AUTOMATICALLY ASSESSING ACOUSTIC MANIFESTATIONS OF PERSONALITY IN SPEECH

Tim Polzehl, Sebastian Möller

Quality and Usability Lab,
Technische Universität Berlin /
Deutsche Telekom Laboratories
10587 Berlin; Germany
Email: {tim.polzehl|sebastian.moeller}@telekom.de

Florian Metze

Language Technologies Institute
School of Computer Science
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

1. INTRODUCTION

There is currently a trend to go beyond "text" in speech technology, with significant work being dedicated to the recognition and synthesis of speech with certain characteristics of age, gender, and emotion, in order to advance the human machine interface [1]. The age and gender classification problems are straightforward, because target classes are well defined, and labeled data is available. Salient features can be determined, and classifiers can be built. Emotion is more difficult [2], because, beyond the classic call-center "angry" vs. "non-angry" case, strictly categorial approaches compete with almost vector-like representations, without consensus on the appropriate number of categories, or dimensions. Despite many attempts at systematization [3], the choice of representation in speech research depends strictly on task and data, often pragmatically adapting established representations.

Early work used "acted" emotional speech, which was produced by professional actors, who were given specific instructions. As a next step in this line of work, we present results from an exploratory study of how an established personality description from psychology can be used to model vocal manifestations of perceived personality traits in speech. As voice-based human machine interaction expands beyond directed dialog and simple command and control interfaces, we believe that a better, and automatic, analysis of voice qualities, and the personality connotations that are being conveyed, will help improve the acceptance of speech technology in man machine interaction.

Similar to the work on emotions, 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 [4], which attempts to describe the personality of a person using five factors. The values of the individual factors, which are normally being referred to as "scales", constitute a personality profile and describe general tendencies of a person's personality. Depending on social environment, situation, etc., values show variations. However, overall profiles are presumed to be relatively invariant after adolescence.

Traditionally, factor values for a person are being generated by asking a person who knows the subject well to fill out a 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, about 20 s in our case. From "persona" work in voice user interface (VUI) design, it is known that humans associate personality even with casual acquaintances. In our experiment, human raters assess the personality they attribute to different speech samples, which were produced by a professional speaker, who had been given instructions to "act" personalities. Raters do not know the speaker, they could only hear his speech.

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 [5]. We also derive own factors from our ratings, and compare them to the factor struc-

ture incorporated in the NEO-FFI, to guarantee valid results. Finally, we build and test an automatic classifier using a large number of acoustic and prosodic features computed on the speech data, and compare our results to the human baseline.

Very little work examines acoustic manifestation of personalities using signal analysis. Although there are many studies on the relationship between personality and speech, little empirical work exists. [6] analyzes prosodic features such as pitch and intensity, and observes that extroverted speakers speak louder, and with fewer hesitations. They conclude that extroversion is the only factor that can be reliably estimated from speech. Mairesse [7] also finds that prosodic and acoustic features are important for modeling extroversion, and that extroversion can be modeled best, followed by emotional stability (neuroticism) and openness to experience. Prosodic features include intensity and pitch only.

Personality also influences the text of a communication: Gill [8] investigates the relationship between the personality of an author of short emails and blog texts, generated by self-assessment, and their language. Using co-occurrence techniques, he observes insufficient correlations, but concludes that personality will be represented in text using more complicated features. Oberlander [9] 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 POS can be characteristic.

A human-like future speech dialog system (or a "companion") would therefore need to take these factors into account, when generating output, and measure these factors, when listening to human input.

2. 'BIG FIVE' AND NEO-FFI INVENTORY

Following the concept of the *Big Five* personality traits, as presented by [4], we use the German version of the NEO-FFI personality inventory [5]. Accordingly, we describe the personality conveyed by a speech sample using 5 scalar values, which describe the degree to which the corresponding trait is absent (value is 0), or present (value is 48). The values are ratings, which are derived from the answers to a questionnaire.

The 5 NEO-FFI traits are characterized as follows:[1]

Neuroticism (*N*): People with high 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 ratings for extroversion indicate a sociable, energetic, independent personality, while introverted personalities are presumed to be rather conservative, reserved and contemplating.

