9-2010

Automatically Assessing Personality from Speech

Tim Polzehl
Technische Universitat Berlin

Sebastian Moller
Technische Universitat Berlin

Florian Metze
Carnegie Mellon University, fmetze@andrew.cmu.edu

Published In
Proceedings of IEEE International Conference on Semantic Computing (ICSC), 134-140.

Automatically Assessing Personality from Speech

Tim Polzehl and Sebastian Möller
Quality and Usability Lab der Technischen Universität Berlin /
Deutsche Telekom Laboratories
10587 Berlin; Germany
Email: {tim.polzehl|sebastian.moeller}@telekom.de

Florian Metze
School of Computer Science, LTI/ InterACT
Carnegie Mellon University
Pittsburgh, PA 15213; U.S.A.
Email: fmetze@cs.cmu.edu

***Abstract*—In this paper, we present first results on applying a personality assessment paradigm to speech input, and comparing human and automatic performance on this task. We cue a professional speaker to produce speech using different personality profiles and encode the resulting vocal personality impressions in terms of the *Big Five* NEO-FFI personality traits. We then have human raters, who do not know the speaker, estimate the five factors. We analyze the recordings using signal-based acoustic and prosodic methods and observe high consistency between the acted personalities, the raters' assessments, and initial automatic classification results. This presents a first step towards being able to handle personality traits in speech, which we envision will be used in future voice-based communication between humans and machines.**

***Index Terms*—personality recognition; acoustic and prosodic modeling; semantics of speech**

I. Introduction

In this paper, we present results from an exploratory study of how an established personality description "language", stemming from psychology, can be used to model vocal manifestations of personality traits in speech-based communication. As the scope of voice-based human machine interaction expands beyond directed dialog and simple command and control type interfaces, machines will need to be able to interpret input and produce output, according to a specific context which is determined by many factors including the quality of the voice.

During communication with virtual agents, for example, a full semantic frame, in which a certain information is to be interpreted, will include an assumed personality, which will also manifest itself in both the voice of the human and automatic speaker. To illustrate, saying "okay" supports

"I totally mean this", i.e. resolutely supporting a fact and actively bringing the fact into account, but it also supports "I would follow if someone takes the lead", e.g. not offering any initiative. This example shows how expressive intonation patterns influence the interpretation of what has been said. Depending on the degree of the "Extroversion" of the speaker's utterance, the reaction of the virtual agent can be to expect more explanation and wait until the user has finished his argument or to take the initiative and offer additional help.

We employ an established psychological categorization of personalities as a general framework, which we hope will allow us to use our results for describing both human and automatic assessment of vocal expressions based on speech in both recognition, and synthesis. Relying on established principles and "languages" should hopefully allow us to use machine learning methods to map between observable features and suitable personality categories effectively. Also correlations between signal properties and continuous personality ratings should enable us to generally estimate the strength of each of the personality traits given a discrete user utterance.

Different inventories have been proposed to describe the personality of a person. Many of them base on the concept of the *Big Five* personality traits [1], which attempts to describe the personality of a person using five independent factors, almost as a vector. The factors, which are also referred to as "scales", build up a complex personality profile and can be understand as an overall image of the general tendencies of personality. Depending on several factors, e.g. social environment, immediate situations, age, etc., these profiles show variations. However, the overall profiles are presumed to be relatively determined after adolescence.

Normally, the factor values for a person are being generated by asking a relative, or a person who knows the subject to fill out a pre-compiled questionnaire. In our work, we are not aiming for such a personality assessment based on a long-term relationship, but in a rating of perceived personality, based on a relatively short sample of speech. further, our raters do not know the speaker, they only observing his speech. If personality inventories can be utilized to structure general human motives, we hope this structure can also be useful to assess instantaneous indications of personality. We therefore perform a baseline experiment, in order to determine if such an approach is feasible. In this, human raters assess the personality of a speaker by listening to different samples, which were produced by a professional speaker, who had been given instructions to "simulate" personalities. Our work is therefore similar in spirit to early work in emotion recognition from speech, which was also based on "acted" emotions, and professional speakers, to simplify the labeling problem.

In addition, we analyze the ratings by conducting a reliability study, i.e. we calculate consistencies and correlations of the obtained ratings and compare our results to those given for the full NEO-FFI questionnaire provided by [2]. A direct comparison between results from acted speech utterances and results from "realistic" speech utterances can be found in [3].

