INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Perception of social speaker characteristics in synthetic speech

Sai Sirisha Rallabandi[1], *Abhinav Bharadwaj*[1], *Babak Naderi*[1], *Sebastian Möller*[1,2]

[1]Quality and Usability Lab, Technische Universität Berlin, Germany,
[2]Speech and Language Technology, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Berlin, Germany

s.rallabandi@tu-berlin.de, abhinav.bhardwaj2011@gmail.com, babak.naderi@tu-berlin.de, sebastian.moeller@tu-berlin.de

Abstract

With the improved computational abilities, the usage of chatbots and conversational agents has become more prevalent. Therefore, it is essential that these agents exhibit certain social speaker characteristics in the generated speech. In this paper, we study the perception of such speaker characteristics in two commercial Text-to-Speech (TTS) systems, Amazon Polly and Google TTS. We carried out a 15-item semantic differential scaling test. The factor analysis provided us with three underlying dimensions that can be perceived from synthetic speech, warmth, competence, and extraversion. Our results show that we can perceive both interpersonal relationships and also personality traits from synthetic voices. Additionally, we observed that the female participants perceived male voices to be more responsible, energetic, relaxed, and enthusiastic. In comparison, male participants found female voices to be more reliable, accessible, and confident. A discussion on the comparison of our results with that of the studies on natural speech is also provided.

Index Terms: Social speaker characteristics, Perceptual dimensions, Exploratory Factor Analysis, Semantic differential scaling test.

1. Introduction

Artificial speech generation has a lot of applications such as customer service, screen readers, personal digital assistants and many more [1, 2, 3]. The advent of end-to-end neural speech synthesizers in the last decade has lead to tremendous improvements in the fidelity of speech generation [4, 5, 6]. Currently, the research community is focused on improving the style of speech generation [7, 8, 9] and the efficiency of neural end-to-end TTS systems [10, 11]. Although a considerable amount of research already exists on speech generation, there is not much work on user satisfaction [12]. The increase in demand for human-computer interaction has enhanced the need for socially acceptable synthetic voices. Therefore, it is necessary to analyse the social acceptance of these speech synthesizers by deriving the additional perceptual dimensions underlying the synthetic voices.

There has been a significant amount of research on perception of a person [13, 14, 15, 16]. The BIG-FIVE questionnaire was developed in a similar fashion and is now widely utilised in studies related to personality [17, 18, 19]. The impressions of another person were previously studied through lists of adjectives that describe various traits and social behavior of a person [20, 21, 22]. These adjectives were further factorized and grouped into clusters. Such clusters were termed as the perceptual dimensions of a person. There have been some works on adapting the BIG FIVE questionnaire to zero acquaintance scenarios [18, 19]. In [18], the dimensions of interpersonal circumplex (warmth, confidence, compliance) and interpersonal attraction (attractiveness, maturity) are derived. Further, in [19] they provide perceptual dimensions of person attribution, (social and physical) attractiveness, confidence, apathy, serenity, and incompetence.

In addition, there are various studies on deriving emotions and paralinguistic phenomenon from speech [23, 24, 25]. In [23], speech perception is studied with a three dimensional model. They analyse subjective responses for the dimensions namely, valence, energy and tense and predict the responsible acoustic features. In [24] authors compare the self rated personality traits and intelligence in verbal and non-verbal speech with that of the ratings by strangers. [25] investigates the users' expressions and emotions in human-computer interactions. The perception of factors such as age, gender, human likeness were studied from IBM Watson TTS voices [26, 27, 28]. However, there is little or no known research on the social acceptance of different speech synthesizers.

To the best of our knowledge, this is the first attempt to analyse the synthetic speech for social speaker characteristics. Following the works [18, 19], we conduct a 15-item semantic differential scaling test on two commercial synthetic systems. The factor analysis performed on the subjective ratings has provided us three clusters (underlying perceptual dimensions) warmth, competence, and extraversion. The final goal is to bridge the gap between synthetic speech perception and the generated speech features. Nevertheless, we restrict our current work to analyze perceived social speaker characteristics in synthetic voices. Through this work, it is evident that we can perceive both interpersonal relationship and also personality traits from synthetic voices.

In this paper, we will be addressing three research questions: 1) What speaker attributes can be perceived from synthetic speech? 2) Can the social speaker characteristics be perceptually recognized by listening to the speech sample? 3) How can the social speaker characteristics be quantified in subjective evaluations?

