alone microphone recordings, did not cause relevant changes in perception with respect to distribution, correlation or consistency in personality perception. Finally, results show, that for further processing overall text-independence as well as time-independence can be assumed for the text-independent data.

References

Bortz J (2005) Statistik für Human- und Sozialwissenschaftler, vol.57. Springer, Heidelberg, p 66

Costa PT, McCrae RR (1992) Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) manual. Psychological Assessment Resources

Cronbach LJ (1951) Coefficient alpha and the internal structure of tests. Psychometrika 16(3):297–334

Polzehl T, Möller S, Metze F (2010a) Automatically assessing acoustic manifestations of personality in speech. IEEE. In: Workshop on spoken language technology, Berkeley

Polzehl T, Möller S, Metze F (2010b) Automatically assessing personality from speech. In: Proceedings of international conference on semantic computing (ICSC 2010), IEEE pp 1–6

Polzehl T, Möller S, Metze F (2011) Modeling speaker personality using voice. In: Proceedings of the annual conference of the international speech communication association (Interspeech 2011), Florence, Italy, ISCA

Zinbarg R, Revelle W, Yovel I, Li W (2005) Cronbach's, Revelle's, and McDonald's: their relations with each other and two alternative conceptualizations of reliability. Psychometrika 70:123–133

Chapter 5
Automatic Personality Estimation

This chapter describes how personality cues can be estimated from speech using an automated system. In order to enable a machine to make a decision on perceived personality from speech, a comprehensive processing chain needs to be developed and applied. First, the recordings need to be preprocessed and a selection of promising audible cues needs to be extracted from the signal. At the same time, identifying different perceptual cues from speech is a complex task even for the human listeners. To date, there is no standardized or commonly acknowledged inventory of perceptual impressions that even linguists or speech experts could potentially resort to when describing speech impressions related to personality.

Moreover, as explained in Chap. 2, the few works outlined before that actually use automatic means apply very few and rather common prosodic features only. Potentially, more ideas on how to capture more information and new ideas on how to extract these information are needed. As a kind of boundary condition, there is a rather limited availability of known audio characteristics that can be extracted. These characteristics are sometimes also named *low-level descriptors*, especially when they are extracted in a continuous way. Essentially, the author tried to cover all major aspects of known acoustic and prosodic audio extraction while adding new ideas on how to draw meaningful features out of the descriptors. Note, there are certainly more features and more algorithms in order to extract audio characteristics proposed in the literature than the number of implemented features in this work. However, many of these features or algorithms have proven to be helpful in specific analyses or when applied to one or the other experiment only. Ultimately, the chosen selection can—according to the best knowledge of the author—be seen as representative and comprehensive for all major approaches with regard to acknowledged state-of-the-art speech systems of related fields like emotion recognition and automatic speech recognition.

In this order, the following sections explain the extraction process, the process of finding relevant segments in the signal,[1] the process of extracting acoustic descriptors

[1] Here, the term *relevant* refers to the search for active speech parts and speech signal segmentation thereafter.

T. Polzehl, *Personality in Speech*, T-Labs Series in Telecommunication Services,
DOI 10.1007/978-3-319-09516-5_5

from the segments, the process of capturing information from the descriptors by means of statistical analysis, the process of estimating goodness of these features in terms of ranking and evaluation, as well as the process of piping these features into a classifier or regression algorithm while searching for optimal algorithm parameters at the same time. This chapter concludes with the presentation of results from automatic personality modeling for both classification and regression tasks. A discussion of these findings along with a consideration of factors of influence can be found in Sect. 6.1.

This chapter presents the learning task with a perspective of a semi-automatic fully data-driven system. In particular, the system is able to automatically segment any incoming speech signal. It extracts a comprehensive feature set that can be understood as feature repertoire. These features also include features more robust to noise. A generic ranking process serves as the first step out of a two step feature selection strategy. This ranking is able to adapt to any new speech classification task autonomously and proposes the subset out of the feature repertoire at hand that should be used for modeling. The second selection step is done jointly with the modeling part. This final part of the processing chain is the actual classification or regression using incremental numbers of top-ranked features and still requires supervision with respect to basic algorithms setup. Details of each processing step will be given in the respective sections below.

5.1 Automatic Extraction of Personality Cues from Speech

The extracted audio descriptors that are included in the process of automatic personality classification can be sub-divided into 7 groups, which are *intensity, pitch, spectrals, loudness, MFCC, formants*, and *other*. All descriptors are extracted using 10 ms frame shift, an analysis window of 30 ms width. Gaussian window functions ensure a smooth continuous overlapping procedure. More information on windowing and the overlap procedure can be found in O'Shaughnessy (2000), Huang et al. (2001). Unless indicated within the following sub-sections, all audio descriptors are extracted using the *praat* analysis toolkit, cf. Boersma and Weenink (2009).

5.1.1 Intensity

Taken directly from the time domain intensity is an often used and fast to extract measure of amplitude magnitude. One has to be careful not to confuse this measure with perceptually motivated measures, such as the loudness of a signal. Klasmeyer (1999) has shown the difference in between these two measures, when she collected user ratings for two distinct stimuli. One stimulus was prototypical for an aspirated voice, i.e. strong spectral decrease towards the higher frequencies and few harmonic multiples of the pitch only. The other stimulus was of low spectral decrease and resembles a rather moderate voice. To make the effect even more obvious the maximal amplitude of the first stimulus was increased, so the resulting physical signal power exceeded the physical signal power of the second stimulus by factor of $+6.2$. Both stimuli were of the same pitch hight. As a result, despite the lower amplitude the second stimulus was consistently perceived louder by a number of test participants. Hence, the physical signal power should not be considered a perceptual measure. In terms of perceptual measure, the *loudness* features will be introduced in Sect. 5.1.4. For the present feature subset on signal power, the squared amplitude values were convert into dB scale relative to a sound pressure level (SPL) of 2×10^{-5} Pa, which resembles the auditory threshold of a just noticeable 1 kHz sine sound, cf. Eq. 5.1. To avoid any DC offset the mean amplitude was subtracted before calculation.

$$I = 20 \cdot \log_{10}\left(\frac{p_1}{p_0}\right) \text{dB} \tag{5.1}$$

with $p_0 = 2 \times 10^{-5}$ Pa

5.1.2 Pitch

Pitch is extracted by means of autocorrelation as described in Boersma and Weenink (2009). In general, the autocorrelation function (ACF) of a discrete speech signal s with length N can be interpreted as similarity of a signal extract $s(n)$ to the original signal. The ACF can be defined as:

$$ACF_s(k) = \sum_{n=0}^{N-1-k} s(n) \cdot s(n+k) \quad \text{for} \quad k = 0, \ldots, N-1 \tag{5.2}$$

Basically, the chosen extraction method uses a normalized autocorrelation function to estimate the amount of harmonicity-related signal components. The algorithm operates with a two-fold strategy. After setting the search range in which pitch is expected, a series of parallel pitch candidates are calculated from the so called *lag domain*. The lag domain represents the amount of periodicity with respect to a certain frequency, as obtained when shifting the signal extract over a certain longer signal window range. Here, pitch values in between 80 until 650 Hz are expected, which can be seen as expanded search range. The reason for setting this expansion is two-fold. In one way, acted datasets are often expected to include exaggerated speech in terms of speech gestures, which could potentially cause artificially high pitch movements. In another way, when recording spontaneous speech and explicitly not asking the speakers for well-formed and calm, official, neutral speaking style, even natural laughter and non-vocalic sounds like signs and cheer can naturally lead to very high pitch ranges.

In order to cope with local quasi-periodicity and small differences in consecutive pitch periods, the fixed window range was chosen to include at least three times the lowest expected pitch period. When iteratively moving and multiplying the window through the whole signal frame, the lag domain shows peaks at possible pitch candidates. In case of harmonic sounds, these peaks are arranged as integer multiples of each other. Candidates are cleaned out by a further threshold that is used to differentiate silence intervals from non-periodical as well as to differentiate non-periodical from quasi-periodical intervals by the absolute magnitude of the lag. After all frames of the whole signal have been processed, i.e. after stepping through the signal with a step size of the chosen frame shift, a path finding algorithm is applied to the space of the individual, i.e. local candidates. This step follows two aims at the same time: (a) the aim to minimize the number of large frequency jumps; and (b) the aim to minimize the scattering of consecutive voiced-unvoiced decision boundaries. In detail the thresholds for the path finder were empirically set as follows:

- Low silence threshold (0.1) to prevent the algorithm from loosing very dynamic segments
- Relative high voicing threshold (0.6) because expressive speech can be expected to also comprise sharp and high energy unvoiced and voiced fricatives which would be vulnerable to be confused with pitch otherwise
- Low octave cost (0.1) for the path finder that would otherwise punish higher pitch registers for an identical frame
- High octave-jump cost (0.85) to prevent the path from jumping too much in between frames, e.g. at octave confusions
- Low voiced/unvoiced cost (0.2) to obtain an overall smooth temporal cohesion for either consecutive voiced or unvoiced segments without smoothing out actual pauses and unvoiced sounds

After extracting the raw pitch and searching for the most favorable path of pitch candidates by the thresholds the following three post-processing steps are commenced:

1. In order to normalize for the absolute height of different speakers the pitch was converted into the semitone domain using the mean pitch as reference value for a whole stimulus, as shown in Eq. 5.3.

$$\triangle\, ST(f, f_{ref}) = \frac{12}{ln(2)} \cdot ln(\frac{f}{f_{ref}}) \qquad (5.3)$$

2. For any non-voiced segments, a piecewise cubic interpolation and smoothing by local regression using weighted linear least squares is applied.
3. The obtained continuous contour is smoothed by local regression using weighted linear least squares and a 1st degree polynomial model. In addition, the method assigns lower weight to outliers and even zero weight to data outside six mean absolute deviations. This is to increase robustness against outliers. The span of regression was set to 7 points, hence the filter locally smooths with 30 ms context inclusion to each side in a sliding window.

While many different ways to interpolate and smooth contours like the pitch are proposed in the literature, the chosen strategy empirically proved to provide smoothed transitions while also allowing for more abrupt changes at a point in time.

5.1.3 Spectrals

In order to calculate spectral descriptors a *Fast-Fourier-Transformation (FFT)* with linear frequency resolution was applied. The applied resolution of 43 Hz accounts for a rather narrow-band resolution. The FFT is a variation of the Discrete Fourier Transform (DFT) which reduces the computational effort from $O(2 \cdot N^2)$ to $O(2 \cdot N \cdot ld(N))$. Let the complex DFT $S_t(k)$ with a discrete frequency k of the signal frame $s_t(n)$ with $0 \leq n \leq N-1$ at the frame t be:

$$\mathrm{DFT}\{s_t(n)\} = S_t(k) = \sum_{n=0}^{N-1} s_t(n) \cdot e^{-\frac{2 \cdot \pi * j}{N} f \cdot n} \tag{5.4}$$

The desired spectrum $S_t(k)$ ranges from frequency $f = 0$ until the Nyquist frequency $f = f_{nyq}$, which resembles half the sampling frequency of the stimulus. The discrete frequencies in the spectrum are distributed by $\Delta f = f_s/N$ with f_s being the sample frequency. In general the original frequency f is of $k \cdot f_s/N$ Hertz and the lowest discrete frequency k is set to 0. For speech applications, it is common to disregard frequencies above 8 kHz. For actual calculation of the FFT the most commonly used Cooley-Tukey algorithm[2] was applied. The main idea behind this algorithm is to divide and conquer. Applying recursive algorithms the calculation brakes down from any composite size to a number of smaller DFTs. A more detailed explanation of the algorithms is out of scope for the presented work but can be found in Cooley and Tukey (1965).

Descriptors are then drawn directly from the unweighted power spectral density $PSD = |S_t(k)|^2$. Calculated descriptors are:

1. The center of spectral mass gravity, also known as spectral *centroid*, as shown in Eq. 5.5
$$Centroid = \frac{\sum_{n=0}^{N-1} PSD(n) \cdot x(n)}{\sum_{n=0}^{N-1} x(n)} \tag{5.5}$$
2. The magnitude of spectral change over time, also known as spectral *flux* or *flow*, as shown in Eq. 5.6
$$S_{flux,t} = \|PSD_t - PSD_{t-1}\| \quad \text{mit} \quad t = 1, \ldots, T-1 \tag{5.6}$$
3. The 95 % *roll-off point* of spectral energy under the spectral slice

The first and third descriptors capture aspects related to the spectral slope, which is also called the spectral tilt, and correspond to perceptual impression of sharpness

[2] The algorithm known by this name was popularized by a publication of Cooley and Tukey (1965) but it was actually known to Carl Friedrich Gauss around 1805 already. Other algorithms exist, e.g. Prime-factor FFT algorithm, Bruun's FFT algorithm, Rader's FFT algorithm, or Bluestein's FFT algorithm.

of sounds, cf. Fastl and Zwicker (2005). The higher these points, the sharper the perception of the sounds. The second descriptor captures the smoothness of spectral transition. The more abruptly changes in the spectrum occur the higher the magnitude of this descriptor.

5.1.4 Loudness

Loudness is calculated as perceptively motivated psychoacoustic measurement as defined by Fastl and Zwicker (2005). This measurement operates on a Bark-filtered version of the spectrum, which can be obtained by applying Eq. 5.7 for discrete signals. The basic idea is to subdivide the bandwidth into critical sub-bands *(z)* that correspond to meaningful bins when compared to human hearing processes. The resulting bandwidth of the filters equals 1 Bark. Finally, the filter coefficients are integrated into a single loudness value per frame by summation.

$$Bark(z) = 13 \cdot \arctan(0.67 \cdot f) + 3.5 \arctan(\frac{f}{7.5})^2 \tag{5.7}$$

Here, f is frequency in kHz, and z is a defined Bark filter $1, \ldots, 24$. The filter 24 reaches an upper limit of approximately 16 kHz.

5.1.5 MFCC

The abbreviation *MFCC* corresponds to *Mel-Frequency-Cepstral-Coefficients*. The *Mel* scale can be used to transform the linear frequency scale into a perceptually corrected scale representing the perceived hight of tones better than in Hertz units. Originally, the Mel scale was introduced using a reference point defined by assigning a perceptual pitch of 1,000 Mel to a 1,000 Hz tone, 40 dB above the auditory threshold. Equation 5.8 shows the transformation.

$$Mel[f] = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \tag{5.8}$$

To obtain MFCCs, first the signal needs to be transformed into the spectral domain using a FFT as described in Sect. 5.1.3. Next, the scale is transformed from Hertz to Mel. Finally, a discrete cosine transformation (DCT), cf. Eq. 5.9, gives the values of the Mel frequency cepstral coefficients (MFCC).

$$y(k) = w(k) \cdot \sum_{n=1}^{N} x(n) \cdot \cos\left(\tfrac{\pi(2n-1)(k-1)}{2N}\right) \quad \text{with} \quad k = 1, 2, \ldots, N$$

where

$$w(k) = \begin{cases} \frac{1}{\sqrt{N}} & \text{for} \quad k = 1 \\ \sqrt{\frac{2}{N}} & \text{for} \quad 2 \leq k \leq N \end{cases} \tag{5.9}$$

16 Coefficients are calculated and the zero coefficient is appended. In general, the DCT can be seen as spectral decomposition operating on the real part of the FFT. The higher the number of the coefficient, the higher the respective frequency in the spectrum. Likewise, the higher the magnitude of this coefficient, the higher the energy that this frequency is observed with in the spectrum. The term *cepstral* accommodates the fact, that analyzing the frequency spectrum of the spectrum itself, the target domain cannot be a frequency scale again. But the correspondence of the scale targets to a frequency-related domain. Hence the word “spec” was simply reversed and introduced so the domain after applying the DCT is called “ceps”. Although MFCCs are most commonly used in speech recognition tasks they often give good performance in emotion recognition as well, as reported by, e.g., Nobuo and Yasunari (2007), Polzehl et al. (2011). Eventually, MFCCs can be interpreted as a perceptually weighted version of a heavily compressed spectrum.

5.1.6 Formants

Formants are maxima in the spectrum which can be defined by a center frequency and a respective bandwidth. Their position in the spectrum is principally independent from the pitch as discussed above. In case of a low first formant on the one hand, which can be expected in between 500–1,500 Hz, and a high pitch on the other hand, which can be expected between 80–700 Hz, these measures are oftentimes mixed up. Formants are caused by resonances in the vocal tract and are most strongly influenced by the position and movements of the articulators such as throat, mouth, lips, tongue. They show a relatively clear pattern when voiced sounds are articulated. Also for voiced fricatives or plosives, the formants or formant transitions have shown consistent patterns. The first two formants are commonly seen as the most important, as experiments have shown that they are sufficient to perceptually differentiate between vowels. Also the third formant is oftentimes seen as contributing to the basic character while higher formants are seen as contribute to the timbre of the sound, rather than to substantial character.

To counteract the natural absorption of higher formants, a pre-emphasis of 6 dB/octave is applied as shown in Eq. 5.10.

$$s^*[n] = s[n] - \alpha_{pre} \cdot s[n-1] \tag{5.10}$$

When producing voiced sounds, the absorption within nose and mouth regions decreases higher formant frequencies much more than it decreases lower formant frequencies. The pre-emphasis increases the higher frequencies against that trend.

For the present experiments the first 5 center frequencies and respective bandwidths are extracted. Looking for formants within the frequency range of up to approximately 8 kHz, a theoretical number of up to 8 formants could be expected in that range. However, initial screening experiments showed an increased failure and confusion pattern when trying to extract more than 5 formants. Examining closer the segments where extraction fails, one possible reason could be stated. In traditional linguistics formants are almost exclusively extracted from vowels, i.e. voiced segments. In this work, segment groups are determined automatically from the speech signal, in more detail from the magnitude in the lag domain. At the same time, the amount of friction in these automatically detected segments differentiates between pure vowels and voiced consonants. The implemented logic is fuzzy. However, a phonetic transcription lies beyond the scope of this work. Eventually, the number of extracted formants is decreased to an empirically determined number of 5 formants.

In order to calculate coefficients *linear prediction coding (LPC)* is applied. The basic idea behind the technique is to predict current signal values from previous ones. When regarding the error between predicted and observed values, the problem can be formulated as recursive filter in the complex domain. The filter coefficients are then examined for local maxima by calculating zero-poles in the complex plane. These poles are interpreted as functions of peaks and respective bandwidths accordingly. Further details about the rather complex calculation cannot be presented here. They can be found in, e.g., Press et al. (1992), who describe and extend the original algorithm proposed by Burg.

5.1.7 Other Descriptors

Referred to as *other* descriptors the *Harmonics-to-Noise Ratio (HNR)* is calculated. Similar to pitch candidate estimation as explained in Sect. 5.1.2, this measurement is taken from the autocorrelation domain. In general, the HNR estimates the amount of harmonicity from the lag domain by means of periodicity detection. Accordingly, this measurement is calculated for voiced segments only. Finally, the *Zero-Crossing-Rate (ZCR)*, extracted very early in the process chain, is treated as descriptor and is normalized by the frame length. More details on the Zero-Crossing descriptor will be explained in Sect. 5.2 when outlining the automatic segmentation of the speech signal.

5.2 Voice Activity Detection and Segmentation

Already after having extracted intensity and pitch contours the voice activity detection (VAD) is applied. Extending Rabiner and Sambur (1975), who have proposed an algorithm for isolated word recognition, the present VAD also detects intermediate pauses. Intensity based double-threshold detection is applied to estimate the part of speech that is highly likely to contain active speech in the first place. The thresholds are therefore chosen conservatively. As a result a tentative global start point and a tentative global end point are set. Looking at the Zero-Crossing-Rate (ZCR) within this area and 400 ms before the start point and after the end-point, a search for unvoiced sounds containing much friction, e.g. fricatives or stops, is commenced. Such sounds are characterized by relatively high zero-crossing rates. If found, the global boundaries are extended respectively. All consecutively extracted contours explained above are then taken from the active speech part only, which saves a considerable amount of computational effort.

After having extracted the audio descriptors, and having found global start points and global end points for active speech the segmentation within this active part is done. Taking into account the obtained pitch contour, the intensity, and the ZCR, all indicators of voice, power and friction are available to logically differentiate between voiced, unvoiced and silence sounds in a fuzzy way. In order to be able to look at individual strength and weakness of certain descriptors, and to account for the needed presence or absence of voice when extracting certain descriptors, all features are calculated from the following three groups individually: (a) voiced sounds only; (b) unvoiced sounds only; and (c) the whole utterance at hand.

5.3 Feature Definition

After segmenting the audio signal and extracting audio descriptors from the segments, individual features are calculated as explained in the following subsections. In general, a multitude of statistics can be applied to obtain features. These statistics mostly comprise means, moments of first to fourth order, extrema and ranges. To also include statistics that have shown to behave more robust with respect to noise in the data, also medians and measurements from the distribution of feature characteristics are included. The features are now presented in the same order as the audio extracts above. All in all, 1,359 features are calculated. Table 5.1 shows the different audio descriptors and the number of features calculated from them.

Table 5.1 Overview and number of features in signal-based feature groups

Feature group	Number of features
Intensity	171
Pitch	143
Spectrals	75
Loudness	171
MFCC	612
Formants	180
Others	7
Sum	1,359

5.3.1 Intensity

From the following statistics, 171 intensity features are derived:

mean This feature is the first order moment calculated as arithmetic average for an estimation of overall intensity. Pauses are included.

median This feature can be seen as a more noise-robust measure of the mean in case of very corrupted contours. Pauses are included.

max This feature gives the global maximum. It will always correspond to one concrete maximal occurrence, regardless of context.

std This feature is the second order moment, i.e. the standard deviation. Pauses are included.

iqr This feature is the interquartile range, which is the difference between the 75th and 25th percentiles. In other words, the range without the upper and lower quartiles of the distribution histogram. This is motivated because these quartiles are expected to hold outliers if they occur.

skewness This feature is the third order moment drawn from the distribution histogram. A zero skewness indicates that values are evenly distributed on both sides of the mean. A negative skew indicates that the tail on the left side of the probability density function is longer than the tail on the right side. Hence, the bulk of the values (possibly including the median) lies to the right of the mean. A positive skew indicates vice versa.

kurtosis This feature is the fourth order moment drawn from the distribution histogram, also sometimes called the *peakedness*. At the same time, it also has the title *heaviness of distribution tail*, since the tails and the peak are correlated to each other. It is seen as a further measure of how outlier-proof a distribution is. Here, the kurtosis of the normal distribution is 3. Higher values state a more outlier-sensitive distribution, which means the observed variance is to a greater part the result of outliers. Lower values indicate an opposite behavior.

lin.reg This feature calculates a linear regression fit according to Du Mouchel and O'Brien (1989). The main difference to an ordinary least-square regression algorithm is that this algorithm uses iteratively reweighted least-squares with a special weighting function. In their paper the authors explain and show increased robustness towards outliers. Many different weighting functions have been proposed in the literature, such as *Andrew, Cauchy, Fair, Huber, Welch* as well as combination of these proposals. For applying a one-dimensional regression the bi-square weighting function according to Du Mouchel and O'Brien (1989) showed most consistent results for different audio descriptor input data with regard to residual observation. This feature carries the slope coefficient.

The smaller this feature the flatter the contour when seen in terms of linear fit.

err. lin.reg This feature carries the standard error of the coefficient from linear regression analysis *lin.reg*. Subsidiary to the previous feature, the lower this feature the better the linear fit above. This feature can also be interpreted as indication for contour flatness with respect to just one direction, since flat contours are expected to exhibit small error coefficients. Note, this could also be a monotonously falling slope. In opposition to the previous feature the coefficient can also signal a non-flatness in case of high error coefficients.

DCT coeff. 1–30 Following Eq. 5.9, DCT coefficients 1 until 30 are calculated. Here, high magnitude of lower DCT coefficients corresponds to strong elements of slow moving contours, while high magnitude in higher coefficients corresponds to fast moving contours. The coefficients are normalized to account for frequencies of 1–30 Hz. This range was chosen because normal speaking rates occur between zero (in case of speech pause) and up to roughly 10 sounds per second on average. Given that in short conversational phrase and in affective speech this range can be expected to increase drastically and instantaneously, for example when speaking extremely fast for only half a second the upper bound of 30 was chosen.

v/uv After calculation of all the above mentioned features the same features are calculated on basis of voiced (*v*) and unvoiced (*uv*) speech frames exclusively and respectively. These times all the features exclude speech pauses.

D(Δ), DD($\Delta\Delta$) In order to capture dynamics of the contour the first and second order derivatives are calculated. The statistics drawn from the first derivative tell about the gradient or slope of the contour at a certain point in time. The second derivative signals the increase or decrease of the gradient, i.e. the acceleration at a certain point in time. All above mentioned features are calculated on basis of the first order and second order derivative respectively.

ratio mean uv/v This feature is inspired by phonetic analysis, which also includes the annotation of sound strength or forcefulness in close annotation procedures. Here, the ratio between vowels and adjacent consonants is sometimes determined in order to obtain an estimation of force independently from signal power level and intensity level. Since a phonetic transcription, be it manual or automatic, is not in the focus and in any case out of scope for this work, the ratio between the mean of all unvoiced sounds and the mean of all voiced sounds is appended as a related measure.

ratio median uv/v Analogously to *mean uv/v* this measure is believed to behave more robustly when encountering very noisy data.

ratio max uv/v This measure is designed to *not* level out observations over time. It captures the one occurrence of maximum unvoiced intensity and the one occurrence of maximum voiced intensity and calculates the ratio of both.

5.3.2 Pitch

For calculating the derivations and for smoothing the contour the pitch is interpolated and smoothed as described in Sect. 5.1.2. Unless indicated with the individual features, all interpolated points were excluded when drawing the statistics. Eventually, 143 pitch features are generated, as presented and explained in the following paragraph:

mean This feature is the first order moment calculated as arithmetic average.

median This feature can be seen as a more noise-robust measure of the mean.

max This feature gives the global maximum. It will always correspond to one concrete maximal occurrence, regardless of context.

min This feature gives the global minimum. It will always correspond to one concrete minimal occurrence, regardless of context.

range This feature is calculated as global maximum minus global minimum.

std This feature is the second order moment, i.e. the standard deviation. Pauses are included.

iqr This feature is the interquartile range, i.e. the difference between the 75th and 25th percentiles. See *intensity* features above for more explanations.

skewness This feature is the third order moment drawn from the distribution histogram. See *intensity* features above for more explanations.

kurtosis This feature is the fourth order moment drawn from the distribution histogram, also sometimes called the *peakedness*. See *intensity* features above for more explanations.

lin.reg This feature calculates a linear regression fit according to Du Mouchel and O'Brien (1989). See *intensity* features above for more explanations. Pauses are excluded.

err. lin.reg This feature carries the standard error of the coefficient from linear regression analysis *lin.reg*. See *intensity* features above for more explanations. Pauses are excluded.

lin.reg (w. pauses) This feature calculates a linear regression fit according to Du Mouchel and O'Brien (1989) keeping the original temporal structure, i.e. including unvoiced segments and pauses in the regression fit time line. Because long pauses can strongly influence the regression fit and might potentially cause effects on the perception of pitch contour as well, this feature keeps the "gaps" and pauses in the pitch contour.

err. lin.reg (w. pauses) This feature carries the standard error of the coefficient from linear regression analysis *lin.reg (w. pauses)*.

mean pos. slope This feature calculates the arithmetic mean of all positive slopes, i.e. rising pitch points. The feature is not defined for contours that do not show any positive slope.

mean neg. slope This feature calculates the arithmetic mean of all negative slopes, i.e. falling pitch points. The feature is not defined for contours that do not show any negative slope.

median pos. slope This feature calculates the median of all positive slopes, i.e. rising pitch points. The feature is not defined for contours that do not show any positive slope.

median neg. slope This feature calculates the median of all negative slopes, i.e. falling pitch points. The feature is not defined for contours that do not show any negative slope.

DCT coeff. 1–30 Analogously to DCT features from *intensity*, coefficients 1 until 30, are calculated. Calculations are done on basis of the interpolated and smoothed signal. This feature is also expected to relate to the auditive measure named *jitter* or *pitch perturbation*. While the jitter is defined as the difference in pitch frequency from any one pitch period to the adjacent period, the proposed method here estimates the coefficients on basis of all voiced points in general. See *intensity* features above for more explanations.

D(Δ), DD($\Delta\Delta$) In order to capture dynamics of the contour the first and second order derivatives are calculated. The statistics drawn from the first derivative tell about the gradient or slope of the contour at a certain point in time. The second derivative signals the increase or decrease of the gradient, i.e. the acceleration at a certain point in time. All above mentioned features are additionally calculated on basis of the first order and second order derivative respectively.

absolute mean This feature is inspired by the potential use of the so called *vocal register*. The term *vocal register* can be defined as a particular series of tones in the human voice that are produced by one particular vibratory pattern and therefore possess a common quality Large (1972), e.g. modal voice, vocal fry, falsetto, or whistle. Normally, the absolute hight of a speaker should not determine the speakers personality perception per se. It is assumed, that the core information is rather expressed by movement patterns. However, if a speaker changes registers, that could also have an effect on the perceived personality impression. Since all other pitch features are transformed into semitone scale values with a reference of the mean of the segment itself as explained in Eq. 5.3, the information on absolute hight of the pitch would be lost. This feature therefore captures this information on basis of the interpolated and smoothed contour. A dedicated feature indicating discrete registers, however, is out of scope for this work and remains as future work.

detrended interp. std To calculate this feature the linear fit as calculated by the feature *lin.reg* is subtracted from the pitch contour. Next the standard deviation is drawn from it. The reason for this procedure is that the *std* is strongly effected by general increasing or decreasing slopes. Removing the overall trend (with respect to as much a linear fit can effectively remove), the resulting standard deviation is believed to capture more of the local movements.

5.3.3 Spectrals

From the spectral descriptors the following 75 features are derived:

mean This feature is the first order moment calculated as arithmetic average for an estimation of overall contour.

max This feature gives the global maximum. It will always correspond to one concrete maximal occurrence, regardless of context.

min This feature gives the global minimum. It will always correspond to one concrete minimal occurrence, regardless of context.

range This feature is calculated as global maximum minus global minimum.

std This feature is the second order moment, i.e. the standard deviation of the overall contour.

v/uv After calculation of all the above mentioned features the same features are calculated on basis of voiced (*v*) and unvoiced (*uv*) speech frames exclusively and respectively.

D(Δ), DD($\Delta\Delta$) In order to capture dynamics the first and second order derivatives are calculated. See *intensity* features above for more explanations.

5.3.4 Loudness

The loudness features use the same feature set as the intensity features, which results in 171 features. See *intensity* features above for more explanations.