Related to our work, Mairesse [4] reports on experiments comparing self-reports with observers' ratings of personality. Evaluating various linguistic models he finds that models of

observed personality will outperform models of self-assessed personality. He hypothesizes, that this may be due to objective observers using similar cues in the models, while self-reports of personality may be more influenced by factors such as the desirability of the trait.

After analyzing the reliability of our ratings, we use signal-based measurements to train acoustic personality profiles for automatic recognition. These basically comprise low-level descriptors, e.g. pitch, intensity, MFCCs etc. Although there are many studies on the relationship between personality and speech, few of them are of empirical nature. In [5] Scherer analyzes prosodic features such as pitch and intensity and observes that extroverted speakers speak louder and with fewer hesitations. In our experiments, we calculate perceptual loudness and use durational features to model disfluency.

Scherer further designates extroversion to be the only factor that can be reliably estimated from speech. Mairesse [4] also concludes that prosodic and acoustic measurements are important for modeling extroversion and that extroversion can be modeled best. In his experiments, extroversion is followed by emotional stability (neuroticism) and openness to experience. His prosodic measurements include intensity and pitch only.

As in the related field of emotion recognition, initial results were produced using few acoustic measurements only. Best combinations of features were discovered afterwards, when more features were taken into account and subsets were compiled by diverse methods [6].

The present study first applies a large scale acoustic and prosodic feature extraction, and determines relevant features using an entropy based feature selection. This system has successfully been applied to recognize emotions as in [7]. Preliminary results on the present recognition task show that classification into ten classes can successfully be applied for our experimental set up, using few acoustic features only.

There are also experiments conducting personality assessment from written texts. In [8] Gill investigates the relations between the personality of an author of short Emails and Blog texts, generated by self-assessment, and their language. Modeling these relations by means of co-occurrence techniques he observes insufficient correlations only. He concludes that knowledge beyond counting co-occurrences of words must be provided in order to generate the same high-level links that humans use to render personality expectations from text. Oberlander [9] examines the relation between part-of-speech (POS) distributions in Email texts and two distinct personality traits, i.e. neuroticism and extroversion, of their authors. He concludes that POS can be characteristic because some personality groups are likely to use it more pervasively then others.

The paper is organized as follows: Section II describes the applied personality test. Section III explains the recording of our database. Section IV estimates the reliability of the ratings. Section V specifies the acoustic and prosodic feature definition. After presenting the overall feature selection, classification and prediction results in Section VI and VII we report on limitations and factors of influence in Section VIII.

II. 'Big Five' and NEO-FFI Inventory

Following the concept of the Big Five personality traits, as presented by [1], we apply the German version of the NEO-FFI personality inventory [2]. Accordingly, we estimate the vocal impression of speech as separate contributions with respect to 5 essential traits. The following paragraphs illustrate general tendencies attributed to these personality traits, as given by the NEO-FFI inventory:

Neuroticism (N): People whose ratings generate a high score in neuroticism are presumed to be emotionally unstable and easily shocked or ashamed. They are easily overwhelmed by feelings or nervousness and are generally not self-confident. On the contrary, people with low ratings are presumed to be calm and stable. They work well under pressure and are not easily agitated.

Extroversion (E): High scores in extroversion indicate a sociable, energetic, independent personality while introverted personalities are presumed to be rather conservative, reserved and contemplating.

Openness (O): The scores of the openness factor estimate the degree to which a person considers new ideas and integrates new experiences in everyday life. High rated persons are presumed to be visionary, curious. They perceive what's happening in the surrounding and are open to venturesome experiments. On the opposite side, people with low scores are generally conservative. They prefer common-knowledge to avant-garde.

Agreeableness (A): High scores in agreeableness suggest that people are rather sympathetic. They trust other people and are being helpful. Non-agreeable personalities are presumed to be egocentric, competitive and distrustful.

Conscientiousness (C): People of high scores in the conscientiousness factor are presumed to be accurate, careful, reliable and effectively planning while people of low scores are presumed to act carelessly, not thoughtfully and improperly.

The raters generate a person's profile by giving answers to 60 propositions from the NEO-FFI questionnaire using "strongly disagree", "disagree", "neutral", "agree", and "strongly agree", which are translated into numeric values 0-4. Every factor score can be in the range of 0 to 48. All 5 factors together generate an overall personality profile of a person.

Intra-scale consistency coefficients, given by Cronbach's Alpha, are constantly above 0.8 which represents overall good cohesion of the ratings. Correlations (after Pearson) between the factors are generally below 0.2 absolute. Two exceptions are the correlations between N and E (0.36) and the correlation between N and C (0.26). The collection of the German NEO-FFI comprises 11 724 samples.

