11th ISCA Speech Synthesis Workshop (SSW 11)
26-28 August 2021, Budapest, Hungary

Identifying the vocal cues of likeability, friendliness and skilfulness in synthetic speech

Sai Sirisha Rallabandi[1], *Babak Naderi* [1] *and 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, babak.naderi, sebastian.moeller}@tu-berlin.de

Abstract

The advent of neural Text-to-Speech (TTS) synthesizers has enhanced the expressivity of synthetic speech in the recent past. However, there is very little work on understanding the acoustic correlates of paralinguistic traits, emotions, speaker attributes and characteristics from synthetic speech. This paper investigates the acoustic correlates of the speaker attributes: likeability, friendliness, and skillfulness. Our study was carried out on the voices derived from the two commercial TTS systems, Amazon Polly (9 voices) and Google TTS engine (10 voices). In our previous study, we performed a crowd-sourcing-based evaluation to collect the subjective ratings for the desired speaker attributes. In this work, we perform the acoustic feature prediction using the backward elimination algorithm. We show that the level of loudness, spectral flux, fundamental frequency, its formant frequencies, and their combinations contribute to the desired speaker attributes. We further combine the ratings of friendliness and likeability to investigate the characteristic, warmth in synthetic speech and correspondingly, skilfullness represents the characteristic, competence.

Index Terms: Synthetic speech, acoustic correlates, linear regression, likeable, friendly, skillful

1. Introduction

Artificial speech generation is predominant through its applications such as navigation [1], language learning systems [2], customer service [3], personal assistants [4] and many more. The neural speech synthesizers have facilitated the expressivity in the generated speech in the recent past [5, 6, 7, 8, 9, 10]. The fidelity of these TTS systems can be further enhanced through the generation of paralinguistic traits, various emotions, and social speaker characteristics. In order to achieve this, there is a need for the investigation of the acoustic correlates of various speaker attributes in synthetic speech. In this paper, we identify the vocal cues responsible for the social speaker characteristics, warmth and competence in synthetic voices.

Perception of a person from their behavior has been researched extensively in 1940s and 50s [11, 12]. Various studies were carried out similar to the BIG FIVE personality traits to categorise and understand the human behavior [13, 14, 15, 16]. In [12], the sociology researchers stated that the perception of a person is done based on two criteria: social norms (warmth) and task accomplishment (competence). In [17] psychology researchers state that the characteristics, warmth and competence can be termed as the universal dimensions of social perception. This is because they include both interpersonal relationships and the social behavior of a person. The attributes that describe the characteristics, warmth and competence in humans were: likeability, friendliness, and skilfulness respectively [18]. Following [18], in our current work, we utilise these 3 dimensions (likeability, friendliness and skilfulness) to interpret warmth and competence in synthetic voices. Similar to BIG FIVE [13, 14] and [18, 19], we conducted a subjective evaluation with two commercial TTS systems (Google TTS engine, Amazon Polly) in [20]. The evaluation was carried out with 15 different adjectives describing various speaker attributes of TTS voices. As an extension, in the current work, we are interested in identifying the acoustic correlates of the speaker attributes that contribute to the social characteristics, warmth and competence in synthetic speech. Inspired by [18] we utilise a subset of the subjective ratings (likeability, friendliness, skilfulness) collected in [20].

The acoustic correlates of various emotions, moods, attitudes, personality traits have been researched previously in both natural and synthetic voices [21, 22, 22, 23, 24, 25]. Speech rate, intensity, speech pauses, pitch, duration and their combinations were commonly identified as the vocal cues responsible for various emotions, personality traits and speaker characteristics [21, 24, 25, 26, 27, 28]. Literature suggests that the acoustic correlates of various emotions and expressions can be divided into 3 categories: voice quality, timing and pitch parameters [21, 22, 28].

