Modeling Speaker Personality using Voice

Tim Polzehl[1]*, Sebastian Möller*[1]*, and Florian Metze*[2]

[1]Quality and Usability Lab, Technische Universität Berlin / Deutsche Telekom Laboratories; Germany
[2]Language Technologies Institute, Carnegie Mellon University; Pittsburgh, PA; USA
tim.polzehl@telekom.de, sebastian.moeller@telekom.de, fmetze@cs.cmu.edu

Abstract

In this paper, we validate the application of an established personality assessment and modeling paradigm to speech input, and extend earlier work towards text independent speech input. We show that human labelers can consistently label acted speech data generated across multiple recording sessions, and investigate further which of the 5 scales in the NEO-FFI scheme can be assessed from speech, and how a manipulation of one scale influences the perception of another. Finally, we present a clustering of human labels of perceived personality traits, which will be useful in future experiments on automatic classification and generation of personality traits from speech.

Index Terms: extra-linguistic speech properties, perceived personality modeling, speaker characteristics

1. Introduction

Within the last decade, researchers have conducted a substantial number of experiments on assessment, extraction and analysis of non-linguistic properties of speech, such as age, gender, emotion, affect, and related speakers states and properties [1].

In this work we continue our research on assessing "personality" from speech. Humans assign personality rapidly and automatically [2], which allows us to quickly construct a model of a person we meet. In interactive settings, we expect that "natural" communication accommodate our personality models and expectations. Human-like interaction therefore requires both: systems that recognize and synthesize personality and designers who understand how humans assess and react to personality. This work contributes in methodology and understanding on how much of a personality can be inferred from speech.

Nass and Lee established that humans could infer personality impressions even from automatically synthesized speech [3]. They find that humans tend to have positive feelings toward a person they encounter and assume to have similar personality characteristics, and that it is possible to generate extroverted and introverted synthetic voices, which people will recognize.

In non-interactive settings, some companies are seeking to establish brand image monitoring for expressive speech-to-text solutions, where guidelines on how to design acoustic, prosodic or even language generation aspects will be paramount.

In [4] we analyzed and classified personality portrayals using a pre-fixed text passage. We now open up the text passage to spontaneous speech utterances. We assign personality along the "Big 5" personality traits [5] by conducting listening tests. We analyze our data in terms of the speaker's and listeners consistencies by means of univariate and multivariate statistics, and compare the results to previous findings.

2. Assessing Personality

In psychology, many attempts to define the concept of personality have been postulated. Following Ryckman [6], personality can be defined 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." Accordingly, researchers have worked on diverse methods to assess such characteristics. Following the so called "trait theory of personality", which sees personality as a defined set of habitual patterns of behavior, thoughts, and emotions, that manifests itself in term of measurable traits, we apply the NEO-FFI [7] scheme. We chose this theory, because it is broadly seen as empirical observation, not a fundamental theory. The NEO-FFI questionnaire measures personality along five bipolar "scales", which are explained 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 a 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.

Human raters generate values on all scales by rating 60 items from the questionnaire, which are coded using a 5-point Likert scale ranging from 'strongly disagree' to 'strongly agree'. A coding scheme transforms the ratings into 5 scale values, ranging from 0 to 48. All 5 scales together generate an overall personality profile of a person. As the highest correlation is -.36, the individual scales are seen as independent from each other. The assessment scheme has been validated with high consistency, including translations, cross-cultural experiments and retests. In particular, we used the German NEO-FFI,

which has been validated with more than 12.000 test persons using various stimuli in various situations.

While NEO-FFI can be applied to both, self-assessment and observer's-assessment, we focus on the context of a voice-based communication, i.e. the scales are to capture vocal manifestations of personality in speech as perceived by a listener.

For application in automated interfaces, systems need to estimate the user's personality within seconds, and from speech only. Presenting long questionnaires to the user is mostly inapplicable. Also, in traditional application, raters have a multitude of cues on which to base their judgment, for example previous or external knowledge, if they know the person to be assessed.