Openness (*O*): Openness describes the degree to which a person considers new ideas and integrates new experiences in everyday life. Highly rated persons are presumed to be visionary and curious. They perceive what is happening in their surroundings and are open to venturesome experiments. On the other side, people with low ranking are generally conservative. They prefer common knowledge to avantgarde.

Agreeableness (*A*): High agreeableness scores suggest that a person is rather sympathetic. He or she trusts other people and is being helpful. Non-agreeable personalities are egocentric, competitive and distrustful.

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

Raters generate a person's profile by answering the NEO-FFI questionnaire's 60 propositions, using "strongly disagree", "disagree", "neutral", "agree", and "strongly agree" as possible answers, which are then translated into numeric values 0-4. The overall value for a factor is then generated from these values, with a range from 0 to 48 for each factor. Intra-scale consistency coefficients for the NEO-FFI questionnaire, given by Cronbach's α [10], are constantly above 0.8, which represents overall good cohesion of the observers' ratings. Correlations (using Pearson's r) between the generated 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.

As an example, paraphrased questions from NEO-FFI include (using the observer's perspective):

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

3. SPEECH DATABASE

As there is only very limited experience available in this field, we decided to conduct speaker-dependent experiments, on acted personalities, using a limited domain, in order to establish baselines for human and automatic performance.

We recorded a professional speaker, who had previously recorded voice prompts in speech dialog systems, and was used to working with voice coaches. We initially recorded his "natural", i.e. non-acted voice. We then presented him the

[1] The original material contains longer descriptions, here we just want to convey a general idea, and re-phrase in our own words.

original descriptions of the 5 NEO-FFI personality traits as given by the NEO-FFI manual, as presented in Section 2. We asked him to prepare 10 voice personalities which represent voice impressions of 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". While we assume that acting on any one scale will also influence the values on the other scales, we did not measure or work with these differences yet.

The spoken text is designed to resemble a neutral, complete phrase, as could be expected in an IVR system, or a hot-line. We chose to record the following text (English translation provided), which typically lasts about 20 s.

> *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 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.*

Because this paragraph consists of many constituents, e.g. of positive ("you can redeem your voucher") and of negative ("unfortunately you cannot activate ...") intent, of activation ("please call again") function, of reception and farewell parts, which are not semantically related to the content, we believe it can be produced well using various personality profiles.

During scheduled recording sessions, in which the speaker would work in a recording studio with a voice coach, we recorded at least 20 takes of each of the conditions, more than an hour of speech in total.

All speech samples were then annotated by two labelers (speech transcriptionists) for "artificiality", and we chose to retain for our human rating experiments the three least artificial takes for each condition. This would restrict our analysis to "natural", acted personalities, and also limit the cost of obtaining the human ratings.

We recruited 87 raters (mostly students at Berlin Universities, mean age 29 years, 60 % male). Every rater would rate 8 takes from different conditions on average. Every take would be rated by 20 different raters. 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. This procedure generated over 600 questionnaires for all 5 scales. Every questionnaire covers all 5 scales, although only one of them is being "acted" deliberately.

4. PERSONALITY RATINGS FROM SPEECH

Figure 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 label the acted personalities quite well, as nearly all the conditions were perceived as intended by the "acting" speaker. Note that natural speech is not necessarily at the middle of the scale, but it expresses the speaker's "normal" personality.

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.

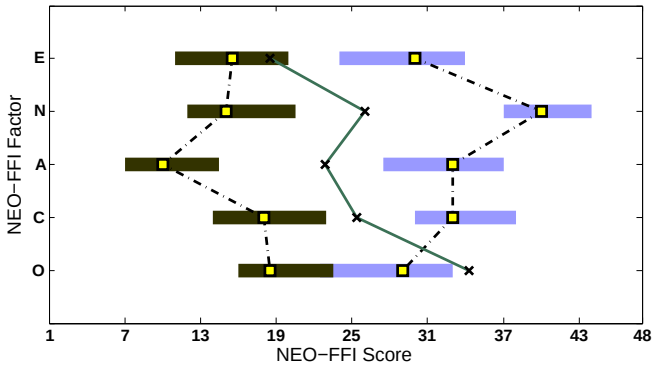

Fig. 1. NEO-FFI ratings of the recorded database. Brown bars (left) represent inter quartile ranges of ratings from variation towards low values, light blue bars (right) towards high values. Vertical lines connect the medians, i.e. solid line for "normal personality", dashed for acted variations.