5.3.5 MFCC

The MFCC features use the same feature set for each individual coefficient (out of 16 coefficients plus the zero coefficient) as described for the *Spectral* features except from the *range* statistics. Finally, this results in 612 individual features. See *Spectrals* features above for more explanations on the statistics.

5.3.6 Formants

The formant features use the same feature set for each individual formant (out of 5 formants extracted) and for each individual bandwidth (out of 5 bandwidths estimated) as described for the *Spectral* features. However, since formants are only defined for voiced sounds they are calculated on basis of voiced sounds only, resulting in 180 individual features. See *Spectrals* features above for more explanations on the statistic feature definition.

5.3.7 Other Features

The group of *other* features comprises 7 individual features. From the *HNR* descriptor the mean in terms of arithmetic average, the second moment, i.e. the standard deviation, and the maximum value are defined as individual features. Next, a single coefficient for the correlation between pitch and intensity is added as an individual feature. This is motivated by the fact that in most cases the pitch is expected to rise when increasing the intensity. This phenomenon is also known as the *Lombard effect*. But that could also hypothetically mean that a voice might cause unnatural perceptions when these measures do not correlate. This experimental feature therefore uses Spearman's correlation and calculates the correlation coefficient magnitude. In order to capture aspects related to rhythmic behavior the following statistics are calculated:

tot. duration This feature tells the overall speech duration including pauses after Voice Activity Detection (VAD) as explained in Sect. 5.2. This is inspired by the hypothesis, that in specific situations the very expression of being either short, i.e. brief, or being outspoken could be of interest.

pcnt voiced points Inspired by the auditory observation that slurred or mumbled speech seems to have a lower frequency of pauses or less prominent pauses, cf. Sect. 4.1, this feature experimentally expresses the ratio of voiced points to the overall contour after VAD in percent.

speech-to-pause This feature calculates the ratio of mean duration of speech parts (voiced and unvoiced sounds) to the mean pause duration. This is inspired by the hypothesis, that the insertion rate of pauses might be deliberately set in order to cause a certain perception, cf. Sect. 4.1.

5.4 Feature Selection

After having defined the number of features as explained in Sect. 5.3 the question of how much explanatory power each individual feature carries arises. As explained before, some of the features are designed purposely to be of experimental character. Others resemble more noise-robust versions of concurrent features. Also, the exploration of different segments that the features account for can lead to redundant information. For example, when very few pauses and few unvoiced segments occur in the signal, the calculation of features on basis of voiced segments might become similar to the calculation of features from the whole utterance. Therefore, a feature selection is the recommended next step in line.

To perform a reasonable selection a two-fold approach is applied in this work.

1. Step: bring the features in a ranked order according to their information contribution
2. Step: determine how many of the top-ranked features are needed to obtain optimal overall system performance

Amongst other ranking methods, a filter-based ranking scheme as proposed here can be applied in a fast way. While there is no general statement on the advantage or disadvantage of any method over other methods, the actual best method strongly depends on the data and problems at hand. In particular, the *Information-Gain-Ratio (IGR)* Duda et al. (2000) filter is chosen for the present ranking, because it has proven to provide good results in similar tasks, cf. Metze et al. (2009), Polzehl et al. (2011). The chosen method evaluates the gain of information that a single feature contributes. In addition to the good result from prior own works, also Kotsiantis and Kanellopoulos (2006) conclude that the chosen method is of overall best quality when comparing multiple methods for a number of standard datasets taken from the *Machine Learning Repository* from the University of Irvine.[3]

Since the term *information* is congruent to the term *uncertainty* or *purity* in decision theory for the present context, the basic idea of IGR is to estimate a contribution of a single feature in adding up to an average amount of information or alternatively the reduction of uncertainty a single feature can contribute with respect to an average uncertainty of a sample population. The underlying concept of *information* is based on the *Shannon Entropy* H, cf. Shannon (1948).

[3] www.archive.ics.uci.edu/ml/.

Let there be a class distribution $P(p_1, \ldots, p_K)$ of a finite set of samples containing K classes. H is then measured in bit units and defined as follows:

$$H = -\sum_{i=1}^{K} p_i \cdot log_2(p_i) \tag{5.11}$$

Now let Ψ be the totality of all samples and $\Psi_i \in \Psi$ the subset of elements that belongs to class index i. The average information needed in order to classify a sample out of Ψ into a class $i_1 \ldots i_K$ is given by

$$H(\Psi) = -\sum_{i=1}^{K} p_i \cdot log_2(p_i) \qquad with \qquad p_i = \frac{|\Psi_i|}{|\Psi|} \tag{5.12}$$

To estimate the contribution of a single feature its unique values are considered. Given a non-discrete distribution the feature span needs to be partitioned into bins, i.e. discretization has to be executed. Let $\Psi_{x,j}$ with $j = 1 \ldots J$ bins be the partition blocks of Ψ_x, holding values of a single feature x, the amount of information contributed by this feature is given by

$$H(\Psi|x) = \sum_{j=1}^{J} \frac{|\Psi_{x,j}|}{|\Psi|} \cdot H(\Psi_{x,j}) \tag{5.13}$$

The Information Gain (IG) of a feature is then given as its contribution to reach the average needed information or alternatively the reduction in uncertainty of the subset when subtracting the uncertainty proportion of the feature at hand.

$$IG(\Psi, x) = H(\Psi) - H(\Psi|x) \tag{5.14}$$

The Information-Gain-Ratio *(IGR)* accounts for the fact that IG is biased towards features with high number of individual values in their span. IGR normalizes IG by the scalar amount of total information that can be drawn out of J splits. This information is known as the *split-information* or *intrinsic information*. It does not account for any class membership but accumulates the amount of information that can be drawn by splitting Ψ using the splits of the feature x at hand.

$$IGR(x, \Psi) = \frac{IG(\Psi, x)}{H(\frac{|\Psi_{x,1}|}{|\Psi|}, \ldots, \frac{|\Psi_{x,J}|}{|\Psi|})} \tag{5.15}$$

When computing the ranking for non-discrete features, supervised[4] discretization following the entropy minimization heuristic requirement is performed in order to

[4] The term *supervised* refers to a machine learning class of methods where meaningful meta-information in terms of labels or annotation is present. In opposition, *unsupervised* methods cannot resort to such information and may introduce additional metrics to generate meta-data independently.

obtain partition blocks. In the chosen method, the number of blocks is determined by the algorithm using recurring partition and a *Minimum Description Length (MDL)* stopping criterion. In principle, this criterion sets an upper bound to the number of splits at hypothesized cut points by regarding the minimum number of bits required to specify the observed feature distribution. More details are out of scope for the present work and can be found in Fayyad and Irani (1992) in terms of the cutting point detection and Fayyad and Irani (1993) for a detailed explanation and empirical study on the stopping criterion. In order to estimate the amount of information inherent in a feature, the algorithm tries to minimize the resulting class entropy after splitting the feature span into a number of individual bins at the hypothesized cut points.

In case of no class information available, as is the case when considering the continuous NEO-FFI ratings rather than the induced class memberships, discretization needs to be carried out in an unsupervised manner. The labels space needs partitioning in this case in order to be able to calculate the probabilities in Eq. 5.12. Most popular unsupervised discretization strategies divide the span into equal frequency bins or equidistant bins, amongst others. Beneficial parameters for these unsupervised discretization methods need to be determined empirically.

Note that in this work the actual IGR magnitude is used to impose a lower threshold of IGR magnitude. Values of less than 0.001 bit on average were disregarded and set to zero after seeing all folds. Increasing or decreasing this threshold would lead to an decrease or increase of ranked features for subsequent modeling steps. Thus, features which show *negligible* information contribution in terms of IGR can be controlled and filtered out. For analyzing whether this threshold needs adjustment all plots on classification and regression results will also show a number of negligible features unless the actual algorithm performance is of higher information for the number of features to include in modeling.

Many other strategies in feature selection exist. The chosen method is based on the information in terms of entropy. Other filter-based methods base on ratios of inter-class and intra-class distributions or local distances. All of these filter-based methods share one characteristic, i.e. they evaluate individual features by certain criteria other than the subsequent classification success. Wrapper-based methods follow a different strategy and include the actual classification success into the evaluation and actual selection, which also increases the complexity drastically.

Yet other approaches seek selection strategies to compile subsets of features. Typically these approaches use different inclusion and exclusion steps to incrementally admit and suspend feature candidates into the eventual feature set. These methods drastically increase the number of evaluations required. Most economical sub-set selection methods use floating windows and a combination of forward- and backward search for their inclusion and exclusion strategies. Very frequently filters are still applied to rank the feature set before starting the selection process. Although these methods obviously apply a more systematic and comprehensive strategy, they show high vulnerability to over-fitting in reality quite frequently. Oftentimes, this leads to increased results on the training sets while incurring a loss when applied to unknown data.

Further, the choice of selection strategies for wrappers also strongly depends on the chosen classification algorithm in the loop. Certain requirements imposed by the classifier may as well lead to drastic decrease of available strategies, e.g. when applying generative classification methods like Bayes classifiers or Gaussian Mixture Models. As the latter depends on a high degree of independence in the feature space, other transformations like principal component analysis have to be executed. The chosen classification algorithms applied in the presented experiments are predominantly insensitive to correlation between features, which will be explained in more detail in Sect. 5.6.

After all, the chosen selection strategy comprises a fast filter followed by a single incremental forward pass without inclusion or exclusion look-around steps. Features are admitted in to the final feature set by starting with the feature occupying the highest rank and incrementally adding the subsequent ranks until the ranking offers negligible features only. The aim is to find the global maximum in the performance in terms of classification success. This way, the found maximum accounts for the global maximum.

Finally, to obtain a robust and independent estimation of the ranking a 10-fold cross validation as described in Sect. 5.7.2 is performed. All results listed below present the average ranks, i.e. the arithmetic average of the ranks in the individual folds, and the standard deviation around this average, in ascending order.

Results from the rankings will be displayed in the following section. For an analysis of classification success given a certain ranking the actual term *success* as well as the classification and regression algorithms must be introduced first. This will be given in Sects. 5.6.1, 5.6.2 and 5.7 respectively.

5.5 Normalization

Normalization is typically done for the following reasons: (1) in order to compare observations from different distributions; (2) because some classification algorithms may depend on equal scaling in the feature space. The classification algorithm selected in this work has proven to be sensitive to normalization, cf. Chang and Lin (2011), Herbrich and Graepel (2001). In general, there are two ways of applying normalization with Support Vector Machines, i.e. normalize the feature space or normalize the kernel itself. At this point it can only be said that in the present work the feature space is normalized. Details about the kernel will be introduced later in Sect. 5.6. In more detail, a *z-normalization* as given in Eq. 5.16 is applied to the features. This transform is also known as *z-scores, normal scores* or *standardized variables*, cf. Duda et al. (2000). The transformed features have a zero mean and a unit standard deviation.

Let $\mu(x)$ be the mean of a feature population and $\sigma(x)$ be its standard deviation. A standardized feature $\widetilde{x}$ is then given by

$$\widetilde{x} = \frac{x - \mu(x)}{\sigma(x)} \tag{5.16}$$

The resulting distance is also called *Mahalanobis distance* and measures the distance in between two data points on x in units of standard deviations.

Note, there are also normalization or transformation steps acting as normalization when extracting the audio descriptors. For example, pitch is converted in semitones relative to the turn's pitch mean value. This implies a normalization of the different heights caused by different normal pitch of speakers' voices. Also intensity is normalized by the auditory threshold in dB. MFCCs represent the cosine filter coefficients in a Mel domain, which uses a reference point between Mel and Hertz as defined by assigning a perceptual pitch of 1,000 Mel to a 1,000 Hz tone, 40 dB above the auditory threshold.

5.6 Modeling for Personality Classification and Trait Score Prediction using SVM

By definition, the task of a classification algorithm is to assign one class, out of a number of classes in the training data, to a sample. Here, the algorithm seeks to determine exactly one best personality class out of 10 possible personality classes. When presenting a stimuli, e.g. **High Openness**, the algorithm will find the *best match*, which can include that although the algorithms detects much similarity to the class pattern of **High Openness**, the class decision might still be **High Extroversion**, if the pattern for **High Extroversion** is a better match than the pattern for **High Openness**. A more detailed description of the term *best match* will be given in the next section. As mentioned before, classification is only applicable when class labels are present. At least two classes must be present, giving examples of two alternate characteristics. Eventually, the chosen set-up of a classification task evaluates whether or not it is feasible to automatically differentiate one distinct personality class from a finite set of alternative classes. Also the question of which classes can be differentiate form other classes best can be answered by looking at class-wise statistics.

On the other hand the database also provides the continuous ratings of the labelers for all the data. Another task at hand is to predict the labeler's scores in terms of the actual trait score value. Since there is no integrated score for all the five personality traits, each trait needs to be predicted individually. This task can be incomparably different and is very often also considerately more difficult. Classifiers utilize class membership information that in regression does not exist. Frequently, classes are designed to comprise alternate characteristics. In regression there is no such supervision, the actual target space opens up a whole range of predictable values for both potential predictions and miss-predictions.

For both tasks, the classification and the regression task, *Support Vector Machines (SVM)*, as initially introduced by Vapnik and Cortes (1995), are applied. This choice is motivated by following advantages:

- SVMs have proven to yield good results for small data sets
- SVMs have proven to provide a high degree of generalization through the large margin principle

- SVMs have proven to yield good results for related emotion recognition tasks, e.g. Schuller et al. (2009), Metze et al. (2010), Schmitt et al. (2010), Polzehl et al. (2009, 2010, 2011).
- SVMs are extensible to non-linear mappings by the use of kernel function

In the following sections, the SVM algorithms for classification and regression tasks are explained in more detail.

5.6.1 Personality Classification Using SVM

SVMs view data as sets of vectors in a multi-dimensional space. The task of the algorithm is to find the hyper-plane in that space that separates the classes and provides a maximal margin in between the hyper-plane and vectors from different classes. By maximizing the corridor between the hyper-plane and the data points from the different classes SVMs provide a high degree of generalization.

Let L be a totality of n-dimensional training instances $\{(x_l, y_l) \mid l = 1, \ldots, L\}$ with $x_l \in \mathbb{R}^n$ and $y_l \in \{+1, -1\}$. The classes are seen as positive and negative instances in this respect, with definitions $y_l = +1$ for positive and $y_l = -1$ for negative classes. The hyper-plane $H(\mathbf{w}, b)$ can be defined by a vector $\mathbf{w}$ and an off-set b as

$$H(\mathbf{w}, b) = \{\mathbf{w} \in \mathbb{R}^n \mid \mathbf{w}^T\mathbf{x} + b = 0\} \qquad \text{with } \mathbf{w} \in \mathbb{R}^n \text{ and } b \in \mathbb{R} \tag{5.17}$$

The term $\mathbf{w}^T\mathbf{x}$ refers to the *dot product*, which is also called the *inner product*, of the two vectors. The task is to find suitable w and b so that the hyper-plane divides the positive from the negative class. Given that these parameters can be found directly for all instances, the solution is a perfect linear division. However, for the vast majority of classification problems a slack variable ξ needs to be inserted, which allows data points to spread to both sides of the hyper-plane, i.e. to be on the “wrong” side, potentially being noise-corrupted or outliers. Thus the required class membership rule for a *soft-margin* SVM is the following:

$$\begin{aligned} y_l = +1 &\Rightarrow \mathbf{w}^T\mathbf{x} + b \geq +1 - \xi \\ y_l = -1 &\Rightarrow \mathbf{w}^T\mathbf{x} + b \leq -1 + \xi \end{aligned} \tag{5.18}$$

In fact, instead of maximizing the margin between the hyper-plane and the support-vectors a minimization of the vector $\mathbf{w}$, which is a strictly convex problem, can be done. The resulting minimization problem subjected to Eqs. 5.18 and $\xi \geq 0$ can then be given as

$$\frac{1}{2}\|\mathbf{w}\|^2 + C \cdot \sum_{i=1}^{L} \xi_i \qquad \text{with C being a weighting factor} \tag{5.19}$$

The two terms in Eq. 5.19 refer to the two-fold minimization aim to keep the hyperplane as smooth as possible, which correspond to the minimization of the

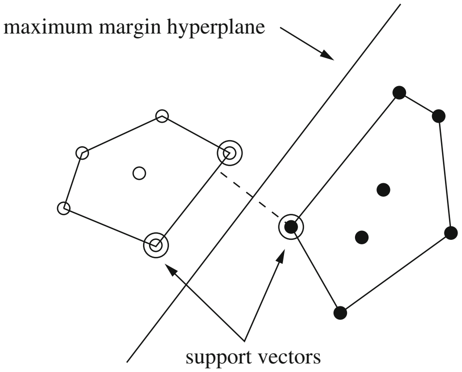

Fig. 5.1 Two-attribute SVM hyperplane and support vectors according to Witten and Frank (2005)

Euclidean norm of **w**, and keep the error small at the same time. The weighting factor C is known as the *complexity* parameter. The factor of 1/2 accounts for the fact that the margin extends to both sides of the hyperplane.

SVMs define the hyper-plane by means of support-vectors in the feature space, which are to be selected out of all data vectors. Support vectors are also called critical boundary instances. Technically, the convex hull of any set of instances can be described as the tightest enclosing convex polygon. It can be obtained when connecting every instance of the set to every other instance. Out of all possible hyperplanes that separate the classes, the maximum-margin hyperplane is the one that has the biggest distance from both convex hulls of the binary classes. Finally, the instances that have the minimum distance to the maximum-margin hyperplane are called support vectors. All other vectors do not influence the hyperplane construction and may even be deleted. At least one support vector is mandatory for each class. Figure 5.1 shows a schematic sketch using a two-attribute case. Filled and non-filled circles indicate the class membership. Big circles are support vectors.

Finding the support vectors for the training instances and determining the parameters of the off-set b and the weights **w** belongs to a standard class of optimization problems known as constrained quadratic optimization. More specifically, the so called *Sequential-Minimal-Optimization, (SMO)* algorithm as implemented in the WEKA toolkit Witten and Frank (2005) is applied in this work. In principle a solution is obtained by using Lagrangian multipliers in the dual form of the problem. SMO always optimizes two Lagrange multipliers at every step, thus breaking down a large optimization problem into a series of smallest possible optimization steps, which can then be solved analytically. Eventually, SMO scales in between linear and quadratic in the training set size for various test problems. More details are out of scope for the current presentation and can be obtained from Platt (1999). By the use of support vectors, overfitting is more unlikely to occur because the maximum-margin hyperplane is relatively stable. It only changes if encountering a new instances that is determined to be a support vectors. Moreover, overfitting is also caused by too much flexibility in the decision boundary, but the boundary is optimized for flatness at the same time.

In essence, the algorithm is capable of separating two classes from each other, but it can be extended to multiple class decisions by applying it sequentially to different combinations of classes. For optimization of this step different strategies have been proposed, e.g.

- 1-vs.-*all*, which results in k binary class decisions
- Pairwise 1-vs.-1 and subsequent majority voting as described in Hastie and Tibshirani (1998), which results in $0.5 \cdot k \cdot (k-1)$ binary decisions
- Arrangement of all decisions along entropy-based tree structures, which results in $(k-1)$ decisions
- Multi-layer arrangements, which positions several SVMs with different feature sets in an tree ensembles

For tree based-arrangements, only the branch leading to the final decision need to be processed in the test run. On the other hand tree-based arrangements show high sensitivity to the actual tree structure, which can lead to increased results but incurs a loss of robustness and generalization power oftentimes. The multi-layer strategy shows the highest complexity, as different data is processed using different attributes, which need to be estimated by extensive experimentation. According to Niemann (2003) the increased computational effort of pairwise 1-vs.-1 over simple 1-vs.-*all* is justified by the over all increased performance that goes along with robust and generalizable estimation—a view that is generally acknowledged in the machine learning community. In the following, all experiments are carried out using a 1-vs.-1 strategy.

5.6.2 Trait Score Prediction Using SVM

The concept of a maximum margin between a hyper-plane and members of different classes only applies to a classification problem. With regression the basic task is to find a function that represents all training data optimally. Let L be a totality of n-dimensional training instances $\{(x_l, y_l) \mid l = 1, \ldots, L\}$ with $x_l \in \mathbb{R}^n$ and $y_l \in \mathbb{R}$. The data is now to be approximated by a linear function of the form

$$f(\mathbf{w}) = \mathbf{w}^T \mathbf{x} + b \qquad \text{with } \mathbf{w} \in \mathbb{R}^n \text{ and } b \in \mathbb{R} \tag{5.20}$$

Again, the concept of a soft-margin from the SVM is transfered to SV regression by insertion of slack variables ξ, ξ^*. In addition, a precision variable ε is introduced with the tacit assumption, that a function f exists that approximates all data pairs (x_l, y_l) with the given precision. Incorporating slack and precision, the resulting minimization problem can be given as defined by Vapnik (1999):

$$\frac{1}{2}\|\mathbf{w}\|^2 + C \cdot \sum_{i=1}^{L} \xi_i + \xi_i^* \text{ with C being a weighting factor}$$

$$\text{subjected to} \qquad \begin{cases} y_i - (\mathbf{w} \cdot \mathbf{x}) - b & \leq & \varepsilon + \xi_i \\ \mathbf{w} \cdot \mathbf{x} + b - y_i & \leq & \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* & \geq & 0 \end{cases} \tag{5.21}$$

Here, all deviations from that function up to a threshold ε are simply discarded, which results in the so called ε-*insensitive* loss function. The threshold ε defines a tube around the function to be optimized. In case of a linear support vector regression, the tube is of an cylindric shape (Fig. 5.2).

In contrast to the classification model, the support vectors for regression models are all vectors that do not fall within the tube—that is, the points outside the tube and on its border. All other instances are computationally irrelevant. Schölkopf et al. (1998) proposed a method to integrate ε into the minimization problem in order to automatically estimate this parameter. The minimization problem can be solved using an improved SMO algorithm as proposed by Shevade et al. (2000) and Smola and Schölkopf (2004), and is applied as implemented in the WEKA toolkit.

Overfitting is avoided by simultaneously minimizing the prediction error while trying to maximize the flatness of the function as expressed by the minimization of the Euclidean norm. Because of the decision to discard all errors within the tube, the resulting error estimation is zero if all data points can be fit into the tube completely. On the one hand the tube could be extended so that all data points fit in it. But this tube would have the worst flatness. On the other hand, if the tube is chosen too small, some training points will have non-zero error. Hence there is a trade-off between the prediction error and the tube's flatness, which is again controlled by a complexity parameter C. The larger C, the more closely the function fits the data, which is congruent to the complexity parameter in support vector classification.

For the consecutive steps of search range exploration and kernel application are analogous to support vector classification.

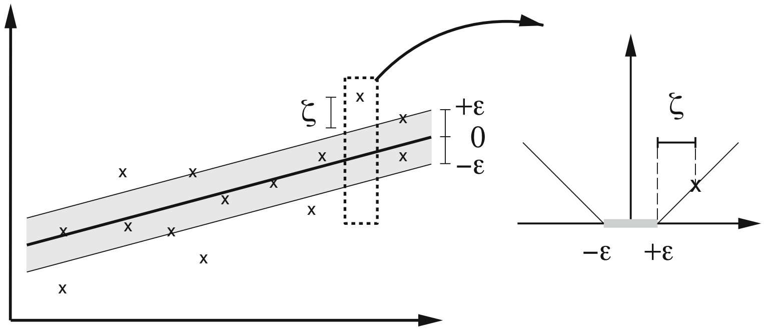

Fig. 5.2 ε-insensitive soft-margin loss function for support vector regression. Image taken from Smola and Schölkopf (2004)

5.6.3 Non-Linear Mapping and Parameter Tuning

In many cases, non-linear mapping can be suspected to outperform algorithms using linear decision boundaries just by the mere fact that linear functions might under-represent data exactness of actual class boundaries and over-generalize with a simple linear fit. Non-linear mappings, in theory, can adapt to the data to be learned in more flexible and various forms. But regarding computational complexity, every time a new instance is processed by non-linear mapping, its dot product, with all support vectors must be calculated which involves one multiplication and one addition for each attribute. Hence, computing a single kernel product requires *O(n)* operations, where n is the input dimensionality. Especially with regard to non-linear mapping, the number of attributes in the new space can be huge, resulting in a rather expensive non-linear mapping. Here, the so called *kernel-trick* can be applied. In essence, the trick allows to calculate the dot product before non-linear mapping is performed. This can be achieved by using a so-called kernel function based on the dot product. Kernel functions are of the form:

$$K^{\phi}(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^T \cdot \phi(\mathbf{y}) \tag{5.22}$$

Also new kernel functions can be designed, under the boundary conditions that the transforms satisfy the criteria of symmetry, Cauchy-Schwarz inequality holds true, and the transform is of positive semi-definite form. More details are out of scope for this work and can be found in, e.g., Platt (1999), Schölkopf and Smola (2001), and Smola and Schölkopf (2004). Applying the kernel trick, SVMs use linear models to implement non-linear class boundaries, i.e. a straight line in the new space does not look straight in the original instance space.

Most popular kernel functions, amongst others, are

- Polynomial kernels $K^{\phi}(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T\mathbf{y} + 1)^k$, with p being the polynomial order and $p = 1$ presenting a linear kernel
- RBF-kernel[5] $K^{\phi}(\mathbf{x}, \mathbf{y}) = e^{\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}}$ with σ resembling the standard deviation of the assumed normal distribution.

The choice of the best kernel function and its parameters can only be done experimentally. Oftentimes, the kernel parameter and the parameter of the SVM need to be explored by a grid search, i.e. systematically jointly altering the parameters at hand. Although SVM training procedures can be very costly, especially when using non-linear kernels and respective multi-class strategies, in the test runs only those vectors that have been selected as support vectors are computationally relevant (Fig. 5.3).

In the present support vector classification and regression experiments the following parameters need to be jointly set in optimal ways

[5] The abbreviation *RBF* stands for *Radial-Basis-Function*. Note, there are different radial-basis functions available. The proposed kernel function presents a Gaussian form.

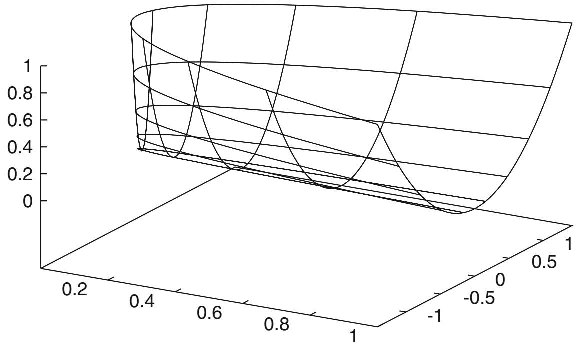

Fig. 5.3 Second degree polynomial feature map. Image taken from Burges (1998)

1. Number of features given to the SVM classifier
2. Complexity parameter *C*
3. Kernel parameter, i.e. either linear, 2nd or 3rd degree for the polynomial kernel, as well as kernel width in case of the RBF-kernel

The number of features given to the model is controlled by the incremental feature selection. This is a global search. The optimal number of features will be determined by the empirical achievement of maximal classification success. Note that for each classification or regression task a different number of relevant features are proposed by the IGR ranking. Since the IGR is data-driven, this number also varies with respect to the database at hand.

For the complexity parameter the search range was set to [0.0001, 0.001, 0.01, 0.1, 1, 5], which corresponds to a wide search space. Degenerated complexity parameter settings have a huge impact on the models capabilities, as will be apparent when looking at the result plots presented in Sect. 5.8. If the complexity is too high, the model fits the data to a maximum amount including all data points. On the other hand, if the complexity is too low, there might not be enough individual data points considered when minimizing, hence the hyperplane becomes under-fitted.

Two polynomial kernels, i.e. second and third degree kernels, have been evaluated in addition to a linear kernel function. For the RBF-kernel an additional parameter presenting the kernel-width need to be determined. The search range for this parameter was set to the same range as for the complexity search.

For each dataset, i.e. the text-dependent dataset, the text-independent dataset, and the multi-speaker datasets this search lead to a computation of $6 \times 6 = 36$ SVM configurations for the RBF-kernel, as well as $3 \times 6 = 18$ SVM parameter configurations for the linear kernel. The total of 54 evaluations were obtained by 10-fold cross validation, hence 540 evaluations for each feature inclusion step were calculated. Depending on the data set, the feature expansion was executed until the feature space reached the number of non-negligible IGR-ranked features, cf. Sect. 5.4. Features were included in chunks with an inclusion step size of minimum 5 features at a time. For illustrating the classifiers response when adding irrelevant features the step size was set to 10 when further expanding the features space. Eventually, for the text-dependent data more than 230 k SVM evaluations have been executed. For the text-independent data more than 120 k SVMs have been evaluated. For the multi-speaker set, the procedure was applied to each individual trait. This resulted in more than 300 k SVM evaluations. Overall, the number of calculated SVM configurations results in more than 0.6 million.

5.7 Evaluation

In terms of evaluating the success of the above described features, two questions should be taken into account. First, the optimal evaluation measurement has to be determined. Most favorably, there exist just one criterion that indicated an all-encompassing modeling success. In other cases more than one criteria need to be inspected at the same time. The choice can strongly be influenced by many factors, e.g. the class distribution, the weight of different error types, or the very meaning of different error types. Apart from the evaluation metrics, also the way to get to reliable estimates for the chosen metric is an important question, which can also differ in accordance to the amount of data available and inherent data structures. The following sub-sections therefore explain the chosen approaches.

5.7.1 Evaluation Metrics

Throughout this work, results from modeling experiments are given by the following evaluation metrics:

- *Accuracy* for classification tasks as overall measure of classification success
- *F-Measure* for individual classes in order to give insights into class-wise performance
- *Precision* and *Recall* for individual classes in order to show the classifier's bias
- *Average Correlation* between human ratings and automatic predictions for trait score regression tasks
- *Root-Mean-Squared Error* (*RMSE*) for trait score prediction off-set

The overall *accuracy* of a classification result is calculated by simply taking the ratio of the correctly classified samples and the number of samples available. The accuracy then gives an intuitive figure on the overall classification success. On the other hand, it is just one figure, while one might want to explore class-wise performances when trying to understand or interpret results or mistakes on a deeper level.

