A Review of Personality in Voice-Based Man Machine Interaction

Florian Metze[1], Alan Black[1], and Tim Polzehl[2]

[1] LTI, Carnegie Mellon University; Pittsburgh, PA; USA
[2] Q&U Lab, Technische Universität Berlin; Berlin; Germany
fmetze@cs.cmu.edu

Abstract. In this paper, we will discuss state-of-the-art techniques for personality-aware user interfaces, and summarize recent work in automatically recognizing and synthesizing speech with "personality". We present an overview of personality "metrics", and show how they can be applied to the perception of voices, not only the description of personally known individuals. We present use cases for personality-aware speech input and/ or output, and discuss approaches at defining "personality" in this context. We take a middle-of-the-road approach, i.e. we will not try to uncover all fundamental aspects of personality in speech, but we'll also not aim for ad-hoc solutions that serve a single purpose, for example to create a positive attitude in a user, but do not generate transferable knowledge for other interfaces.

Keywords: voice user interface, paralinguistic information, speech processing.

1 Introduction

Every speech act transmits not only a linguistic message ("text"), but it also encodes additional information in "how" things are being said. This is true for human-human communication, as well as for man-machine exchanges. In this paper, we will discuss the role of "personality" in voice-based user interfaces, and the state-of-the-art for implementing systems that can recognize and synthesize personality features. We will discuss ways to analyze speech data recorded from humans, advanced speech synthesis methods, and ways to encode personality information in the original text message, which is being transmitted. We believe that future research on voice-enabled human computer interfaces must go beyond the analysis of "what" is being said, and should include aspects of "how" it is being said. This capability is needed in order to adapt to an unknown user, or a known user's changing state of mind. The system must then send consistent messages across all channels used, i.e. it would express the same message by using different words, and by modifying the voice characteristics used.

Personality is certainly the richest means for characterizing, and even classifying people [16], and people assign it rapidly and automatically [18]. This instinct allows us humans to quickly construct a model of a person we meet, and predict a wide range of attitudes, behavior, and other properties, which we expect to encounter during interaction. Personality can for example be used to differentiate the introverted from

J.A. Jacko (Ed.): Human-Computer Interaction, Part II, HCII 2011, LNCS 6762, pp. 358–367, 2011.

the extroverted, the shy from the exuberant, the egoist from the altruist, or the conservative from the adventurous. We will assume that extroverted persons talk more than they listen, and use strong language, while introverted persons listen more, and use qualifiers such as "maybe", or "perhaps". The concept of personality gives us cues on what to expect from others, and how to behave ourselves. Descriptions such as "extroverted" or "introverted" serve as a shorthand descriptor for a bundle of traits, which we attribute to persons. Interestingly however, self-reporting of personality traits often leads to different results then attribution by others.

2 Personality in Voice-Based Man Machine Interaction

In an advanced voice user interface, the computer should be aware of the human's personality and tailor its response accordingly. Similarly, the user's behaviour will be influenced by his perception of the system's personality, conveyed by what the system says, and how it is being said. In the "Computers as Social Actors" (CASA) paradigm, Nass and Brave [16] postulate that humans communicate with machines just as they would with another human. Generally, when people encounter someone who seems to have a personality like their own, they tend to have positive feelings toward that person [24]. They conclude that designers of user interfaces should therefore seek to manipulate the speech characteristics of any technology that can produce speech, and thereby give it a personality. If one wants to be able to adapt to unknown users, possibly in real-time, the human's speech should also be analysed for personality traits, as should other input channels. In an automated voice user interface, the assessment of a user's personality must be done within seconds, and on the basis of the speaker's voice only. Methods established and verified in psychology, like the use of long questionnaires [7], are therefore inapplicable, or at least cumbersome.