This paper is organised as follows: We present the overview of our work in Section 2, the evaluation setup in Section 3, followed by the analysis in Section 4. We provide a comparison of our results with other works in the Section 5. Finally we provide a discussion on the analysis, and future work in the Sections 6 and 7 respectively.

Throughout this work, we will be using the terms voices/speakers/systems to refer the TTS voices and participants/raters/listeners for the subjects who participated in the evaluation. The terms attributes/adjectives/questionnaire refer to the questions we used in the subjective evaluation. We use the terms clusters/factors/perceptual dimensions/social speaker

http://dx.doi.org/10.21437/Interspeech.2021-1229

characteristics to address the underlying perceptual aspects of synthetic voices.

2. Overview

We performed a 2-stage evaluation on the synthetic voices, a) attribute selection (questionnaire preparation), b) semantic differential scaling test. Figure 1 displays the flowchart we have employed for predicting the perceptual dimensions underlying various synthetic voices. Through the evaluation stage 1, we have obtained different speaker attributes that can be perceived from a variety of synthetic voices. The attributes finalised in the evaluation stage 1, were used as the questionnaire in the evaluation stage 2. The subjective ratings obtained for different speaker attributes were factorised to obtain the underlying perceptual dimensions. The details on the TTS systems, number of voices used, subjective test setup and the analysis are further provided.

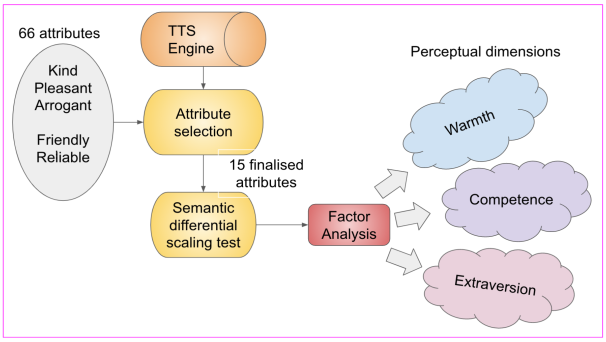

Figure 1: *Flowchart displaying the prediction of underlying perceptual dimensions of synthetic voices.*

3. Evaluation Setup

3.1. Dataset preparation

The commercial TTS systems, Google TTS[1] (Wavenet) and Amazon Polly [2] (Neural) were utilized for our studies. We chose the US native speakers for our work. There were 10 female (Google voices = 5, Amazon polly = 5) and 9 male voices (Google voices = 5, Amazon polly = 4). We study the perceptual dimensions underlying these synthetic voices in case of read speech. The speech samples were generated for 32 sentences in the Harvard database [3]. For each TTS voice, we combined the individual speech clips, into a speech segment of around 20 sec length. Each speech segment had 8 sentences. To avoid any order effect, we randomized the order of speech clips. We finally had 4 speech segments for each of 19 TTS voices.

3.2. Evaluation stage 1: Attribute selection

In the Evaluation stage 1, we derive the questionnaire (list of adjectives) that can be perceived from synthetic voices. This evaluation was carried out with both native (15 participants) and non-native speakers (15 participants). Their ages are in the range of 25 to 36 (mean = 27.76, std =3.23).

Table 1: *Adjectives finalised for the 15-item semantic differential scaling test*

Adjectives			
relaxed	confident	enthusiastic	energetic
friendly	arrogant	pleasant	likeable
responsible	reliable	accessible	sympathetic
skilful	kind	extrovert	

We have prepared a list of 66 adjectives obtained from [13, 14, 19, 18]. The participants were instructed to listen to a speech clip and select as many adjectives as they found relevant from the provided list. The participants were free to listen to the speech samples multiple times. The total number of questions presented to the participants were 4 speech segments × 19 TTS voices × 66 adjectives. The participants were also allowed to suggest new adjectives. They could take breaks in between to avoid any fatigue. All the participants of the study were compensated for their contribution.

The adjectives with highest frequency of ratings (15 adjectives) were selected and the remaining were discarded. The only significant difference we observed between the native an non-native speakers is in the selection of the adjective ”responsible”. The native speakers have preferred to select this adjective while non-native speakers did not. The finalised adjectives after this stage of evaluation are displayed in the Table 1.