In these cases the use of other evaluation criteria are suggested, which can be briefly introduced by explaining two basic concepts from the field of *Information Retrieval*, namely *precision* and *recall*. The recall of a class measures how many examples—out of all examples of a class that exist—are effectively classified into the right class. The precision on the other hand accounts for the classified examples and counts how many of these examples—that were already classified into the class at hand—actually really belong to that class. Note that one can always reach a recall of 100 % by simply collecting all examples into one class. But this class would have the worst precision. On the other hand one could classify over-cautiously and only assign an example to a class when being absolutely sure about the decision. In this case one results in a high precision with regard to the class at hand, but misses a lot of examples that actually belong to this class too, hence decreasing the recall. In this respect, precision and recall can be used to measure classification success while giving an intuitive estimation on the bias of the classification model at the same time. Eventually, the aim is to achieve a classification of both high recall and high precision. The *F-Measure* is one measurement capable of dealing with this balancing problem Witten and Frank (2005). F-measure of a class is given by Eq. 5.23:

$$\text{F-Measure} = \frac{2 \cdot recall \cdot precision}{recall + precision} \tag{5.23}$$

Note, the amount of training data can have an influence on the performance estimation. This becomes obvious when there is a considerable difference in amount of data for different classes. In this regard, the accuracy measurement can be biased since it is influenced by the majority class to a greater extent than by other classes. That means, if one has an imbalanced class distribution, and the model of the majority class yields better results than other models of other non-majority classes, the resulting accuracy measurement gives overestimated figures. Also in these cases the use of precision and recall are suggested in order to calculate the unbiased *f1-Measurement*, which is defined as the arithmetic (unweighted) mean of F-measures from all data classes.

The data actually used for the classification tasks presented in this work is the text-dependent subset and the text-independent subset. The text-independent data shows no imbalance, the text-dependent data was carried out including all usable data which results in negligible class imbalance only, cf. Sect. 3.1. All overall results are therefore given by accuracy in this work. Class-wise performance is given by precision, recall and F-Measure.

In case the data misses a class structure, like for the multi-speaker dataset, other evaluation measurements need to be taken into account. Here the *average correlation*, which is calculated by taking the arithmetic mean of all correlations between the mean of all labelers ratings on the one hand and the result from automatic prediction on the other hand. Like explained in Sect. 4.2 when looking at the characteristics of the consistency in terms of Cronbach's α, there is a major drawback on the use of correlation as performance indicator. Correlation is sensible to relative relations, not absolute. In the present case, there would be a high correlation when the algorithm predicts the personality systematically higher than the humans rated it. The more consistent this off-set, the higher the correlation. Perfect consistency results in perfect correlation although absolute estimates might show an off-set. Therefore, another measure estimating the absolute off-set is needed, such as the *Root-Mean-Squared Error, (RMSE)*. Note, many other measurements exist that account for such a bias, e.g. the *mean-absolute error*. However, the RMSE seems to have the widest distribution in the research community and is chosen in order to enable an intuitive comparison of the present results and other works from the literature. Equation 5.24 shows the calculation, with y being the mean value of all reference ratings from humans for one stimulus n out of a set of stimuli of size N, and $\hat{y}$ the predicted value:

$$\text{RMSE} = \sqrt{\frac{\sum_{n=1}^{N}(y_n - \hat{y_n})^2}{N}} \tag{5.24}$$

Finally, all results from regression experiments are given and plotted by correlation first, and respective RMSEs are indicated thereafter. This is because the levels of correlation can be intuitively grouped into *weak, moderate, good*, and *very good* correlation, as explained in Sect. 4.2. Note, also the other way around would be possible. But for interpreting the actual distance measured by RMSE one would need to postulate which differences should be considered as *very small* and which as *too large*. Such an understanding can differ due to the desired application of the prediction models, as will be discussed in Sect. 6.3.2.

5.7.2 Evaluation Method

Normally one strives to obtain results that are not only valid for current data at hand but also for any unknown data, i.e. one strives to avoid over-fitting towards one particular dataset. To avoid over-fitting in classification or ranking algorithms a 10-fold cross validation on the whole training set is applied for all evaluations. In more detail, after randomizing the instances, all the training data is partitioned into 10 fixed equally sized splits, each approximately mirroring the class distribution of the whole training set. Now we split the overall evaluation into evaluation of these 10 folds. Each fold treats 9 out of 10 splits as training material and uses the 10th split for testing. For each of the 10 folds a different split is chosen as test split. Since data from test splits is never included in the training splits within a fold, the 10 estimations provide independent estimations. After all folds have been processed the resulting estimations are averaged. For the multi-speaker dataset, partitions are designed in a speaker independent way, i.e. speakers in the test split of a fold never occur in the training material of that fold. This procedure gives advantage over the traditional definition of one global train set and one global test set, where all the material in the train set served for building one global model and the evaluation is done processing the unseen global test set once. By applying cross validation we get a better impression about how successful the classifier operates when encountering different test splits while processing more of the data at hand for training. The traditional way gives only one evaluation estimation which is based on just one particular test set, and training has been processed with less of the overall data.

5.8 Results from Text-Dependent Data

5.8.1 Results from Automatic Classification

Having generated a ranking for the 10-class task as explained in Sect. 5.4 allows for first insights into the value of individual features and feature groups. Table 5.2 presents the 30 top-ranked features for the text-dependent data according to IGR. The first two columns show the average rank and the standard deviation of the rank from the cross-validation procedure. A small standard deviation is desirable since this would mean that the respective feature has proven to be of a comparable rank in all the data splits.

Besides a few exceptions, the standard deviations shown in the table result relatively low proving a rather general validity of the features in their ranks. Less stable ranks seem to occur with the pitch features, indicating that they are of very high IGR in many but not all data splits. As for the pitch feature capturing the rising slope, this can be assumed to account for the fact, that not all utterances necessarily consist of a rising slope at all. The high rank of the IQR, being a measure of variation more robust to outlier occurrences, here measured for the pitch dynamics, indicates the importance of pitch gestures. Moreover, it indicates the importance of the speed of the gestures. Again, this pitch behavior proves to be of high importance, but not for all data splits. A tentative conclusion from the table is that heterogeneity seems to be beneficial. Various statistics from various descriptors including dynamics are present in highest ranks. When looking at the last column, the examination of voiced segments or the whole utterances seem to be more promising than examining unvoiced segments.

Figure 5.4 shows the overall segments (cf. Sect. 5.2) and audio descriptor (cf. Sect. 5.3) distributions when moving along the IGR top-ranked features beyond the 30 top-ranks. Plots are given as absolute counts to the left and relative shares among the counts to the right. The scale of the x-axis correspond to the number of top-ranked features with non-negligible contribution in terms of IGR. Audio-descriptor groups are coded by color.

Looking more closely at the x-axis limit of Fig. 5.4a, roughly half of the defined features show non-negligible contributions for the fixed-text data. The most dominant color seems to be green, hence the most important features are MFCCs, populating roughly 40 % of the top-ranks throughout the incremental expansion. The remaining 60 % of top-ranks seem to be evenly distributed to the remaining 5 descriptor groups except group *others*. Note that the more features a group has, the more of them can potentially show up in the top-ranks. Given this perspective, it is remarkable that the few features in group *others* are populating highest ranks, so they can be seen as highly relevant for the given dataset.

Figure 5.4b shows the same plot for the distribution of segments in the ranks. The general tendency follows the trends from the observations of the 30 top-ranks. Overall, voiced segments seem to be the most promising source for feature definition. Unvoiced segments seem to account for 20 % of all features only, also starting to contribute beyond the 50 top-ranked features only.

Table 5.2 Top-30 ranked IGR features for text-dependent dataset

Avg. rank	Std	Feature group	Derivation	Statistics	Segment
1.7	±0.78	spec	Δ	meanRollOffpoint	voicedSeg
2.2	±0.75	loudness	$\Delta\Delta$	err.Lin.Reg	wholeUtt
4.7	±0.9	other		pcnt	voicedSeg
5.5	±0.5	MFCC	Δ	coeff$_2$ std	wholeUtt
6.6	±13.82	loudness	Δ	err.Lin.Reg	wholeUtt
7.2	±0.6	MFCC		coeff$_{14}$ std	voicedSeg
9.3	±10.11	loudness	$\Delta\Delta$	std	wholeUtt
10.2	±3.09	intensity	Δ	kurtosis	voicedSeg
12.8	±11.44	pitch	Δ	medianPositiveSlope	voicedSeg
13.1	±5.47	MFCC		coeff$_0$ std	wholeUtt
17.7	±5.75	MFCC		coeff$_7$ mean	wholeUtt
17.9	±2.77	intensity	Δ	err.Lin.Reg	wholeUtt
18.8	±9.02	MFCC		coeff$_7$ std	wholeUtt
19.1	±4.66	other		speechToPause	wholeUtt
19.4	±13.19	MFCC		coeff$_5$ std	wholeUtt
19.9	±6.67	MFCC		coeff$_2$ mean	voicedSeg
20.2	±4.45	other		durationAfterVAD	wholeUtt
26.2	±9.05	loudness	$\Delta\Delta$	kurtosis	wholeUtt
26.4	±5.14	MFCC		coeff$_{12}$ mean	voicedSeg
27.3	±8.26	MFCC		coeff$_9$ mean	wholeUtt
27.7	±23.45	pitch	Δ	iqr	voicedSeg
28.5	±36.69	pitch	Δ	meanPos.Slope	voicedSeg
28.5	±17.26	pitch		meanAbsolute	voicedSeg
30.5	±10.73	spec		meanCentroid	voicedSeg
31.6	±7.68	formants		#5 centerMedian	voicedSeg
31.8	±7.7	formants		#5 bandwidthMedian	voicedSeg
32.4	±14.56	loudness	$\Delta\Delta$	max	wholeUtt
32.7	±7.52	MFCC		coeff$_6$ mean	voicedSeg
32.9	±7.54	MFCC		coeff$_{11}$ std	voicedSeg
33.9	±13.03	MFCC		coeff$_6$ std	wholeUtt

Figure 5.5 shows the plot for the best classification accuracy, which was obtained by using a linear kernel. Random guessing would result in an accuracy of 10 % on average. This plot shows a very steep increase up to almost perfect classification success. As a first observation, these excellent results are reached with far fewer features that have been predicted as meaningful, i.e. non-negligible, in terms of IGR ranking. Also beyond the shown 300 top-ranks the accuracy suffers only marginally when including features up to the predicted number of 700 non-negligible IGR features. After this number, the accuracy decreases. Hence, the ranking proves reasonable.

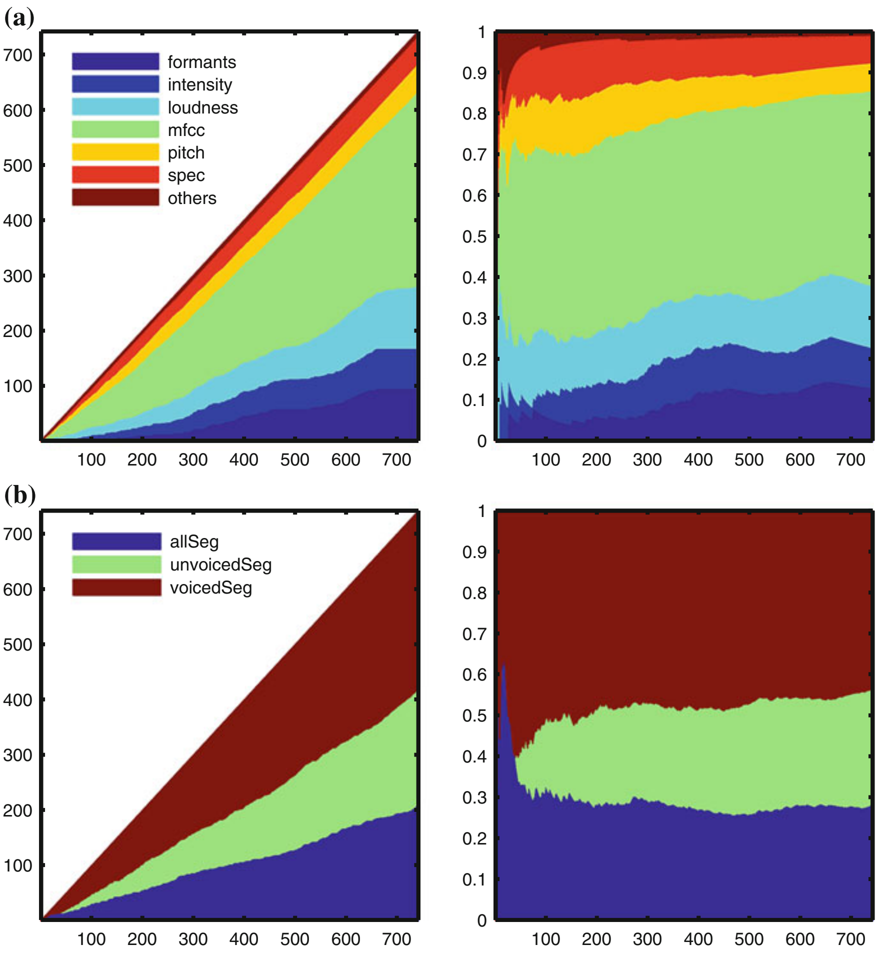

Fig. 5.4 Stacked audio descriptor group distributions (cf. Sect. 5.3, panel a), and stacked segment distributions (cf. Sect. 5.2, panel b) on y-axes along expanding number of top-ranks on x-axes for the text-dependent database. Plots are given as absolute counts to the *left* and relative shares among the counts to the *right*. **a** Audio descriptor group distribution (*left* absolute, *right* relative). **b** Audio segments distribution (*left* absolute, *right* relative)

In terms of SVM complexity, including only few support vectors seems to be of disadvantage. When extending the linear kernel to a polynomial of second and third order this trend can be counteracted, but this would incur drastically higher computation demand. Eventually, also the application of RBF-kernels did not result in an even steeper incline. Hence, non-linear extension seems to be unnecessary.

When including the top-16 IGR features the classification reaches an accuracy of 96.3 % already. Doubling this amount of features leads to accuracies of higher than 99 %. However, the aim of these experiments is not primarily to push up results until obtaining 100 % accuracy, but to prove the feasibility of automatic personality

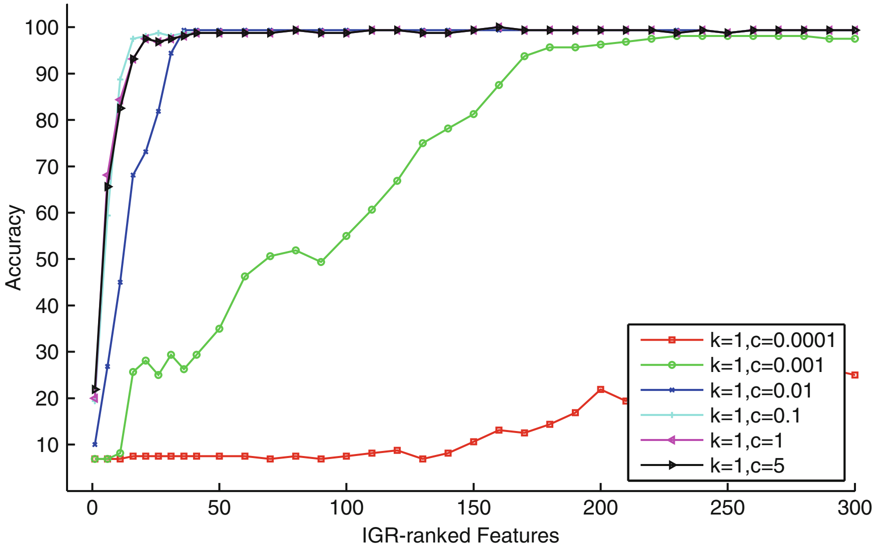

Fig. 5.5 Classification accuracy from incremental SVM-IGR feature selection using a linear kernel for the text-dependent dataset. X-axis shows number of features up to 300 top-ranks, y-axis shows accuracy

classification. For the given dataset, these results fully supplies the evidence. Table 5.3 presents the respective confusion matrix for the SVM classifier using a linear kernel and the top-16 IGR ranks. Table 5.4 shows the recalls, precisions and F-measure for this classifier. Because all of the presented figures are very high, a deeper analysis of class-wise statistics will not lead to new insights at this time. However, these figures will have to be compared with statistics from the text-independent data, cf. Sect. 5.9.1.

Summing up, very few features are sufficient to classify the text-dependent dataset using SVM and the top-ranked features as presented in Table 5.2 up to an almost perfect accuracy. Features are of various type as described in Table 5.2. The SVM shows best results when allowing the algorithm to use a fair or high number of support vectors. Note, when using a high number of support vectors the model also becomes more adapted to the presented data. But at the same time, the presented data consists of few features only. In addition, using cross-validation prevents the algorithm from adapting to a specific set of examples. Overall, the classification task seems to be a very consistent task, which has been indicated by the high consistency in the ratings and original data design, i.e. by controlling the linguistic content and including a single speaker only. Although this data is not of every-day realistic character for naive speakers, it shows that the classification of personality from voice can be feasible, especially when the speech is of blueprint-like, staged characteristics and when the system is limited to known linguistic content. Ultimately, these excellent results are certainly as unexpected as encouraging for future experiments.

Table 5.3 Confusion matrix showing absolute numbers of instances from SVM-IGR using top-16 ranked features and a linear kernel on text-dependent data

		Predicted									
		a+	a−	c+	c−	e+	e−	n+	n−	o+	o−
Actual	a+	15	0	0	0	0	0	0	0	0	0
	a−	0	15	0	2	0	0	0	0	0	0
	c+	0	0	15	0	0	0	0	0	0	0
	c−	0	2	0	14	0	0	0	0	0	0
	e+	0	0	0	0	17	0	0	0	0	0
	e−	0	0	0	0	0	17	0	0	0	0
	n+	0	0	0	0	0	0	17	0	0	0
	n−	0	0	0	0	0	0	0	15	0	0
	o+	0	0	1	0	1	0	0	0	14	0
	o−	0	0	0	0	0	0	0	0	0	15

Rows show the actual class membership, columns represent the classification result

Table 5.4 Class wise statistics for classification of text-dependent dataset

Class	Precision	Recall	F-measure
a+	1	1	1
a−	0.882	0.882	0.882
c+	0.938	1	0.968
c−	0.875	0.875	0.875
e+	0.944	1	0.971
e−	1	1	1
n+	1	1	1
n−	1	1	1
o+	1	0.875	0.933
o−	1	1	1

5.8.2 Results from Automatic Trait Score Prediction

The task of the support-vector regression machine is to predict the actual numerical trait score, as referenced by the mean value of the labelers ratings. Since there is no overall personality value to be predicted, each trait is predicted individually. Figures 5.6, 5.7, 5.8, 5.9 and 5.10 present the results from individual trait prediction using SVM regression and the IGR ranking based on the continuous ratings on each individual scale. As described in Sect. 5.4, IGR can only be applied after discretizing the label space. Discretization was carried out by splitting the labels into binary classes, i.e. one below the average and one above the average. Thus, a high characteristic and a low characteristic is introduced, which can be seen in analogy to the *high* and *low* targets of the class structure. Note, that the original class structure

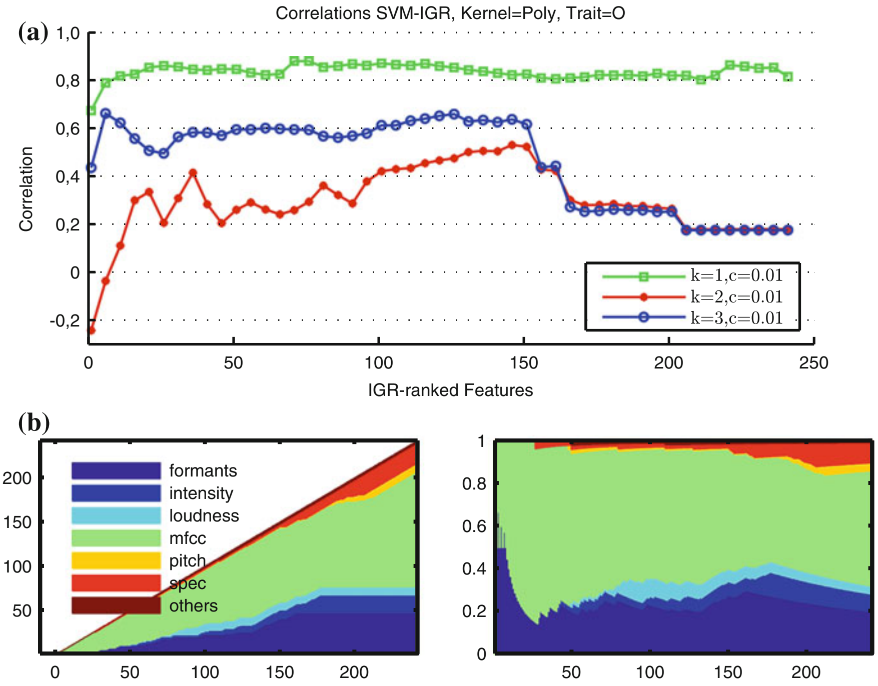

Fig. 5.6 Panel a: Correlation between automatic prediction and human ratings (y-axis) along IGR ranking (x-axis) for openness. Panel b: Stacked audio descriptor groups distribution (y-axis) along IGR ranking for the text-dependent database after discretization openness ratings into high and low splits

could have been used for splitting the data. But this structure is only present for the speaker-dependent datasets. The main goal of this split is to provide a basis for comparison with the text-independent and multi-speaker ranking, the latter demanding unsupervised discretization, which will be presented in Sect. 5.10. In addition, using the class structure did not show to increase the overall correlation between human labels and automatic prediction. Results and plots are omitted for this reason.

Panel a: Correlation between automatic prediction and human ratings (y-axis) along IGR ranking (x-axis) for openness. Panel b: Stacked audio descriptor groups distribution (y-axis) along IGR ranking for the text-dependent database after discretization openness ratings into high and low splits.

Figure 5.6 shows the results for openness prediction. Best correlation between human openness ratings and SVM prediction reached 0.88, achieved when including 84-top ranks and using a linear kernel as well as a fair number of support vectors. RMSE resulted in 1.87 points on a scale ranging between 0 and 48 points. Non-linear extension did not lead to better results, as can be clearly seen from Fig. 5.6a. Overall, this result accounts for good and almost very good correlation. The shape of the correlation curve is overall smooth, indicating a reasonable ranking outcome.

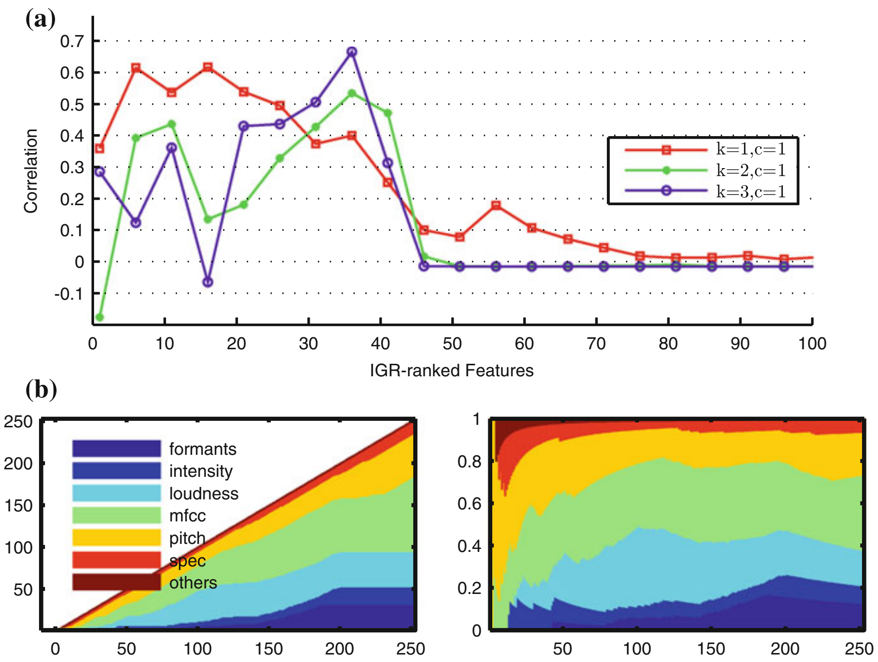

Fig. 5.7 Panel a: Correlation between automatic prediction and human ratings (y-axis) along IGR ranking (x-axis) for conscientiousness. Panel b: Stacked audio descriptor groups distribution (y-axis) along IGR ranking for the text-dependent database after discretization conscientiousness ratings into high and low splits

When looking at the ranking in Fig. 5.6b, more than 240 features have been estimated as meaningful by IGR. Best results are obtained when using one third of these features only. There are clearly two dominant feature groups, namely MFCC and formant-derived features. MFCCs are almost exclusively derived from the unvoiced segments or whole segments. While the statistics are drawn from the coefficients and their derivations to the same amounts, most frequently high coefficients are basis for features capturing the standard deviation of these coefficients. For the formant features no such tendency becomes apparent, i.e. various statistics can be found in the top-84 ranks.

A quite different picture can be obtained from Fig. 5.7 showing selected results for conscientiousness prediction. Best correlation between human ratings and SVM prediction reached 0.68, achieved when including 36-top ranks and using a polynomial kernel of third degree. This result appears as spike in an unsmooth graph in Fig. 5.7a and may be interpreted with caution. The more smooth red curve in this figure showing the linear kernel function resulted in 0.6, which after all also resembles moderate correlation only. RMSE resulted in 4.29 points. Figure 5.7b shows the respective ranking. 250 ranks have been predicted as meaningful, but the performance drops drastically after including more than 36 top-ranks.

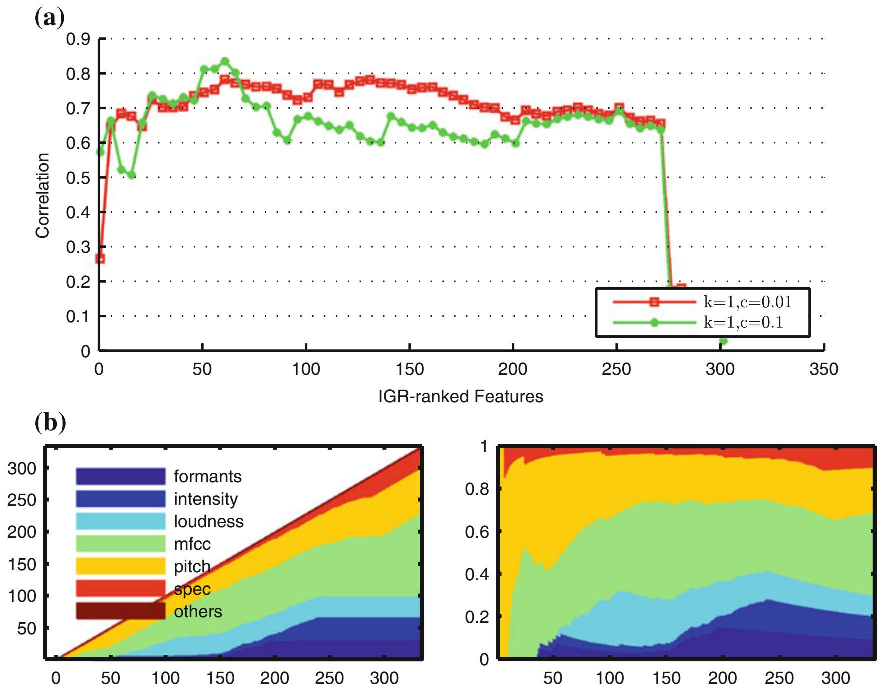

Fig. 5.8 Panel a: Correlation between automatic prediction and human ratings (y-axis) along IGR ranking (x-axis) for extroversion. Panel b: Stacked audio descriptor groups distribution (y-axis) along IGR ranking for the text-dependent database after discretization extroversion ratings into high and low splits

When looking at the ranking plot, a colorful picture can be obtained. However, the information present in the ranks could not be modeled by the SVM beyond the 36 top ranks. Until this point, the distribution, the range and the slopes of pitch are dominant. More systematic results cannot be drawn.

Figure 5.8 shows the results for extroversion prediction. Best correlation between human ratings and SVM prediction reached 0.82, achieved when including 61-top ranks and using a linear kernel. RMSE resulted in 5.15 points. As non-linear extension did not lead to better results the respective curves were omitted in Fig. 5.8a. The shape of the correlation curves shown are overall smooth, with more support vectors leading to smoother curves. Overall, this result accounts for a good result from correlation. More than 300 features have been estimated as meaningful by IGR. Best results are obtained when using only a fraction of it. Drastic loss of correlation is incurred when including more than 270 features.

From Fig. 5.8b it can be seen, than there are two dominant feature groups, namely MFCC and pitch-derived features. These pitch features capture slopes, ranges, variation predominantly, in lower ranks also the dynamics of the movements. There is

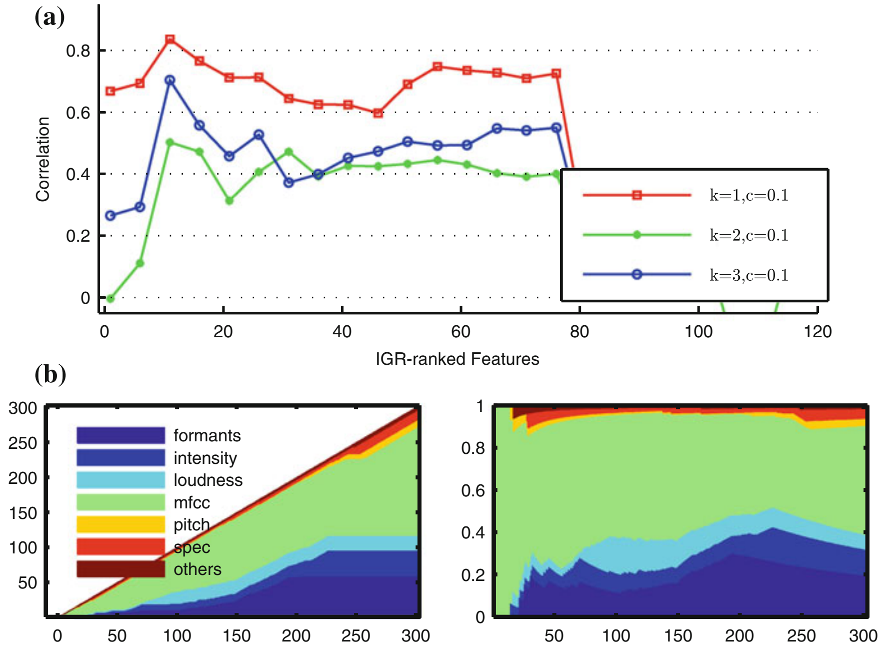

Fig. 5.9 Panel a: Correlation between automatic prediction and human ratings (y-axis) along IGR ranking (x-axis) for agreeableness. Panel b: Stacked audio descriptor groups distribution (y-axis) along IGR ranking for the text-dependent database after discretization agreeableness ratings into high and low splits

no clear structure with regard to the MFCC statistics. Almost all MFCC features are derived from voiced segments or the whole utterances.