The NEO-FFI is designed for self-assessment and assessment by observers. The test questionnaire is particularly designed to be carried out by raters who know the person being profiled. In our experiments, raters will not assess the overall personality of a person, but instead assess the instantaneously

perceived vocal impression of an unknown speaker's voice. The basic assumption is that raters will give reproducible ratings after having heard only a few seconds of speech from a person. Example questions from NEO-FFI include (in the observer-assessment form):

- The speaker likes to have a lot of people around him.
- The speaker often feels inferior to others.
- The speaker laughs easily.

III. Database

In order to collect a suitable speech corpus for experiments, we recorded a professional speaker, who was given the task to immerse himself into the NEO-FFI description of different profiles resembling different personalities. After recording his 'natural', i.e. non-acted version of a predefined text passage, the speaker imitated 10 personality variations yielding the required speech. The performance of the actor was directed towards factor extremes. Instructions were formulated as given by described in the NEO-FFI manual and Section II.

To provide an example using the factor neuroticism, after recording the non-acted take we instructed the speaker to imitate the speech of both, a more neurotic person and a less neurotic person. Consequently, the required impressions produced by the speaker are expected to account for two distinct areas of either higher or lower factorial score, not to cover the whole range of possible factor scales. As every variation was repeated at least 20 times the full audio database contains 220 recordings. The spoken text passage is designed as artificial IVR (Interactive Voice Response) prompt of minimum 20 seconds length, representing a random gift voucher redeemer service which translates as follows:

> *Hello, and welcome to your voucher redeemer service! This is where you can redeem your voucher and credit your points to your account. Unfortunately, you cannot activate your credit points from the line you are using just now. Please call again from the line you want to charge your credits to. Thank you, and goodbye.*

As labeling of all recordings would be very expensive we chose six recordings from every factor manipulation resulting in three examples of low-targeted values and three examples of high-targeted values. In order to minimize the distortions and artifacts of acted speech recordings two labelers annotated the recordings for naturalness in an initial listening test. As only three out of the 20 variation samples could be passed to perceptive assessment using the full NEO-FFI scheme we chose samples that were rated low in artificiality.

Verifying the performance of variation we conducted NEO-FFI personality tests, where the chosen takes were now assessed by 20 different naive raters. Note, that every rating yields values for all 5 factors of the NEO-FFI inventory. Eventually, approximately 600 NEO-FFI questionnaires were filled out, resulting in over 7 200 ratings for each factor. The 87 test raters were mostly students of 29 years of age on average; three out of five raters were male. Every rater filled out 8 NEO-FFI questionnaires on average.

Figure 1 shows the distributions of the raters' assessments for both the acted and natural speech samples for the five personality factors. The cross-markers connected by the solid line show the medians of the speaker's natural personality ratings. Ratings of the acted variations are given by the yellow markers (medians) connected with the dotted lines. Bars show the respective inter-quartile ranges of the ratings, i.e. blue bars represent the acting towards higher scores, and brown bars show acting towards lower scores.

The raters attest an overall good recognition of variation. Most of the variations were perceived in accordance with the instructions. Interpreting the distributions one has to be aware of a number of issues. The natural speech is not necessarily of statistic mean characteristic. It does not have to be at a value of 24, which would be the mean point of each scale. Such a profile would rather be set as targeted variation or taken as criterion when selecting a speaker for further recordings. The natural speech assessments must be seen as distinct profile, depending on the speaker's non-acted personality. Consequently, the acted variations need to be analyzed with respect to the natural profile.

As can be seen from the yellow median markers in Figure 1 the speaker, by acting, successfully manipulated the factors *N, A* and *C*. Acting in order to manipulate the score of factors *E* and *O* seems more difficult. While the attempt to lower the perceived extroversion in speech had only little effect, the attempt to raise the impression of openness in fact lead to a lower perceived score. Presumably, this could be due to three main reasons. First, the scores of the natural speech of these factors were already very low in *E*, respectively very high in *O*. Manipulations here could therefore be limited in scope. Second, the speaker was not able to successfully perform the desired manipulations. Other speakers might be more successful. Both of the preceding hypotheses need to be tested empirically by repeated experiments with different speakers. Third, this could indicate that the factors cannot be assessed reasonably by hearing speech samples only. In that case, the ratings and performances are expected to show incoherence. Section IV therefore analyzes the reliability of the obtained NEO-FFI ratings.