In a more general setting, personality is also a property of embodied conversational agents (ECAs) [4]. Cassell et al. show that perceived personality of the agent is a major factor in the perception of such user interfaces [2]. Catrambone et al. list the personality of the user as one of the factors to be included in an evaluation of ECAs [5], arguing that an understanding of the mechanisms involved will eventually allow the design of appropriate personalities. [6] presents on-going work which uses a Wizard-of-Oz paradigm, because personality can not currently be analysed and synthesised satisfactorily using fully automatic means, as would be required.

Given the above it is clear that personality must be modelled properly in the audio channel of any speech-enabled multi-modal user interface.

While we have used the term "personality" many times already in this paper, we have not formally defined it yet. Following Ryckman [25], 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.*" In our own work, we follow the trait theory of personality [12], and see personality as a defined set of habitual patterns of behaviour, thoughts, and emotions. We can then apply an assessment scheme using the "Big Five" NEO-FFI [7] personality traits. We chose this scheme, because the traits are seen as empirical observations, not a fundamental theory, which aims to fully explain personality. [9] gives an overview of different schools of describing the rich concept of personality.

The NEO-FFI describes personality traits along five ordinal dimensions, which are called "scales", namely *Neuroticism (N), Extroversion (E), Openness (O), Agreeableness (A),* and *Conscientiousness (C).* Human raters generate another person's profile by giving answers to 60 propositions from the NEO-FFI questionnaire (called "items") using a 5-point Likert scale ranging from "strongly disagree" to "strongly agree". A self-report form is also available, but was not used for the experiments described here. The 60 items are then aggregated into numeric values for the 5 scales using the NEO-FFI coding scheme. The questionnaires and the resulting scales and factors have been validated with high consistency, including translations, cross-cultural experiments and retests, confirming the reliability of this approach for a large number of conditions. The German NEO-FFI, which was used in our experiments described below, has been validated with more then 12.000 test persons.

In the context of a voice-based communication, the scales correspond to vocal manifestations of *perceived* personality traits, unless the participants have other cues on which to base their judgement, for example previous, external knowledge, or the transmission of a message in another, conflicting personality (see Section 0). Attribution happens on basis of auditory impressions, and it is questionable how this compares to the conventional assessment methods, where raters know the person to be rated. In studies, Nass et al. find low, but significant correlation for synthetic speech [17]. This shows that personality impressions can be generated by the choice of a voice, and that they can influence the perception of other information presented at the same time. Our results below confirm this conclusion.

3 Recognizing Personality in Speech

Apple et al. [1] found that pitch and speaking rate influence the perception of speakers' voice with regard to factors such as truthfulness, empathy, "potency", amongst others. They also observed interplay between the message (the text which was spoken) and the effect of a manipulation of the above factors towards the attribution. Scherer & Scherer [26] analysed prosodic features such as pitch and intensity, and observes that extroverted speakers speak louder, and with fewer hesitations. They suggest that extroversion is the only factor that can be reliably estimated from speech. Mairesse [15] also finds that prosodic and acoustic features are important cues for recognizing extroversion, and that extroversion can be modelled best, followed by emotional stability (neuroticism) and openness to experience. Finally, Nass and Lee established that humans could infer personality impressions even from automatically synthesized speech [17]. They find that humans are attracted more to a voice that is similar to their own, and that it is possible to generate extroverted and introverted synthetic voices, which people will recognize as such. Our recent, more systematic work roots in previous experiments on emotion recognition [23], which also showed the benefit of relying on multiple information sources, acoustic and linguistic cues in this case. In [22], we present results of an automatic assessment of all five NEO-FFI traits.

3.1 Database and Human Perception of Personality in Speech

We recorded the "natural" voice of a professional speaker, who had previously recorded voice prompts in speech dialog systems, and was used to working with voice

coaches. We then presented him the original descriptions of the 5 NEO-FFI personality traits from the NEO-FFI manual. He prepared 10 voice personalities, representing persons with either high or low values on each of the five scales. We therefore have 11 different recording conditions: 2 extremes on each of the five scales, plus "normal".