3.3. Evaluation stage 2: Semantic differential scaling test

A crowd-sourced subjective test was carried out using Amazon Mechanical Turk (AMT). We extended the P.808 Toolkit [29] to include the attribution task. We used continuous 100-point scale[4] with the adjective and its antonym at each end. The P.808 Toolkit provides tools for conducting subjective speech quality test according to the ITU-T Rec. P.808 [31]. The toolkit includes participants eligibility, system and environment suitability tests within the crowdsourcing task to ensure reliability and validity of ratings collected in the crowdsourcing test [29]. In each session, participants should listen to 4 speech clips and rate them on 15 attributes selected in previous stage. Each speech clip was 20 seconds long and played in the loop until the participant finished by rating all adjectives. In addition we repeated one randomly selected adjective for each clip and used it as a hidden quality control mechanism [32, 33]. We have recruited native English speakers for our study.

4. Data Analysis

4.1. Pre-processing of data

An initial pre-processing of data was carried out to remove the participants whose ratings were not reliable. We have rejected 41 responses based on the environment suitability test. We calculated the Pearson correlation coefficient for the repeated attribute's subjective ratings and the original attribute in the questionnaire. The participants who displayed inconsistent ratings (correlation coefficient < 0.5) were removed (five). Overall, we have accepted 90% of the responses obtained from our subjective test. The number of participants that remained after the

[1] https://cloud.google.com/text-to-speech/

[2] https://aws.amazon.com/polly/

[3] https://www.cs.columbia.edu/ hgs/audio/harvard.html

[4] We used The Fragebogen [30] implementation for presenting the scale

pre-processing were 87 (female participants = 43 and male participants = 44). Their ages are in the range of 19 to 77 (mean = 40.31 and std = 12.57) .

We calculate the intraclass correlation coefficient ICC(1,k) to understand the inter-rater reliability over the collected subjective ratings. The average raters absolute value was 0.974 with a 95% confidence interval in the range of 0.95 to 0.99.

4.2. Exploratory Factor Analysis

We conducted a factor analysis to find out the underlying factorial structural. A gender dependent exploratory factor analysis (EFA) was carried out on the collected subjective ratings. We chose EFA, as the underlying perceptual dimensions of synthetic voices are unknown. Using Horn's parallel analysis, we determined the number of factors to be three. We performed an EFA with promax rotation and minimum residual factoring method. Further, we retained the attributes whose i) main loading > 0.5 and ii) difference between the main loading and the cross-loading was > 0.2. Later on, the attributes whose communality value was < 0.4 were removed. Based on the first criteria, we have removed the attribute, "accessible" (main loading = 0.47), from the female speakers and "arrogant" (main loading = 0.49) from the male speakers. Based on the second criteria, the attributes, "arrogant" (communality = 0.36 for female, 0.30 for male) and "relaxed" (communality = 0.39 for female, 0.38 for male) were removed from both male and female speakers. A second factor analysis was performed on the remaining attributes which explained a variance of 63% for female speakers and 60% for male speakers. From this analysis, we observe that the synthetic voices can express the social speaker characteristics (warmth, competence) and also the personality traits (extraversion). The final results of the factor analysis are provided for female and male speakers in the Tables 2, 3 respectively.

Table 2: *Factor loading for female speakers*

Attributes	Warmth	Competence	Extraversion
friendly	0.72		
kind	0.82		
likeable	0.79		
pleasant	0.77		
sympathetic	0.78		
confident		0.79	
reliable		0.87	
responsible		0.89	
skillful		0.88	
energetic			0.89
enthusiastic			0.83
extrovert			0.71

The internal consistency was examined for each of the derived factors through Cronbach's alpha. The cronbach's alphas obtained for each of factors are provided in the Table 4.

4.3. Observations

The 2 sample t-test has revealed statistically significant differences in the participants' ratings for different speaker genders ($p<0.05$). We observed that the female participants perceived the male voices to be more relaxed (p=0.01), enthusiastic (p = 0.03), energetic (p = 0.03), and responsible (p = 0.00) compared to that of female voices. While, male participants perceived the female voices to be more accessible (p = 0.01), confident (p = 0.04), and reliable (p = 0.02).