IV. Reliability Analysis

In order to gain insight into the ratings we look at their distributions. The distributions of the targeted manipulation factors are given in Figure 2. As in Figure 1 brown bars show the ratings for the low-score directed manipulations, blue bars show the high-score directed manipulations, lines represent a fitted normal distribution for the respective manipulations.

As a first observation, the ratings generally seem to relate to a normal distribution. We expect a general separability, with the exception of factor *O*. Here, ratings are very close to each other. Note, that since every speech variation was assessed using a full NEO-FFI test, Figure 2 shows only one a part of the information available for each speech recording. Only the targeted variations are depicted in Figure 2, i.e. when the instruction to the speaker was to lower the impression of

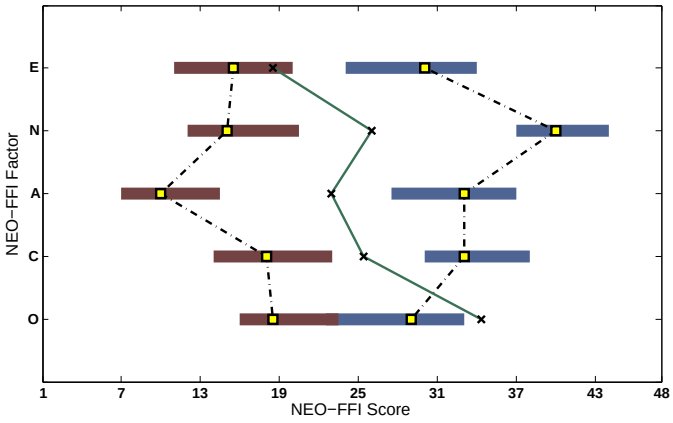

Fig. 1. NEO-FFI ratings of the recorded database. Brown bars (left) represent low-targeted manipulations, blue bars (right) high-targeted manipulations. Yellow markers represent distribution medians. The solid line in the middle connects the median ratings for the "natural" personality for the speaker.

factor *N* we show the distribution of 20 ratings for each of 3 examples. We do not show the distribution of all obtained *N* rating from all variations, although every variation is always assessed in terms of all 5 factors. Further, we observe that the variation of one factor is always accompanied by alternation of all other 4 factors as well. Consequently, every targeted factor variation exerts influence on all other factors, too.

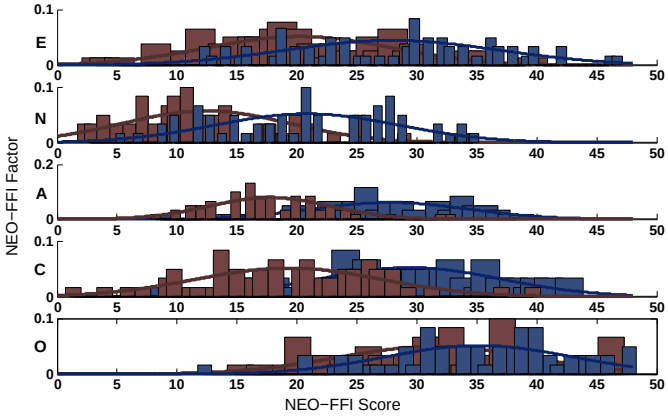

Fig. 2. Probability density distribution of NEO-FFI ratings of acted speech. Brown bars (left) show the ratings for the speech acted towards low scores, blue bars show the ratings for the speech acted towards high scores directed manipulation. The lines indicate a fitted normal distribution.

In order to estimate the consistency of our recorded data we calculate inter-factor consistencies and between-factor correlations analogously to the figures provided in the NEO-FFI test manual. Table I shows correlations between the factors and, on the diagonal, the coefficients for consistency, i.e. Cronbach's Alpha. Since we did not attempt to record the whole range of possible factor values, we actually compare two rather separated regions, e.g. one of higher than natural values, one of lower than natural values. The figures of consistency given in Table I account for the average consistency given the two separated targeted profile directions.