The spoken text is designed to resemble a neutral, complete phrase, typical for a hotline, which comprises a welcome, information on a voucher redeemer service, and a short goodbye. Each recording lasts about 20s. We recorded at least 20 takes of each of the conditions, more than an hour of speech in total. All speech samples were annotated for "artificiality" by two experienced labellers, and we retained for our human rating experiments the three least artificial takes for each condition.

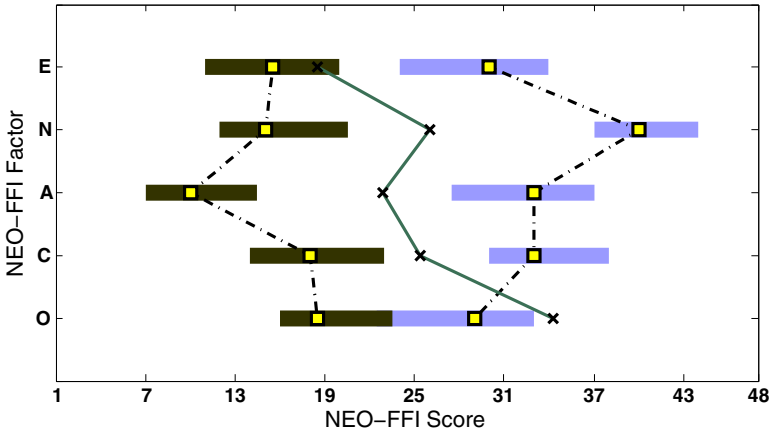

Fig. 1. NEO-FFI ratings for speech: brown bars (left) represent inter-quartile ranges of ratings from variation towards low values, light blue bars (right) towards high values. Vertical lines connect the medians, i.e. solid line for "normal personality", dashed for acted variations.

87 raters rated 8 takes from different conditions on average. 20 different raters rated every take. Raters could listen to the takes through high-quality headsets up to 5 times, while completing a NEO-FFI questionnaire about their impression of this take's speaker. Overall, over 600 questionnaires for all 5 scales were generated.

Fig. 1 shows the distribution of the raters' assessments for both the acted and the natural speech samples for the 5 factors. Each data point represents 60 ratings from 3 different takes. Overall, raters were able to label the acted personalities quite well, as nearly all the conditions were perceived as intended by the actor. In our recordings, the speaker successfully varied the values of the factors *N*, *C*, and *A*, while *E* and *O* seem 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 lowered the perceived score. This could be due to the "natural" value for this speaker being quite extreme already for *E* and *O*, an inability of our particular speaker to act these traits, or a general difficulty in perceiving and assessing these modifications from speech, or our speech sample. Further experiments will be needed to answer these

questions. These findings coincide with [26] and [15], and show that impressions of extroversion and neuroticism can be distinguished using speech. Furthermore, we show for our speaker that all 5 traits can be varied and recognized. We also observe low differences between the distribution variances of ratings on different conditions.