In order to confirm the underlying structure of our data, we conducted an exploratory factor analysis [11], hypothesizing the presence of 5 latent factors in the user ratings, using the maximum likelihood method for component extraction. The aim is to reveal any latent variables that cause the assumed factors to correlate. The observation to item ratio of our data is 10:1, which is appropriate for this type of analysis. Only 5 of 60 items show a commonality less than $0.4\,h^2$, so over 90 % of items are reliable, even with our amount of data. Commonality measures the percent of variance in a given variable explained by all the factors jointly, and may be interpreted as the reliability of an item. All other items have commonalities between 0.4 and 0.7, which represents low to moderate commonality.

Analyzing the factor loading, we find a 1:1 mapping between original factors and extracted latent factors F_m. Table 1 shows that, due to the low number of cross-loadings, and the equal number of total items loading on the latent factors, every latent factor can clearly be associated with exactly one original factor.

Although the structure of our factors is congruent with

Table 1. Number (#) and identity of original factors loading with a coefficient above 0.4 on latent factors.

Latent Factor	F_1	F_2	F_3	F_4	F_5
# of items total	14	13	14	14	9
# of cross-loadings	3	2	2	2	0
Factor with max #	*A*	*E*	*N*	*C*	*O*

the NEO-FFI factors, our model fails to explain the overall variance in the data as well as the NEO-FFI, which has been validated using 11 724 items. We believe that using the NEO-FFI scheme to measure personality impressions from speech does not cover all the variance, which is revealed by applying the NEO-FFI because, it was originally designed to assess the overall personality. Thus its items' scores show more variance, than can be covered by assessing speech only. We plan more detailed analysis of influencing factors and factor structure in future work.

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

5. SIGNAL-BASED SPEECH ANALYSIS

The previous section established that humans can deliberately produce and recognize speech with distinct personality profiles, and that an existing personality assessment scheme can be used to quantify these variations, using human listeners. This permits us to create an automatic personality assessment of speech signals, and compare it to a human baseline, using speech samples only.

In our first experiments, we attempt to identify which personality was acted, i.e. we have a ten class problem, neglecting the neutral speech. For our classifier, we will rely on prosodic and acoustic features, in line with findings of salient features reported in related work [6, 7]. We leverage from previous work on emotion recognition [12], and extract audio descriptors such as 16 MFCC coefficients, 5 formant frequencies, intensity, pitch, perceptual loudness [13], zero-crossing-rate, harmonics-to-noise-ratio, center of spectral mass gravity (centroid), the 95 % roll-off point of spectral energy and the spectral flux, etc, using a 10 ms 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 behavior we append first and second order finite differences.

A more detailed description of the total of 1450 features is beyond the scope of this paper, and can be found in [12].

6. FEATURE SELECTION AND CLASSIFICATION

In order to select salient features from our pool of features, we rank them according to Information Gain [14], so that f_1 is the most salient feature, f_2 the next, etc. We determine the optimal feature set $\mathcal{F}_{opt}$ by successively training and testing classifiers with an increasing number of features n, on feature sets $\mathcal{F}_n := \{f_1, f_2, ..., f_n\}$, until the performance saturates. We use 10-fold cross-validation and all recordings in the database for these experiments. For classification, we use Support Vector Machines (SVM) with linear kernel functions.

Classifying using $\mathcal{F}_1 = \{f_1\}$ (i.e. the most salient single feature) alone, we observe an accuracy of approx. 28 %, which is about three times chance level already. Using $\mathcal{F}_{10}$, we achieve about 50 % accuracy. Using $\mathcal{F}_{20}$, we obtain another improvement of 8 %. Using even more features, the accuracy starts fluctuating at about 60 %.