In our experiments, attribution happens on basis of auditory impressions exclusively. Although there are many studies on the relationship between personality and speech, little empirical work exists. Apple [8] finds that prosodic speech characteristics, such as pitch and speaking rate, can influence the attribution of truthfulness, empathy, and 'potency'. Scherer [9] analyzes personality traits and observes that extroverted speakers speak louder, and with fewer hesitations. He concludes, that extroversion is the only factor that can be reliably estimated from speech. Mairesse [10] also confirms prominence of prosodic properties for modeling extroversion, and that extroversion can be modeled best, followed by emotional stability (neuroticism) and openness to experience. Of course, personality is also expressed with linguistic cues [11, 10].

In [4], we presented results of an automatic assessment and classification of all five NEO-FFI traits on a subset of very restricted settings of speech, i.e. acted speech consisting of a prefixed text passage. Factor analyses showed , that for the speaker at hand, the NEO-FFI indeed captures vocal personality in terms of N,E,A, and C scales. Openness, however, caused most confusions by human and machine classification.

3. Data Collection

In our initial data collection [4], our speaker had been given a fixed text paragraph, that he portrayed in various personality profiles. Opening up to more realistic settings, we now asked the same actor to speak freely and spontaneously about associations and/or descriptions, while presenting him a series of images. We asked him to prepare and perform 10 voice personalities, representing persons with either high or low values on each of the five NEO-FFI scales, which will be referred to as conditions. The conditions were again defined using the original, textual NEO-FFI personality trait descriptions. Every portrayal starts with a very brief description of the image presented. It is followed by associations and feeling that last up to a minute of speech. In total, we presented 7 images. We selected the images in order to create an opportunity for verbal descriptions in diverse styles, i.e. the images function as subject of interpretation when acting a personality-driven perspective. Two black-and-white images are borrowed from the Thematic Apperception Test (TAT) [12] (Nr. 2, 4), which is a projective psychological test designed in 1935 for clinical application. The pictures are claimed to trigger the subject's unconscious to reveal aspects of personality, however, some psychologists nowadays see this theory as highly controversial and outdated. We therefore included other images, e.g. a rather violent situation of a masked person holding a long bloody knife, a rather peaceful situation showing the Dalai Lama folding his hands and smile, the face of an old rich elegant woman, a schematic drawing of a young attractive woman, as well as surrealistic line drawings of abstract faces or body parts. Our choice being subjective, the impact on personality perception will be discussed in Section 4.1.

Table 1: Consistencies and correlations between NEO-FFI scale ratings from speech.

	N	**E**	**O**	**A**	**C**
N	(.93)	-.67	-.34	-.14	-.43
E		(.91)	.47	.28	.19
O			(.87)	.53	.34
A				(.90)	.22
C					(.93)

Our material was recorded in three sessions, spread over three months. Each session comprised recordings of each of 10 conditions for each of 7 images. The effective speech data amounts to more than three hours of recordings. To generate human personality labels, we followed the procedure described in [4]. Approximately 100 raters (mostly students at Berlin Universities, mean age 29 years, 53% male) listen to the stimuli through high-quality headsets as often as they wanted to, while completing a series of NEO-FFI questionnaires about their impression of the speaker's recordings. Lilliefors tests, which resemble a Kolmogorov-Smirnov test for normality with mean and variance unknown, attest overall normal distributions for all conditions' ratings ($p<0.05$), i.e. labelers did not rate randomly.

Table 1 shows Pearson's inter-scale correlations calculated on ratings from all images and recording sessions. While inter-scale correlations in traditional NEO-FFI application results very weak on average (0.14), the relative correlation pattern generally matches our findings in speech application. Highest correlation occurs between N and E, as well as between N and C, e.g. the more neurotic, the less extroverted and reliable people are expected to be. The correlation between O and A is probably characteristic for either our speaker or the assessment of our speaker from speech, as it reconfirms our finding from [4].All correlations are significant ($p<0.01$), and weak on average (0.36). The diagonal shows excellent intra-scale consistencies (Cronbach's Alpha [13]), ratings are coherent, systematic.

