Towards Situation- and User-Adaptive Voice Output: Classifying Driver Personality in Context

Daniela Stier
Mercedes-Benz AG &
Ulm University, Germany
daniela.stier@daimler.com

Wolfgang Minker
Ulm University
Germany
wolfgang.minker@uni-ulm.de

Ulrich Heid
Hildesheim University
Germany
heidul@uni-hildesheim.de

Abstract

The current trend towards natural, intuitive speech interaction requires intelligent dialogue systems, that flexibly react according to an individual *user* in the interaction *context*, *e.g.*, by adapting the language style of voice output. A precondition for the efficient realization of such adaptive system behaviour is the reliable identification of these parameters. Against this background, this work serves as a basis for the development of user-adaptive in-car voice output and presents our classification results based on acoustic features.

1 Introduction

Intelligent Spoken Dialogue Systems (SDSs) are expected to provide an increasingly natural and intuitive human-machine interaction, *e.g.*, by flexibly adapting their voice output to the requirements of an individual user. An interaction as efficient as possible based on the human model (*cf.* interactive alignment model; Pickering and Garrod, 2004; Branigan et al., 2010) is particularly relevant in situations where voice-based communication with an SDS runs as a secondary task. Here, speech needs to be produced and processed in parallel and should not distract from the primary task, such as driving a car. In such dual-task scenarios, the development of an adaptive SDS depends both on individual user characteristics and the context of interaction. From a user's point of view, for instance, the characterization of user language in the interaction context is essential for the design of adaptive voice output strategies. Studies on the language used by drivers, *e.g.*, revealed clear differences in complexity depending on the manifestation of their personality traits (Stier et al., 2020a) and the driving situation (Stier et al., 2020b). From the SDS perspective, a requirement for user- and situation-adaptive voice output is the reliable identification of both the user and the driving context. While the automatic identification of personality traits in interpersonal or human-machine interaction has already been investigated extensively (*e.g.*, Mairesse and Walker, 2007; Sidorov et al., 2014; Subramaniam et al., 2016), to our knowledge no research exists which extends this aspect by the interaction context of a primary task. In this paper we present our results of this fundamental work. Our approach can be summarized in the following steps:

1. Collect spoken data for diff. driving contexts
2. Extract acoustic features
3. Create user clusters based on personality traits
4. Build and test statistical models for driving situations and user clusters based on features

2 Data Collection and Processing

We collected data as WoZ experiment in a fixed-base simulator (Mercedes E-Class) with 180° monitor. After an initial training in the car (5min), each participant followed a lead vehicle at distance of 100m and drove manually with SAE 0 on a highway (100km/h) and in a city (50km/h; 8min each). There was light oncoming traffic and additional traffic lights placed each km in the city, which changed to green when the participant approached to prevent motion sickness. In order to collect as much data as possible, the simulated voice-based interaction was limited to system-initiated small talk questions of the WoZ (*e.g.*, "What is your favourite hobby and how did this hobby come about?"). Each participant was instructed in advance to answer spontaneously and in his or her own language style (i.e. without overemphasis, as with a human interlocutor). Each participant was recorded on video with a camera installed in the vehicle (A-pillar).

Overall, 44 subjects (28 male, 16 female) with an average of 43 years (*sd* 13.24) participated and self-assessed their Big Five Personality traits on five-point Likert scales (German questionnaire; Rammstedt and Danner, 2016). Our data collection comprises 376 answers for the highway (*mean* 62.67, *sd* 18.67) and 331 for the city (*mean* 55.17, *sd* 16.93).

Table 1: Overview of extracted features and their variations.

Spectral Centroid	mean and standard deviation of spectrum
Energy Difference	difference in energy between low (<500Hz) and high (>500Hz) frequency
Intensity, Pitch	maximum, mean, minimum and standard deviation of intensity and pitch
MFCCs and deltas	mel-frequency cepstrum coefficients (16 features) and changes (16 features)
Tempo	mean speaking rate (in beats per minute)

Table 2: Distribution of participants and answers.

UC	**1**	**2**	**3**	**4**	**5**	**6**	$\sum$
Sub.	6	6	8	4	12	8	44
Age	41.7	41.3	45.6	41.8	43.5	41.9	42.9
O	3.72	2.90	3.14	4.00	3.75	3.63	3.52
C	3.96	3.85	4.46	4.69	3.94	3.92	4.09
E	3.56	3.54	3.73	4.75	3.78	3.70	3.78
A	3.17	3.53	4.05	3.63	3.81	4.15	3.77
N	2.83	2.46	1.89	1.59	2.00	2.75	2.26
H	50	60	65	40	95	66	376
C	47	51	53	37	87	56	331

Note: O-N= Big Five traits, H= number of answers on highway, C= number of answers in city

2.1 Acoustic Feature Extraction

The voice recordings were tailored to the individual user responses ranging from 3 to 400s (*mean* 41.78, *sd* 40.19) without further processing. Based on Landesberger et al. (2020), we extracted 43 features (Table 1) using `Praat` (Boersma and Weenink, 2020) and `librosa` (McFee et al., 2015).