Note that the absolute height of consistency does not directly compare to the consistencies given in the NEO-FFI test manual since those figures are generated from more data, spreading over a wider range of possible factor values. Our sample basis, which consists 20 ratings of each of 3 selected recordings for a desired factor direction, results in 120 ratings for each factor only. The ratings cover three discrete regions of the factor values. The comparison of our figures to the figures provided with the NEO-FFI sample basis therefore yields indicative results only. However, when rebuilding the NEO-FFI factor structure, we observe overall high consistencies within the factors and lower correlation between the factors. Generally, cross-correlations among our ratings are higher than the figures provided with then NEO-FFI. Due to our small amount of data and the assumption, that some variance generated by the NEO-FFI cannot be explained by listening to speech only, this increased correlation level was expected. Factors N, A, C and E show overall moderate correlations, Factors N and A are even strongly correlated, i.e. these vocal impressions seem to be similar. Factor O shows overall weak correlation with other factors, which leads to the hypothesis, that this factor can be seen as relatively independent. Reconsidering the low separability between ratings from this factor and its relative independence on the one side, and looking at its high consistency on the other, we observe an ambivalent situation. Future experiments will therefore focus on factor structure and variance distribution in the factor structure. We hypothesize, that a lower number of factors will lead to an reduced structure, showing less cross-correlations. This will help to understand, how humans assess personality from speech. To sum it up, our rating of perceived personality impressions from speech using the NEO-FFI test structure seems promising. Raters are indeed able to differentiate between the recorded variations, the ratings show high consistency. Factor structure improvement has to be scheduled for future work. Also note, that at this point, these results do not address the issue of speaker independence, reproducibility over time, or scope of personalities that can be acted successfully. Section VIII discusses these and further limitations.

TABLE I
CONSISTENCY (DIAGONAL) AND CORRELATIONS OF THE NEO-FFI RATINGS.

	E	N	A	C	O
E	(.80)	-.55	-.59	.51	-.20
N		(.85)	.78	.67	.46
A			(.84)	.60	.41
C				(.88)	.30
O					(.83)

V. SIGNAL-BASED SPEECH ANALYSIS

In order to estimate the separability of the recorded vocal personality impressions by automatic means, we build a prosodic and acoustic feature classifier. We extract audio descriptors using a 10ms frame shift, and derive statistics from

these descriptors at the utterance level. Overall, we generate about 1450 features, which we have successfully used in our previous work on emotion recognition [6]. Features are extracted using Praat[1].

A. Extraction of Audio Descriptors

The audio descriptors can be sub-divided into 7 groups of extraction origin, e.g. *intensity, pitch, loudness, formants, spectrals, MFCC* and *other* features.

Taken from the time domain we extract the contour of *intensity* in decibel. *Pitch* features are calculated by means of autocorrelation. After converting pitch into the semitone domain we apply piecewise cubic interpolation and smoothing by local regression using weighted linear least squares. Taken from the spectral domain we extract perceptual *loudness* as defined by [10]. This measurement operates on a Bark filtered version of the spectrum and finally integrates the filter coefficients to a single loudness value in sone units per frame. We extract 5 *formant* frequencies using PLP and estimate the respective bandwidths. Further features from the spectrum are the center of spectral mass gravity (centroid), the 95% roll-off point of spectral energy and the spectral flux. These features will be referred to as *spectrals* in the following experiments. After filtering into the Mel domain, a discrete cosine transformation (DCT) gives the values of the Mel Frequency Cepstral Coefficients (MFCC). We extract a number of 16 coefficients and keep the "zero" coefficient. Referred to as *other* features we calculate the Harmonics-to-Noise Ratio (HNR), the Zero-Crossing-Rate (ZCR) and features related to speech rhythm.

B. Definition of Statistic Features

After finishing the extraction of audio descriptors the statistical unit derives means, moments of first to fourth order, extrema and ranges from the respective descriptors' contours in the first place. Special statistics are then applied to pitch, loudness and intensity, i.e. we estimate their spectral composition applying a DCT transformation. We subdivide the speech signal into voiced, unvoiced and silenced segments and calculate features on them alike. To model temporal behavior we append first and second order finite differences.

VI. Feature Selection and Classification

The performance of individual features given the vocal personality impressions as target classes is estimated by means of information theoretic consideration, i.e. we apply ranking by Information Gain [11]. Modeling uncertainty of the overall information and uncertainty reduction obtained by an individual feature this entropy-based filter generates a ranking. We determine an optimal feature set size by appending an increasing number of high gain features into the feature space. The global maximum of obtained classification scores determines the optimal set size. To obtain general estimates, we generate the ranking using 10-fold cross-validation and the full database. Figure 3 shows the development of accuracy scores when expanding the feature space.