4. Experiments

To further validate the data and approach, we perform three experiments: first, we investigate the differences between portrayals generated in different recording sessions, spread over three months. Next, we look at scale interplay, as acting on any one scale can also influence the values on the other scales. Finally, we perform clustering of our data. Given that dependencies between scales exist, earlier work showed that some distinctions are very weak. This gives another perspective on which properties can be estimated robustly from speech, and which cannot.

4.1. Time and Text Dependency

In order to analyze reproducibility, we recorded speech in three distinct sessions, several weeks apart. Each portrayal was assessed by at least 15 raters. To see whether the mean perceived personality attribution depends on the recording date, we applied Tukey's honestly significant difference criterion for post-hoc tests of one-way analyses of variance. Although few anomalies (4 out of 40 comparisons) exist, the vast majority of ratings on stimuli from different recording sessions do not differ significantly ($p<0.05$). While the image seems to be the main cause of anomalies, our speaker generally produces consistent portrayals. Pooling all images, most anomalies occurred

on openness ratings.

Pooling all sessions, we analyze the influence of the images, i.e. the spoken text. In general, results show the expected significant difference ($p<0.01$) between the variations. Assessments do not vary significantly within the high or low variation groups, i.e. personality portrayals were not significantly different between the images. On the E and A scales, one out of seven images does not show significant difference, but the expected tendency. On the O scale, however, three out out of seven images show exceptions. Given the slightly lower labeler consistency (cf. Section 3), O seems to be most difficult to assess and/or portray. These results confirm our findings on a fixed text [4]. Also personality impressions from spontaneous speech can generally be produced and perceived consistently for all scales, O being most challenging.

4.2. Cross-Scale Interference

Naturally, any acted speech will change the perception not only along the intended scale, but also on all other scales. We therefore use the term "primary induction" for the targeted condition, while "secondary induction" refers to the unintentional effect on all other scales. Pooling all images and recording sessions, i.e. obtaining sample sizes greater than 150 for each condition, we partition the ratings of each scale into three groups: high primary induction, low primary induction and no intentional manipulation of this (but of other) scales. Applying the same methodology as in Section 4.1 for each individual scale, all ratings on low conditions (with the exception of N resulting a clear trend) show significantly lower values than the collections of non-manipulated samples, which are again significantly lower than the high conditions' ratings ($p<0.01$).

Table 2 shows that some scales are more affected by secondary inductions than others. Arrows represent significant influences ($p<0.01$). For instance, the perception of C is influenced only by a deliberately increased N, for which C decreases. When deliberately increasing C we observe no secondary effect, but when deliberately lowering C, we cause decrease of perceived E, O, and A. Overall, C seems to be robust and relatively independent from other scales, while the O and E scales are least robust. Including results from above, E shows susceptibility to secondary inductions, but not to a extent that causes confusions of condition groups. Further, N and E actings are perceived reciprocally, i.e. raising N causes falling E, and vice versa, which also manifests itself in higher negative inter-scale correlation, cf. Section 3.

Similarly, decreasing A, e.g. becoming more egocentric, causes a less open impression, increasing it acts the opposite way. Following previous experiments on our fixed text setting, C shows least affection by secondary inductions, followed by N. E, O and A however are more affected. A detailed comparison of interplay on fixed and spontaneous texts is ongoing.

4.3. Cluster Analysis

While a 5-dimensional NEO-FFI personality profile is a rich and powerful information, it may in many situations be too hard to estimate robustly from speech. A speakers' perceived "personality", however, consists of all 5 scales, jointly. We therefore analyze how similar the induced personalities are being perceived as joint profiles.