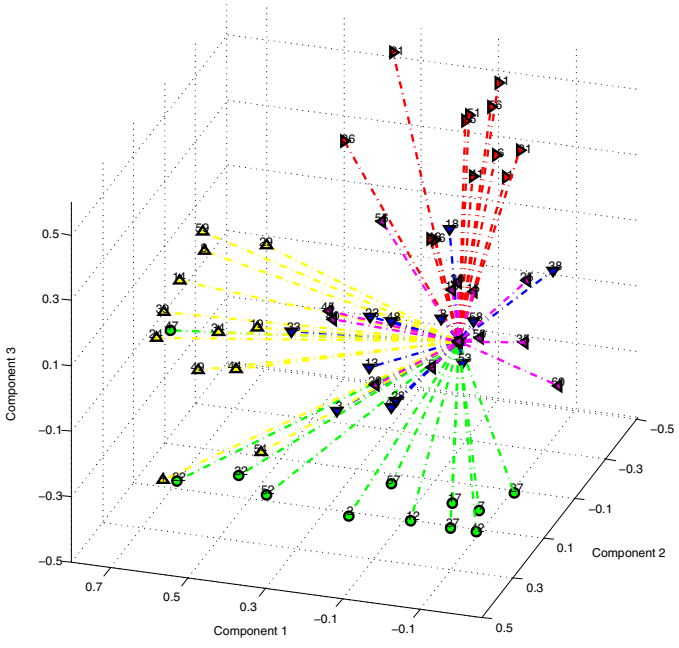

Fig. 2. NEO-FFI loadings of the factor scores, shown in a three-dimensional projection on the user-test assessment space. The original loadings, established using the questionnaire for known persons (marked by different colors), are well reproduced in the assessments of acted speech, as the colored "bundles" generally point in the same directions.

In order to compare the factors in the traditional NEO-FFI coding scheme with the underlying structure of the ratings on our recorded speech data, we conducted an exploratory factor analysis [8], hypothesizing the presence of 5 latent factors in the user ratings. Figure 2 shows the 3 most dominant (latent) factors, in the given NEO-FFI coding space. Lines correspond to directions of item loadings in our data. Colours correspond to item loading membership in the NEO-FFI coding scheme. The coding found in our data correlates quite well with the NEO-FFI coding scheme, as most factors with the same colour (original loadings) point in a similar direction (new factor loadings). We also observe a low number of cross-loadings between factors, and low to moderate commonality for almost all items [22].

In sum, this experiment shows that the NEO-FFI scheme can not only be used for assessment of known persons' personalities, but also to create profiles of perceived personality, from listening to short samples of speech.

3.2 Automatic Recognition of Personality from Speech

Our automatic system computes and classifies prosodic and acoustic speech properties. The feature set leverages previous work on emotion recognition [23]. We extract audio descriptors such as 16 MFCC coefficients, 5 formant frequencies, intensity, pitch, perceptual loudness, zero-crossing rate, harmonics-to-noise-ratio, centre of spectral mass gravity (centroid), the 95% roll-off point of spectral energy and the spectral flux, etc., using a 10ms frame shift. From these descriptors, we derive statistics at the utterance level, separate for voiced and unvoiced regions, on speech parts only. These statistics include means, moments of first to fourth order, extrema, skewness, kurtosis, and ranges from the temporal contours over one utterance. To model temporal behaviour, we append first and second order finite differences. In total, 1450 features are being computed. Before classification or regression using a Support Vector Machine (SVM) with linear kernel functions [30], the most salient features are being selected by ranking them according to Information Gain Ratio (IGR) using an Sequential Floating Forward Selection (SFFS) wrapper, and only retaining the top N features in a given set.

Using 10-fold cross-validation, we obtain an accuracy of approximately 60% on our balanced ten-class classification task, consisting of the high and low targets for the 5 personality traits, which is six times the chance level. Because humans have only been asked to fill in NEO-FFI questionnaires, and did not perform a classification task, a human baseline cannot be computed. Interestingly, high and low neuroticism and conscientiousness, as well as high extroversion can be recognized better than the other classes, with class-specific *F*-measures between 0.70 and 0.89, while the other classes perform between 0.32 and 0.54. Best results were achieved when using about 40 features, although very little change occurs as soon as at least 20 features are being retained. High extroversion (*E*) can be classified well, which is in line with observations by [26, 15]. Most problematic are the *O* and *A* factors. Different from separability by humans (see Figure 1), automatic classification gives poor results for *A*. *O* seems to be hard for both human and automatic classification.

Analysing the most salient feature types, 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. Features capturing dynamics of unvoiced speech parts are generally sensitive to strength of fricatives and plosives. Features capturing pitch variation and derivatives are generally sensitive to intonation movements. Along with cepstral features, which also partly capture spectral sharpness and tilt, the importance of these features seem to be in line with results from auditive analyses presented in earlier work.

In many applications, however, we may not be interested in automatic classification of personality, or in a personality assessment by a machine, but in the reproduction of a human rating by a machine.