Figure 5.9 shows selected results for agreeableness prediction. Best correlation between human ratings and SVM prediction also reached 0.82, but this time achieved when including 11-top ranks only. Extension to non-linear kernels did not improve the performance, as can be observed in Fig. 5.9a. RMSE resulted in 3.83 points. Overall, this result accounts for good correlation. The shape of the correlation curves show the same trends, and in general indicate an overall smooth ranking outcome.

The ranking plot in Fig. 5.9b suggest 300 non-negligible features to be used. While the performance is not harmed until a number of approximately 80-top ranks in the feature space, the performance drops drastically beyond this number. When looking in the actual 11 top-ranks only MFCC statistics from voiced segments or the whole utterance can be found. Since various statistics also drawn from the Delta coefficients are included a more systematic analysis is not possible.

Figure 5.10 shows selected results for neuroticism prediction. Best correlation between human ratings on neuroticism and SVM prediction reached 0.92, again when including 11-top ranks only. For extroversion, using a polynomial kernel of second degree proved to be beneficial, as shown in Fig. 5.10a. RMSE resulted in 2.93 points. These results accounts for very good correlation. Note that this result

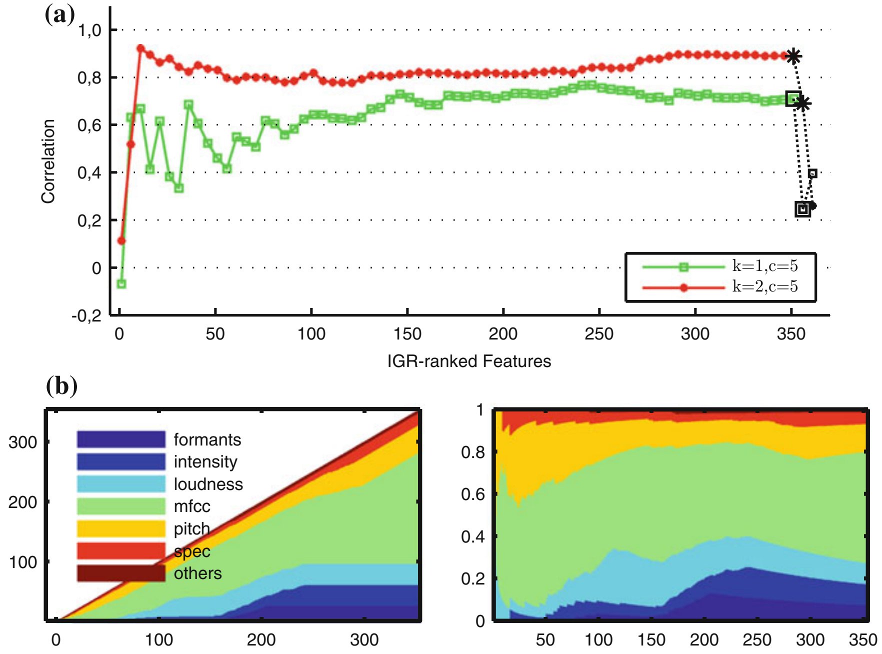

Fig. 5.10 Panel a: Correlation between automatic prediction and human ratings (y-axis) along IGR ranking (x-axis) for neuroticism. Panel b: Stacked audio descriptor groups distribution (y-axis) along IGR ranking for the text-dependent database after discretization neuroticism ratings into high and low splits

could only be achieved at the costs of high complexity in the model. The shape of the correlation curves shown are smooth.

The ranking plot in Fig. 5.10b suggest to use up to 350 top-ranks to be included in the model. The symbols plotted in increased size and the dashed lines show the decrease in performance when adding more than the IGR suggested number of top-ranks into the feature space. In the current figure, the suggested number exactly matched the 350 features, which illustrate that for the present extroversion prediction the IGR provided both, a reasonable ranking and a reasonable suggestion about how many of the top-ranks should be included in the models. While for agreeableness, extroversion and conscientiousness the performance dropped before this suggested number, for openness and neuroticism prediction this suggestion proves beneficial. Also, the models trained for these two traits archived very good correlations, unlike the models for the remaining traits. When looking in the actual 11 top-ranks loudness ranges of unvoiced sounds and the rising slope of pitch features can be found in highest ranks, while the majority of features are drawn from MFCCs. This time, these features are equally based on voiced segments, unvoiced segments or the whole utterance.

5.9 Results from Text-Independent Data

5.9.1 Results from Automatic Classification

Table 5.5 shows the 30 top-ranks from 10-class IGR ranking for the text-independent data. The most relevant information seems to be bound to formant #5, its shape and variation. As a first observation, it can be stated that information about this formant seems to contribute much in terms of IGR, which is a rather unexpected result. In more detail, when looking into the class distributions within the bins after supervised discretization, it becomes apparent that the splits mostly separate extroversion high targets from extroversion low targets, and in addition to this there is a separation in between both of these targets and the rest of all other classes. This relative purity can be assumed to be the main cause for the high IGR.

From the phonetic point of view, the importance of formant #5 can be explained by the observation that the shape of formants and the corresponding bandwidth are influenced by articulatory settings, which in turn influence the sharpness of the speech or the air pressure used to stimulate the vocal folds. Formant #5 is widely acknowledged to be contributing to the timbre of a speech sound in the linguistic community. This observation can be understood as indication for the importance of the correlates of perceptual sound quality including burstiness or pressure of the vocal fold stimulation. Still, the question why the importance of the formant is higher for this database than for any other rankings presented in this work is left unanswered. Note, as explained before, the formant tracker that extracts this formant's contour is bound to the frequency range between 4,500 and 5,500 Hz . Eventually, even if the tracker did not capture the actual formant peak within this range, still a maximum in the magnitude of energy in this frequency range shows high information gain value.

From the acoustic point of view this result can also partly be substantiated by referring to the importance of MFCC features, which were shown to be useful in many former studies, cf. Sect. 5.1.5, and which constitute the second biggest feature group in to-ranks. While MFCC features capture the spectral magnitude of energy at a fixed frequency range (by a triangular filter), also the formant peaks are calculated in a fixed range around a center frequency. The search range for formant #5 corresponds to three MFC coefficients roughly. However, the occurrence of the high number of formant features capturing various characteristics specifically of formant #5 is not mirrored by a respective finding from the three corresponding MFC coefficients. Eventually, this finding cannot fully be explained from neither phonetic nor acoustic point of view.

Also few statistics on the distribution of intensity and pitch show a high information gain. Again, statistics drawn from voiced segments or the whole utterance prevail. Further insights can be obtained when comparing Figs. 5.11 and 5.4. Here, one finds a different picture for the text-dependent dataset. First, only roughly 200 features show a non-negligible amount of information in terms of IGR. Second, formant estimation seems to be most promising also beyond the 30 top-ranks, while MFCCs constitute a second main feature group especially beyond the 60 top-ranks, as

Table 5.5 Top-30 ranked IGR features for text-independent dataset

Avg. rank	Std	Feature group	Derivation	Statistics	Segment
1.9	±1.45	formants	DD	#5 center std	voicedSeg
2.7	±1.27	formants	DD	#5 bandwidth std	voicedSeg
3	±1.61	formants	D	#5 center std	voicedSeg
3.6	±1.28	formants	D	#5 bandwidth std	voicedSeg
7.4	±3.07	formants		#5 bandwidth std	voicedSeg
7.6	±3.2	formants		#5 center std	voicedSeg
8	±5.42	formants		#5 center mean	voicedSeg
8	±5.02	formants		#5 bandwidth mean	voicedSeg
12.3	±8.37	formants	D	#5 bandwidth max	voicedSeg
12.5	±2.73	formants	D	#5 bandwidth range	voicedSeg
12.5	±8.41	formants	D	#5 center max	voicedSeg
12.5	±2.5	formants	D	#5 center range	voicedSeg
13.1	±4.46	formants	D	#5 center min	voicedSeg
13.3	±4.27	formants	D	#5 bandwidth min	voicedSeg
16.3	±3.66	formants	DD	#5 center max	voicedSeg
16.7	±3.13	formants	DD	#5 bandwidth min	voicedSeg
16.7	±2.65	formants	DD	#5 center min	voicedSeg
16.9	±3.67	formants	DD	#5 bandwidth max	voicedSeg
17.2	±2.89	formants	DD	#5 bandwidth range	voicedSeg
17.4	±2.62	formants	DD	#5 center range	voicedSeg
21.5	±5.5	intensity		skewness	wholeUtt
25.4	±5.54	pitch	D	median of negative slope	voicedSeg
26.6	±4.03	formants	D	#5 bandwidth mean	voicedSeg
26.8	±4.26	formants	D	#5 center mean	voicedSeg
27.4	±4.74	intensity		kurtosis	wholeUtt
33.9	±6.7	MFCC		coeff$_{11}$ std	wholeUtt
37.5	±3.77	spec		mean Centroid	voicedSeg
43.4	±7.54	intensity		std	unvoicedSeg
43.5	±17.3	pitch	D	iqr	voicedSeg
43.6	±7.23	MFCC		coeff$_8$ mean	wholeUtt

shown in Fig. 5.11a. All other feature groups seem to be of minor importance, including intensity or loudness features. The segment distribution plot shown in Fig. 5.11b looks similar to the plot on the text-dependent dataset. Here, it also becomes apparent, that the bulk of the aforementioned MFCCs seem to be drawn from voiced segments or the whole utterance, but not from unvoiced segments.

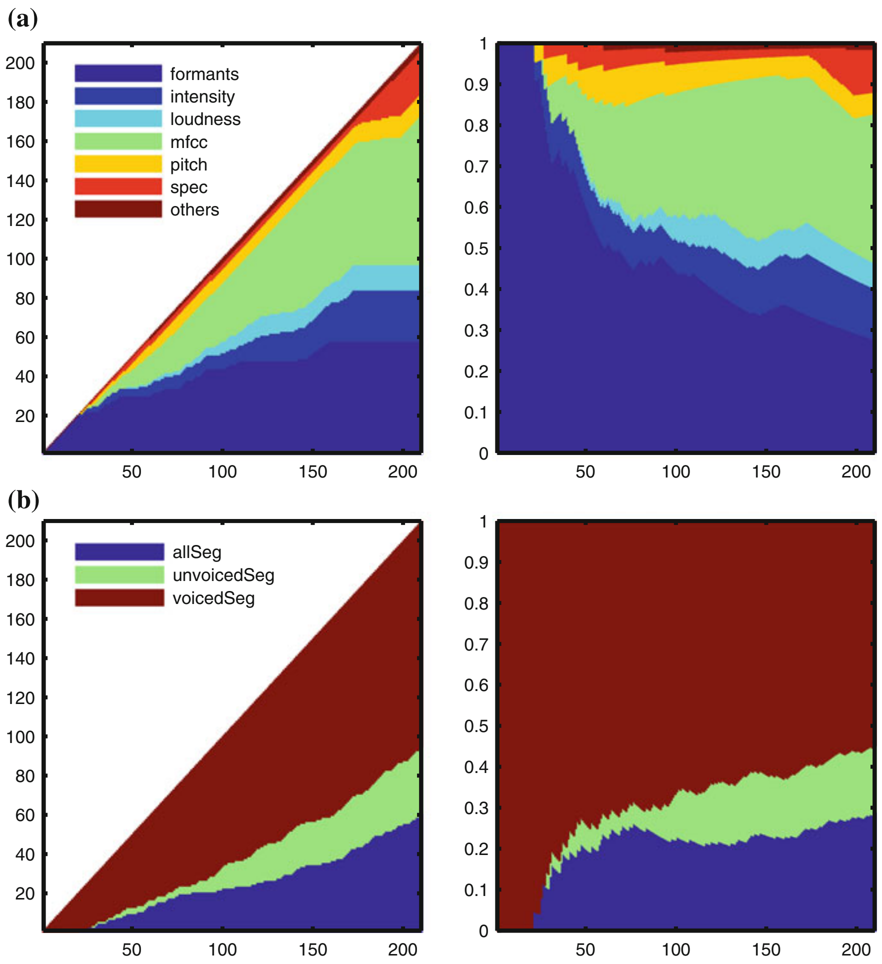

Fig. 5.11 Stacked audio descriptor group distributions (cf. Sect. 5.3, panel a), and stacked segment distributions (cf. Sect. 5.2, panel b) on y-axes along expanding number of top-ranks on x-axes for the text-independent database. Plots are given as absolute counts to the *left* and relative shares among the counts to the *right*. **a** Audio descriptor group distribution (*left* absolute, *right* relative). **b** Audio segments distribution (*left* absolute, *right* relative)

Figure 5.12 shows the results of the text-independent classifier using a linear kernel function. In comparison to Fig. 5.5 showing the text-dependent performance, this figure reveals that any inclusion of features that have been estimated as irrelevant by IGR ranking contributes in a decrease of overall performance. In the figure this becomes obvious when looking at the dashed lines, which correspond to negligible features. The number of non-negligible features as suggested by IGR resembles a very good match to the number of features that actually lead to best results. Overall, good classification results have been obtained. Random guessing would result in an accuracy of 10 % on average. Best results reach 85.2 % accuracy when including

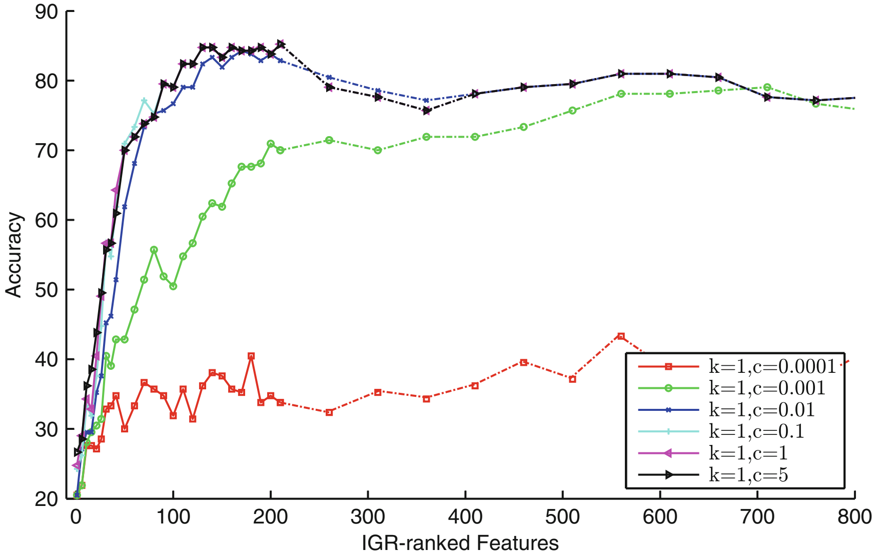

Fig. 5.12 Classification accuracy from incremental SVM-IGR feature selection using a linear kernel for the text-dependent dataset. X-axis shows number of features up to 300 top-ranks, y-axis shows accuracy

all 210 non-negligible IGR top-ranked features. The complexity parameter setting shows that complexities of higher than mid-range values, i.e. higher than 0.01, are beneficial.

Results from the SVM with RBF-kernel function look very similar. Overall, the RBF-kernel shows more sensitivity to non-optimal settings, resulting in a greater loss of performance when choosing non-optimal kernel settings. High complexity and smaller kernel width are beneficial for the classification task. Best results reach 85.7 % accuracy, which is about half a percent better than results with a linear kernel. To reach this accuracy 180 top-ranked features were included. Figure 5.13 shows the plot for selected complexities and kernel widths. Here, the combination of mid-range or low kernel width and high complexity proves to be beneficial for the classification task, but also very small complexities reach reasonable accuracies with respect to a 10 % accuracy base line from random guessing. Because kernel width and complexity settings show mutual influence for the classification performance these parameters need to be determined jointly, here by executing a grid search. The two parameters are plotted into a contour heat map image in Fig. 5.14. From the picture it becomes apparent that highest performance is obtained when combining high complexity with mid-range RBF-kernel width. Follow-up experiments have extended the search range to include even higher and even smaller complexities but did not result in improvements. Note, when increasing the complexity there is a point where all existing data points are included in the model. Further extension does not include new information in these cases.

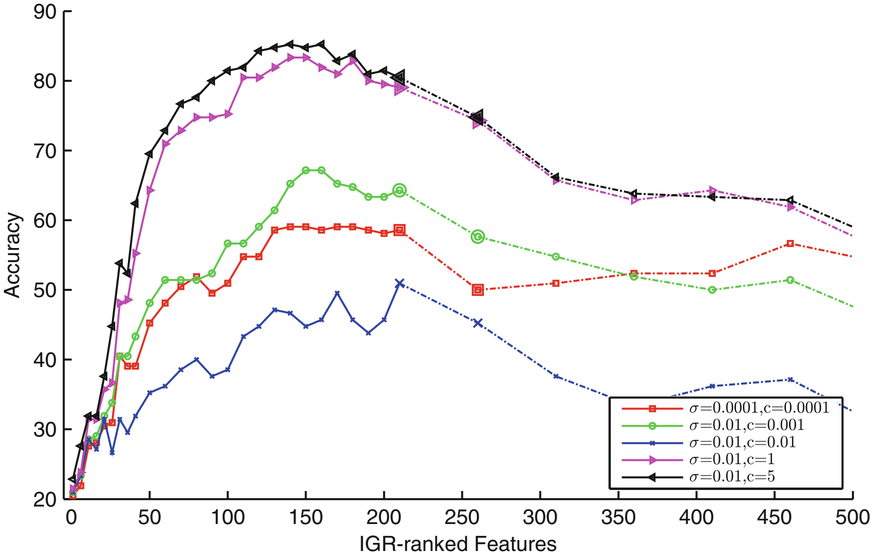

Fig. 5.13 Incremental SVM-IGR feature selection on 500 top-ranked features using RBF-kernels for the text-independent dataset. X-axis shows number of features, y-axis shows accuracy

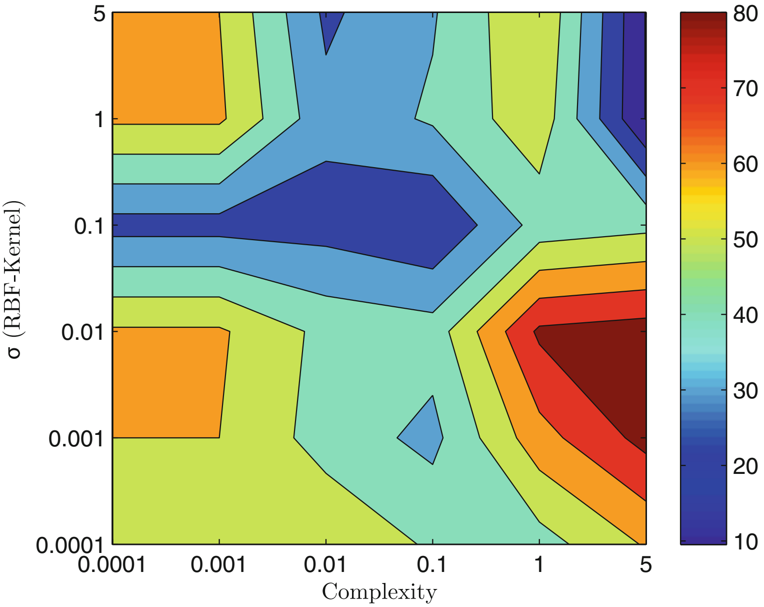

Fig. 5.14 Contour heat map plot using 180 top-ranked IGR features and RBF-kernel for the text-independent dataset. X-axis shows the setting of the complexity parameter c, y-axis shows kernel width σ. Accuracy is indicated by color

Table 5.6 Confusion matrix showing absolute numbers of instances from SVM-IGR using top 210 ranked features and polynomial kernel on text-independent data

		Predicted									
		a+	a−	c+	c−	e+	e−	n+	n−	o+	o−
Actual	a +	15	0	1	0	0	0	2	1	1	1
	a −	0	19	0	0	0	0	0	0	1	1
	c +	3	0	16	1	0	0	1	0	0	0
	c −	0	0	0	21	0	0	0	0	0	0
	e +	0	0	0	0	20	0	0	0	1	0
	e −	0	0	0	0	0	21	0	0	0	0
	n +	3	0	0	1	0	0	15	0	1	1
	n −	1	1	0	1	0	0	0	17	1	0
	o +	0	1	0	0	0	0	0	1	19	0
	o −	3	0	1	0	0	0	0	1	0	16

Rows show the actual class membership, columns represent the classification result

Table 5.7 Class wise statistics for classification of text-independent dataset

Class	Precision	Recall	F-measure
a+	0.6	0.714	0.652
a−	0.905	0.905	0.905
c+	0.889	0.762	0.821
c−	0.875	1	0.933
e+	1	0.952	0.976
e−	1	1	1
n+	0.833	0.714	0.769
n−	0.85	0.81	0.829
o+	0.792	0.905	0.844
o−	0.842	0.762	0.8

Table 5.6 presents the respective confusion matrix for the SVM classifier based on the top 210 IGR ranks and a linear kernel. Table 5.7 shows the recalls, precisions and F-measures for this classifier. Accordingly, **Low Agreeableness**, **Low Conscientiousness** as well as **Low** and **High extroversion** could be classified best. Worst results were obtained in terms of **High Agreeableness**, as these stimuli were confused with neuroticism or openness. Also stimuli containing **High Neuroticism** were misclassified rather frequently followed by **Low Openness**. All other classes reveal good or very good classification results. Finally, also the results from automatic classification of personality based on the text-independent dataset reveal good performance and good general feasibility.

5.9.2 Results from Automatic Trait Score Prediction

In order to generate a ranking of individual features unsupervised discretization as applied for the text-dependent dataset was carried out as described in Sect. 5.8.2. Figure 5.15 shows selected results for openness prediction. Best correlation between human openness ratings and SVM prediction reached 0.42 only. Extension to non-linear kernels did not improve the performance, as can be observed in Fig. 5.15a. RMSE resulted in 4.34 points. Although the shape of the correlation curves indicate an overall smooth ranking, the performance of the SVM regression results in weak correlation only.

Note that the results in terms of correlations can be somewhat independent from results in terms of RMSE. They do not necessarily increase and decrease to the same degree. SVM models trained on text-dependent data showed considerably higher correlation with comparable RMSE, e.g. conscientiousness with RMSE of 4.29 points and correlation of up to 0.68 as well as extroversion with RMSE of 5.15 points and correlation of up to 0.82.

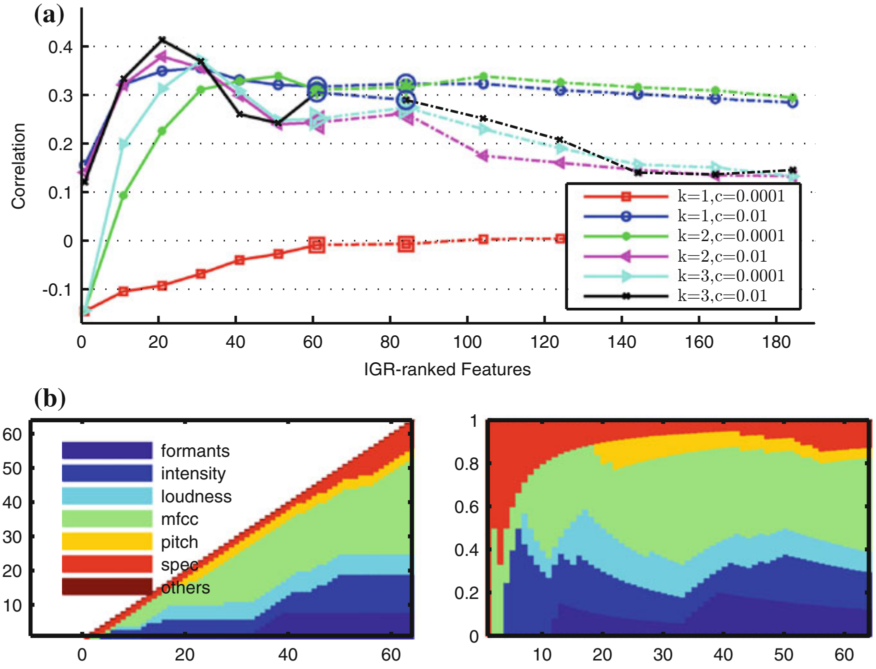

Fig. 5.15 Panel a: Correlation between automatic prediction and human ratings (y-axis) along IGR ranking (x-axis) for openness. Panel b: Stacked audio descriptor groups distribution (y-axis) along IGR ranking for the text-dependent database after discretization openness ratings into high and low splits

In the present case resulting in weak correlation between human ratings and automatic SVM predictions, the ranking plot in Fig. 5.15b should not be interpreted. The question of whether a potentially bad ranking has lead to the weak prediction performance or if the ranking should be taken as accurately and the SVM regression algorithm should be blamed for the weak performance, remains open. Throughout this work, both components, i.e. ranking and SVM modeling, have proven to be well functioning in many cases. Therefore, further experiments need to be carried out in order to analyze this problem to a deeper understanding. After all, if the problem remains, it could well be attributed to the speech properties themselves or to the specific kind of speech generation used to generate the text-independent database. For a more detailed analysis including all results presented in this work please refer to Sect. 6.3.

Basically, results prove similar for conscientiousness, extroversion, agreeableness, and neuroticism prediction. Best results reached correlations of 0.30, 0.41, 0.50, and 0.46 with corresponding RMSEs of 4.48, 4.50, 4.22, and 4.22 points,

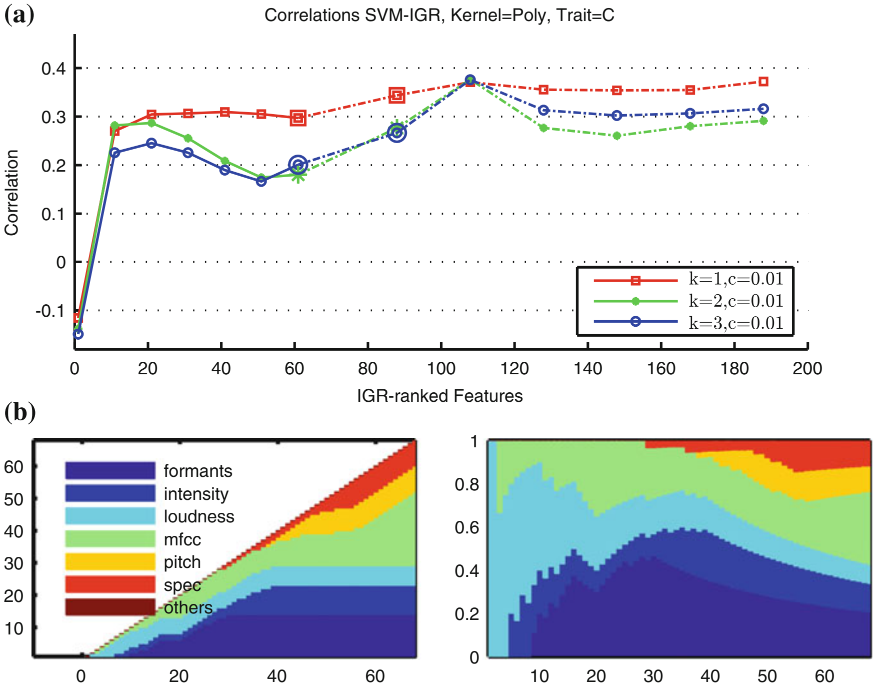

Fig. 5.16 Panel a: Correlation between automatic prediction and human ratings (y-axis) along IGR ranking (x-axis) for conscientiousness. Panel b: Stacked audio descriptor groups distribution (y-axis) along IGR ranking for the text-dependent database after discretization conscientiousness ratings into high and low splits

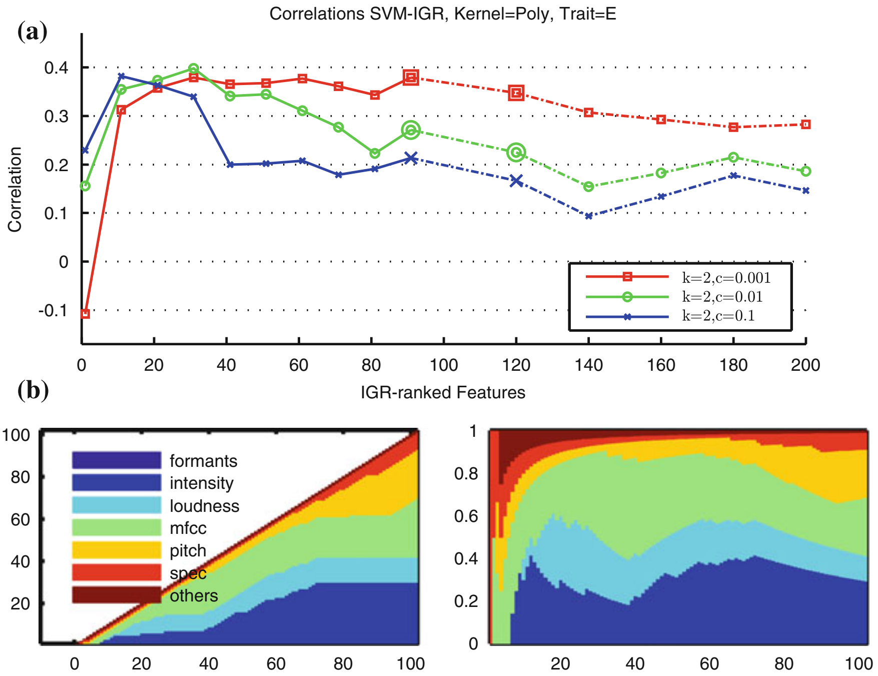

Fig. 5.17 Panel a: Correlation between automatic prediction and human ratings (y-axis) along IGR ranking (x-axis) for extroversion. Panel b: Stacked audio descriptor groups distribution (y-axis) along IGR ranking for the text-dependent database after discretization extroversion ratings into high and low splits

respectively. Figures 5.16, 5.17, 5.18 and 5.19 show the respective plots. Hypothetical explanations for these observations will be given in Sect. 6.3.

5.10 Results from Multi-Speaker Data

When running the experiments for the multi-speaker dataset the results suggest that the influence of the recording condition, i.e. close-talk versus stand-alone capture, must be dealt with separately. While the analysis of human personality assessment did not dismantle any significant difference in the perception of the close-talk and stand-alone microphone recordings, cf. Sects. 4.2 and 4.4, the performance of the automatic classifiers and regression models deteriorate in almost all cases when blending the different recordings conditions into just one group. Therefore, results from the rankings as well as results from prediction experiments are presented separately throughout the next sections. An interpretation of this finding will be given in Sect. 6.3.