To the best of our knowledge, this is the first attempt to analyse the vocal cues of speaker attributes, friendliness, likeability and skilfulness in order to understand the social speaker characteristics, warmth and competence from synthetic speech. Inspired by [25], we perform an acoustic feature prediction over the OpenSMILE features [29] extracted for the synthetic voices. Through our work, we present that the spectral flux, formants (F1, F2, F3), slope of the voiced segment are responsible for warmth in female voices. While, first and second formants (F1, F2), slope of Unvoiced segment, and loudness contribute to warmth in male voices. Competence in female voices is perceived through slope of voiced segment, spectral flux and mfcc. While, the competence in male voices is due to fundamental frequency (F0) and voiced segment length. Later, we perform an automatic prediction of warmth and competence using linear regressor and Support Vector Regressor (SVR).

This paper is organised as follows: In Section 2, we describe the evaluation setup followed by the system performance in Section 3. In Section 4 we present the acoustic feature prediction. Automatic prediction of warmth and competence is presented in Section 5 followed by a discussion in Section 6.

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 items/attributes/adjectives/questionnaire refer to the questions we used in the subjective evaluation. We

10.21437/SSW.2021-1

use the terms dimensions/speaker attributes to refer likeability/friendliness/skilfulness.

2. Evaluation setup

2.1. Speech data preparation

The commercial TTS systems, Amazon Polly [1] (Neural) and Google TTS engine [2] (Wavenet) have been explored for the study of speaker attributes. There were 4 male and 5 female voices from Amazon Polly and 5 male, 5 female from Google TTS engine. In total, we had 19 different US native speakers' voices. The speech samples generated for each of these voices were from the Harvard database [3]. The number of sentences generated were 32. We have combined the individual speech samples and created speech segments each of length 20 seconds (approx.). Finally, there were 4 speech segments for each of the TTS voices (4 speech segments * 19 voices).

2.2. Subjective evaluation

The social perceptions of synthetic voices were studied in [20]. In [20], we performed a 15-item semantic differential scaling test in a crowd-sourcing subjective test setup. The 15 items were: relaxed, confident, enthusiastic, energetic, friendly, arrogant, pleasant, likeable, responsible, reliable, accessible, sympathetic, skilful, kind, extrovert. For our test, we have included the speaker attribution task in the P.808 toolkit [30]. We have utilised the continuous 100-point scale with the adjective-antonym pairs at its extreme ends. We used The Fragebogen [31] implementation for presenting the questionnaire during the subjective tests. The test was conducted on Amazon Mechanical Turk (AMT). We have included the eligibility and the environment suitability tests in out evaluation setup. We have recruited only US native speakers for the task. The participants were instructed to use headphones throughout the test without fail. The following are the instructions provided to the participants during the test.

For each question, please listen to the audio sample and give your opinion about the voice you hear on the following scales. You will find the positive adjective at the extreme left and a negative adjective at the extreme right of the scale. You can listen to each audio sample multiple times during the test. There is no right or wrong answer as long as you listen to the audio files and give your opinion.

In each session, participants were provided with 4 speech clips and 15 attributes. Additionally, we have repeated one attribute randomly for every question in a session. We have used this repeated attribute as the hidden quality control mechanism [32, 33]. Each speech clip played in a loop until the participant rated all the adjectives.

2.3. Data processing

We performed a pre-processing of the subjective data to remove the participants whose ratings were not reliable. Based on the environment suitability tests, we have rejected 41 responses. Later on, we calculated the Pearson correlation coefficient for the repeated attributes in the test and the original attribute. The ratings were rejected if the correlation coefficient was below 0.5 (5 participants were removed). In total, we have accepted 90% of the subjective data. The number of participants that were retained after the pre-processing were 43 female and 44 male (87 participants). Their ages ranged between 19 and 77 (mean = 40.31 and std = 12.57). On the retained subjective data, we have calculated the intraclass correlation coefficient ICC(1,k) for inter-rater reliability. The average raters absolute value was 0.974 with a 95% confidence interval in the range of 0.95 to 0.99. For our current study, we have utilised the subjective ratings of the scales, friendliness, likeability and skilfulness.