Table 3: *Factor loading for Male speakers*

Attributes	Warmth	Competence	Extraversion
accessible	0.63		
friendly	0.72		
kind	0.75		
likeable	0.71		
pleasant	0.67		
sympathetic	0.77		
confident		0.65	
reliable		0.81	
responsible		0.93	
skillful		0.84	
energetic			0.81
enthusiastic			0.81
extrovert			0.79

Table 4: *Cronbach's alphas*

Factors	Male	Female
Warmth	0.84	0.88
Competence	0.73	0.71
Extraversion	0.72	0.74

- ***Warmth:*** The attributes loaded within the cluster, warmth for both the genders were, "kind", "likeable", "pleasant", "friendly", and "sympathetic". The attribute "accessible" loaded under same warmth for male speakers.
- ***Competence:*** The attributes in the cluster, competence for both the genders are, "confident", "reliable", "responsible", and "skilful".
- ***Extraversion:*** The third cluster had the attributes, "energetic", "enthusiastic", and "extrovert" in both male and female speakers.

Figures 2, 3, 4, 5 show the performance of different TTS voices used in the study along with the 95% confidence intervals. We calculated the mean of all the adjectives' ratings loading on each factor (warmth/competence/extraversion). The male voice, Kevin (Amazon Polly) has the highest averaged ratings on all the dimensions. The male speaker, D (Google's TTS) has the lowest averaged rating on the dimension, warmth (39.39). The female voice, A (Google's TTS) has the lowest averaged rating on the dimension, competence (34.68). The male voice, J (Google's TTS) has the lowest averaged rating on the dimension, extraversion (35.42). We observed that, except the TTS voice J, all the others had the highest averaged ratings for the dimension, extraversion and least on the dimension, competence.

5. Comparison with other studies

5.1. Studies on natural speech from literature

We compare our results with that of the studies on natural speech [18, 19]. We observe that the attributes "friendly","sympathetic", and "likeable" are commonly found in both natural [18] and synthetic voices under the characteristic warmth. The attribute "confident" was found in the cluster, attractiveness, in natural speech. While in synthetic speech, it

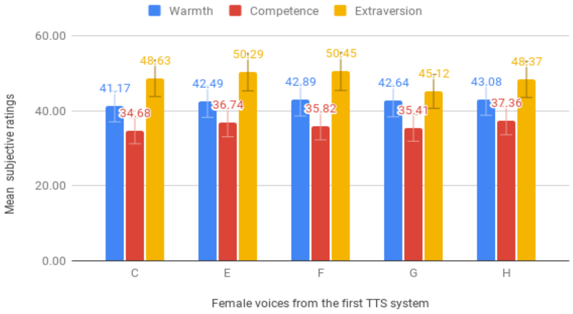

Figure 2: *Mean subjective ratings calculated over three perceptual dimensions for Google's female voices.*

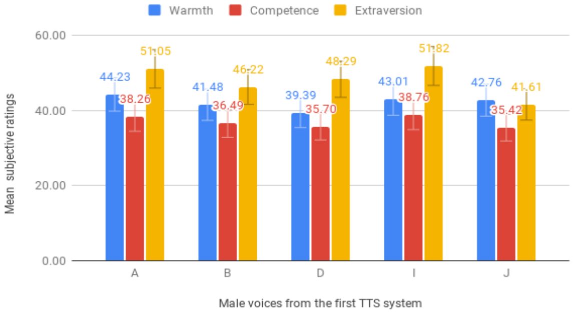

Figure 3: *Mean subjective ratings calculated over three perceptual dimensions for Google's male voices.*

is in the cluster, competence for both male and female speakers. The attributes that loaded on competence in synthetic speech were completely different from the ones in natural speech [19]. Infact, the attribute list we obtained after the evaluation stage 1, did not contain the attributes that were found in natural speech within the cluster, competence [19]. We compare the attributes under extraversion with that of the ComParE 2012, speaker trait challenge [17]. The attributes describing extraversion in natural speech were "active", "assertive", "energetic", "outgoing", and "talkative". The attribute "energetic" was loaded within the cluster of extraversion in both natural and synthetic speech [17].

5.2. Studies on social speaker characteristics

We compare the warmth and competence dimensions in our work with the studies on social cognition [20, 21, 22, 34]. These works propose that the dimensions, warmth and competence can be considered as the universal dimensions of social perception. The adjectives "friendly" and "likeable" in our work are correlated with the adjective list defined for warmth in [34, 22]. Similarly, the adjectives, "confident", "skilful" are correlated with the adjectives list for competence in [34].