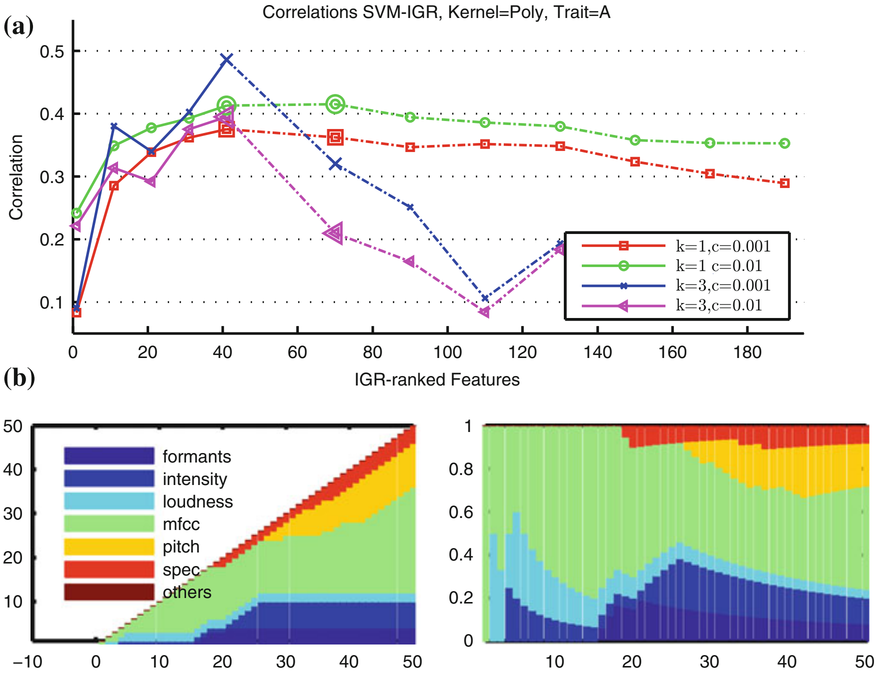

Fig. 5.18 Panel a: Correlation between automatic prediction and human ratings (y-axis) along IGR ranking (x-axis) for agreeableness. Panel b: Stacked audio descriptor groups distribution (y-axis) along IGR ranking for the text-dependent database after discretization agreeableness ratings into high and low splits

Furthermore, the experiments on multi-speaker data cannot be executed as presented for the speaker-dependent data sets due to the lack of category information. Here, it remains unclear which samples to subsume into what class structure. For this dataset, a classification task of one out of 10 classes is not feasible, neither can a supervised IGR be calculated. In order to generate a ranking of individual features unsupervised discretization as applied for the speaker-dependent datasets was carried out as described in Sect. 5.8.2. Still, having binarized each individual scale separately does not tell which of the 5 resulting labels, i.e. accounting for either high or low characteristic of the 5 traits individually, should serve as the one global label for that sample.

However, the strength of this dataset is to provide a realistic basis for trait score prediction experiments. The results from these experiments are presented in the following section.

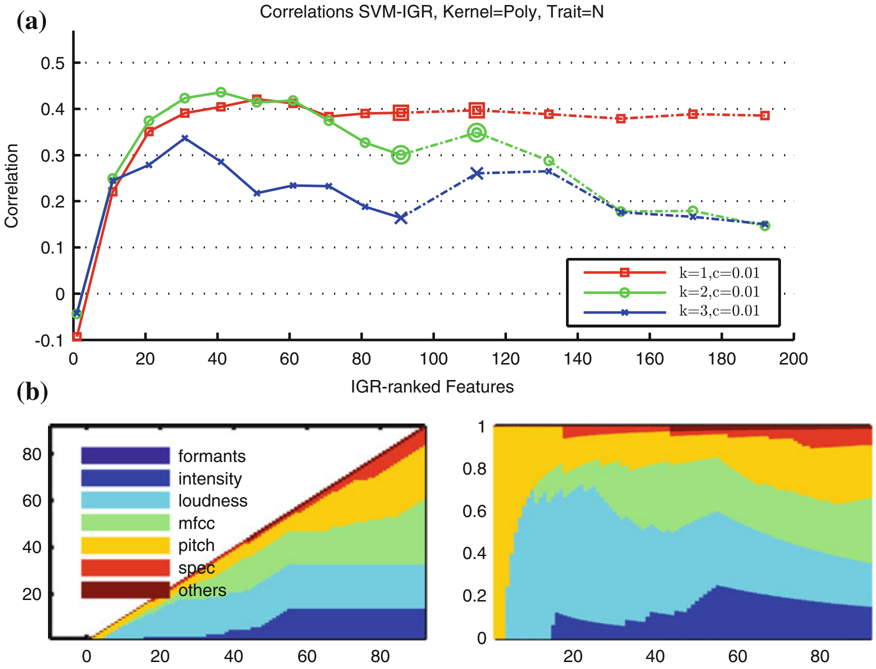

Fig. 5.19 Panel a: Correlation between automatic prediction and human ratings (y-axis) along IGR ranking (x-axis) for neuroticism. Panel b: Stacked audio descriptor groups distribution (y-axis) along IGR ranking for the text-dependent database after discretization neuroticism ratings into high and low splits

5.10.1 Results from Automatic Trait Score Prediction

Figure 5.20 shows the results for openness prediction. Best correlation between human openness ratings and SVM prediction reached 0.71 when including 91 top-ranked features for the close-capture condition. For the stand-alone microphone condition best results show a correlation of 0.63 using 120 top-ranks. RMSEs resulted in 1.85 and 1.95 respectively. In both cases, non-linear kernel function proved to be beneficial, as can be seen in Figs. 5.20a, b. Overall, these results account for moderate and good correlation respectively. The shapes of the correlation curves are overall smooth, indicating a reasonable ranking outcome for both subsets. Also the IGR estimation on number of reasonable features seems to provide a reasonable match. For the close-capture condition the highest performance was obtained exactly at the suggested number of non-negligible features. For the stand-alone microphone recordings the number does not exactly match but results within reasonable deviation. Performance suffers beyond the suggested number of features in general.

When comparing the rankings in Fig. 5.20c, the basic difference in feature composition becomes apparent. The close-capture subset shows few more spectral features in highest ranks. In particular the roll-off point calculated from unvoiced segments

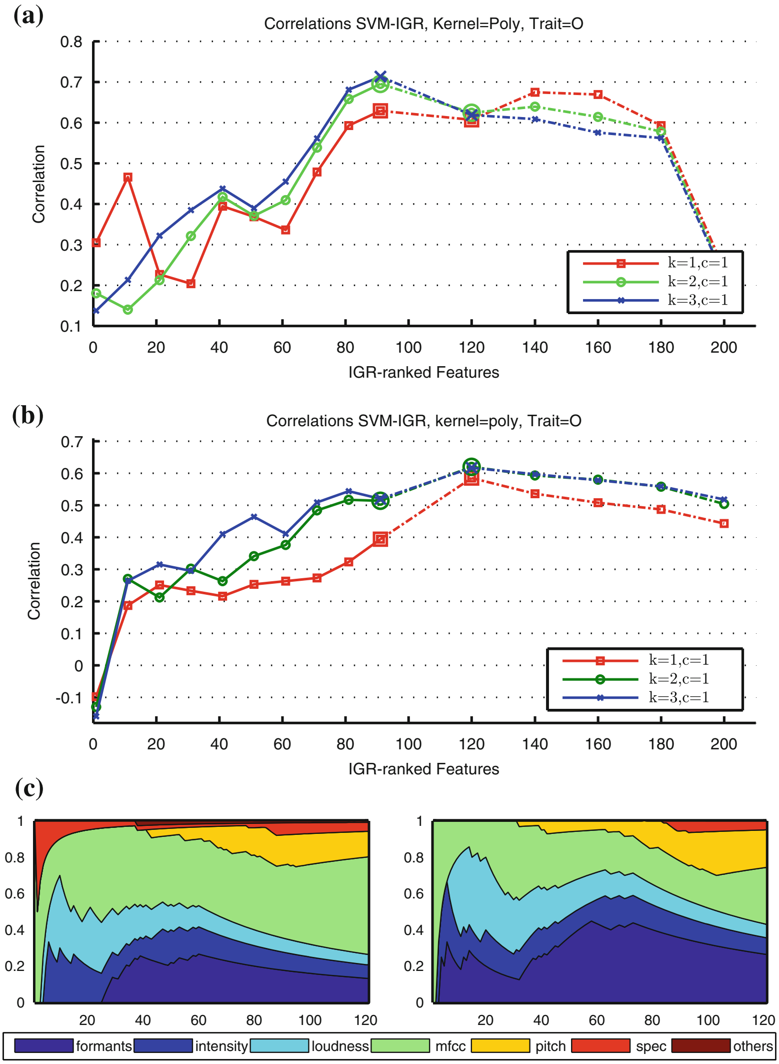

Fig. 5.20 Correlation between automatic prediction and average human ratings (y-axis) along IGR ranking (x-axis) for openness from close-capture (panel a) and stand-alone microphone recordings (panel b). Part c: Stacked audio descriptor groups distribution (cf. Sect. 5.3) on y-axis along expanding number of top-ranks on x-axis for the multi-speaker database (*left* close-capture, *right* stand-alone recording) when discretizing ratings into high and low targets for individual scales

proves to be of high information gain. All other feature groups seem to be comparable in their ranks. MFCCs, loudness and formant statistics are most frequently found in high ranks.

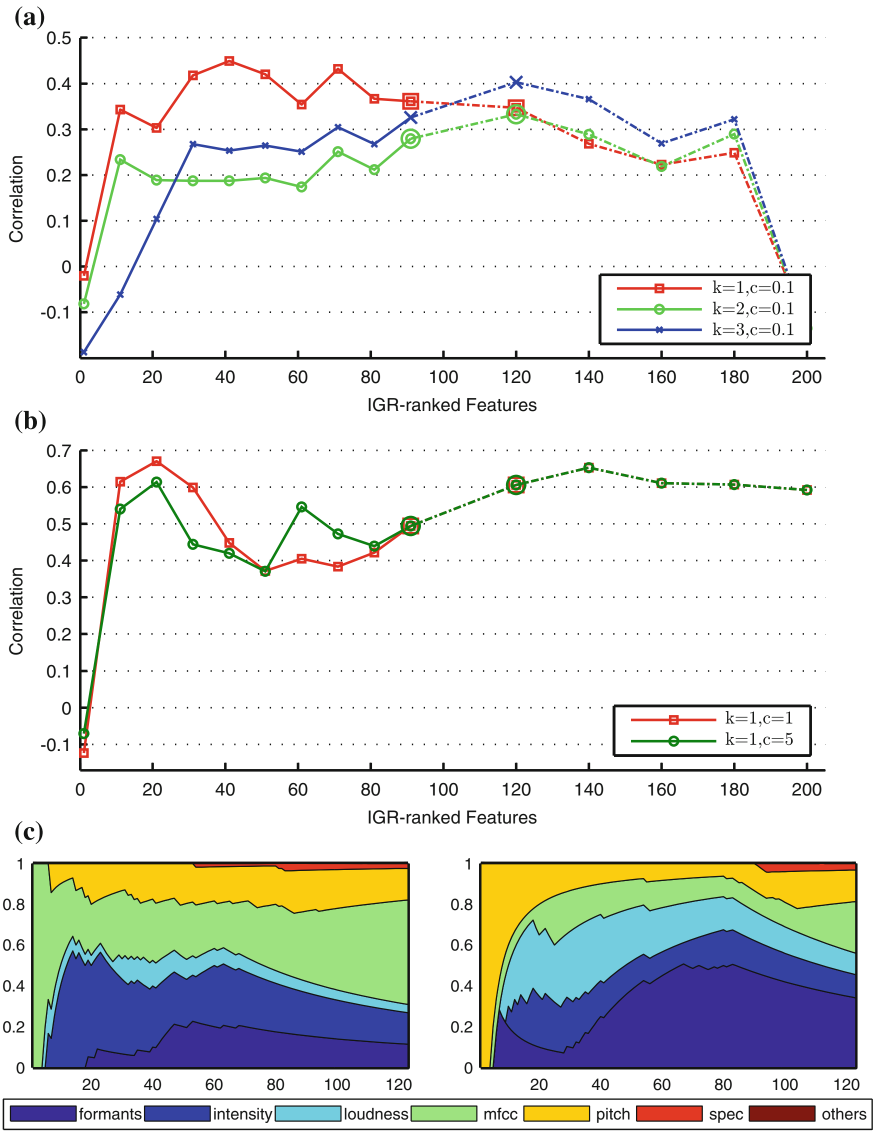

Fig. 5.21 Correlation between automatic prediction and average human ratings (y-axis) along IGR ranking (x-axis) for conscientiousness from close-capture (panel a) and stand-alone microphone recordings (panel b). Part c: Stacked audio descriptor groups distribution (cf. Sect. 5.3) on y-axis along expanding number of top-ranks on x-axis for the multi-speaker database (*left* close-capture, *right* stand-alone recording) when discretizing ratings into high and low targets for individual scales

Figure 5.21 shows the results for conscientiousness prediction. Best correlation between human ratings and SVM prediction reached again moderate and almost good results, i.e. 0.67, when including 21 top-ranked features only. But this is valid only for the stand-alone microphone recordings. Figure 5.21b shows the plot. The corresponding RMSE resulted in 3.66. A very different picture can be obtained when looking at the close-capture condition in Fig. 5.21a. Best results reach a correlation of 0.46 only, using 41 top-ranks. RMSEs resulted in 4.22. In both cases, non-linear kernel function proved to be beneficial, but the overall performance of the SVM model on the close-capture data proves weak. The shapes of the correlation curves are again overall smooth, but this time the suggested number of features did not match the determined optimal number of features.

When looking at the rankings in Fig. 5.21c, the stand-alone microphone data shows pitch features in highest ranks as well as a substantial number of loudness and intensity-related features. The pitch features mostly capture dynamics and also include the error coefficient from linear regression fit. Also all the loudness and intensity-based features capture dynamics, hence slopes and acceleration plays an important role for the prediction of the present data subset.

Results from extroversion prediction are shown in Fig. 5.22. Best correlation between human extroversion ratings and SVM prediction reached 0.79 when including 180 top-ranked features for the close-capture condition. For the stand-alone microphone condition best results show a correlation of 0.60 using 91 top-ranks. RMSEs resulted in 3.36 and 3.36 respectively. After all, extroversion could be modeled with good prediction results on basis of the close-capture recordings. Modeling on basis of stand-alone microphone recordings reach moderate correlations. Non-linear kernel functions proved to be beneficial for the latter only. Plots using different kernels can be obtained from Figs. 5.22a, b. The shapes of the correlation curves are rather smooth, indicating a somewhat reasonable ranking outcome for both subsets. The IGR estimation on the number of reasonable features provided good results for the stand-alone recordings, but did not lead to optimal results for the close-capture condition.

When comparing the rankings in Fig. 5.22c, the most visible difference is the higher proportion of formant-related features in the ranks of the close-capture data. However, most relevant for extroversion prediction are the pitch features, which in turn capture dynamics of pitch in terms of statistics on pitch derivatives and DCT coefficients from pitch analysis. Also MFCCs are of high information gain.

Figure 5.23 shows the results for agreeableness prediction. Best correlation between human ratings and SVM prediction reached again moderate and good results. Models trained on close-captured data reach a correlation of 0.70, when including 71 top-ranked features. The corresponding RMSE resulted in 2.79. Figure 5.23a shows curves using different kernel functions. Here, non-linear extension proves helpful. When looking at the stand-alone microphone condition shown in Fig. 5.23b it can be seen that best results reach a correlation of 0.57 only, this time using 71 top-ranks. RMSEs resulted in 3.56. While the overall performance is lower than the model trained on the close-capture data, also non-linear modeling did not help to improve the performance. In general, the curves show a rather smooth behavior. The suggested

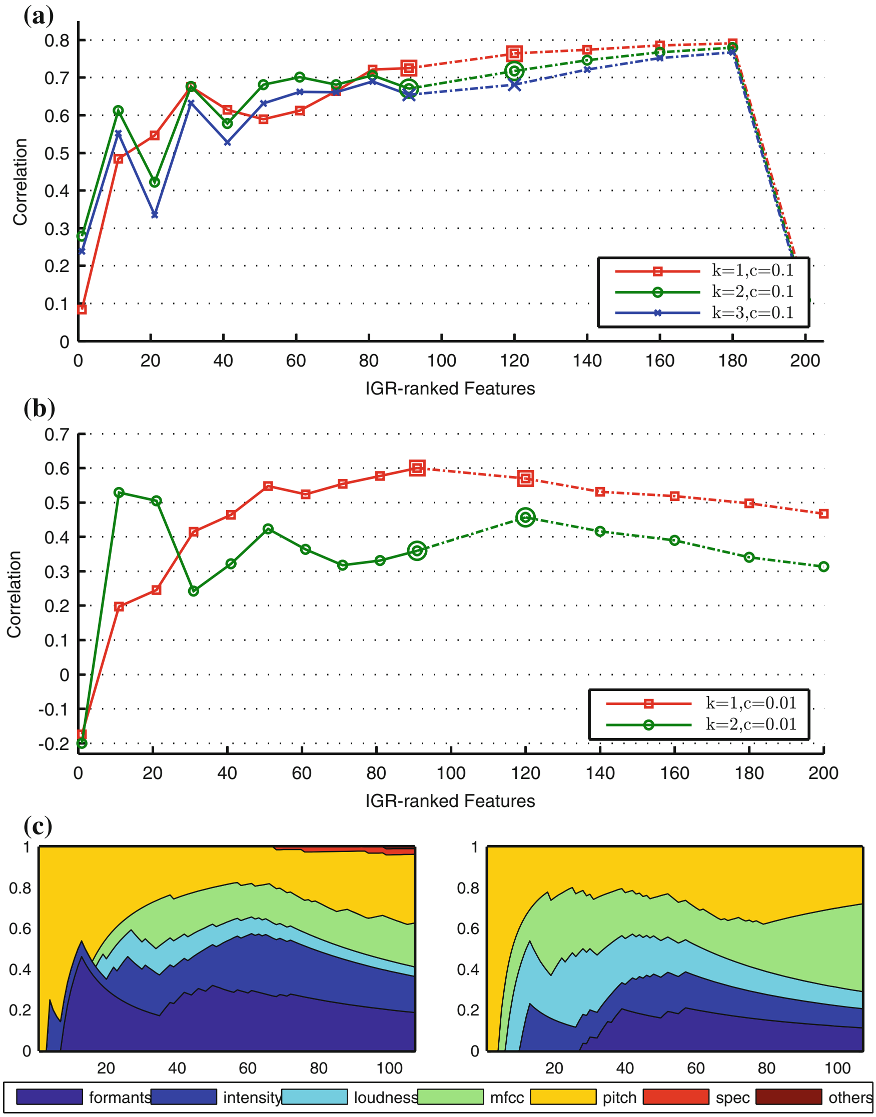

Fig. 5.22 Correlation between automatic prediction and average human ratings (y-axis) along IGR ranking (x-axis) for extroversion from close-capture (panel a) and stand-alone microphone recordings (panel b). Part c: Stacked audio descriptor groups distribution (cf. Sect. 5.3) on y-axis along expanding number of top-ranks on x-axis for the multi-speaker database (*left* close-capture, *right* stand-alone recording) when discretizing ratings into high and low targets for individual scales

number of features after IGR matches the determined optimal number of features quite well.

But when looking at the ranking plots in Fig. 5.23c one clearly sees a huge difference in feature composition in both models. The stand-alone microphone recordings

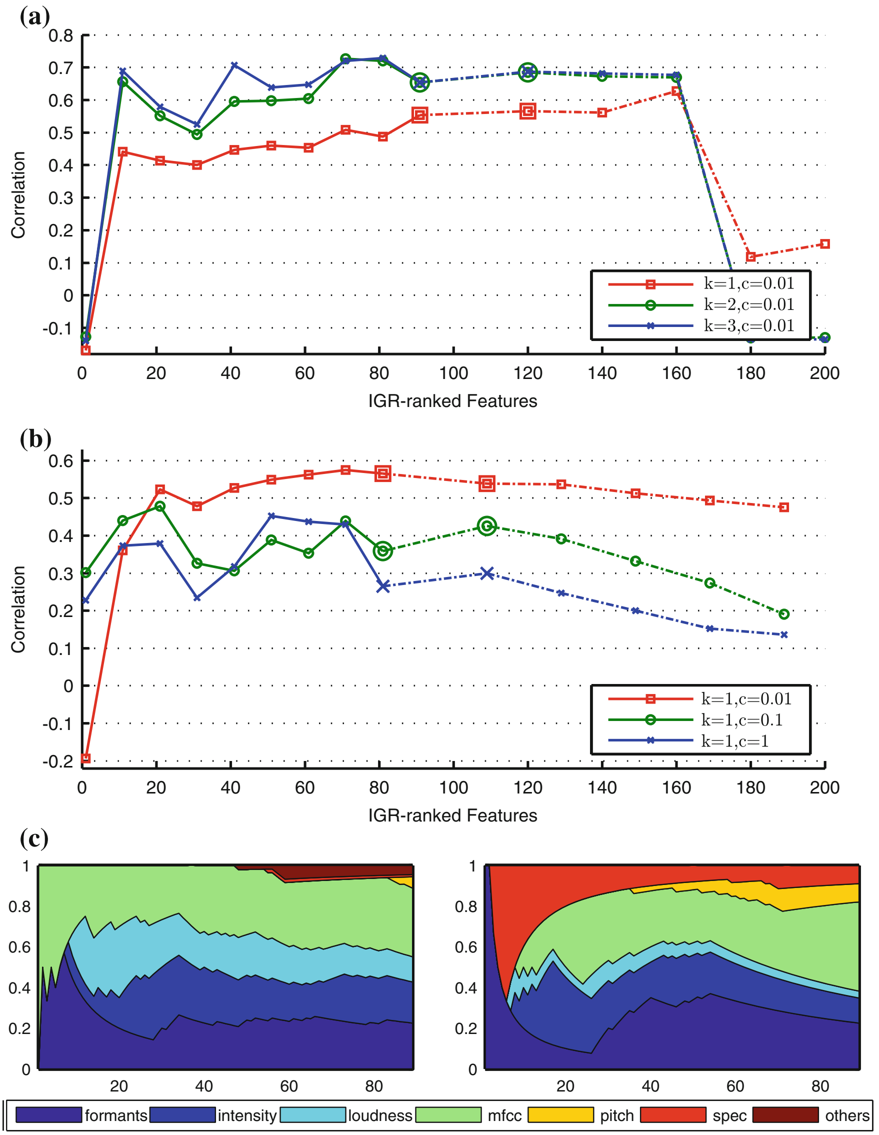

Fig. 5.23 Correlation between automatic prediction and average human ratings (y-axis) along IGR ranking (x-axis) for agreeableness from close-capture (panel a) and stand-alone microphone recordings (panel b). Part c: Stacked audio descriptor groups distribution (cf. Sect. 5.3) on y-axis along expanding number of top-ranks on x-axis for the multi-speaker database (*left* close-capture, *right* stand-alone recording) when discretizing ratings into high and low targets for individual scales

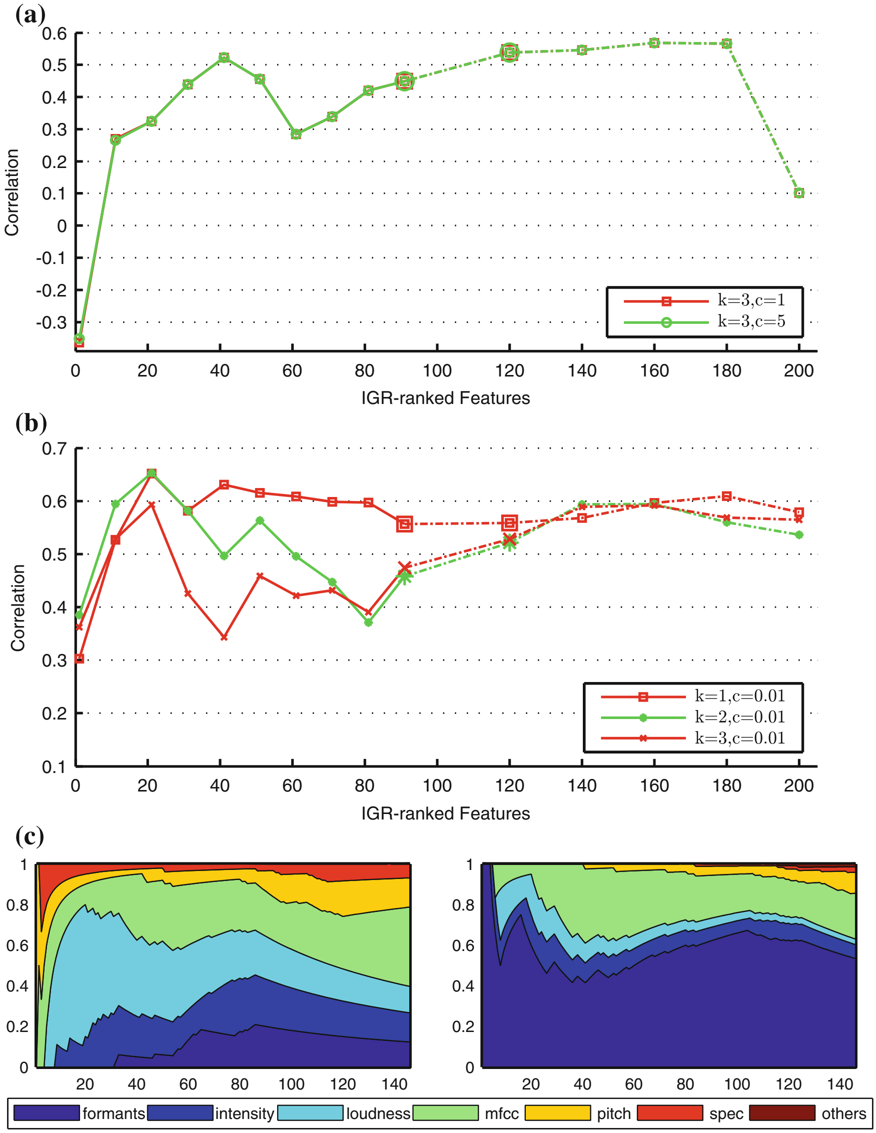

Fig. 5.24 Correlation between automatic prediction and average human ratings (y-axis) along IGR ranking (x-axis) for neuroticism from close-capture (panel a) and stand-alone microphone recordings (panel b). Part c: Stacked audio descriptor groups distribution (cf. Sect. 5.3) on y-axis along expanding number of top-ranks on x-axis for the multi-speaker database (*left* close-capture, *right* stand-alone recording) when discretizing ratings into high and low targets for individual scales

include a substantial higher number of spectral features, most of which capture statistics regarding the maximum change in the position of the roll-off point in time from unvoiced segments or from whole utterances. However, these statistics did not provide enough information as to obtain good overall correlation to human labels. Most important for the close-capture models are MFCCs, followed by loudness, intensity and formant information.

Finally, the results for neuroticism prediction are presented in Fig. 5.24. Best correlation between human neuroticism ratings and SVM prediction reached 0.57 when including 160 top-ranked features for the close-capture condition. For the stand-alone microphone condition best results show a correlation of 0.65 using as few as 21 top-ranks only. RMSEs resulted in 3.69 and 3.44 respectively. For the former model a polynomial kernel of degree three proved to be beneficial, for the latter model a linear kernel obtained best results, cf. Fig. 5.24a, b. Both models obtained moderate prediction correlation. The shapes of the correlation curves are generally smooth but show a certain degree of ripple at the same time. Here, the IGR estimation on number of reasonable features would not lead to optimal results since the empirically determined maxima are far away from the suggested points.

When comparing the rankings in Fig. 5.24c an essential difference becomesbreak obvious. While the close-capture subset shows many different feature groups in high ranks, the stand-alone microphone data reveals many more formant-related features in highest ranks. Here, over 70 % of top-ranks carry mostly statistics on average bandwidth or the center frequencies of the higher formants. Also the error coefficients from linear regression fit of loudness and intensity results in high ranks.

References

Boersma P, Weenink D (2009) Praat, doing phonetics by computer

Burges C (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov 2(2):121–167

Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2(3):27

Cooley JW, Tukey JW (1965) An algorithm for the machine calculation of complex Fourier series. Math Comput 19(90):297–301

Du Mouchel WH, O'Brien FL (1989) Integrating a robust option into a multiple regression computing environment. In: computer science and statistics: Proceedings of the 21st symposium on the interface, Alexandria: VA, USA. American statistical association

Duda RO, Hart PE, Stork DG (2000) Pattern classification, 2nd edn. Wiley, New York

Fastl H, Zwicker E (2005) Psychoacoustics: facts and models, 3rd edn. Springer, Berlin

Fayyad, Irani (1993) Multi-interval discretization of continuous-valued attributes for classification learning. In: Proceedings of the international joint conference on uncertainty in AI, 1022–1027

Fayyad UM, Irani KB (1992) On the handling of continuous-valued attributes in decision tree generation. Mach Learning 8:87–102

Hastie T, Tibshirani R (1998) Classification by pairwise coupling. Annals Stat 26(2):451–471

Herbrich R, Graepel T (2001) A PAC-Bayesian margin bound for linear classifiers: Why SVMs work. NIPS 2000. Advances Neural Inf Process Syst 13:224

Huang X, Acero A, Hon H-W (2001) Spoken language processing. Prentice Hall, Upper Saddle River

Klasmeyer G (1999) Akustische Korrelate des stimmlich emotionalen Ausdrucks in der Lautsprache, Dissertation, TU Berlin. In: Forum Phoneticum, volume 67. Hector Verlag, Frankfurt am Main

Kotsiantis S, Kanellopoulos D (2006) Discretization techniques. A recent survey

Large J (1972) Towards an integrated physiologic-acoustic theory of vocal registers, vol. 28. The NATS Bulletin

Metze F, Batliner A, Eyben F, Polzehl T, Schuller B, Steidl S (2010) Emotion recognition using imperfect speech recognition. In: Proceedings of the annual conference of the international speech communication association (Interspeech (2009)1–6. Makuhari, Japan, IEEE

Metze F, Polzehl T, Wagner M (2009) Fusion of acoustic and linguistic speech features for emotion detection. In: Proceedings of International Conference on Semantic Computing (ICSC (2009) vol. 1. Berleley, USA, CA, IEEE

Niemann H (2003) Klassifikation von mustern, 2nd edn. Springer, Berlin

Nobuo S, Yasunari O (2007) Emotion recognition using mel-frequency cepstral coefficients. Info Media Technol 2:835–848

O'Shaughnessy D (2000) Speech communications: human and machine, 2nd edn. Institute of Electrical and Electronics Engineers, New York

Platt J (1999) Advances in kernel methods: support vector learning, chapter fast training of SVMs using sequential minimal optimization. MIT Press, Cambridge

Polzehl T, Schmitt A, Metze F (2010) Spoken dialogue systems technology and design., Salient features for anger recognition in German and English IVR portalsSpringer, Berlin

Polzehl T, Schmitt A, Metze F, Wagner M (2011) Anger recognition in speech using acoustic and linguistic cues. Sens Emot Affect Facing Realism Speech Process 53:1059–1228

Polzehl T, Sundaram S, Ketabdar H, Wagner M, Metze F (2009) Emotion classification in children's speech using fusion of acoustic and linguistic features. In: Proceedings of the annual conference of the international speech communication association (Interspeech 2009), 340–343, Brighton, England. ISCA

Press W, Teukolsky W, Vetterling W, Flannery B (1992) Numerical recipes in C, 2nd edn. Cambridge University Press, Cambridge

Rabiner L, Sambur MR (1975) An algorithm for determining the endpoints of isolated utterances. Bell Syst Tech J 56:297–315

Schmitt A, Polzehl T, Minker W (2010) Facing reality: simulating deployment of anger—recognition in IVR systems. In: Spoken dialogue systems for ambient environments—lecture notes in computer science, volume V. 6392, 23–48. Springer, Makuhari

Schölkopf B, Smola A (2001) Learning with kernels: support vector machines, regularization, optimization, and beyond (Adaptive computation and machine learning), 1st edn. MIT Press, Cambridge

Schölkopf B, Bartlett P, Smola A, Williamson R (1998) Support vector regression with automatic accuracy control. In: Proceedings of ICANN'98, perspectives in neural computing, 111–116. Springer Verlag

Schuller B, Metze F, Steidl S, Batliner A, Eyben F, Polzehl T (2009) Late fusion of individual engines for imrpoved recognition of negative emotion in speech—learning vs. democratic vote. In: International conference on acoustics, speech and signal processing (ICASSP). IEEE

Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27(3):379–423

Shevade SK, Keerthi SS, Bhattacharyya C, Murthy KRK (2000) Improvements to the SMO algorithm for SVM regression. IEEE Neural Netw 11(5):1188–1193

Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. Stat Computing 14(3): 199–222

Vapnik V, Cortes C (1995) Support vector networks. Mach Learning 20:273–297

Vapnik V (1999) The nature of statistical learning theory (Information science and statistics), 2nd edn. Springer, Berlin

Witten I, Frank F (2005) Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Diego

Chapter 6
Discussion of the Results

Sections 5.8–5.10 presented the results from personality modeling out of the perspective of exploiting different data sets with different inherent characteristics. In order to conclude on general tendencies across differences in data structure, all results that can contribute to either personality classification or trait score prediction are now presented and discussed along a personality-centered perspective.