We therefore conducted a regression experiment, in which we use the (numeric) ratings of the labellers as ground truth. In this experiment, we use all available ratings for the speech recordings, and SVM regression. Correlation analysis shows how different the predictions by humans and machines are, for the various factors. As in

the classification experiments, there is almost no change when using more than 40 features. Analysing the top ranked features, we see that for factors *O* and *C*, predominantly MFCC features are being used. For the other factors, the picture seems much more diverse. For factors *E* and *A*, features that capture dynamics of pitch are given high ranks, e.g. standard deviation, slopes, ranges, derivatives. For *N*, loudness and intensity features are prevalent, using statistics describing the distribution, 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 focus on a detailed interpretation of these findings. Generally, our findings are again in line with previous work on signal-based analysis [26, 15].

Comparing results from classification and regression analysis, we observe that predicting factors values and classifying for binary classes can be applied with good results for factors neuroticism (*N*) and extroversion (*E*). While classifying into high and low variations along the conscientiousness (*C*) dimension also yields reasonable classification scores, our models poorly predict the value humans would assign to that factor. Relatively poor results are achieved for openness (*O*) and agreeableness (*A*).

4 Synthesizing Voices with Personality

The messenger influences our perception of the message [16]. For isolated experiments in lab settings, it may be enough to manually control volume, pitch, pitch range, and speech range in a desired way, but by now a significant body of work in "expressive" synthesis exists, which allows to systematically modify voice properties associated with emotions, or personality. In practice, volume is very hard to control in real-world settings, for example over a telephone line.

A number of groups have been successful at synthesizing different emotional states, which is similar to synthesizing speech with personality. This is most often done by using acted data of different states, and using models trained from that data to impose prosody (intonation, duration and phrasing) on synthetic output. It is possible to classify such synthesis attempts into two forms following the current two major techniques in speech synthesis.

Using unit selection [13] technology, where appropriate sub-word units are selected from large natural speech database, requires that the database itself contains examples of the desired emotional state. [10] achieved this by deliberately recording different versions of their database with the range of desired emotional states. This did have some success, but it was of course limited to the types of data recorded, and to some extent the domain of the recorded data. [28] however take a more indirect approach. For a spoken dialog system, instead of recording prompts in isolation with explicitly stated emotional variation, they recorded the prompts within an actual simulated spoken dialog, where the voice talent plays the machine end. In their work they find the voice talent modified their prosody naturally given the dialog context. Also, by adding a prosody feature based on the classification of different dialog states, they could improve synthesis. But again this technique is very targeted toward the particular application and dialog context that the synthesizer would be used in.

The second synthesis technique is Statistical Parametric Speech Synthesis [31], where speech data is modelled in a generative fashion, which as a first approximation can be viewed as averaging, in contrast to clustering instances of data in the unit selection case. Statistical Parametric Speech Synthesis is typically using smaller amounts of data, thus can produce a wider range of output than a unit selection system could on the same data. [3] describes using statistical models of F_0 power and duration to model different emotional states. But even though statistical parametric speech synthesis allows for more control of the modelling, it is still hard to get the distinctions significant enough to allow the listener to perceive the intended state.

[29] propose an evaluation strategy for expressive speech, but admit that it is hard to evaluate all the subtle variations. It is possible to make general remarks about prosody use in stylistic speech. Angry and happy speech typical has higher F_0s and larger dynamic range. Sad speech typically has lower F_0 and small dynamic range and longer durations. However more subtle differences are harder to explicitly describe, and synthesis from them equal harder to do well. Though in the similar problem of voice conversion, which can be used for both conversion to new voices but also conversion to new styles from the same speaker, we have found it useful to train Speaker ID classifiers and use them to evaluation metrics for our voice conversion [14]. Assuming that the speaker ID (or personality ID) classifiers are trained on human speech, if synthesized examples can be classified in the intended way then we have a good evaluation technique (and optimization measure) for building synthesis models. However it has been noted that even when objective measures are satisfied, that does not mean human evaluator necessary agree.