6. Summary

In this work, we are interested in analysing the social perceptions of synthetic voices. We devised a 2-stage subjective evaluation setup, to i) determine the attributes that can be perceived from synthetic voices, ii) derive the underlying perceptual dimensions of synthetic voices. Our studies were conducted on read speech (Harvard sentences), synthesized by two commercial synthetic systems (19 voices). Through this work, we answer three research questions, i) the list of attributes that can be perceived from synthetic voices, ii) plausibility of perceiving the social speaker characteristics through listening tests, iii) a subjective evaluation setup for the quantification of social speaker characteristics in synthetic voices

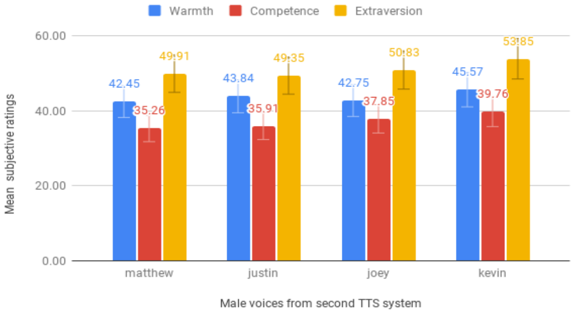

Figure 4: *Mean subjective ratings calculated over three perceptual dimensions for Amazon Polly's male voices.*

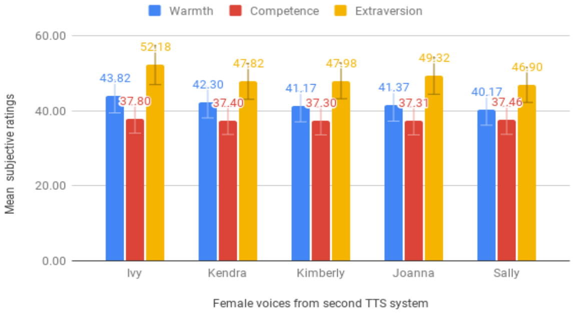

Figure 5: *Mean subjective ratings calculated over three perceptual dimensions for Amazon Polly's female voices.*

7. Future Work

Our analysis was carried out on sentences alone. Analyzing the speaker attributes, characteristics, and personality traits with other data (such as audiobooks, conversations, news reading) might lead to more interesting results. Additionally, exploring the perceptual dimensions underlying a wide variety of TTS systems could be future work. A comparison between the perceptual dimensions derived in the case of multiple TTS systems and a subset of them can be studied. On the other hand, identifying the acoustic correlates of the derived characteristics and traits in synthetic voices is another future work.

8. Acknowledgements

Authors would like to thank Benjamin Weiss for his valuable time and feedback. This work is being supported by the German Research Foundation (DFG), under funding MO 1038/29-1, TU PSP-Element: 1-50001062-01-EF. We also thank all the participants of our subjective tests.

9. References

[1] Potluri, Venkatesh and Rallabandi, SaiKrishna and Srivastava,

Priyanka and Prahallad, Kishore, “Significance of Paralinguistic Cues in the Synthesis of Mathematical Equations,” in *Proc. of the 11th International Conference on Natural Language Processing*, 2014.

[2] A. Wilkinson, A. Parlikar, S. Sitaram, T. White, A. W. Black, and S. Bazaj, “Open-Source Consumer-Grade Indic Text To Speech,” in *Proc. SSW*, 2016.

[3] Yaniv, Leviathan and Yossi, Matias, “Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone,” in *Google AI Blog*, 2018.

[4] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “Wavenet: A generative model for raw audio,” *arXiv preprint:1609.03499*, 2016.

[5] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, and et al., “Tacotron: Towards End-to-End Speech Synthesis,” *in Proc. Interspeech*, 2017.

[6] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanevsky, and Y. Jia, “Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation,” 2019.

[7] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, “Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis,” *arXiv preprint:1803.09017*, 2018.

[8] D. Stanton, Y. Wang, and R. Skerry-Ryan, “Predicting expressive speaking style from text in end-to-end speech synthesis,” *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 595–602, 2018.

[9] Y.-J. Zhang, S. Pan, L. He, and Z. Ling, “Learning latent representations for style control and transfer in end-to-end speech synthesis,” *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6945–6949, 2019.

[10] R. Prenger, R. Valle, and B. Catanzaro, “Waveglow: A flow-based generative network for speech synthesis,” 2018.

[11] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu, “Efficient neural audio synthesis,” 2018.

[12] Gabriel, Mittag and Sebastian, Möller, “Deep Learning Based Assessment of Synthetic Speech Naturalness ,” *in Proc. Interspeech*, 2020.