The interpretation and discussion of the results includes human consistencies in annotation, cf. Sect. 4.2, ranking performance and smoothness of the incremental forward-pass, cf. Sect. 5.4, feature space composition, Sects. 5.8–5.10, and finally, whenever applicable, references and comparisons to other findings from the literature as presented in Sect. 2.2.

In addition, each individual feature in the top-ranks was tested for significant differences between the high and low personality targets on a 5 % level using ANOVA as explained in Sect. 4.4. In case of non-normal feature distribution, a Kruskal-Wallis test with a significance level of 5 % was executed. In comparison to the classical ANOVA, the underlying assumption that the populations to be compared have normal distributions is replaced with the less strong assumption that all samples come from populations having the same continuous distribution, apart from possibly different locations due to group effects. The test compares the medians of the samples, rather than their means and is known as a non-parametric version of the classical one-way ANOVA, and an extension of the Wilcoxon rank sum test to more than two groups, cf. Hollander and Wolfe (1999).

A discussion of influencing factors on the insights and further directions are presented in Sects. 6.3 and 7.2.

6.1 Results from Classification of Personality Targets

The analysis of human labels from the text-dependent and text-independent data suggest that humans are well able to differentiate between the 10 personality classes acted out by the professional speaker, cf. Sect. 2.2. Neither (a) introducing different

T. Polzehl, *Personality in Speech*, T-Labs Series in Telecommunication Services,
DOI 10.1007/978-3-319-09516-5_6

linguistic content, (b) comparing recordings from different recording sessions, nor (c) comparing close-talk and stand-alone microphone recording conditions significantly influences this perception of low and high targets of personality.

Results from automatic classification of personality classes were conducted using a one-out-of-ten multi-class SVM. Thus the random guessing baseline results in 10 % accuracy. The training material consisted of acted speech data, as described in Chap. 3. Experiments included 160 samples of text-dependent recordings and 210 of text-independent recordings. Results were obtained using 10-fold cross-validation.

When keeping the linguistic context, i.e. words and content, constant, the automatic personality classification reaches excellent results. More accurately, when including as few as 16 acoustic and prosodic features only, the classification reaches an accuracy of 96.3 %. Doubling this amount of features leads to accuracies higher than 99 %. Hence, the SVM models was able to perfectly learn the differences between the 10 personalities by looking at an exclusive choice of speech signal features. Also for the text-independent data, which is composed by speech samples containing verbal descriptions of various pictures, good results were obtained. 85.2 % accuracy could be achieved using 210 acoustic and prosodic features. Non-linear extension of the SVM showed only marginal improvement. However, the intention of the experiments was not to push results to absolute peaks, but to explore the feasibility of personality classification from speech. To this end, the results fully supply the evidence that personality can be automatically classified from acted speech.

When looking into the class-wise statistics of both low and high targets from both databases the results can be organized into the following three groups:

1. **Extroversion** shows highest overall F-Measures. The high target stimuli resulted in an average F-Measure of 0.97. The low targets reached a value of 1, i.e. all of the stimuli from both databases were correctly classified and no other stimuli was confused with **Low Openness**. These are excellent results.
2. The group of **neuroticism**, **conscientiousness**, and **openness** reached overall good and very good results. F-Measures result in 0.88 and 0.91 for high and **Low Neuroticism** targets respectively, the corresponding precisions and recalls are balanced. For conscientiousness and openness F-Measures of 0.89 and 0.90 are reached for high and low targets, recalls and precisions are also overall balanced.
3. For **agreeableness** the situation is two-fold. For **Low Agreeableness** the average F-Measure reaches 0.89, precisions and recalls are balanced. But for **High Agreeableness** results are inconclusive. While for the text-dependent data the models perfectly recall all samples without any confusion, for the text-independent data results are only of moderate magnitude, i.e. 0.65. Since both precision and recall are equally moderate, no systematic bias is observable.

Regarding the acoustic and prosodic features, a ranking was obtained by estimating the Information-Gain-Ratio, cf. Sect. 5.4. This ranking suggests two kinds of information: (a) an ordered list of features due to information contribution; and (b) a number of features that actually contribute considerable amounts of information.

When classifying along an increasing number of top-ranks, the results show generally smooth performance curves and a steep incline at the beginning. After reaching the predicted number of non-negligible features the performance is expected to either drop immediately or stay at a constant level for a certain number of additional features and then start to decline. The observed curves, shown in Fig. 5.5, as wells as in Figs. 5.12 and 5.13 depict the expected behavior. Finally, the suggested number of features to include into modeling exactly matches the empirically determined number for the text-independent data. For the text-dependent data very few features are sufficient for the classification task at hand, and the estimated number of 700 non-negligible features reveals too high. However, the classification accuracy does not decrease when including all 700 features, still the computational costs incurred are unnecessary. Eventually, the IGR ranking proves to be very helpful in predicting a useful subset of features out of the presented repertoire of acoustic and prosodic features, cf. Sects. 5.1–5.3, when applied to the speaker-dependent acted databases. Figures 5.4 and 5.11 give deeper insights into the composition of the suggested feature spaces. For a detailed explanation please refer to Sects. 5.8–5.10.

In more detail, the mean increase of positive slopes as well as statistics on variation and ranges, e.g. IQR, were of high information in terms of pitch for both models. Also the flatness of intensity and loudness contours as well as their statistical distribution turned out to be of high information. Further, the spectral centroid as well as the mean roll-off point were found in high ranks in terms of spectral features. In addition, speech-to-pause ratio as well as the duration of speech and mean duration of voiced segments were of high value. Finally, statistics from formant #4 and #5 are of general high information gain as well as mean values of standard deviations from a number of MFCCs. Overall, the examination of voiced segments or the whole utterances proved helpful.

To give an example of the actual feature values and their interpretation, the feature capturing the IQR of the first derivative of the pitch is analyzed using a Kruskal-Wallis test. Accordingly, there exist significant differences between many of the feature values from the 10 personality classes. Interestingly, the group of **High Openness**, **High Conscientiousness** and **High Extroversion** shows significantly higher IQR of derivatives as all other classes apart from **High** and **Low Agreeableness**, which account for intermediate values. The group mean of the former states an IQR of 25 semitones, while the mean of the latter results in 15 semitones. In other words, when looking at the speed of pitch gestures, there is significantly more dispersion within the former group. Most interestingly the highest values are found for high-personality targets. Consequently, and with neuroticism being the exception, all high targets show clearly higher dispersion of pitch gestures speed. More specifically, a related feature capturing the slopes suggest, that it is predominantly the rising slopes, increasing with a median speed of approximately 13 semitones per second, that lead to such high dispersion.

Comparing these findings from data-driven feature selection to the expert's opinion presented in Table 4.1 these findings account for the perceptual impression with the exception of **High Conscientiousness**. Here, pitch variation was judged to be damped, while the above mentioned features suggest that there is high dispersion of pitch gestures. Also perceptually neuroticism was found to be an exception to this tendency. Here both, the perception and data-driven approaches find the pitch variation to be of low or damped magnitude.

As another example, the behavior of the mean value of all calculated spectral centroids taken from voiced sounds of the utterances shall be discussed. According to the Kruskal-Wallis test, **High Extroversion** shows the highest median point, namely 436 Hz, which is significantly higher than all other targets. It is also non-significantly higher than **High Conscientiousness**, **High** and **Low Openness**, as well as **Low Agreeableness**. The mean of the resulting low-centroid group results in 200 Hz roughly. As a hypothesis, this observation can be caused by the exceptionally high variation in both pitch and intensity for **High Extroversion**. Prosodically, this can cause a more bright perception of voiced speech sounds, which manifests itself in a lift-up of the higher speech frequencies. This would potentially also take effect on the spectral centroid. Note that the observation is made for voiced sounds predominantly. Lifting-up unvoiced sounds might lead to a different perceptual impression of sharper articulation. But in order to conclude on trait-specific dependencies individual trait models having features ranked due to annotations on the single trait at hand are needed. This setup corresponds to the individual trait score prediction experiments presented in following Sect. 6.2.

Plots of results from the comparisons between all 10 classes as well as more detailed analyses between the 10 groups and the number of non-negligible IGR-ranked features are out of scope for the current presentation and therefore omitted in this work.

Overall, when comparing the findings from the three major groups above to the literature, also Mohammadi et al. (2010), Mohammadi and Vinciarelli (2011) report that in their binary classification setup containing the Big 5 personality traits extroversion and conscientiousness reached best recognition. Recognizing openness, however, revealed most difficult in her work. The presented experiments in this work also indicate good and very good results for openness classification, as well as for neuroticism classification. However, note that due to essential differences in the data structure and labeling procedure the results cannot be compared directly. Also Ivanov et al. (2011) were able to classify high versus **Low Extroversion** and **High** versus **Low conscientiousness** with significant improvement over a random guessing baseline only. Experiments on other scales failed to exceed the baselines. Also here, a direct comparison is not applicable due to the underlying differences in data at hand and labeling scheme. Finally, also Mairesse et al. (2007) found that extroversion can be modeled best, followed by emotional stability (neuroticism) and openness to experience. Again, the database differed essentially and the personality questionnaire applied was yet another. In addition, the authors collected self-assessments for their work. Details on all mentioned works and databases as well as on the respective labeling schemes are presented earlier in this work. For a more accurate enumeration of differences please refer to Sect. 2.2.4. None of the presented works analyze the features of the models in order to discover significant differences between the chosen personality targets. In the presented work, these results contribute new insights for all the Big 5 traits from an integrated classification task perspective in a systematic approach. Moreover, specifically results on agreeableness are novel. More insights into a trait-wise perspective on personality is presented in the next Sect. 6.2.

6.2 Prediction of Individual Traits Scores from Speech

When designing the speech databases, the degrees of freedom were assumed to be opening up from text-dependent recordings, towards text-independent recordings to a maximum degree of freedom with the multi-speaker recordings. At the same time, results were expected to increase from multi-speaker data, towards text-independent data, and reach highest accuracies for the text-dependent data. Obviously, the results from the prediction experiments do not follow this assumption. The overall results from the text-independent recordings are much lower than the results from the multi-speaker recordings, cf. Sects. 5.8.2, 5.9.2, and 5.10. Hence, other factors must have counteracted the assumed trend. For a consideration of potential factors of influence refer to Sect. 6.3.

Experiments from automatic prediction of personality scores were conducted on all three databases, i.e. text-dependent, text-independent, and multi-speaker recordings. Experiments included 30 text-dependent recordings, 210 text-independent recordings and 64 multi-speaker recordings. All these recordings were annotated with continuous NEO-FFI score values from the listening tests. Modeling results were obtained using 10-fold cross-validation. Due to low modeling performance, the rankings from text-independent data are disregarded from the following analysis. Because the models did not successfully capture enough relevant information to reach at least moderate correlation, i.e. correlation higher than 0.6, it remains unclear if the rankings provide useful information or not. On the one hand, the rankings could in principle still provide useful information, which then might not be learned by the SVM model in a proper way. On the other hand, the ranking might as well be corrupt and the SVM would not be able to learn systematically in any case. This interdependency cannot be made transparent for analysis or interpretation in the current work. A deeper analysis of the degraded results remains future work.

In the next sections, the actual rankings from best-performing models are compared to each other in order to conclude on general tendencies. Also individual models and features are not included in the comparison, if the respective models performed worse that with correlation of 0.6 to human annotations. This will also be indicated in the respective sections.

With regard to interpretation and comparison of the achieved results to other works, the following sections draw comparisons to related works whenever possible. As introduced in Sect. 2.2, most literature can be found for extroversion and neuroticism. These literature almost exclusively delivers descriptive analyses only. To the best of the author's knowledge, no other literature on actual trait score prediction using acoustic-prosodic features and signal-based modeling exist to date. While in Sect. 6.1 very few works on personality classification were available for comparison, here a comparison to other literature is not possible at all. The author provides such a basis for comparison to the research community with the current work.

6.2.1 Prediction of Openness

When trying to automatically predict an openness personality value from speech one has to take into account that this trait has revealed to be most difficult to assess by humans in the first place. Consistencies between the labelers from all datasets, i.e. text-dependent, text-independent as well as multi-speaker recordings, resulted in lowest magnitudes for all the respective datasets, cf. Table 4.2. The intuitive hypothesis that matches this finding would be that if humans are insecure about assessing openness from speech, machines cannot do better.

But in fact, when comparing the results from openness prediction to results from other trait predictions the overall results are of good magnitude. Best results reached correlations of 0.89 for text-dependent data, and 0.71 for multi-speaker data.

The average of the best SVM predictions across the text-dependent and multi-speaker databases results in 0.78. While for text-dependent data a linear kernel function proved to be beneficial, on the multi-speaker data non-linear kernels lead to moderate improvement. The IGR proved overall good results for ranking since the respective curves in Figs. 5.6 and 5.20 are smooth. Especially with regard to smoothness, non-linear mapping seems to be beneficial for the more realistic multi-speaker data.

From the analysis of the joint rankings, which include the text-dependent ranking as shown in Fig. 5.6b, and the two separated rankings from the close-capture and stand-alone microphone recordings from the multi-speaker subset as shown in Fig. 5.20c, two main groups become apparent, namely MFCCs and formant-related features. When looking at the individual features, the majority of the MFCC features capture diverse statistics. No systematic trend can be observed. The formant-related features focus on mean or median statistics, mostly drawn from the bandwidth of the first and fifth formant. Especially when looking into the realistic multi-speaker data also pitch features are of importance. These features capture the speed of change and the variation in pitch, but no significant difference was found by ANOVA.

Comparing these findings from data-driven feature selection to the expert's opinion presented in Table 4.1 two corresponding findings can be stated. First, the perceptual impression of high pitch variation and relative differences in pitch can be matched to the automatically selected pitch features. Second, the perceptual difference in voice quality, i.e. the rather tensed voice impression, can be matched to the features capturing the bandwidths of formants. The perceptual differences with regard to speech intensity did not manifest in the features at hand. Note, all other differences found in perception are of more subtle or more inconsistent nature. These characteristics can be expected to be either difficult to model with the presented SVM models or difficult to capture with the presented acoustic-prosodic features, or both.

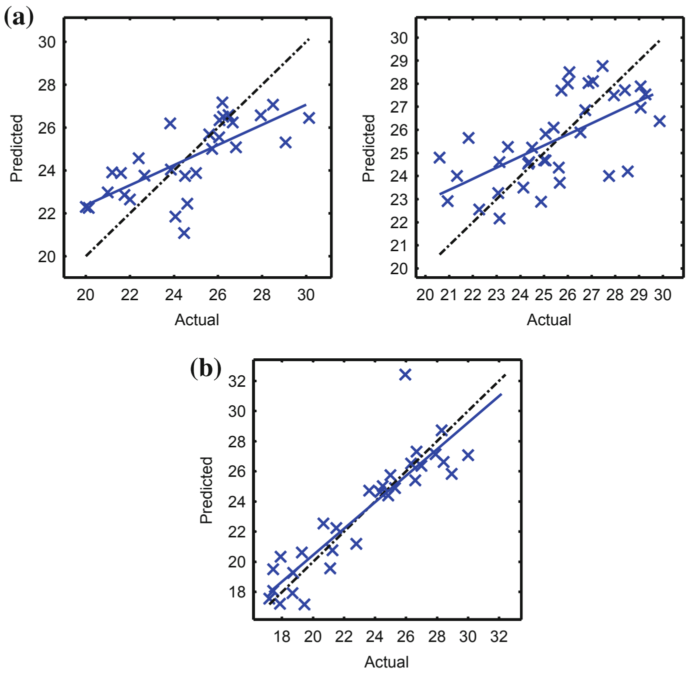

Fig. 6.1 Scatter plot of human ratings (x-axes) and SVM-predicted trait scores (y-axes) on multi-speaker data (*panel a*) and text-dependent data (*panel b*) for openness. *Dashed black line* represents perfect prediction, *blue solid line* gives the actual least-squares regression fit. **a** Human ratings (x-axes) and SVM-predicted trait scores (y-axes) for close-capture recording (*left*) and stand-alone microphone recordings (*right*). **b** Human ratings (x-axes) and SVM-predicted trait scores (y-axes) for text-dependent recordings

When looking at the scatter plots in Fig. 6.1 no systematic bias can be seen. Figure 6.1a shows the plot for close-captured speech recordings to the left, and results for stand-alone microphone recordings to the right. RMSEs resulted in 1.84 and 1.93, correlation to human predictions resulted in 0.71 and 0.62 respectively. For the close-talk dataset samples of lower human ratings are estimated too high, while higher human scores are estimated too low. For the stand-alone microphone recordings this tendency becomes weak. Here, the prediction error seems to be distributed more evenly. Figure 6.1b shows the same plot for the acted text-dependent data. Here, a generally higher dispersion of trait scores for both human labeling and SVM prediction can be seen. RMSE resulted in 1.79, correlation between human annotations and predicted scores resulted in 0.89.

The blue solid lines in the plots present a linear least-squares regression fit. As a post-processing step, these fit lines can be utilized to adapt the SVM output in order to better match the human ratings. The more systematic these deviations from the diagonal, the more promising a post-processing re-alignment appears. For example, for the close-capture data a re-alignment seems to be more promising than for the text-dependent data. For the present scope of experiments, these steps are omitted.

6.2.2 Prediction of Conscientiousness

Human annotations on conscientiousness personality values from speech shows very high overall consistency, cf. Table 4.2. When comparing these results to automatic conscientiousness prediction the models perform on moderate levels only. Best results reached correlations of 0.68 for text-dependent data, and 0.67 for multi-speaker data, each time including around 40 top-ranked features only.

The IGR proves overall applicability but shows unusual nodges and peaks at the same time, cf. Figs. 5.7, and 5.21. This indicates a non-optimal ranking, which can be the cause for the relative low model performance. A deeper discussion on how to improve the IGR ranking can be found in Sect. 6.3.2. With regard to performance robustness, linear mapping seems to be beneficial.

Close-talk recordings are excluded from the analysis of the joint rankings due to the overall weak performance. Eventually, the comparison comprises the text-dependent ranking as shown in Fig. 5.7b and the and stand-alone microphone recordings from the multi-speaker subset as shown in Fig. 5.21c. As a result, three dominating groups can be seen. Most important are pitch features. Here, the behavior of slope movements and statistics on variation turn out to be speaker-dependent to a certain degree. For example, while the difference between **High** and **Low Conscientiousness** became significant for the standard deviation, the pitch range or the median increase of rising slopes for the acted data, this behavior can only be identified as non-significant trend in the multi-speaker data. The same becomes obvious when looking at intensity and loudness-related features, e.g. the median loudness of voiced points. A more detailed analysis of individual features and trends is out of scope for this presentation.

Looking into the expert's opinion in Table 4.1 the perceptual difference in relative pitch and intensity matches the corresponding features from pitch, loudness and intensity. Note, other dominant perceptual differences between **High** and **Low Conscientiousness** effect tempo, rhythm and voice quality. None of these characteristics match high-ranked features. These characteristics could not be captured, eventually.

When looking at the scatter plots in Fig. 6.2 no systematic bias can be seen for the multi-speaker data. The Figure shows the plot for stand-alone microphone recordings to the left, and the text-dependent recordings to the right. RMSEs resulted in 3.94 and 3.73, correlation to human predictions resulted in 0.63 and 0.74 respectively. For the text-dependent data, the scatter shows the same trend like obtained from openness prediction, namely samples of lower human ratings are estimated too high, while higher human scores are partly estimated too low. Specifically for values between 28 and 35, which resemble **High Conscientiousness**, the SVM model shows a bias. It dominantly predicts 30 within this range. Why this bias only appears within this range and how to improve the model remains for future work. Again, a post-processing

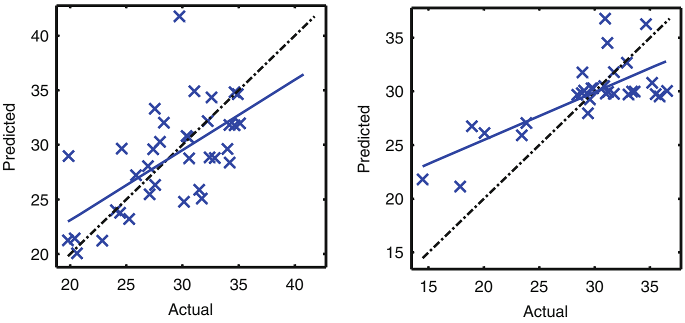

Fig. 6.2 Human ratings (x-axis) and SVM-predicted trait scores (y-axis) for stand-alone microphone recordings (*left*) and text-dependent data (*right*) for conscientiousness. *Dashed black line* represents perfect prediction, *blue solid line* gives the actual least-squares regression fit

step, that aligns the linear least-squared fit (blue lines) seems promising for the text-dependent data, but remains out of scope for this work as well.

6.2.3 Prediction of Extroversion

When looking at human consistencies in extroversion assessment from speech the overall consistencies result in rather moderate magnitudes, cf. Table 4.2. When comparing these results to automatic extroversion prediction the models perform on good levels. Best results reached correlations of 0.82 for text-dependent data, and 0.79 for multi-speaker data.

In fact, the average of the best SVM predictions across the text-dependent and multi-speaker databases results in 0.81, which is the best average trait prediction obtained in the comparisons. The IGR proved overall good results for ranking, cf. Figs. 5.8, and 5.22, although smoothness decreases when using non-linear kernel functions.

The analysis of the joint rankings include the text-dependent ranking as shown in Fig. 5.8b, and again the two separated rankings from the close-capture and stand-alone microphone recordings from the multi-speaker subset as shown in Fig. 5.22c. The plots look very similar for the datasets. When comparing all the rankings, pitch clearly proves dominant. To give an example, the IQR from pitch movement shows significant differences in all datasets. The dispersion and variation of pitch seems to be an important cue. While for acted data **Low Extroversion** shows a mean IQR of 2.8 semitones, **High Extroversion** reaches almost 6 semitones IQR on average. This distance shrinks when looking at the multi-speaker data, but still reaches an average of 3.5 semitones for low targets and 5.6 for high targets. Also many slope features behave in a similar way, for example the median of falling slopes. **High Extroversion** shows steeper pitch slides downwards, i.e. 3.6 semitones

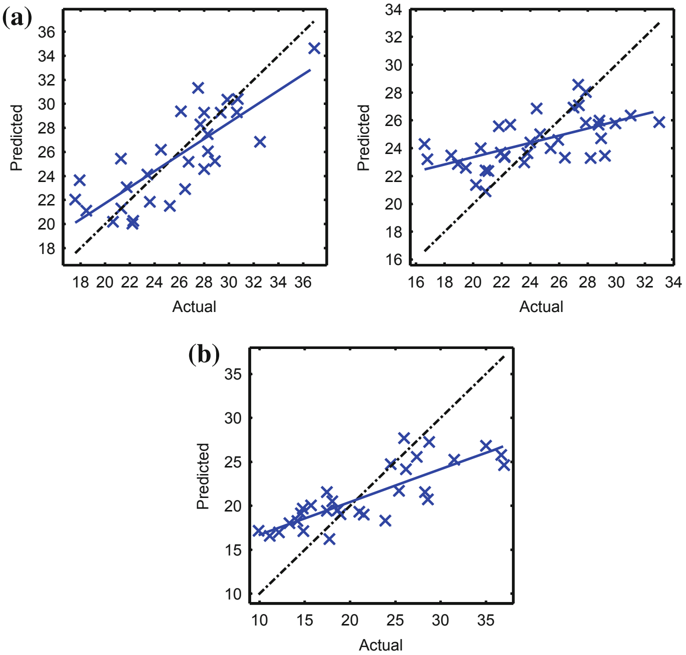

Fig. 6.3 Scatter plot of human ratings (x-axes) and SVM-predicted trait scores (y-axes) on multi-speaker data (*panel a*) and text-dependent data (*panel b*) for extroversion. *Dashed black line* represents perfect prediction, *blue solid line* gives the actual least-squares regression fit. **a** Human ratings (x-axes) and SVM-predicted trait scores (y-axes) for close-capture recording (*left*) and stand-alone microphone recordings (*right*). **b** Human ratings (x-axes) and SVM-predicted trait scores (y-axes) for text-dependent recordings

per second for acted data, and even above 4 for non-acted data, although this difference to the low target does not become significant for the close-talk database. There is no clear structure with regard to the MFCC statistics, apart from the observation that differences and significance generally shrink when comparing acted data to realistic data. Regarding loudness and intensity-related features, most statistics capture the flatness of the contours, i.e. in terms of error coefficient from linear regression or the standard deviation and its dynamics.

Comparing these data-driven findings from feature selection to the expert's opinion presented in Table 4.1 overall congruency can be stated. In the table, perceptually relative pitch and pitch variation have been assessed as a dominant feature. Also a rather big difference in intensity level could be perceived, specifically with regard to the variation of the perceived intensity. Only the difference in perception of rhythm and pauses could not be captured by the features.

When interpreting the scatter plots in Fig. 6.3 the trend from the results of hitherto presented scatter plots continues. In all three figures, i.e. Figure 6.3a showing the plot for close-captured speech recordings to the left, and results for stand-alone microphone recordings to the right, as well as Fig. 6.1b showing the same plot for the acted text-dependent data, samples of lower human ratings are estimated frequently too high, while higher human scores are estimated mostly too low. For the close-capture recordings this tendency is weakest. Here, the prediction error seems to be distributed more evenly. RMSEs resulted in 2.75, 3.35, and 5.14, while correlation to human predictions resulted in 0.79, 0.60, and 0.82, respectively. Again, a bias in the models can be seen, here for text-dependent and stand-alone microphone data. In both cases the predictor's ranges are too narrow. Post-processing in from of linear re-alignment by regression fits seems promising, specifically for the stand-alone microphone recordings.

Comparing these results to the literature, Addington (1968) reached consistencies of 0.64 only when collecting human ratings on four trained speakers while reading a text passage. In those study extroversion ratings were of lowest consistencies. While the actual figure cannot be compared, the relative level of human annotation consistency to other trait's consistencies results moderate or low for both Addington's study and the current work. Regarding modeling, Mairesse et al. (2007) found that extroversion can be modeled best, which meant better than neuroticism and openness in the context of his work. Most prominent features contained pitch mean and intensity variation as well. Further, Mallory and Miller (1958), Scherer (1977) found, that extroversion and high intensity are positively related. Regarding speech fluency constituents, i.e. pauses, American extroverts were found to speak with fewer pauses, also including fewer filled pauses. On the other hand, specifically for the German personality expressions, Scherer (1974, 1979) found that German extroverts are observed to show more silent pauses. Eventually concluding on pitch and intensity, Scherer suggests that extroverts speakers speak louder, and with fewer hesitations. Furthermore, he declares that extroversion is the only factor that can be reliably estimated from speech. To the contrary, the present study provides evidence that other factors can reliably estimated as well. Finally, also Apple et al. (1979) as well as Buller and Aune (1988), Pittam (1994) and Tusing (2000) declare that mean amplitude is positively associated with extroversion.

Note that while in general the results are not refuted by results from the current work, here it is specifically the variation of pitch and intensity, that shows importance from the data-driven analyses and the expert's opinion, along a generally higher pitch and loudness level. Finally, also Aronovitch (1976) reported on correlations between intensity variance and extroversion.

6.2.4 Prediction of Agreeableness

Human assessment of agreeableness from speech results in good overall consistency, which turns out to be rather average when compared to other traits, cf. Table 4.2. When comparing these results to automatic agreeableness prediction the models reach overall fair performance. Best results reached correlations of 0.82 for text-dependent data, and 0.70 for multi-speaker data. The latter accounts for the performance of models trained on close-capture recordings. The performance of models trained on stand-alone microphone recordings did not reach a moderate level and will be disregarded.

Results from the non-linear mapping are inconclusive since it helped to increase performance on the realistic data, but led to a loss on the acted data. For both data, the IGR shows a very steep incline, which is then followed by some ripple, cf. Figs. 5.9, and 5.23. Thus the ranking seems to be non-optimal. In both cases the performance drops drastically when including too many features.

The analyses of the joint rankings include the text-dependent ranking as shown in Fig. 5.9b, and the rankings from the close-capture recordings from the multi-speaker subset as shown in Fig. 5.23c. Clearly, MFCCs build the dominant feature group. With respect to MFCCs mainly voiced and unvoiced segments are captured. Other than that the coefficients account for a fair cross-section of dynamics and statistics from the feature repertoire. No further systematic insights can be given in this work.