Analyzing the ranking, we observe a predominance of MFCC-based features. Most important are the statistics derived from the unvoiced speech parts. Also features from intensity and duration of segments, as well as pitch derivatives are of high importance, e.g. the maximum intensity from unvoiced speech parts or the distribution and percentage of voiced segments overall. Although single MFCC coefficients cannot be interpreted in a linguistic way directly they nevertheless contribute to a complex description of spectral composition. In particular, calculating the first coefficient and drawing statistics from it seems promising. As the first coefficient can be understood to increase in case of correlation to a global overall fall of the spectral slope it can be interpreted as yielding high correlation when the slope is consistent within a class and varying in between classes. Higher coefficients correlate to higher cosine frequencies, i.e. they contribute to shape details in the overall spectral slope. The higher the coefficients the more peaks they represent. The intensity of unvoiced speech parts corresponds to power level when articulating unvoiced sounds, e.g. fricatives. Note that also the Zero-Crossing-Rate (ZCR) of unvoiced sounds is expected to increase, as articulatory power increases. However, the Information Gain Ratio (IGR) ranking scheme does not indicate overall usefulness for this feature. Finally, looking at the dispersion of pitch we see a discriminative character when we compare (average) length of the voiced segments to the (average) length of the unvoiced speech segments. This also suggests a difference in speech-to-silence ratio. Our features include a explicit speech-to silence ratio as an individual feature, however, like ZCR this feature is not given a high rank by the IGR filter.

In order to answer the question if we can generally distinguish between the different acted personality profiles we have carried out an initial classification experiment using Support Vector Machines (SVM) with linear kernel functions. Support vector machines view data as points in a multidimensional space. A hyper-plane in the space is constructed that separates two classes. Binary classification is extended to a multi-class problem using pair-wise classification. Maximizing the margin between the hyper-plane and the data SVMs offers a high degree of generalization, even when there are only small training data sets, as given in the present study. We evaluate classification results using 10 fold cross-validation.

Figure 3 shows the development of the classification accuracy when the feature space is increasingly expanded according to the IGR rank. Classifying on basis of a single feature only we observe an accuracy of approx. 28%, which is about three times chance level already. Using 10 features only, we already achieve about 50% accuracy. Including 10 additional feature we obtain another improvement of 8%.

Since our database size is small we presumably see effects of over-fitting when exceeding 40 features. As many of the top-ranked features are of MFCC origin, i.e. the features are not completely independent. The gray dotted line therefore plots results that are most likely "learned by heart" from our data.

[1] http://www.fon.hum.uva.nl/praat/

However, the solid line shows not over-fitted data. The dashed line region supposedly shows best results without losing too much generalization. A more precise statement about over-fitting has to be postponed to future experiments when more data will be available.

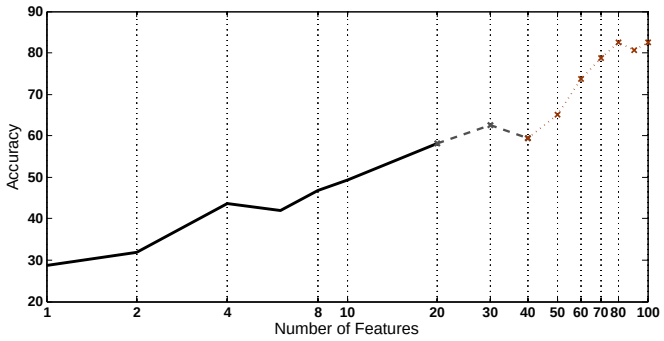

Fig. 3. Classification results using incremental feature space expansion according to IGR ranking. Line styles reflect expected degrees of over-fitting, i.e. dashed as most likely not over-fitted, dotted as most likely over-fitted. However, the dashed line region supposedly shows best results without losing too much generalization.

In sum, we achieve about 60% accuracy in a 10-class task. Given the baseline chance accuracy of 10%, we can clearly conclude that our models are able to capture relevant information for perceived vocal personality recognition. Automatic classification yields promising results for the current database.

Looking at class specific performance we observe that models of factor *N* and *C* are of highest performance. Precisions and recalls of these models show a harmonic mean of approximately 80% or higher. Within these factors both the high and the low variation models are of high performance. Also high extroversion in *E* can be classified to a reasonable extent. Most problematic are the models from classes *O* and *A*. When both, human and automatic *O* factor ratings turned out to be problematic, automatic processing of factor *A* also gives poor results, which is in opposition to the human raters' performance. Hence, our models fail to capture the class-relevant characteristics. A detailed analysis has to be scheduled for future work, which need to inquire about the unique prosodic structure of agreeable voice impressions.