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

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

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

3. TTS performance

Figures 1, 2, 3, 4 display the perceived speaker attributes: friendliness, likeability and skilfulness in both the TTS systems. We have calculated the mean of the subjective ratings for each of these speaker attributes.

3.0.1. ***Google's female voices***

Figure 1 displays the mean subjective ratings calculated over the three desired attributes for Google's female voices along with the 95% confidence intervals. Among the Google's female voices, H displays lowest mean ratings on friendliness (37), likeability (37.6) and skilfulness (32.13). Speaker E displays highest rating on friendliness (59.9) and likeability (54.3). Speaker C displays the highest rating on skilfulness (41.64).

3.0.2. ***Google's male voices***

Figure 2 displays the mean subjective ratings calculated over the three desired attributes for Google's male voices along with the 95% confidence intervals. Among the Google's male voices, J displays lowest mean ratings on friendliness (31), likeability (29.25) and skilfulness (25.53). Speaker B displays highest rating on friendliness (47) and likeability (46.89). Speaker A displays the highest rating on skilfulness (35.66).

3.0.3. ***Amazon Polly's female voices***

Figure 3 displays the mean subjective ratings calculated over the three desired attributes for Amazon Polly's female voices along with the 95% confidence intervals. Among the Amazon Polly's female voices, Ivy displays lowest mean ratings on friendliness (36.9). Joanna display lowest ratings on likeability (35.93) and skilfulness (31.17). Speaker Kendra displays highest rating on friendliness (54.28) and likeability (51.39). Speaker Ivy displays the highest rating on skilfulness (45.6).

3.0.4. ***Amazon Polly's male voices***

Figure 4 displays the mean subjective ratings calculated over the three desired attributes for Amazon Polly's male voices along with the 95% confidence intervals. Among the Amazon Polly's male voices, Justin displays lowest mean ratings on friendliness (30.6), likeability (32.53). Matthew displays the lowest mean ratings on skilfulness (31.6). Speaker Joey displays highest rating on friendliness (46.2) and likeability (48.03). Speaker Kevin displays the highest rating on skilfulness (45.03).

4. Prediction of acoustic correlates

In order to predict the acoustic correlates of the desired characteristics, we initially downsample the speech segments to 16 kHz and derive the OpenSMILE [29] features. We have employed the Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) configuration [34], since this was built to capture affective speaker characteristics. We have therefore derived

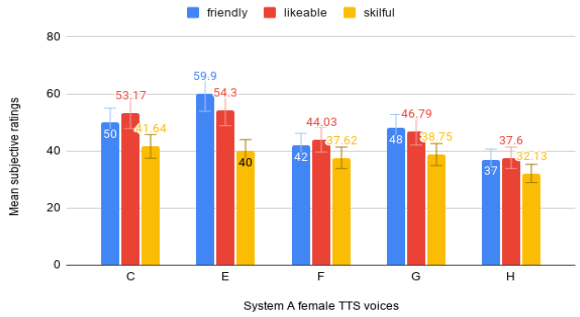

Figure 1: *Mean subjective ratings calculated over friendliness, likeability and skilfulness for Google's female voices.*

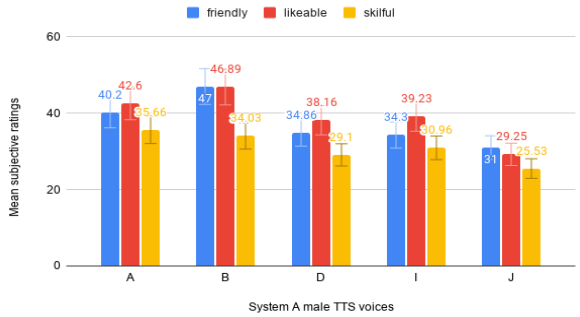

Figure 2: *Mean subjective ratings calculated over friendliness, likeability and skilfulness for Google's male voices.*