But also some formant, loudness and features from the group of *other* occur in the joint ranks. For example, the speech-to-pause ratio shows significant difference between a ratio of roughly 11:1 for **Low Agreeableness** and 17:1 for **High Agreeableness** when looking at the acted data. On realistic data the same trend can be observed with 15:1 on **Low Agreeableness** and 20:1 for **High Agreeableness**, but no significance is found on the 5 % level. The same behavior can be seen when looking at the harmonics-to-noise ratio. Also the total duration (after cropping initial and final silence) shows this behavior with almost 18 s for **High Agreeableness** and 10 s for **Low Agreeableness** as significant difference on acted data. This suggests that agreeable personalities speak more than non-agreeable personalities. On realistic data the respective trend results in 16 s for the low target and 20 s for the high target. The opposite situation can be found for the feature capturing the correlation between pitch and intensity movement. While on acted data this feature shows a strong trend only, on realistic data the difference becomes significant, i.e. **Low Agreeableness** shows a higher correlation of above 0.50, when compared to **High Agreeableness** with 0.39. Similar behavior can be found for, e.g., the ratio of mean loudness from voiced to unvoiced sounds, the range of the roll-off points in the frames, and the overall flatness of intensity as captured by the error coefficient from linear regression modeling of intensity contours. With respect to the formant features, the mean center frequency and mean bandwidth of low agreeable speech samples prove significantly higher than for **High Agreeableness**, valid for both acted and realistic data. This observation becomes significant for the first formant and results in trends for other formants as well. As a hypothesis, the lengthening of the throat as achieved by, e.g., lowering the larynx[1] can be expected to cause a general decrease of all formant frequencies, cf. Neppert and Pétursson (1986). On the other hand, a smile can cause all formant frequencies to increase, cf. Szameitat et al. (2011), Scherer (1995), Quené et al. (2012), which is expected to happen more frequently with agreeable, softhearted, trusting, and helpful motivating speech.

[1] The larynx is an organ situated at the bottom of the throat.

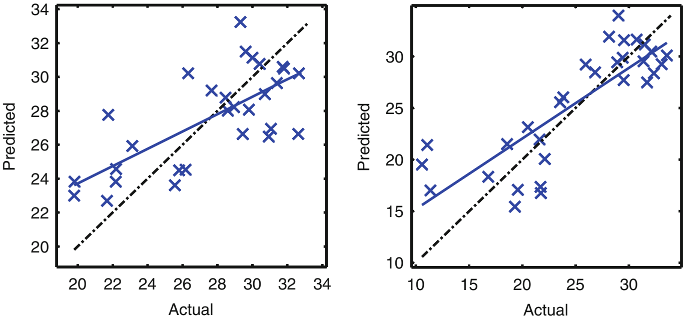

Fig. 6.4 Human ratings (x-axes) and SVM-predicted trait scores (y-axes) for stand-alone microphone recordings (*left*) and text-dependent data (*right*) for agreeableness. *Dashed black line* represents perfect prediction, *blue solid line* gives the actual least-squares regression fit

From the expert's opinion in Table 4.1 the most influential characteristics are expected to be related to voice quality, pitch or pitch harmonics. Results from data-driven signal-based analyses suggest a similar constellation. The difference in harmonics-to-noise ratio is expected to correspond to the perceptual impression of a less sonorous, more tensed and more sharp voice quality. The importance of formant features also supports the importance of voice quality. Finally, also the range of the roll-off points is expected to correlate to voice quality in terms of sharpness. Note that the table suggests a dynamic perceptual impression for both high and **Low Agreeableness**. Although not resulting in a significant difference, the features capturing the ratio of mean loudness from voiced to unvoiced sounds, as well as the features capturing the difference in the overall flatness of intensity, i.e. the error coefficient from linear regression modeling, indicate such a trend nevertheless. Accordingly, high agreeable speech is more flat and shows a lower voiced to unvoiced difference in loudness.

When looking at the scatter plots in Fig. 6.4 no systematic bias can be seen other than the trend seen for openness. Samples of lower human ratings are again estimated too high, while higher human scores are partly estimated too low. Once more, a post-processing step, that aligns the linear least-squared fit (blue lines) seems promising for the text-dependent data, but remains out of scope for this work. The figure shows the plot for close-captured recordings to the left, and the text-dependent recordings to the right. RMSEs resulted in 2.78 and 3.83, correlation to human predictions resulted in 0.70 and 0.82 respectively.

6.2.5 Prediction of Neuroticism

Human assessment of neuroticism from speech shows overall high consistency, which turns out to be above the average when compared to other traits, cf. Table 4.2. When comparing these results to automatic extroversion prediction the models reach different levels. Best results reached correlations of 0.92 for text-dependent data. Results from multi-speaker data remain moderate, e.g. 0.65 for the stand-alone microphone recordings and below 0.57 for the close-capture recordings. The latter is therefore discarded from further analyses.

Results from non-linear mapping indicate that linear kernels perform more robust. For both data, acted and realistc recordings, the IGR ranking shows an overall smooth curve when linear kernels and high complexities in the SVM models are set up, cf. Figs. 5.10, and 5.24. Thus the ranking seems to be applicable but also not optimally composed.

The analyses of the joint rankings include the text-dependent ranking as shown in Fig. 5.10b, and the rankings from the stand-alone microphone recordings from the multi-speaker subset as shown in Fig. 5.24c. Overtly, when comparing these rankings the much diversity becomes apparent. The difference in feature composition is up to a level, where common features are almost non existent. Many of the formant features populating the rankings on the realistic data capture characteristics of the higher formants, such as mean and standard deviation of center frequencies and bandwidths. For pitch, the significant difference in average rising slope found in acted data continues as a trend for the multi-speaker realistic data. Accordingly, **High Neuroticism** leads to a median of 1.3 semitones per second, while **Low Neuroticism** allows for more elaborated pitch gestures resulting in 3 semitones per second for the acted data. On realistic data the respective median slopes result in a trend only. Note, that although this result is presented as important for the SVM model, the overall importance for human auditive perception might not be affected at all, since the absolute difference in slopes is rather small. With regard to spectral features the variation of the roll-off point from voiced segments shows significantly lower values for **High Neuroticism**, while the variation results higher for low target stimuli. This could indicate to a situation, where less neuroticism allows for more variation in terms of voice quality related characteristics, e.g. sharpness. Finally, this behavior can be found in realistic data as a trend as well.

When looking at the scatter plots in Fig. 6.5 a clear systematic bias can be seen for the close-talk dataset depicted to the left. While the human ratings range from 14 to 31 roughly, the predictions only cover a range from 19 to 24. The SVM prediction turns out to be systematically too narrow, which might also be due to the overall moderate prediction correlation to human ratings only. RMSE results in 3.44. For the text-dependent data the achieved performance is much better. RMSE results in 2.91, correlation to human ratings results in 0.91. Re-alignment of the prediction seems unnecessary unless the primary evaluation metric changes to, e.g., RMSE.

Table 4.1 containing the expert's opinion suggests, that the influential perceptual differences in impression are also bound to voice quality. Differences in pitch variation were not perceived to a prominent level. Eventually, features capturing the frequencies and bandwidths of the formants can potentially correlate to the perceived difference in voice quality.

The few studies from the literature mostly concentrate on voice quality and the relation to neuroticism perception. Addington (1968) concludes that increased throatiness leads to a more neurotic impression, but only for female speakers. In

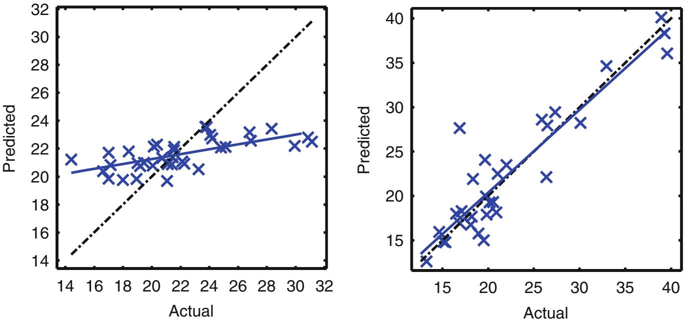

Fig. 6.5 Human ratings (x-axes) and SVM-predicted trait scores (y-axes) for stand-alone microphone recordings (*left*) and text-dependent data (*right*) for neuroticism. *Dashed black line* represents perfect prediction, *blue solid line* gives the actual least-squares regression fit.

terms of pitch, Aronovitch (1976) reported on correlations between the mean pitch and the impression of emotionality and unemotionality, cf. Table 2.2. This perception can be affirmed by the expert's suggestion in Table 4.1, but becomes visible for the acted dataset only in the current experiments. In Schröder and Grice (2003) results show a diverse outcome, still some of his results from speech synthesis indicate that higher pitch corresponds to emotionality and nervousness. Eventually, the found bias in the model, the non-optimal ranking and the big differences within the feature compositions in the models suggest that the current results need to be interpreted with care. Results from the literature suggest that pitch features could be missing in the model trained on realistic data.

6.3 Discussion of Influencing Factors

The answer to the question which factors can potentially have had influence on the results from the presented experiments becomes quite comprehensive. Given the novelty and the low number of literature available many factors of potential influence have been left untackled to date. In the next section, a list of influencing factors with presumably major impact is explained in detail.

6.3.1 Design of Speech Database

The most apparent factor of influence is the usage of acted speech in comparison to non-acted speech. On the one hand good actors are able to control many variables in the speech deliberately. They are able to alternate speech with respect to certain characteristics while keeping other characteristics constant, which is beneficial for targeted manipulations. On the other hand, personality expressions generated by actors are expected to differ from naive every-day realistic laymen expression by the fact that actors are believed to express themselves more clearly and more enunciatedly, while naive speakers are expected to exhibit perceivable characteristics only to a diminished degree. Hence, better results in terms of increased separability and increased classification accuracy could be expected from professional speakers. But these speech actings are often criticized as unnatural, over-enunciated, and staged. This cleavage is well known from neighboring research fields such as speech recognition or emotion recognition. Eventually, the generated personality expressions can be believed to extend the range of non-acted personality expressions also in acoustic parameters.

But there are also other perspectives on acted speech. As results from Sect. 4.2 show, the raters were well able to perceive the targeted personality expressions as intended, hence the personality meaning was conveyed nevertheless. Consequently, the personality expression was not corrupted by the acted interpretation. Rather, another layer containing cues for acted speech was superimposed. Therefore, results from acted speech are not *wrong* or *invalid*. It is the goal of the experiments at hand that renders the usage of acted speech as beneficial or derogatory.

In the present case of exploration of acoustic and prosodic personality gesture in speech the usage of acted data is of high benefit, since no systematic approach for personality expression in German language exist to date. The first and essential question to answer is: *Do ways of vocal expression of personality exist that we can model automatically at all?* Acted data gave the unequivocal answer: Yes, automatic modeling is possible. Now the next logical questions is: *Under which situations can it be applied successfully?* Since we do not know the exact answer, different scenarios have been designed and recorded and experimented with. Whether there exist also situations where the naive speaker's expression may or may not become close to professional expression, or whether there exist scenarios that prohibit any personality expression at all, we do not know by now. In order to explore this space of situations the current three databases were introduced.

Table 6.1 gives a comprehensive overview of speaking situations and discusses the chosen subsets recorded for the presented work with respect to degrees of freedom and relation to each other. Starting with the most restricted design, shown in row #1, the classifiers operate on known linguistic content. Words are kept identical in all samples. In addition, the 10 personality classes can be learned from a sufficient number of examples. Because the system only sees data from one speaker, it is speaker-dependent. In other words, the system is useful for decoding the speech personality gestures from this one speaker only. Results from modeling were shown to be almost perfect, i.e. modeling proved successful.

While the recorded data does not include a subset of many speech actors performing a fixed text, as would be presented in row number #2, the one speaker was asked to perform spontaneously without giving him a pre-fixed text passage to portray, as is described in row #3, i.e. the *text-independent* subset. Asking many speech actors to perform the personality targets with their own words, i.e. #4, would lead to a further increase of expected variance.

Note, it is still unclear whether the introduction of more speakers or more linguistic contexts would lead to more diversity. And subsequently, also the question whether

Table 6.1 Discussion of assumed degrees of freedom in speech and recorded subsets

#	Control	Database subset	Design factors		
			Acted/ non-acted	Fixed/ spontaneous	Speaker-dependent/ speaker-independent
1	Restricted	*Text-dependent*	A	F	D
2			A	F	I
3	↓	*Text-independent*	A	S	D
4			A	S	I
5			N	F	D
6			N	F	I
7			N	S	D
8	Unrestricted	*Multi-speaker*	N	S	I

more diversity leads to better or worse results is unclear. However, acted recordings will always coin the resulting models to be **non-realistic**. Realistic settings in speech analysis would require recordings from laymen, as would be the case in row #5 until row #8.

Corresponding to row #5, there can be a situation when all speakers have been seen during training, which resembles a kind of personality identification problem as known from the neighboring field of speaker identification. Moreover, models would know what will be said. This can be beneficial specifically for dialog systems or in other scenarios of a known user group, where the spoken content is fixed or can be sensed in advance. Training the models to predict unknown speaker's personality would then lead to a situation depicted in row #6.

Eventually, further decreasing constraint in the design, more diverse data can be expected when each speaker talks about different content, which is resembled in row #7 and #8. This situation could be expected in any HCI scenario in principle. The data collection therefore comprises the most unconstrained condition, i.e. many laymen speaking about different linguistic content. This choice is on purpose to give a sort of lower bound on trait score estimation.

In order to build speaker-independent models in general, recordings from many diverse speakers are needed. Still, in order to reliably estimate each individual speaker's personality a sufficient number of recordings for each individual speaker must be seen by the models. These models could then be thought of as **robust** or **saturated** personality models. In case there is not enough speech material the models might fail to capture relevant information. At the same time, seeing many linguistic situations is enhancing the text-independence. How many situations and which situations exactly would be needed are just two out of many questions that remain unanswered. On the one hand, personality can be expected to be relatively stable in many situations, but on the other hand instantaneous samples, especially when being of short duration, can be corrupted by emotions or situational effects. Thus, a personality print of a person is believed to be both, relatively stable and within confinement of reasonable variation.

Hypothetically, the following list of other items could have influenced the results in general:

1. Applying other labeling schemes than NEO-FFI might lead to slight or clear divergence in ratings and mean values. Comparisons between different questionnaires mostly show moderate to very good correlations, indicating that many of them measure related constructs, not necessarily identical constructs. Depending on many factors like labeling situation, time constraints, domain and level of detail other questionnaires might be more beneficial in other situations. To date, there does not exist any such manual or common practice guiding or pondering on the application of different assessment schemes given different application scenarios.
2. Although the NEO-FFI is widely acknowledged for assessment of personality in psychology, it might not be optimal for assessment from speech. Again, no manual or common practice can be applied to compare the NEO-FFI to other assessment questionnaires. The presented work is pioneering work. However, factor analyses as presented in Sect. 4.3 revealed excellent applicability in terms of congruent item coding structure in between NEO-FFI applied in traditional psychology and NEO-FFI applied to speech input only. While lose items have been identified and the inventory of items could be decreased to better fit the speech input, its overall applicability could be verified.
3. In particular cases, the speech samples could be too short, so that personality expression is cropped. But in many cases the good results do not indicate such a problem. Within this work, the utterance length was kept nearly constant for all labeled data. The exact determination of optimal lengths of personality analyses windows, maybe even given personality traits, is just one out of many interesting follow-up questions unanswered to date.
4. Since the acted data are produced by a single actor only, the actor's ability to perform good personality actings might differ with respect to traits or targets. From the rater's consistencies such a difference could be suggested for openness and partly for extroversion, cf. Table 4.2. But with the exception of low neuroticism actings on the text-independent data, the actor proved his acting qualities through significant differences in subsequent ratings on the actings, cf. Figs. 4.4 and 4.6.
5. Not every personality target or trait might be equally well perceivable from vocal impression only. Table 4.2 suggests such an insight and indicated that openness is hardest to assess. The respective consistencies for the multi-speaker data result relatively low. But correlations achieved between human ratings and SVM predictions from trait score modeling result within 0.65 and 0.79 for all traits, with openness showing second best results of 0.71.
6. In Sect. 4.4 time- and text dependency was analyzed. Time dependency of personality expressions was tested by recording personality expressions at different times which were several weeks apart. Text-dependency was tested by eliciting spontaneous speech in response to the exhibition of several images. The actor spoke freely composing own words and expressed descriptions and associations towards the exhibits. Still, larger time gaps between sample recordings as well as referencing to other exhibits or completely other elicitation scenarios can be

expected to influence the way of speaking. However, the question if also personality ratings will be affected by such a potential influence remains to be answered in follow-up experiments.

7. The number of speakers in the multi-speaker data is relatively low. The scales' spans might not be covered to a sufficient extent and/or by a sufficient number of examples. More speakers and more examples are needed in this respect. A more comprehensible database is fully desirable. Ultimately, the presented results should be interpreted as indicators. They indicate feasibility and potential magnitudes for the exemplified situations in the database. More speech material will certainly lead to more robust insights and/or to more fine-grained analyses.
8. Personality can be thought of a holistic characteristic, which, amongst other things, allows or prohibits the display of spontaneous emotions. With regard to vocal expression, emotions can be expected to overlay emotional signature over personality-related characteristics. But at the same time, emotions are believed to emerge and cease more spontaneously while also enduring only a short period of time. Eventually, an emotional overlay cannot be prevented completely. But from another perspective it must not be prevented at all because it can well be integral part of trait expression like extroversion. Finally, the potential mix of personality and emotion overlay was assessed with high consistencies and modeled with good and excellent results in case of the acted data classes. Future work will need to focus on how to demultiplex this overlay. More consideration of follow-up experiments are presented in Sect. 7.2.
9. All data analyzed throughout this work is of German language. Analyses for other languages might result in other directives due to cultural differences. While in one language a certain gesture might have no common meaning and expresses individual personality, in other cultures it might be inappropriate due to norms or habits. Fast talking or loud speaking are only two intuitive examples that might cause different interpretations with respect to personality in different cultures. Also Allport and Odbert (1936), Allport (1937), cf. Sect. 1.2.1, as well as Rolland (2000), cf. Sect. 1.2.5, examined data from different cultures and concluded that extraversion and agreeableness are most sensitive to cultural context. McCrae and Terracciano (2005) compare 51 cultures and declare that personality can even be used to examine cultural differences.

At least two more specific observations from the experiments should be discussed: As said before, when designing the databases, the degrees of freedom were assumed to be opening up from text-dependent recordings, towards text-independent recordings to a maximal degree of freedom with the multi-speaker recordings. Modeling results were expected to follow this assumption with respect to accuracy and prediction correlation in the results. Obviously, the results from the prediction experiments from text-independent data do not follow this trend. The overall results from these recordings are much lower than the results from the multi-speaker recordings, cf. Sect. 5.9.1. This results in the first question:

1. Why is there such a big difference in trait score prediction performance on text-dependent and text-independent data?

Of course, text-diversity might be a potential reason why the trait score prediction performance on text-independent data turned out to be so bad, but then this trigger would probably also take effect on the multi-speaker data and its diverse texts to a certain extent. But the obtained results on the multi-speaker dataset do not indicate such a strong influence. At the same time, classification of the text-independent data showed very good results.

After all, two hypothetical factors are expected to have influenced these results predominantly. First, the number of labels assigned for one sample was lower for the text-independent data than for the text-dependent data. Out of all mean label values taken from the 22 image-session pairs, cf. Tables 3.2 and 3.3, only 5 were drawn from a number of 15 labeler's annotations. Other averages were drawn from 3 or 5 labelers annotations only. Thus, the association between the average label and the acoustic properties extracted might become weak or, alternatively, not robust enough. For classification, noise remains inactive as long as it does not cause the data to alternate across the margin around the separating hyperplane. For regression fit, also the direction of small deviations can become influential.

A second factor becomes more obvious when comparing the database to the multi-speaker database. In this respect the databases differ much in speaking style. While the actor might have consciously acted out a rich repertoire of diverse cues in order to signal different personalities, laymen in the multi-speaker set spoke out of a conversational perspective. This conversational perspective might as well be more restricted and normalized when compared with the actings. But Figs. 4.2 and 4.3 do not suggest a strong trend for diminished scales spans in general. Hence, the speech gestures used by the actor might be more difficult to model because they might be acted out in less predictive ways. In this case, the artificial character of the actings turn out to be of disadvantage.

More experiments are needed for a better understanding of this observation. At the same time, next experiments should focus on realistic-data, since it is of less common use to clarify the observation for just one professional speaker than clarifying situations for multi-speaker non-acted conditions.

2. How come the differences between performance on closed-talk and stand-alone microphone recordings?

Speaking style can also be a major difference within the multi-speaker sub-data. Given a headset, the intuitive choice and extent of speech gestures might be different from the situation where the users communicate via a narrow-band telephone device. The current data and especially the results from prosodic capturing of pitch and intensities do not provide conclusive results on whether these conditions lead to increased or decreased ranges or gestures. Also the overall performance is not better for just one condition but alternates between the conditions and traits. Finally, the speaking style can also be expected to depend on the behavior of the interlocutor, which opens up a dependency matrix of conversation partner mixtures to be randomized in order to deeper understand this observation.

With regard to speaking style, users might have been affected by a overall bias on *good* or more *concentrated* behavior when realizing that they are being monitored and analyzed. Eventually, this might apply to all laboratory testing in general. To counterpart such a restrained behavior unconstrained field tests are needed.

6.3.2 Signal-Based Feature Extraction and Model Set-up

Also a relatively high number of parameters set-up during the feature extraction, ranking and modeling procedures can potentially have influenced the results and interpretations. The following list gives an overview of potential errors and suggest strategies to avoid them in sequential order.

1. One of the first steps applied is the *Voice Activity Detection*, cf. Sect. 5.2. Here, active speech parts are segmented. In addition to the segment of whole active speech, the signal is sub-indexed into voiced, unvoiced, and pause parts. Acoustic and technical noise such as background sounds, cross-talk, off-talk, zero-signal parts, spikes caused by physical or electric shocks in the microphone etc. can lead to erroneous detection. Because this segmentation is the basis for further processing, it is important that the detection works within reasonable accuracy. Many different methods for voice activity detection have been proposed, cf. Ramirez et al. (2007) for an overview. Especially for emotional and presumably also for personality-containing speech these detection algorithms need to be set-up with care. The implemented VAD was inspected manually, ensuring a reasonable detection outcome.
2. The feature extraction, cf. Sect. 5.1, could have failed to capture relevant information. This can be due to noise corruption, with noise capturing a wide range of interferences as described in the previous point of VAD. Audio descriptors might fail in these cases, oftentimes signaling no clear sign of failure. Most errors of pitch trackers result in octave-errors, meaning that the pitch is detected with double or half the actual values. Also frequently, thresholds for voicing or silent intervals need to be set with care. Again, this is specifically true for emotionally colored or personality-related speech. Also all other audio descriptors are prone to noise corruption to a larger or smaller scale. The presented implementation of audio features was set-up with great care and inspected manually.
3. Feature selection chosen could have produced non-optimal results for specific conditions, e.g. if entropy concept was adverse. Other methods exist, cf. Sect. 5.4 for a discussion. Since the proposed method showed good results in previous works and leads to overall smooth performance curves with few exceptions only, its applicability can once more be validated.
4. The number of data points could have been low. The learning algorithms might fail due to too much unsystematic noise and too less systematic variation in the data. More data might provide a more robust data base for, e.g., ranking, modeling

of hyperplane, support vectors and regression functions. More data is expected to prove beneficial but its acquisition and labeling was out of scope for the presented work. For a reasoning of the selected data in the collection please refer to the Sect. 6.3.1. Eventually, the excellent results for the acted data do not suggest an overall data shortage. For the multi-speaker experiments more data is desirable, but results achieved with the present dataset proved sufficient for indication of trends and feasibility statements.

5. The modeling parameters exert a huge influence on the performance. Parameter tuning was perused with great care using systematic variation and empirical validation. The following list gives an overview of the parameters and it's values.

    - For unsupervised discretization as prerequisite for IGR application to continuous label spaces the chosen number of bins was set to two, i.e. binary discretization. Experiments with a larger number of bins, i.e. 3, 4, and 5, showed inconclusive results. While the prediction correlation to human annotations improved by 0.2 for openness, all other traits showed overall inferior results. When discretizing into 5 bins scores for neuroticism raised by 0.2, but again other trait models lost correlation to human annotations. Because of fuzziness of these observations overall binary discretization was retained for model training. In case of supervised discretization as applied for the classification tasks, the number of bins was determined in accordance to Fayyad and Irani (1992). Again, other methods exist. For a discussion of the applied method please refer to Sect. 5.4.
    - The number of features in the model was determined empirically by incremental feature inclusion for each classification and regression task individually, cf. Sect. 5.4. Optimal settings have been determined by the procedure. The lower bound of absolute IGR value imposed an upper bound of maximum number of non-negligible features to include in the models. Only in 2 out of the 18 presented and discussed results in Sects. 5.8.1–6.2 best results were obtained when exceeding the upper bound of features. Overall, the number of features in the model remains one of the most influential parameters in the overall system. The actual number strongly depends on the data at hand.
    - Support vector models include a complexity parameter, which handles the number of support vectors included in the hyperplane estimation as a threshold. This parameter was explored by systematic search over an extended search range jointly with the number of features. Thus, it can be expected that optimal settings have been determined. The actual optimal complexities depend on the data at hand.
    - The usage of non-linear kernel functions introduces new parameters which need to be tuned jointly with the complexity and number of features in a grid-search. These parameters were also explored by systematic search over an extended search range, cf. Sect. 5.6. It can be expected that also here optimal settings have been determined. Only for the multi-speaker database the prediction of closed-captured openness and agreeableness trait scores resulted in clearly better results when using non-linear models. After all, linear mod-

eling most frequently outperformed non-linear modeling or the results were at equal level while incurring lower costs. Note that other kernel functions exist, which might change the actual results. The current selection comprises the most widely applied transforms only. More detailed analyses need to be transfered to future experiments when other kernel functions are included in experiments.

6. While support vector classification proves relatively robust against remaining correlations between the features, support vector regression shows more sensitivity in this respect. As explained in Sect. 5.4, the ranking does not consider correlations among the ranked features. Thus, the remaining dependency between the features can potentially influence the results in a negative way. However, additional experiments applying de-correlation by principal component analyses (PCA), did not lead to improvement either.
7. Although analyses of human annotations by means of ANOVA did not suggest overall significant influence with respect to time- and text-dependency, cf. Sect. 4.3, non-significant differences can still influence the modeling performance. Follow-up experiments need to explore these dependencies by collecting more data of diverse temporal and textual conditions in the first place. It cannot be completely excluded that while human perception proves not to be dependent on these conditions in significant ways, machine classification might still depend on a division of presumably oppositional situations or linguistic contexts. After all, the same hypothesis became observable when looking at the closed-talk and stand-alone microphone recordings of the multi-speaker data. Also here, human assessment did not result in significant differences after ANOVA, but modeling results improves much when imposing this division in the training phase.
8. The overall evaluation metric of choice for trait score prediction experiments throughout this work is the correlation between human ratings and SVM predictions. Focusing on error margins, e.g. optimizing for RMSE, can result in different insights. Note, such experiments could also consider human error-margins and a clearing of these error margins from the error metric obtained from modeling. Such a change resembles a change in perspective on results as well, as overall satisfactory results need to be re-defined on the same error metric, cf. Sect. 5.7.
9. The speech quality as captured by the microphones have shown to exert a strong influence on the model performance as well. As said before, it has proven beneficial to model close-talk speech separately from stand-alone microphone recording speech. This could also imply a certain dependency of the results to microphone quality, which would add another technical condition into the line of strong influences and is not desirable. But such a statement can only be given as hypothesis from the current results. Eventually, it is impossible to determine whether the observed effect is caused by microphone quality or speech behavior, as explained in Sect. 6.3.1. Using a headset might cause a different usage of speech gestures when compared with using telephone devices. More data recorded under controlled recording conditions covering both scenarios need to be collected in order to disentangle the individual dependencies behind this observation.