Analyzing the salient features in $\mathcal{F}_{40}$, 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. Single MFCC coefficients however can not easily be interpreted in a linguistically meaningful way.

Table 2 shows precision, recall and accuracy for the individual classes. Factors *N* and *C* can be classified best. 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 perform well. High extroversion (*E*) can also be classified well, which is in line with observations by [6, 7], and Figure 1. 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 cases.

Table 2. Classification scores for individual classes.

Class	Precision	Recall	F-Measure
High neuroticism	0.84	0.94	0.89
Low neuroticism	0.92	0.71	0.80
High extroversion	0.61	0.82	0.70
Low extroversion	0.60	0.40	0.48
High openness	0.50	0.60	0.54
Low openness	0.30	0.35	0.32
High agreeableness	0.42	0.33	0.37
Low agreeableness	0.50	0.56	0.53
High conscientiousness	0.86	0.75	0.80
Low conscientiousness	0.85	0.73	0.79

As the NEO-FFI questionnaire is validated only for generation of personality profiles, not personality classification, we do not have a baseline for human emotion classification.

Automatically generating a classifier on top of the human-generated profiles did not yield satisfactory results so far.

Still, we believe that being able to achieve 60 % classification accuracy on acted personality data in a 10-class task is an encouraging result, which shows that the *Big Five* personality traits have acoustic correlates, that can be identified automatically in speech.

7. CORRELATION ANALYSIS

For most purposes, however, we are not interested in a classification of personality, or in a personality assessment by a machine, but in the reproduction of a human rating by a machine. In dialog systems, or for voice assessment, the machine should be able to predict the impression humans will have, not the instructions that were given to the speaker.

We therefore 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, and 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 minimized by keeping the function as flat as possible.

Figure 2 shows the correlations between the human ratings for the 5 factors and automatic factor prediction from speech, when expanding the feature space from $\mathcal{F}_1$ to $\mathcal{F}_{70}$. Feature space ranking was obtained by IGR evaluation as discussed in Section 6.

Correlation analysis shows how different the predictions by humans, and machines are, for the various factors. As in the classification experiments, there is very little change as soon as at least 20 features are being used, and almost no change when using more than 40 features.

Looking at the distributions of features in the top ranks of $\mathcal{F}_{30}$, 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 in line with previous work on signal-based analysis [7, 6].

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 *N* and *E*. While classifying into high and low variations along the conscientiousness (*C*) dimension also yields reasonable classification scores, our models poorly predict the actual value of that factor. Relatively poor results are achieved for openness and agreeableness.

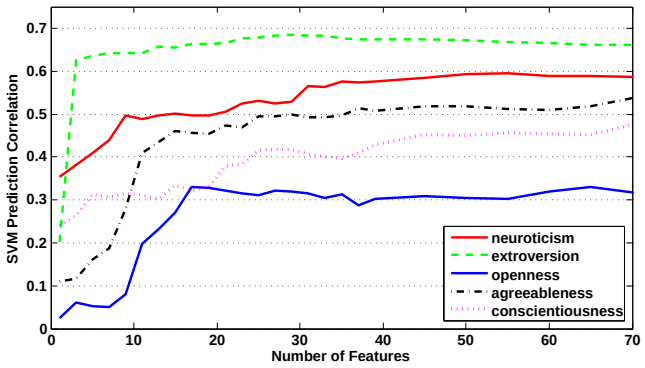

Fig. 2. Correlation between human (NEO-FFI) and automatic rankings over number of features used for automatic classification.

8. LIMITATIONS OF SPEECH ASSESSMENT

Our attempt at assessing a perceived personality in speech meets methodological difficulties, as the questionnaire was not originally designed for this purpose. However, in clinical phoniatrics the NEO-FFI questionnaire is used successfully to examine the correlation between vocal disorders and personality of patients. In our approach, we can avoid risks associated with self-assessment, or assessment of a familiar person. Still, collecting suitable data for modeling or recognition of personality traits will be even more difficult than it was for early work on emotion recognition. Acting is clearly not feasible for large corpora.