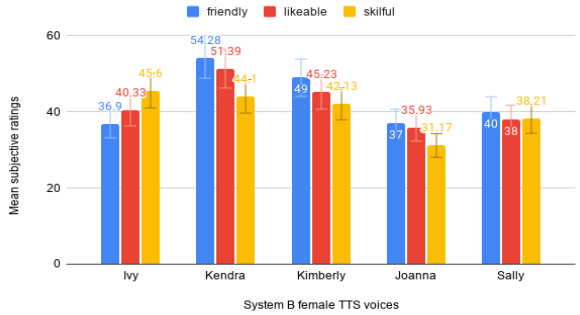

Figure 3: *Mean subjective ratings calculated over friendliness, likeability and skilfulness for Amazon Polly's female voices.*

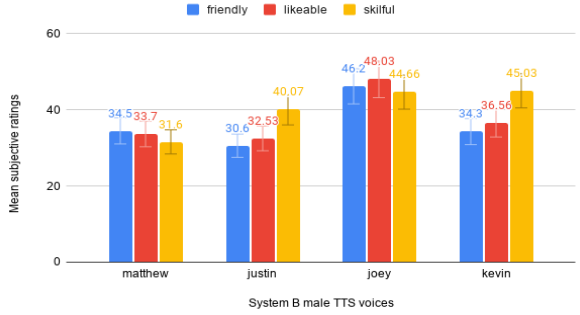

Figure 4: *Mean subjective ratings calculated over friendliness, likeability and skilfulness for Amazon Polly's male voices.*

88 acoustic features corresponding to each speech segment of the TTS voices. Later, we have employed a linear regression based backward elimination algorithm for each of the speaker attributes, friendliness, likeability and skilfulness. Tables, 1, 2, 3 display the derived acoustic features (for each of friendliness/likeability/skilfulness respectively), their corresponding coefficients along with their variances for female voices. We derived these vocal cues using the linear regression on 4 speech segments per each female voice (4 speech segments * 5 Google female voices + 4 speech segments * 5 Amazon Polly female voices). Tables, 4, 5, 6 display the derived acoustic features (for each of friendliness/likeability/skilfulness respectively), their corresponding coefficients along with their variances for male voices. We derived these vocal cues using the linear regression on 4 speech segments per each male voice (4 speech segments * 5 Google male voices + 4 speech segments * 4 Amazon Polly male voices).

Here, the derived acoustic features are the independent variables, the speaker attributes: friendliness, likeability and skilfulness are the dependent variables. A positive coefficient indicates that for a 1-unit change in the acoustic feature (independent variable), there will be an increase in the perception of the speaker attribute (friendliness/likeability/skilfulness) from that voice (increase in the mean of the dependent variable by that coefficient value) and vice versa.

4.1. Friendliness in female voices

Table 1 displays the acoustic correlates of friendliness in female voices (Google and Amazon Polly female voices). We observed that the attribute, friendliness was dependent on the acoustic features, spectral flux, and formants F1, F2 and F3 in female voices. We have also presented the explained variance (R squared = 0.829) obtained during the acoustic feature prediction. We observed that with the change in the spectral flux and the second formant (F2), the friendliness in female voices decreases by the value of 8.7546 and 0.0052 respectively. Accordingly, with the change in the first (F1) and third formants (F3) the friendliness in female voices increase by 0.0174 and 0.0037 respectively.

Table 1: *Acoustic correlates of friendliness in Amazon and Google female voices. The explained variance for female friendliness is 82.9%.*

Acoustic features	Coefficients
Spectral Flux	-8.7546
F1 mean	0.0174
F2 mean	-0.0052
F3 mean	0.0037

4.2. Likeability in female voices

Table 2 displays the acoustic correlates of likeability of female voices (Google and Amazon Polly female voices). We observed that the attribute, likeability was dependent on the acoustic features, spectral flux, formants F1, F2 and Voiced segment Slope (500-1500) in female voices. We have also presented the explained variance (R squared = 0.812) obtained during the acoustic feature prediction. We observed that with the change in the spectral flux, the second formant (F2), and the slope the likeability in female voices decreases by the value of 9.0631, 0.0086 and 57.3852 respectively. Accordingly, with the change in the

first (F1), the likeability of female voices increase by a factor of 0.0241.