5 Consistency

Outside of speaker identification and verification, there will hardly ever be a voice-based user interface, in which the voice itself is the only information being required and exchanged. It will therefore be important to convey the same notion of personality in these modalities during synthesis.

Examining distinctive linguistic and lexical choices in 2007, Gill [11] investigates the relationship between the personality of an author of short emails and blog texts, generated by self-assessment, and their language. He observes weak correlations, but concludes that personality is represented in text using more complicated features. Oberlander [19] examines the relation between part-of-speech (POS) distributions in email texts and two distinct personality traits, neuroticism and extroversion, of their authors. He concludes that part-of-speech information can be characteristic.

During generation of text using a binary extroversion/ introversion distinction, Nass et al. show that relatively simple rules, for example replacing weak adjectives and quantifiers (“quite rich”) by strong language (“absolutely sensational”), will change the perception of the text, particularly when paired with a corresponding synthetic voice [17]. In fact, the voice properties seem to have a stronger effect than the text properties, in particular when matched to the reader/ listener.

6 What to Do? What Next?

While there won't be a simple recipe to go by for quite some time yet, we believe that speech technology will soon be able to automatically recognize and generate more complex personality structures than just a binary "introverted vs. extroverted" distinction. More detailed, and reliable, automatic analysis of voice signals, be they synthetic, natural, or concatenated, will be necessary for this type of technology to leave the lab. While it is reasonable to expect to be able to manipulate the "volume" of a voice prompt in a lab setting, it can be quite difficult to do the same thing over the telephone, where a variety of network transmission channels and handsets are waiting, with no control on the part of the service provider.

While the use of personality in user interfaces may well be guided by relatively simple rules ("similars attract") for quite some time, a company might for example want to make sure that voices that have been casted for brand image are perceived the same on high-quality head-sets and on low-grade equipment with different transmission characteristics (imagine long-distance lines or VoIP connections). Once the correlation between human ratings (perception) and factors, which can be varied during synthesis (generation) are known, simple rules can be replaced by algorithms, with parameters that can be fitted to specific situations, ensuring appropriate service.

Of course, the speech features presented here will only be a fraction of the information, which can be extracted from a multi-modal interface; we have neglected text, timing, video/ graphical, or haptic information, if available. All of these have to be analysed together, and synthesised consistently. Still, meta-data extraction from speech is currently a very active field of research with challenge-style evaluations [27], as is the "social signal processing" community [20]. We hope that the significant effort directed towards automatic analysis of meta-data in speech will soon uncover relevant factors and features in more detail, which can then be used to automatically generate appropriate speech. We believe that the ability to model personality in speech recognition and synthesis will be a big and necessary step towards natural-like man machine interaction, as promised by the vision of "affective computing" [21].

References

[1] Apple, W., Streeter, L.A., Krauss, R.M.: Effects of pitch and speech rate on personal attributions. Journal of Personality and Social Psychology 37(5), 715–727 (1979)

[2] Bickmore, T., Cassell, J.: Social Dialogue with Embodied Conversational Agents. In: Natural, Intelligent and Effective Interaction with Multimodal Dialogue Systems. Kluwer Academic, New York (2004)

[3] Bulut, M., Lee, S., Narayanan, S.: A statistical approach for modeling prosody features using postags for emotional speech synthesis. In: Proc. ICASSP, Honolulu, HI (2007)

[4] Cassell, J., Sullivan, J., Prevost, S., Churchill, E.F. (eds.): Embodied Conversational Agents. MIT Press, Cambridge (2000)

[5] Catrambone, R., Stasko, J., Xiao, J.: Anthropomorphic agents as a user interface paradigm: Experimental findings and a framework for research. In: Proc. 24th Annual Conference of the Cognitive Science Society, Fairfax, USA (August 2002)

[6] Chen, Y., Naveed, A., Porzel, R.: Behavior and preference in minimal personality: A study on embodied conversational agents. In: Proc. ICMI-MLMI. ACM Press, New York (2010)