VII. Correlation between Low-Level Speech Features and Vocal Personality Ratings

In order to gain insight into the quality of numeric prediction of the 5 factors from speech, we conduct a regression experiment in which we use the ratings of the labelers as ground-truth. In this experiment, we use all available ratings for the speech recordings, i.e. 20 ratings for each of 30 stimuli representing 10 target personality instructions. For analysis, we again use SVM regression. The algorithm for SVM regression tries to approximate a function that represents the training vectors and minimizes the prediction error at the same time. The risk of over-fitting is encountered by keeping the function as flat as possible. The algorithm establishes a tube around the function. Prediction errors that fall within the tube are being ignored, while all others are weighted due to their distance from the tube. Striving to find a trade-off between the function flatness and the tube width an upper bound of the vector weights is imposed and controlled by a threshold.

Figure 4 shows the correlations between the human ratings from the 5 factors and automatic factor prediction from speech when expanding the feature space. Feature space ranking was obtained by IGR evaluation as explained in Section VI. In opposition to categorical computation the continuous ratings had to be converted to discrete values. We chose a number of bins that represents half the value of the span of the ratings. Rankings were generated for each factor individually.

As a first result we clearly see a difference in prediction quality in between the factors. Starting with a correlation of 0.2 when using only a single MFCC coefficient to model acoustic properties we can increase the figure up to 0.67 by expanding the feature space to 20 features for factor *E*. We reach moderate to good correlation. We observe the steepest incline and overall best correlation at the same time. Also the correlation for factor *N* shows moderate scores. On the other hand, predicting the degree of openness seems to be most difficult, since the coefficients stay at a weak level. All other coefficients are in between weak and moderate correlation.

Looking at the distributions of features in the top ranks we see that for factors *O* and *C* predominantly MFCC features have been given top ranks. For the other factors the picture seems much more diverse. In terms of factors *E* and *A* also pitch features play an important role. More precisely, features that capture dynamics of pitch are given high ranks, e.g. standard deviation, slopes, ranges, derivatives. For *N* factor also loudness and intensity features contribute in great numbers to the top ranks. Here, foremost statistics describing the distribution are in high ranks, e.g. skewness or kurtosis. Interpreting our results, degrees of extroversion and agreeableness seem to be conveyed much more by tonal expression than degrees of other factors. In addition, intensity and loudness levels can be exploited to gain indications of vocal impression of neuroticism. Further research will need to focus on a detailed interpretation of these findings.

Comparing results from classification and regression analyses we generally see that predicting factors values and classifying for binary classes can be applied obtaining good results for the factors *N* and *E*. While classifying into high and low variations along the conscientiousness dimension also yields reasonable classification scores our models poorly predict the actual value of that factor. Results turned out to be inferior in terms of openness and agreeableness.

VIII. Limitations of Speech Assessment

Assessing the perceived vocal personality implicates methodological difficulties. In clinical phoniatrics the NEO-FFI questionnaire is used to examine the correlation between vocal disorders and personality of patients. Our method has to face a complexity that derives from various factors: The Big Five questionnaires in general are often criticized as their evidence relies on self report questionnaires and self report

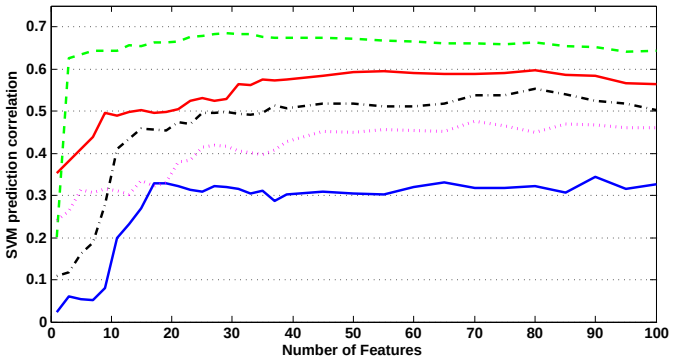

Fig. 4. Correlation between human (NEO-FFI) and automatic rankings over number of features when using incremental IGR ranking for the automatic classification. Key: Neuroticism: solid red line; Extroversion: dashed green line; Openness: solid blue line (below red line); Agreeableness: dash-dotted black line; Conscientiousness: dotted magenta line.