It is well known in speech science that the spoken text has an influence on the perception of voice quality. Future experiments need to verify the present results using different speakers, and investigate speaker independent, and "non-acted" personality traits, recorded not from professional speakers, but presumably from "real" speakers.

Finally, we point out that we do not primarily aim at classifying the personality of a speaker. Rather, we target a perceived personality impression, i.e. acoustic correlates of personality, taken from voice and speaking style, and correlate automatic measurements with human profiles generated on the same data.

9. CONCLUSIONS AND OUTLOOK

In this paper, we investigated the applicability of personality assessment, as established in psychology, to expressive speech. We created "acted" speech samples, which were targeted towards certain personality traits, namely the extrema

of the NEO-FFI personality scales. We established that (a) the quality of our data meets criteria for successful analysis, and the inherent factor structure resembles the structure found in the original NEO-FFI questionnaire data, (b) human listeners can recover the intended personality traits in a listening test, and (c) automatic recognition of prominent personality traits is possible, and good correlation between automatic and human assessments can be found for some of the five factors.

Although our raters did not know the speaker, the consistency and correlation analysis implies the applicability of the test scheme to vocal input. The accuracy of our automatic personality classification experiments reaches approximately 60 % for a ten class task, consisting of isolated, acted productions of high and low targets for the 5 personality traits, by a single speaker. Acted personalities along the neurotic and extroverted scales could be classified best. We also observe the highest correlation between human and automatic analysis for these factors. Conscientiousness can be recognized well, although correlation with human ratings is always less than 0.5.

In ongoing work, we collect and annotate more speech data from a single speaker, in order to provide a corpus for text-independent experiments, and for speech synthesis with personality. We are also refining our class definitions and analysis methods. In future work, we would like to investigate the links between text-based and voice-based personality assessment, for example using co-training, and proceed to speaker independent personality assessment.

While our results are clearly preliminary, we hope to be able to eventually create and validate a framework for the description of personality as conveyed by voice. This framework could be used for the analysis of user (input) speech in speech dialog systems, and provide more context for analysis and adaptation. It could also be used for the generation of appropriate output speech, or for automatic validation of properties attributed to output speech, for example in brand image monitoring.

Acknowledgements

The first author was funded by Deutsche Telekom AG, Laboratories. We would like to thank Joachim Stegmann and Claus Cramer for supporting this research in kind and spirit.

10. REFERENCES

[1] Clifford Nass and Scott Brave, *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*, The MIT Press, 2005.

[2] Anton Batliner, Kerstin Fischer, Richard Huber, Jörg Spilker, and Elmar Nöth, "Desperately seeking emotions: Actors, wizards, and human beings," in *Proceedings of the ISCA Workshop on Speech and Emotion*, Belfast; UK, SEP 2000, ISCA.

[3] Robert Plutchik, *The Psychology and Biology of Emotions*, Harper Collins College, New York, U.S.A., 1994.

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

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

[6] Klaus R. Scherer and Ursula Scherer, "Speech Behavior and Personality," *Speech Evaluation in Psychiatry*, pp. 115–135, 1981.

[7] François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger 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.

[8] Alastair J. Gill and Robert M. 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] Jon Oberlander and Alastair J. 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] Richard Zinbarg, William Revelle, Iftah Yovel, and Wen Li, "Cronbach's, Revelle's, and McDonald's: Their relations with each other and two alternative conceptualizations of reliability," *Psychometrika*, pp. 123–133, 2005.

[11] Anna B. Costello and Jason W. Osborne, "Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis," *Practical Assessment, Research & Evaluation*, vol. 10, no. 7, July 2005.

[12] Tim Polzehl, Alexander Schmitt, and Florian Metze, "Comparing Features for Acoustic Anger Classification in German and English IVR Portals," in *Proc. First International Workshop on Spoken Dialogue Systems Technology (IWSDS)*, Irsee; Germany, Dec. 2009.

[13] Hugo Fastl and Eberhardt Zwicker, *Psychoacoustics: Facts and Models*, Springer, Berlin, 3rd edition, 2005.

[14] Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*, John Wiley & Sons, 2nd edition, 2000.