[13] J. S. Wiggins, P. Trapnell, and N. Phillips, “Psychometric and Geometric Characteristics of the Revised Interpersonal Adjective Scales (IAS-R),” *Multivariate Behavioral Research*, 1988.

[14] R. McCrae and O. John, “An Introduction to the Five-Factor Model and its Applications,” *Journal of Personality*, 1992.

[15] V. P. Rosenberg S, Nelson C, “A multidimensional approach to the structure of personality impressions,” *Journal of Personality and Social Psychology*, 1968.

[16] S. E. Asch, “Forming impressions of personality,” *The Journal of Abnormal and Social Psychology*, 1946.

[17] B. W. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. J. J. H. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, “The INTERSPEECH 2012 Speaker Trait Challenge,” in *Proc. Interspeech*, 2012.

[18] L. Fernández Gallardo and B. Weiss, “The Nautilus Speaker Characterization Corpus: Speech Recordings and Labels of Speaker Characteristics and Voice Descriptions,” in *Proc. Eleventh International Conference on Language Resources and Evaluation (LREC)*, 2018.

[19] L. Fernández, Gallardo, and B. Weiss, “Towards Speaker Characterization: Identifying and Predicting Dimensions of Person Attribution,” *in Proc. Interspeech*, 2017.

[20] A. Cuddy, S. Fiske, and P. Glick, “The BIAS Map: Behaviors from Intergroup Affect and Stereotypes,” *Journal of personality and social psychology*, vol. 92, pp. 631–48, 05 2007.

[21] A. Abele, N. Hauke, K. Peters, E. Louvet, A. Szymkow, and Y. Duan, “Facets of the fundamental content dimensions: Agency with competence and assertiveness—communion with warmth and morality,” *Frontiers in Psychology*, vol. 7, 2016.

[22] S. T. Fiske, “Stereotype Content: Warmth and Competence Endure,” *Current Directions in Psychological Science*, 2018.

[23] W. Thompson and G. Ilie, “A comparison of acoustic cues in music and speech for three dimensions of affect,” *Music Perception*, vol. 23, p. 319, 04 2006.

[24] P. Borkenau and A. Liebler, “Convergence of stranger ratings of personality and intelligence with self-ratings, partner ratings, and measured intelligence,” *Journal of Personality and Social Psychology*, vol. 65, pp. 546–553, 09 1993.

[25] P. Laukka, D. Neiberg, M. Forsell, I. Karlsson, and K. Elenius, “Expression of affect in spontaneous speech: Acoustic correlates and automatic detection of irritation and resignation,” *Computer Speech Language*, vol. 25, no. 1, pp. 84–104, 2011, affective Speech in Real-Life Interactions.

[26] A. Baird, S. Jørgensen, E. Parada-Cabaleiro, N. Cummings, S. Hantke, and B. Schuller, “The perception of vocal traits in synthesized voices: Age, gender, and human likeness,” *Journal of the Audio Engineering Society*, vol. 66, pp. 277–285, 04 2018.

[27] A. Baird, E. Parada-Cabaleiro, S. Hantke, F. Burkhardt, N. Cummins, and B. Schuller, “The perception and analysis of the likeability and human likeness of synthesized speech,” 09 2018, pp. 2863–2867.

[28] A. Baird, S. Jørgensen, E. Parada-Cabaleiro, S. Hantke, N. Cummins, and B. Schuller, “Perception of paralinguistic traits in synthesized voices,” 08 2017, pp. 1–5.

[29] B. Naderi and R. Cutler, “An open source implementation of itu-t recommendation p.808 with validation,” to appear in INTERSPEECH. ISCA, 2020.

[30] D. Guse, H. R. Orefice, G. Reimers, and O. Hohlfeld, “Thefragebogen: A web browser-based questionnaire framework for scientific research,” in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2019, pp. 1–3.

[31] ITU-T Recommendation P.808, *Subjective evaluation of speech quality with a crowdsourcing approach*. Geneva: International Telecommunication Union, 2018.

[32] B. Naderi, I. Wechsung, and S. Möller, “Effect of being observed on the reliability of responses in crowdsourcing micro-task platforms,” in *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, 2015, pp. 1–2.

[33] B. Naderi, *Motivation of workers on microtask crowdsourcing platforms*. Springer, 2018.

[34] S. T. Fiske, A. J. Cuddy, and P. Glick, “Universal dimensions of social cognition: warmth and competence,” *Trends in Cognitive Sciences*, vol. 11, no. 2, pp. 77–83, 2007.