bias and falsification of responses are impossible to deal with completely. The NEO-FFI has been developed to do a self rating or a rating of a familiar person, but we let the listeners rate a voice that is not familiar to them. The vocal gender and vocal age of the voice could influence the perceived vocal personality. Also personal features of the listener could have influence on the ratings. It is well known in speech science that the spoken text has an influence on the perception of listeners. Further experiments need to compare the present results using different speakers. Our database comprises one speaker only. Furthermore, this speaker is a professional speaker. Although the ratings indicate, that our listeners indeed perceived the manipulated speech recordings as intended, we nevertheless need to be cautious when generalizing the results to naive speakers. There is no guarantee that his patterns of expression are in line with what untrained speakers do. After all, it is very important to point out that we do not aim to rate the personality of a speaker. We target the perceived personality, i.e. acoustic correlates of personality, taken from voice and speaking style.

IX. Conclusions and Outlook

In this paper, we investigated the application of personality assessment, as established in psychology, to expressive speech input. This creates a powerful framework, which we hope to use in the future to understand how semantic information is conveyed by the paralinguistic parts of speech. An understanding of how personality is encoded in spoken communication will be paramount to proceed from transcription of speech to interpretation of speech, and will be needed for future personalized speech synthesis systems.

We recorded an experimental database with a professional speaker and generated the Big 5 factor scores for the recordings by conducting listening test using the NEO-FFI personality inventory. Although our raters did not know the speaker, i.e. they only heard about 20 seconds of speech, the consistency analyses attest an overall applicability of the test scheme to vocal input. Furthermore, we extracted signal-based features capturing prosodic and acoustic speech properties. Initial classification results show overall good recognition accuracies of approximately 60% in a ten class task consisting of isolated, acted productions of high and low targets for the 5 personality traits. In these first experiments, personalities along the neurotic and extroverted scales could be classified best. Comparing results from classification and regression analyses we generally see that models predicting the factors neuroticism and extroversion perform best. Furthermore, high and low conscientiousness can be successfully discriminated. Predicting numeric agreeableness scores seems promising as well. In comparison the factors neuroticism, extroversion, openness to experience, and agreeableness, first results indicate, that factor openness cannot be assessed or predicted from speech using the NEO-FFI test scheme. Future work needs to focus on further factor reduction, since correlations between the remaining NEO-FFI factors are relatively high.

The present database was recorded from a single speaker under laboratory conditions, but we are planning future experiments including more speakers and also more diverse text material. Pooled together with refined methods to analyze speech, future work aims to enable well-known categories for personalities to be used for speech-based communication between man and machines.

Acknowledgements

This work was funded by Deutsche Telekom Laboratories. We would like to thank our colleagues Joachim Stegmann, Bernhard Kaspar, and Claus Cramer for funding this research and supporting it in kind and spirit. We also thank our anonymous reviewers for very detailed and informed feedback.

References

[1] L. R. Goldberg, "The structure of phenotypic personality traits," *American Psychologist*, vol. 48, pp. 26–34, 1993.

[2] P. Costa and R. McCrae, *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) manual*. Psychological Assessment Resources, 1992.

[3] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "Desperately Seeking Emotions: Actors, Wizards, and Human Beings," in *ISCA Workshop on Speech and Emotion*, 2000.

[4] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text," *Journal of Artificial Intelligence Research (JAIR)*, vol. 30, pp. 457–500, 2007.

[5] K. R. Scherer and U. Scherer, "Speech Behavior and Personality," *Speech Evaluation in Psychiatry*, pp. 115–135, 1981.

[6] T. Polzehl, A. Schmitt, and F. Metze, "Comparing Features for Acoustic Anger Classification in German and English IVR Portals," in *International Workshop on Spoken Dialogue Systems*, 2009.

[7] T. Polzehl, S. Sundaram, H. Ketabdar, M. Wagner, and F. Metze, "Emotion Classification in Children's Speech Using Fusion of Acoustic and Linguistic Features," in *Interspeech*, 2009.

[8] A. Gill and R. French, "Level of Representation and Semantic Distance: Rating Author Personality from Texts," in *Proc. of the Second European Cognitive Science Conference (EuroCogsci07)*, Delphi, Greece, 2007.

[9] J. Oberlander and A. Gill, "Individual Differences and Implicit Language: Personality, Parts-of-Speech and Pervasiveness," in *Proc. of the 26th Annual Conference of the Cognitive Science Society*, Chicago, IL, U.S.A., 2004.

[10] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*, 3rd ed. Springer, Berlin, 2005.

[11] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. John Wiley & Sons, 2000.