We again pool sessions and images and perform a multivariate data-driven top-down clustering of corresponding ratings. Figure 1 shows two cluster trees, fixed text data (top) and spontaneous speech data (bottom). The height of the horizontal lines

Table 2: Influences of primary inductions on secondary scales; e.g., increasing E causes decease of N and increases of O.

		primary									
		increase					decrease				
		N	**E**	**O**	**A**	**C**	**N**	**E**	**O**	**A**	**C**
secondary	**N**	·	↓	↓			·		↑		
	E	↓	·	↑	↑			·	↓		↓
	O		↑	·	↑				·	↓	↓
	A	↑			·					·	↓
	C	↓				·					·

connecting the clusters corresponds to the similarity between clusters, as expressed by the distance at which they are split during clustering, i.e. the higher the line, the more dissimilar the conditions or clusters are perceived. To compute the distances, we apply an unweighted average cluster linkage to the matrix of Mahalanobis distances between condition means. While distances between the observations are measured by covariance, linking cluster averages generally tries to balance tightness and elongation effects.

We observe that splits occur at smaller distances in the spontaneous case, indicating that differences are harder to produce and/or perceive. Given the experience of researchers in the related field of emotion recognition from speech, this is an expected effect, and the speaker might be trying to "hyperarticulate" personality in the fixed text case, without the freedom of expressing personality by varying the choice of words. On the fixed text, the actor was given time in advance to prepare distinct versions of the portrayals, repeating the same text all the time, acting disfluencies like pauses and hesitations consciously. In spontaneous speech, these disfluencies might also be caused by cognitive processes.

The basic structure of the clustering trees seems consistent between the two scenarios. Increased neuroticism (N+) and extroversion (E+) are split early in the process, with increased openness (O+) being very similar to E+. Accordingly, the very neurotic personality is perceived clearly different from all others, while very open and very extroverted actings converge. Table 2 also shows that E+ induces O+ and vice-versa.

The A+, C+, E- triplet and the C-, A- pair occur in both data condition. Decreased openness (O-) and neuroticism (N-) on the other hand form a less stable pair, which is grouped with A+, C+, and E- in the fixed text, and C- and A- in the spontaneous data. It is interesting to note that four out of five lowered targets are clustered together for the spontaneous speech condition, E- being the only exception. More experiments, however, are needed to understand how stable these classes are against other influences (text, situation, speaking style, etc.).

5. Summary and Outlook

This paper presents a continuation of recent experiments on assessment of perceived personality. Inspired by the desire to understand human assessment of perceived personality from speech we analyze newly recorded spontaneous speech data, in which a professional speaker systematically generates 10 personality profiles. Our results confirm that the "Big Five" methodology can be used to consistently assess personality impressions from speech. In a user test, our raters show high consistency and overall weak correlation between most NEO-FFI scales when answering the NEO-FFI questionnaire's items.

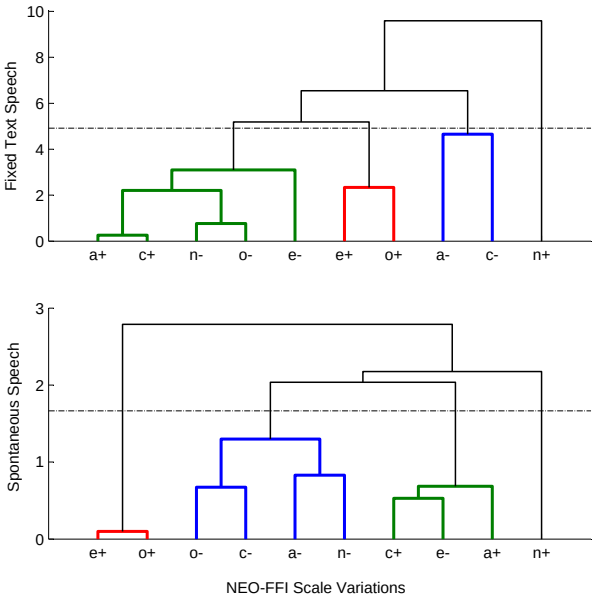

Figure 1: Cluster trees showing similarities between multivariate personality portrayals; clusters (arbitrarily color-coded into 4 clusters, dashed line) are very similar between fixed text (top) and spontaneous text (bottom), only N-/ O- assignments and absolute distance magnitudes differ.

Our findings confirm and exceed finding by Scherer [9] and Mairesse [10], as our systematic approach allows us to understand interdependencies and generate more specific results for all five personality traits. While earlier work concluded that only isolated personality traits can be manipulated successfully, we now show, that our speaker can reproduce results over several months and on basis of different spontaneous texts.