References

Addington DW (1968) The relationship of selected vocal characteristics to personality perceptions. Speech Monogr 35(4):492–503

Allport GW (1937) Personality: a psychological interpretation. Holt, Rinehart & Winston, New York

Allport G, Odbert H (1936) Trait-names: a psycholexical study. Psychological Monogr 47(211)

Apple W, Streeter LA, Krauss RM (1979) Effects of pitch and speech rate on personal attributions. J Pers Soc Psychology 37(5):715–727

Aronovitch CD (1976) The voice of personality: stereotyped judgments and their relation to voice quality and sex of speaker. J Soc Psychology 99(2):207–220

Buller DB, Aune RK (1988) The effects of vocalics and nonverbal sensitivity on compliance a speech accommodation theory explanation. Hum Commun Res 14:548–568

Fayyad UM, Irani KB (1992) On the handling of continuous-valued attributes in decision tree generation. Mach Learn 8:87–102

Hollander M, Wolfe D (1999) Nonparametric statistical methods, 2nd edn. Wiley, New York

Ivanov AV, Riccardi G, Sporka AJ, Franc J (2011) Recognition of personality traits from human spoken conversations. In: Proceedings of INTERSPEECH 2011, pp 1549-1552

Mairesse F, Walker MA, Mehl MR, Moore RK (2007) Using linguistic cues for the automatic recognition of personality in conversation and text. J Artif Intell Res 30:457–500

Mallory EB, Miller VR (1958) A possible basis for the association of voice characteristics and personality traits. Speech Monogr 25:255–260

McCrae RR, Terracciano A (2005) Personality profiles of cultures: aggregate personality traits. J Pers Soc Psychology 89(3):407–425

Mohammadi G, Mortillaro M, Vinciarelli A (2010) The voice of personality: mapping nonverbal vocal behavior into trait attributions. In: Proceedings of the international workshop on social signal processing, pp 17–20

Mohammadi G, Vinciarelli A (2011) Humans as feature extractors: combining prosody and personality perception for improved speaking style recognition. In: Proceedings of IEEE international conference on systems, man and cybernetics, pp 363–366

Neppert J, Pétursson M (1986) Elemente einer akustischen Phonetik. Helmut Buske

Pittam J (1994) Voice in social interaction: an interdisciplinary approach. Sage, London

Quené H, Semin G, Foroni F (2012) Audible smiles and frowns affect speech comprehension. Speech Commun 54(7):917–922

Ramirez J, Górriz JM, Segura JC (2007) Voice activity detection. Fundamentals and speech recognition system robustness. Robust Speech Recognit Underst 6(9):1–22

Rolland JP (2000) Cross-cultural validity of the five factor model of personality. In: XXVIIth International Congress of Psychology, Stockholm, Sweden

Scherer KR (1974) Voice quality analysis of American and German speakers. J Psycholinguistic Res 3:281–298. doi:10.1007/BF01069244

Scherer KR (1977) Effect of stress on fundamental frequency of the voice. J Acoust Soc Am 62:25–26

Scherer KR (1979) Personality markers in speech. Cambridge University Press, Cambridge, pp 147–209

Scherer K (1995) Expression of emotion in voice and music. J Voice 9(3):235–248

Schröder M, Grice M (2003) Expressing vocal effort in concatenative synthesis, pp 2589–2592. Citeseer

Szameitat D, Darwin C, Szameitat A, Wildgruber D, Alter K (2011) Formant characteristics of human laughter. J Voice 25(1):32–37

Tusing K (2000) The sounds of dominance. Vocal precursors of perceived dominance during interpersonal influence. Hum Commun Res 26(1):148–171

Chapter 7
Conclusion and Outlook

7.1 Contributions and Conclusions

This work combines interdisciplinary knowledge and experience from research fields of psychology, linguistics, audio-processing, machine learning, and computer science. To sum up the major contributions of the presented work, the following list of 8 main achievements can be pointed out. During this work, the author has:

1. Systematically explored a novel research topic devoted to automated modeling of personality expression from speech. Former research including references to psychologist's and linguist's literature have been introduced in Chaps. 1 and 2.
2. Conducted several recording sessions and build up a systematically stratified personality speech database as described in Chap. 3.
3. Conducted comprehensive labeling sessions using NEO-FFI questionnaires in order to annotate the speech data and prepare it for automated modeling as discussed in Chap. 3.
4. Conducted analyses and discussed insights from human personality perception and its assessment using NEO-FFI questionnaire including validation of novel NEO-FFI application in speech domain using factor analyses in Chap. 4.
5. Implemented a comprehensive audio-extraction and feature-definition unit in Chap. 5 in order to capture personality expressions and operationalize fully automated modeling.
6. Implemented and discussed feature rankings in order to obtain insights into prominent speech-qualities for individual personality classification or novel trait-dependent score prediction tasks in Chap. 5.
7. Applied modeling by support vector models for classification and regression including model extension for non-linear mapping, systematic parameter optimization and evaluation in Chap. 5.
8. Presented a comprehensive comparison and discussion of individual results from human and automated assessment as well as general trends observable from the individual results in Chap. 6. Comparisons are also drawn to the authors expert's

T. Polzehl, *Personality in Speech*, T-Labs Series in Telecommunication Services,
DOI 10.1007/978-3-319-09516-5_7

hypotheses on perceptual characteristics as presented in Sect. 4.1 since very few references from the literature exist to date.

In order to approach any empirical work connected to personality, it is necessary to understand the concepts of personality. For this work, the most widely acknowledged model of the *Big 5* personality traits is employed. In more detail, the Big 5 consist of openness, conscientiousness, extroversion, agreeableness, and neuroticism. A short historical review on the emergence of the trait models including its characteristics as captured by the most frequently applied NEO-FFI assessment questionnaire can be found in Chap. 1.

Prerequisites for empirical work on personality modeling from speech is a speech database containing personality-varied speech. For this purpose three subsets of data were recorded:

1. The *text-dependent* subset comprises speech from a single professional speaker acting out 10 personality targets. The targets were set to the extremes of the Big 5 personality traits. In addition, the actor repeated the same short text-passage all the time, interpreting it out of the 10 personality targets. Details are given in Sect. 3.1.
2. The *text-independent* subset comprises speech from the same actor. This time, the actor was presented images and drawings showing diverse content. The actor was asked to spontaneously speak about feeling and associations while still acting out of the same 10 personality classes. Because the constraint of keeping the linguistic content constant is discarded, the resulting speech is expected to show more diversity with respect to vocal expressions. Moreover, by conduction repeating tests at several time points and by comparing results from different images presented, the subset offers the opportunity to analyze time and text-dependency of personality expressions. Details are given in Sect. 3.2.
3. The *multi-speaker* subset comprises speech from 64 non-professional speakers. The speakers were not asked to perform any actings with respect to personality. Conversational scenarios were given to them and speech was captured with in a close-capture manner using headsets, and with a stand-alone microphone. Because no constraint on personality was given at all, this subset is expected to mirror realistic exertion of speech gestures. Details are given in Sect. 3.3.

A selection of data was forwarded to extensive labeling procedures using NEO-FFI questionnaires as described in Sect. 3.4. Results from the analysis of the human personality labels prove that raters were well able to assess personality from speech with high consistencies, cf. Sect. 4.2. Factor analyzing the item responses revealed very congruent latent structures when compared with the original NEO-FFI structure as applied in psychology. The questionnaire can therefore be expected to build up the same constructs accounting for personality as known from psychology. Using the item responses to analyze the data showed overall normal distributions in the ratings of each acted target. Also all targets were perceived by the listeners as expected. Moreover, all low targets were perceived lower than targets which were not manipulated on the trait at hand. The latter was perceived lower as high targets in turn.

All differences became significant with the exception of low neuroticism resulting in a trend only. Ultimately, personality assessment from speech could be shown to be applicable using the NEO-FFI. Applying the NEO-FFI the chosen speaker could be shown to be able to produce the desired perceptually different speech acting with respect to personality. The acted speech database therefore is of optimal characteristic for initial empirical experiments.

Furthermore, experiments show that neither (a) introducing different linguistic content, nor (b) comparing recordings from different recording sessions, nor (c) comparing close-talk and stand-alone microphone recording conditions significantly influenced the perception of the designed low and high targets on the traits. Interestingly, labels for all three databases indicate a consistent moderate inverse correlation between neuroticism and extroversion. At the same time, there is a moderate inverse correlation between neuroticism and conscientiousness. Consequently, the more neurotic a speaker is being judged, the less extroverted and conscientious he is being perceived as well. A secondary correlation in between the traits indicates that when a speaker is being perceived as more open this speaker would also be perceived as more agreeable. In other words, the more curiosity and open-mindedness is perceivable, the more commitment and social reflection is attributed as well.

In terms of automatic signal-based personality modeling, a rather comprehensive repository of acoustic and prosodic features were implemented as described in Sects. 5.1–5.5. These features capture various statistics in terms of speech intensity, pitch, loudness, formants, rhythm, spectral characteristics and other. Results from IGR rankings of these features can be used to obtain first insights into the impact of individual acoustic and prosodic features or, in a broader view, feature groups. At the same time, the interpretation of these results depends on the subsequent modeling performance. Modeling included support vector models for classification and regression. Linear kernels were extended to non-linear mapping. Eventually, the joint impact of features, models and databases was systematically evaluated by using grid-search methods and applying cross-validation for evaluation of models and rankings.

With respect to personality classification from the acted speech databases very good and good results were obtained. When keeping the linguistic context, i.e. words and content, constant, the automatic personality classification reaches excellent results of a baseline of 10 % accuracy when randomly guessing the classes. Top-16 ranked features lead to an accuracy of 96.3 % already. Doubling this number of features accuracies increased up to higher than 99 %. Also for the text-independent data good results were obtained. 85.2 % accuracy could be achieved using 210 acoustic and prosodic features. Overall, non-linear extension did not result in considerable improvement. However, the intention of the experiments was not to push results to absolute peaks, but to explore the feasibility of personality classification from speech. To this end, the results fully supply the evidence that personality can be automatically captured and classified from acted speech. Class-wise insights show that extroversion could be modeled best. Also neuroticism, conscientiousness, and openness reached overall good and very good results. When compared to the few works available in the literature, these results contribute new and more systematic insights for all the Big 5 traits. While isolated individual results from, e.g., Mairesse et al. (2007) and Mohammadi and Vinciarelli (2011) suggested that extroversion and conscientiousness can be modeled, the current work provides a common database and the ability for direct comparison between three databases and all the Big 5 traits. Note, works available in the literature are hard to compare to date, cf. Sect. 2.2 due to many differences in labeling and data recording. With respect to agreeableness, high agreeableness results are of moderate magnitudes only. The observed confusions do not allow for deep systematic insights into this finding, since both precision and recall are equally moderate and the confusions do not cluster across other classes.

Throughout all the experiments the IGR ranking proved very helpful. In more detail, pitch slopes, their speed as well as their ranges turned out to be of high value. Also the flatness of intensity and loudness contours as well as their statistical distribution was of high information. Rhythmic features capturing the speech-to-pause ratio as well as the duration of speech and mean duration of voiced points were of high value. In terms of spectral features the centroid and roll-off point were found in high ranks. Also the behavior of formant #4 and #5 including their bandwidths were of general high information gain value. Unexpectedly, an extraordinary high number of features from formant #5 populated highest ranks for the text-independent data. The specific reason for this observation could not be clarified completely within this work. Overall, statistics from unvoiced segments proved to be of minor importance with respect to highest ranks. Examples from the data along a more elaborated discussion and interpretation can be found in Sect. 6.1. Ultimately, these results contribute new and more systematic insights for all the Big 5 traits.

Unprecedented in the literature are the results on automatic trait score modeling using signal-based measures. Results were produced for individual scales. Models which did not produce predictions that correlate to human assessments by 0.6 are discarded from the interpretation because these models capture too less relevant information to be considered as meaningful. As a consequence, also the respective feature spaces do not offer meaningful insights. The main findings can be listed as follows:

- *Openness* is the most challenging trait for human assessment. However, results from automatic trait score predictions revealed relatively high correlations between the predictions and human annotations, i.e. 0.89 for acted data, 0.71 for realistic data. From the analyses of the rankings on the text-dependent subset and the multi-speaker subsets two main feature groups prove helpful, namely MFCCs and formant-related features, especially with regard to bandwidth of the first and fifth formant. In addition, the realistic multi-speaker recordings show pitch features in high ranks, which capture pitch gestures. Overall, these insights seem to be consistent for all analyzed databases. Moreover, findings build a good match to the expert's hypotheses on perceptual characteristics.
- *Conscientiousness* results in very high labeling consistency. However, results from automatic trait score prediction perform on relatively moderate levels only. Best results reached correlations of 0.68 for acted data, and 0.67 for realistic data. IGR ranking results in a more unstable outcome, which can be the cause for the moderate

results. Three dominating feature groups become visible in the rankings, namely pitch slope movements, intensity and loudness-related levels, as well as MFCCs. While this finding matches the expert's perceptual impression, other impressions affecting tempo, rhythm and voice quality were not be included in the IGR ranks. As one possible explanation, feature extraction could have failed to capture these characteristics. A more detailed analysis of factors that can cause the extraction to fail is presented in Sect. 6.3 and specifically in Sect. 6.3.2.

- *Extroversion* shows rather moderate consistencies in labeling while models perform on good levels. Best results reached correlations of 0.82 for acted data, and 0.79 for realistic data. The solid rankings show very congruent results across all databases. Here, pitch clearly proves dominant. More specifically it is the range of pitch movements and the respective dispersion and variation measures that are prominent. This finding matches the expert's opinion very well. Also MFCCs are important. Finally, also loudness and intensity-related features capturing the flatness of the contours or its variation can be found in high ranks, which again matches the auditory impressions. These results match the finding from related works quite well. Unfortunately, once more the perceptual difference in rhythm and pauses perceived by the expert as well as indicated in the literature could not be captured by the features.
- *Agreeableness* assessment results in good overall consistency as well as good overall prediction performance. Best results reached correlations of 0.82 for acted data, and 0.70 for multi-speaker close-captured data. IGR shows an applicable bot non-optimal ranking. Clearly, MFCCs drawn from both voiced and unvoiced segments build the dominant feature group, but also some formant and loudness features as well as features like the speech-to-pause ratio and total duration appear in high ranks. The latter two suggest that high agreeable speech contains more speech. Hypothetically, the usage of lowering the larynx as well as the usage of a smile could account for the observed differences of formant center frequencies and bandwidths in between low and high agreeable. Other high ranked features correlate to voice quality, which also resembles the dominant perceptual trigger in between the targets as perceived by the expert. Eventually, high agreeable speech is more flat in terms of intensity and differences in loudness.
- *Neuroticism* ratings show overall high consistency, models perform on different levels. Best results reached very good correlations of 0.92 for acted data but 0.65 for the realistic stand-alone microphone recordings only and even below for the close-capture recordings. Again IGR seems to be applicable but not optimal, non-linear modeling further decreases robustness. This time, the differences in feature spaces is up to a level, where common features are almost non existent. Formant features, mostly capturing higher formants, are populating almost 80 % of top ranks from realistic data, while they do not account for considerable amounts for acted data. Here, also pitch and MFCC features are frequent. According to the expert's impression, voice quality plays an important role. Differences in pitch variation were not perceived to a prominent level. Eventually, features capturing the formants as well as the high ranked roll-off point can potentially correlate to the perceived difference in voice quality. Also results from the literature are

inconclusive. Most consistence is within the importance of pitch, which have been found by the expert's assessment as well as by works from the literature. Models trained on realistic speech also exert a strong bias. Predictions are within a narrow range only, insights have to be interpreted with care.

Major influences that could have affected the results and interpretations are listed in Sect. 6.3. Most apparent is the usage of acted speech, but results were always compared to realistic speech and interpretations are aligned accordingly. Also the very high classification results are expected to account for the fact that acted speech data was used. Eventually, these high and very high accuracies obtained from the modeling experiments also signal that the tasks presented can be performed successfully. Given this verification, next experiments should concentrate on less restricted situations. In addition, for the presented pioneering work of signal-based trait score modeling the usage of acted speech proves beneficial also for the detection of trends in realistic speech. Ultimately, the purpose of these controlled experiments is to lighten the way towards less controlled, i.e. more realistic experiments. The assumption that realistic personality could be extracted and predicted automatically certainly is as venturous as bold in spirit. Results on realistic speech are therefore not be expected to tackle 100 % accuracies or perfect prediction correlation, not at all. The overall magnitude of at least moderate, good or very good correlation that were actually obtained by the modeling experiments is therefore encouraging and promising. Also few inconclusive observations need to be revisited in follow-up work. Results from the multi-speaker subsets, i.e. the close-talk subset and the stand-alone microphone recordings suggest a clear influence of speech quality on the results, which signals an insufficient robustness to noise and channel characteristics. Moreover, even when relevant characteristic are robustly modeled, the respective feature might only be valid in specific situations. Finally, this work contains a limited amount of data, i.e. speakers, situations, recording qualities. More data will lead to new insights, especially when combined with emotional annotation. More perspectives are given in the next Sect. 7.2

7.2 Outlook

Starting with the labeling, scheme future work should focus on providing dedicated questionnaires for personality assessment from speech. Also comparisons between different questionnaires might lead to better assessments. This requires a sufficient amount of data in the first place. The data also needs to include more speakers and more situations, as personality is believed to be relatively stable over time and situations, but not absolutely stable. There might exist situations where personality expressions become oppressed. Other scenarios might allow for the observation of distinct cues because personality might be overtly exhibited. With this respect also the sampling or recording of these scenarios will be a challenging factor of influence, since distortions by operationalization should be avoided. Unobtrusive field tests need to be carried out.

The material also need to account for emotional expression. Emotions are a factor of influence for all traits. At the same time they are integral part of the neuroticism domain, as neurotic behavior also includes being susceptible to emotions. Emotions and personality certainly also depend on cultural background, at least some parts of it can be expected to be acquired during socialization. International comparisons might therefore be an interesting challenge.

One very concrete future challenge is to answer the open questions discussed in Sect. 6.3: *Why is there such a big difference in trait score prediction performance on text-dependent and independent data?* One indication given in the section suggests that the number of labels assigned for one sample is possibly to low. Another indication focuses on the influence of speaking style. Another open question is: *How come the differences between performance on closed-talk and stand-alone microphone recordings?* Again indications are given that speaking style might be an influence. Given recordings using a headset, the intuitive choice and extent of speech gestures might be different from the situation where the users communicate via a narrow-band telephone device. Finally, the behavior of the interlocutor might also exert an influence. Systematic empirical analyzes might come very comprehensive when trying to include the interlocutors influence. An initial dataset of mutual assessment of personality from 13 speakers have already been recorded and is available for follow-up work.

Speakers might be clustered by their personality characteristics. Human-computer interaction might then align to such clusters in order to make interactions more intuitive and less static. At the same time, the clusters might be used to execute cluster-specific processing which might serve as pre-processing for other systems like speech recognition or emotion recognition. With this regard, the personality overly might be minimized and the recognition systems might perform better.

Recordings should also comprise different speech lengths. The chosen value of 20 s excerpts did show sufficient for acted data, but naive speakers might as well resort to longer or shorter units. Eventually, the units might also depend on the speaker and/or the situation. Ideally, a dynamic segmentation that is able to detect optimal segment boundaries would be of benefit. If enough material for supervision had been collected, segment splitting could explore a splitting scheme similar to the one introduced with the IGR in Sect. 5.4.

With regard to modeling improvement, more features capturing more speech specific characteristics will be helpful. But not the absolute number of features counts, but the ability of the features to capture relevant information. The features need to be robust against acoustic and technical noise such as background sounds, cross-talk, off-talk, zero-signal parts, spikes caused by physical or electric shocks in the microphone etc. Throughout this work, speech characteristics specifically with respect to voice qualities and rhythmic behavior were perceived as relevant and distinguishing by the expert. The respective features, however, did not end up in high ranks. Hence, most likely this information could not be captured properly. Specifically voice qualities accounting for fragile or crumbled impressions might be beneficial. Also advanced rhythm features are expected to be beneficial. At the same time, also non-optimal segmentation can introduce too many irrelevant data points so that the final step of drawing statistics from the speech extracts levels out certain information. In this respect, segmentation, robustness and perceptual relevance are interwoven. Moreover, different speech qualities might require different pre-preprocessing steps. Speech might be captured using various equipment or it is transmitted using various channels. Results from the multi-speaker subsets, i.e. the close-talk subset and the stand-alone microphone recordings, suggest a clear influence of speech quality on the results. Eventually, even when a perceptually relevant characteristic was successfully captured and robustly modeled, the respective feature might only be valid in specific situations. The high ranks for pitch features in combination with the high standard deviation observed during ranking indicated such a situation, where a characteristic is highly relevant when reaching a certain range but barely relevant when falling outside this range. Encountering these pitch gestures is of high information, but one might need to wait for a longer time since such a gesture might occur. Relevant features may depend on certain situations and may be not always present in all situations.

Features must not necessarily be bound to acoustic and prosodic characteristic. Also text-based and semantic features as well as sensor-based features capturing skin conductivity or even body movement can deliver interesting cues. Related unpublished work of the author also found correlations between upper body movements and emotions when affectively speaking on the phone.

Many other methods for feature selection exist. Although the chosen method of IGR proved overall good and very good results, a comprehensive comparison of IGR to advanced and elaborated racing and wrapper-based methods would be as interesting as promising for future improvement. The challenge would be to avoid overfitting at the same time, cf. Sect. 5.4. But also for IGR more extensive experiments on discretization might lead to changes in the results. However, the influence is expected to be of much lower magnitude.

The exploration of other kernel functions or other loss functions might lead to improvement, although improvement is expected to be of low magnitude. Also many other models exist, such as generative models, trees and neural networks. Results might depend on the actual model but, as explained in Sect. 5.6, SVMs have proven to yield good results in speech based tasks and in general. Nevertheless, other models as well as the combination of models to ensembles remain for many interesting future experiments.

Re-scoring the results with respect to, e.g., RMSE or other error metrics can also give valuable insights specifically when an application provides directions for interpreting the error magnitudes.

Such an application might be a personality-based recommendation engine, so that two speakers who presumably match in personality would be connected. Hopefully, these persons might then be able to talk in a more natural or targeted way with each other than other couples would do. Presumably, parameters like the shown speech-to-text, speech rate, harmonic and dynamic excitement in pitch and intensity, gestures frequency and range, as well as many other parameters can deliver valuable cues for estimating specific degree of openness, conscientiousness, extroversion, agreeableness or neuroticism. Repeating the composition and development of the traits from psychology, the recommendation engine would then estimate the tendency of a person to be imaginative, independent, and interested in variety versus practical, conforming, or interested in routine. On the openness trait it would predict the tendency to be organized, careful, and disciplined versus disorganized, careless, and impulsive behavior. With regard to extroversion it would predict the tendency to be sociable, fun-loving, and affectionate versus a retiring, somber, and reserved nature. Agreeableness estimates relate to the tendency to be softhearted, trusting, and helpful versus a tendency to be ruthless, suspicious, and uncooperative on the opposite side. Finally, being able to estimate the neuroticism degree would enable to assess the tendency to be calm, secure, and self-satisfied versus the personality of anxious, insecure, and self-pitying characteristic.

Given a grid of such personality clusters new interlocutors could be matched against already seen speakers and their characteristics. Such a scenario would be applicable and very valuable for both essential research in HCI and science as well as for commercial usage. The latter would be given the opportunity to route callers in call-centers to matching agents, connect pre-selected potential customers with marketing or sales forces, or connect matching profiles for dating and partnership websites, just to point out a few perspectives. Commercial usage of personality need to be monitored with great care, since speech is a personal data. It should not be given away without awareness! It should not be captured without permission! It should not be modeled and applied for usage unknown or unaccepted by the speaker!

Speech synthesis is another research field that would benefit from the existence of acoustically described profiles of personality. Synthesized speech could be given an own character, which then might be able to adapt to the user's personality subsequently. Especially for gaming, these personalities can be expected to be of high interest for the game-gamer interaction. The gamers could be addressed in specific ways to achieve more captivating or excitement arising effects. Current high-quality synthesis systems mostly convey a generic universal expression generated out of small speech chunks and units from selected persons. These speakers are expected to produce speech that matched many situations and will be accepted by many listeners. But with synthesis of personality new virtual speakers could be generated. Technically, this prospective promises to be much more flexible than traditional systems because synthetic personalities could be altered quickly with respect to the human interlocutor.

Also an application in automated dialog systems can be suggested. Here, the domain and potentially also the choice of words or answering options that the callers can choose from when answering to the system prompts can be expected to be limited for special purposes. In extreme cases, these scenarios might even become very close to a text-dependent scenario similar to the one simulated by acted speech in this work. When callers can be allocated a fixed caller identity, the systems can evolve and mature by collecting speech material from a series of calls. Thus the expected personality profile will ideally converge. If it does not converge, interesting follow-up experiments on situation estimation clustering might be advisable.

Yet another application is to estimate a generic acoustic personality profile in order to provide either the model or a kind of normalized version containing personality-filtered speech to other speech processing such as automatic speech recognition or emotion recognition. The personality overlay is seen as disturbing factor in speech and emotion recognition. The aim of the model could be to provide a speaker-dependent enhancement or cancellation with respect to prosodic and acoustic evidence of personality expressions.

Eventually, a continuous data collection with continuous labeling at the same time, as well as a continuous exploration of expressions with regard to different situations are prerequisites for real-life speech application. However, the future will expose ourselves to more human-computer interaction in any case. It is the explicit task and the explicit challenge of speech experts and engineers to make it as enjoyable as possible, which means to observe and learn continuously, as we go along.

References

Mairesse F, Walker MA, Mehl MR, Moore RK (2007) Using linguistic cues for the automatic recognition of personality in conversation and text. J Artif Intell Res (JAIR) 30:457–500

Mohammadi G, Vinciarelli A (2011) Humans as feature extractors: combining prosody and personality perception for improved speaking style recognition. In: Proceedings of IEEE international conference on systems, man and cybernetics, pp 363–366

Appendix A
Label Distributions Text-Dependent Recordings

T. Polzehl, *Personality in Speech*, T-Labs Series in Telecommunication Services,
DOI 10.1007/978-3-319-09516-5

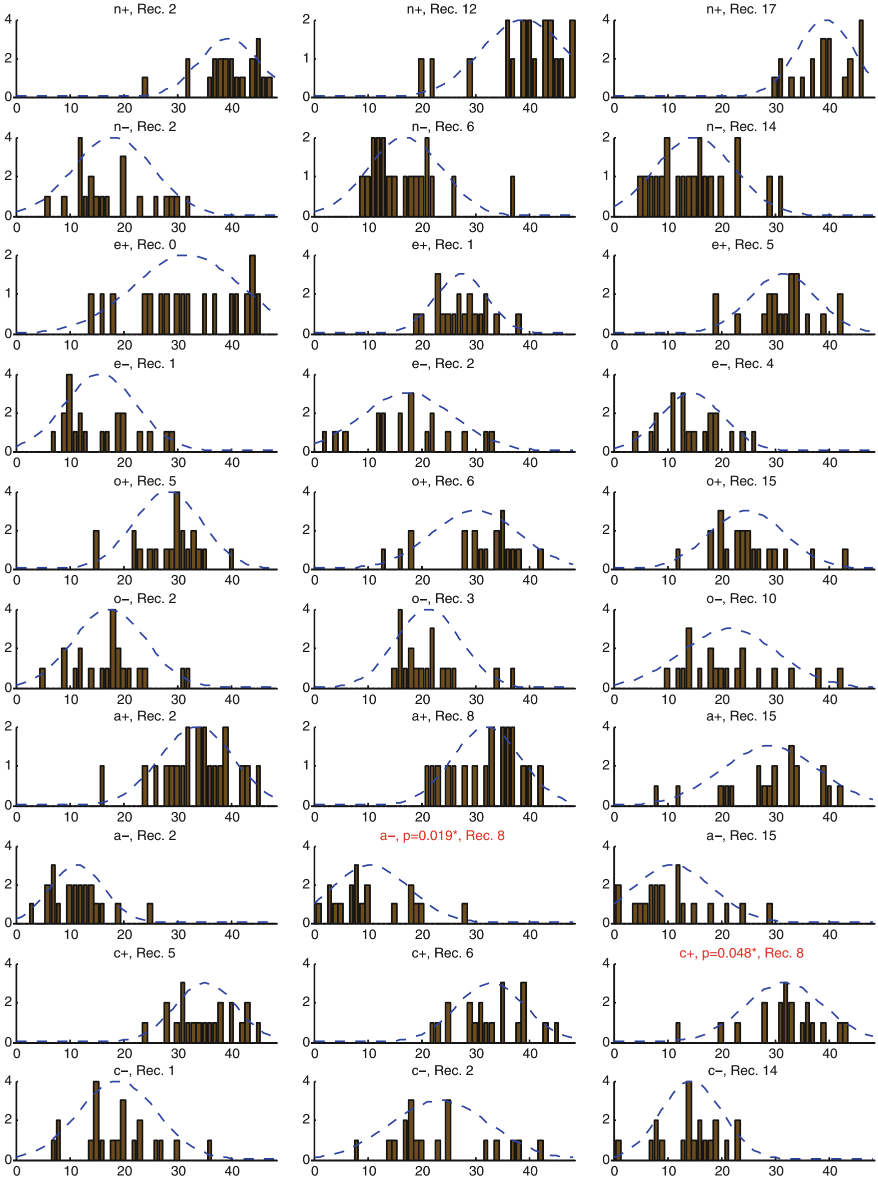

Fig. A.1 Visualization of label distributions. Each row corresponds to one personality target, each column to a recording repetition. Histograms show the distribution of the 20 labeler's assessments, *dashed lines* show a normal distribution fit. *P*-values smaller than 0.05 (5 %-Alpha level, marked with "*") reject the hypothesis of a normal distributions due to Lillifors Tests and are plotted in *red*

Appendix B
Label Distributions Text-Independent Recordings

T. Polzehl, *Personality in Speech*, T-Labs Series in Telecommunication Services,
DOI 10.1007/978-3-319-09516-5

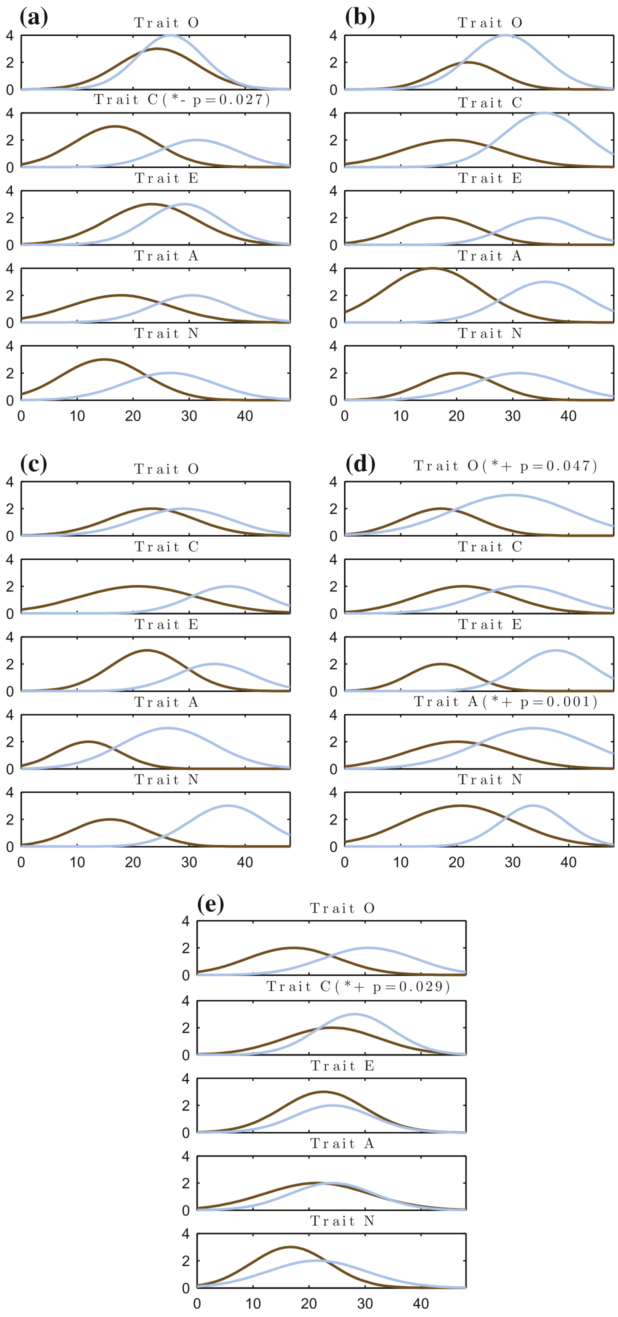

Fig. B.1 Gaussian normal fit (*lines*) for visualization of high (*light blue*) and low (*brown*) target ratings on the text-independent recordings along the Big 5 personality traits. Lillifors test failures are indicated by a "*" and respective *p*-values are given in the headers. **a** Rec. Session 2, Image 9. **b** Rec. Session 3, Image 4. **c** Rec. Session 3, Image 9. **d** Rec. Session 4, Image 4. **e** Rec. Session 4, Image 9