[7] Costa, P.T., McCrae, R.R.: Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) manual. Psychological Assessment Resources (1992)
[8] Costello, A.B., Osborne, J.W.: Best practices in exploratory factor analysis. Practical Assessment, Research & Evaluation 10(7) (July 2005)
[9] Drapela, V.J.: A Review of Personality Theories, 2nd edn. Charles C. Thomas Publ. (1995)
[10] Eide, E., Bakis, R., Hamza, W., Pitrelli, J.: Multilayered extensions to the speech synthesis markup language for describing expressiveness. In: Proc. Eurospeech, Geneva, Switzerland (2003)
[11] Gill, A.J., French, R.M.: Level of Representation and Semantic Distance: Rating Author Personality from Texts. In: Proc. Euro Cogsci, Delphi, Greece (2007)
[12] Goldberg, L.R.: The structure of phenotypic personality traits. American Psychologist 48, 26–34 (1993)
[13] Hunt, A., Black, A.: Unit selection in a concatenative speech synthesis system using a large speech database. In: Proc. ICASSP, Atlanta, Georgia, vol. 1 (1996)
[14] Jin, Q., Toth, A., Black, A., Schultz, T.: Is voice transformation a threat to speaker identification? In: Proc. ICASSP, Las Vegas, USA, NV (2008)
[15] Mairesse, F., Walker, M.A., Mehl, M.R., Moore, R.K.: Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. Journal of Artificial Intelligence Research (JAIR) 30, 457–500 (2007)
[16] Nass, C., Brave, S.: Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship. MIT Press, Cambridge (2005)
[17] Nass, C., Lee, K.M.: Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. Journal of Experimental Psychology: Applied 7, 171–181 (2001)
[18] Nass, C., Moon, Y., Fogg, B., Reeves, B., Dryer, D.C.: Can computer personalities be human personalities? International J. of Human-Computer Studies 43(2), 223–239 (1995)
[19] Oberlander, J., Gill, A.J.: Individual Differences and Implicit Language: Personality, Parts-of-Speech and Pervasiveness. In: Proc. Cogsci, Chicago, IL, USA (2004)
[20] Pentland, A.: Social signal processing. IEEE Signal Proc. Magazine 24(4), 108–111 (2007)
[21] Picard, R.W.: Affective Computing (1995)
[22] Polzehl, T., Möller, S., Metze, F.: Automatically assessing acoustic manifestations of personality in speech. In: Proc. SLT Workshop. IEEE, Berkeley (2010)
[23] Polzehl, T., Schmitt, A., Metze, F., Wagner, M.: Anger recognition in speech using acoustic and linguistic cues. Speech Communication, Special Issue on Sensing Emotion and Affect - Facing Realism in Speech Processing (2011)
[24] Reeves, B., Nass, C.: The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places. Cambridge University Press, Cambridge (1996)
[25] Ryckman, R.M.: Theories of Personality. Thomson/Wadsworth, Belmont CA (2004)
[26] Scherer, K.R., Scherer, U.: Speech Behavior and Personality. Speech Evaluation in Psychiatry, 115–135 (1981)
[27] Schuller, B., Steidl, S., Batliner, A.: The INTERSPEECH 2009 emotion challenge. In: Proc. INTERSPEECH, ISCA, Brighton, UK (September 2009)
[28] Syrdal, A., Conkie, A., Kim, Y., Beutnagel, M.: Speech acts and dialog TTS. In: Proc. SSW 7, Keihanna, Japan (2010)
[29] Türk, O., Schröder, M.: Evaluation of expressive speech synthesis with voice conversion and copy re-synthesis techniques. IEEE Trans. on ASLP 18(5), 965–973 (2010)
[30] Witten, I.H., Frank, E., Trigg, L., Hall, M., Holmes, G., Cunningham, S.J.: Weka: Practical machine learning tools and techniques with java implementations (1999)
[31] Zen, H., Tokuda, K., Black, A.: Statistical parametric speech synthesis. Speech Communication 51(11), 1059–1064 (2009)