Analyzing scale interplay, i.e. the effect of acting any one scale manipulation on other scales, confirms a reciprocal perception of neuroticism and extroversion, i.e. increasing one causes significant decrease of the other. They also show moderate negative inter-correlation when assessed from speech. Similarly, decreasing agreeableness, e.g. becoming more egocentric, causes a less open impression, increasing it acts the opposite way. Conscientiousness reveals most robustness with this respect, openness and extroversion least. Overall, openness seem to be most challenging in acting by and/or perception from speech, also raters' consistency slightly decreases. Further research will focus on dismantling these dependencies.

Integrating individual trait characteristics into an overall personality impression, we conduct cluster analyses. This gives insights into how similar the various personality profiles are perceived and can be used to understand the data structure. Comparing fixed text and spontaneous text data, we observe smaller distances in the spontaneous case, indicating that differences are harder to produce and/or perceive. The basic structure, nevertheless, seems consistent between the two data sets. For instance, portrayals of a very neurotic personality are perceived clearly distinct from all others, while very open and very extroverted actings are very similar.

Analogously to the development and spirit of early works on emotion recognition, and with perspective to automatically measure and maintain a certain perception of voice qualities in a corporate environment, we reduced the complexity of our task by choosing an actor. We emphasize, that current results cannot be generalized to everyday speech or other speakers. To differentiate speaker-specific from assessment-specific effects in our results, we currently record 30 new speakers. Opening up from acted to more realistic speech, these speakers are also non-professionals. When leaving a corporate setup we do not know the application domain or speaking situation, which will lead to more variation and influence overlay, e.g. speaking style, emotions, interlocutors, etc. In particular, future work will need to focus on determining linguistic overlay, i.e. the interplay between a linguistic message, and how it is being delivered in speech.

Finally, we emphasize that we do not primarily aim at assessing the personality of a speaker, as the term is used in psychology. Rather, we target a perceived personality impression in communication, as rendered by interlocutors or listeners. Depending on social environment, situation, etc., personality profiles are known to show variations. However, overall (long-term) profiles are presumed to be relatively invariant after adolescence. In our case, raters assess the speaker on basis of a short utterance only. Analyses of perceived personality stability on longer material and the interplay between speakers and listeners traits as well as the communication situation on perceived personality need to be addressed in future work.

6. References

[1] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The INTERSPEECH 2010 Paralinguistic Challenge," in *Interspeech*, 2010.

[2] C. Nass, Y. Moon, B. Fogg, B. Reeves, and D. C. Dryer, "Can computer personalities be human personalities?" *International Journal of Human-Computer Studies*, vol. 43, pp. 223–239, 1995.

[3] C. Nass and K. M. Lee, "Does computer-synthesized speech manifest personality? experimental tests of recognition, similarity-attraction, and consistency-attraction," *Journal of Experimental Psychology*, pp. 171–181, 2001.

[4] T. Polzehl, S. Möller, and F. Metze, "Automatically assessing acoustic manifestations of personality in speech," in *Workshop on Spoken Language Technology*. Berkeley, U.S.A.: IEEE, 2010.

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

[6] R. M. Ryckman, *Theories of Personality.* Thomson/ Wadsworth, 2004.

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

[8] W. Apple, L. A. Streeter, and R. M. Krauss, "Effects of pitch and speech rate on personal attributions," *Journal of Personality and Social Psychology*, vol. 37, no. 5, pp. 715–727, 1979.

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

[10] 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.

[11] J. Oberlander and A. Gill, "Individual Differences and Implicit Language: Personality, Parts-of-Speech and Pervasiveness," in *Cognitive Science Society*, Chicago, IL, U.S.A., 2004.

[12] H. A. Murray, *Explorations in Personality*. Oxford University Press, 1938.

[13] R. Zinbarg, W. Revelle, I. Yovel, and W. Li, "Cronbach's, Revelle's, and McDonald's: Their relations with each other and two alternative conceptualizations of reliability," *Psychometrika*, pp. 123–133, 2005.