2.2 Labelling User Personality and Context

Personality manifests multiple traits simultaneously. Thus instead of investigating traits individually, we combined the assessed Big Five traits als clustering variables and obtained six user clusters (UCs; `SPSS` 2-step clustering, av. silhouette 0.4, size ratio 3.25). We additionally distinguished the driving context (DC) during which a response was recorded. Table 2 summarizes details of our data.

3 Classifying User and Situation

We performed stratified five-fold cross-validation and compared performance measures (accuracy, precision, recall, f-measure; macro-averaging) for a Multinomial Logistic Regression classifier[1] (MLR), Random Forest Classifier[2] (RFC) and Support Vector Machine[3] (SVM). We classified UC and DC separately before combining them (UC & DC). In either case, the best results were obtained for RFC (Table 3). However, all classifiers performed remarkably well. The application of combined resampling (under-, oversampling, smote) and feature selection (cross-val. recursive elimination) methods did not significantly improve classification results.

Table 3: Classification results.

		acc	**prec**	**rec**	**f**
UC	MLR	0.774	0.770	0.762	0.762
	RFC	**0.979**	**0.982**	**0.974**	**0.977**
	SVM	0.815	0.805	0.793	0.793
DC	MLR	0.939	0.939	0.939	0.939
	RFC	**0.990**	**0.990**	**0.991**	**0.990**
	SVM	0.955	0.955	0.956	0.955
UC & DC	MLR	0.723	0.728	0.718	0.715
	RFC	**0.966**	**0.971**	**0.962**	**0.965**
	SVM	0.799	0.798	0.787	0.784

4 Conclusion and Future Work

In this paper we presented our results for the automated classification of both the user personality and the driving situation based on acoustic features. This investigation serves as a basis for the development of an in-vehicle SDS with user and situation-adaptive voice output. All of our statistical models achieved remarkable classification results. Thus a reliable identification of the driver personality in the driving context seems possible. For this purpose, acoustic features serve as a more than suitable source. They especially led to a precise DC categorization, presumably due to the unprocessed driving noises. They were surprisingly also able to successfully differentiate between our UCs. Due to our small and special data set, we will validate our findings in real driving situations in future work.

[1] C=80, penalty=l1, solver=liblinear

[2] n_estimators=500, bootstrap=false, max_features=log2

[3] C=0.1, degree=1, gamma=0.001, kernel=linear

References

Paul Boersma and David Weenink. 2020. Praat: Doing phonetics by computer [Computer program, version 6.1.09].

Holly P. Branigan, Martin J. Pickering, Jamie Pearson, and Janet F. McLean. 2010. Linguistic alignment between people and computers. *Journal of Pragmatics*, 42(9):2355–2368.

Jakob Landesberger, Ute Ehrlich, and Wolfgang Minker. 2020. Do the urgent things first! – Detecting urgency in spoken utterances based on acoustic features. In *Adjunct Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20 Adjunct, New York, NY, USA. ACM. In press.

François Mairesse and Marilyn A. Walker. 2007. Personage: Personality generation for dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 496–503.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. Librosa: Audio and music signal analysis in Python. In *Proceedings of the 14th Python in Science Conference*, pages 18–25.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.

Beatrice Rammstedt and Daniel Danner. 2016. Die Facettenstruktur des Big Five Inventory (BFI). *Diagnostica*.

Maxim Sidorov, Stefan Ultes, and Alexander Schmitt. 2014. Automatic recognition of personality traits: A multimodal approach. In *Proceedings of the 2014 Workshop on Mapping Personality Traits Challenge and Workshop*, pages 11–15.

Daniela Stier, Katherine Munro, Ulrich Heid, and Wolfgang Minker. 2020a. Personality traits, speech and adaptive in-vehicle voice output. In *Adjunct Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20 Adjunct, New York, NY, USA. ACM. In press.

Daniela Stier, Katherine Munro, Ulrich Heid, and Wolfgang Minker. 2020b. Towards situation-adaptive in-vehicle voice output. In *2nd Conference on Conversational User Interfaces*, CUI '20, New York, NY, USA. ACM. In press.

Arulkumar Subramaniam, Vismay Patel, Ashish Mishra, Prashanth Balasubramanian, and Anurag Mittal. 2016. Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features. In *European Conference on Computer Vision*, pages 337–348. Springer.