Table 2: *Acoustic correlates of likeability in Amazon and Google female voices. The explained variance for female likeability is 81.2%.*

Acoustic features	Coefficients
Spectral Flux	-9.0631
F1 mean	0.0241
F2 mean	-0.0086
SlopeV500-1500	-57.3852

4.3. Skilfulness in female voices

Table 3 displays the acoustic correlates of skilfulness in female voices (Google and Amazon Polly female voices). We observed that the attribute, skilful was dependent on the acoustic features, Voiced segment Slope (0-500),spectral flux, voiced segment mfcc3 in female voices. We have also presented the explained variance (R squared = 0.581) obtained during the acoustic feature prediction. We observed that with the change in the spectral flux, slope and Mel Frequency Cepstral Coefficients (mfcc3), the perception of skilfulness in female voices decreases by the value of 6.4559, 0.1868 and 0.2858 respectively.

Table 3: *Acoustic correlates of skilfulness in Amazon and Google female voices. The explained variance for female skilfulness is 58.1%.*

Acoustic features	Coefficients
SlopeV0-500	-0.1868
SpectralFlux	-6.4559
mfcc3V	-0.2858

4.4. Friendliness in male voices

Table 4 displays the acoustic correlates of friendliness in male voices (Google and Amazon Polly male voices). We observed that the attribute, friendliness was dependent on the acoustic features, first formant (F1), Unvoiced segment Slope (500-1500) and loudness in male voices. We have also presented the explained variance (R squared = 0.685) obtained during the acoustic feature prediction. We observed that with the change in the first formant (F1), Voiced segment Slope (500-1500) and loudness the friendliness in male voices decreases by the value of 0.0117, 176.8888 and 1.1870 respectively.

Table 4: *Acoustic correlates of friendliness in Amazon and Google male voices. The explained variance for male friendliness is 68.5%.*

Acoustic features	Coefficients
F1 mean	-0.0117
SlopeUV500-1500	-176.8888
loudness	-1.1870

4.5. Likeability in male voices

Table 5 displays the acoustic correlates of likeability of male voices (Google and Amazon Polly male voices). We observed that the attribute, likeability was dependent on the acoustic features, loudness, loudness rising slope, formant F1, and unvoiced segment Slope (500-1500) in male voices. We have also presented the explained variance (R squared = 0.731) obtained during the acoustic feature prediction. We observed that with the change in the loudness rising slope, first formant (F1), second formant (F2), and unvoiced slope the likeability of male voices decreases by the value of 0.6164, 0.0101 and 169.6958 respectively. Accordingly, with the change in the loudness, the likeability of male voices increase by a factor of 6.7662.

Table 5: *Acoustic correlates of likeability in Amazon and Google male voices. The explained variance for male likeability is 73.1%.*

Acoustic features	Coefficients
loudness	6.7662
loudness rising slope	-0.6164
F1 mean	-0.0101
SlopeUV500-1500	-169.6958

4.6. Skilfulness in male voices

Table 6 displays the acoustic correlates of skilfulness in male voices (Google and Amazon Polly male voices). We observed that the attribute, skilful was dependent on the acoustic features, fundamental frequency (F0) and voiced segment length in male voices. We have also presented the explained variance (R squared = 0.698) obtained during the acoustic feature prediction. We observed that with the change in the fundamental frequency (F0), the perception of skilfulness in male voices decreases by the value of 8.7332. Accordingly, with the change in the voiced segment length, the perception of skilfulness in male voices increase by a factor of 6.1338.

Table 6: *Acoustic correlates of skilfulness in Amazon and Google male voices. The explained variance for male skilfulness is 69.8%.*

Acoustic features	Coefficients
F0 semitone	-8.7332
Voiced Segment Length	6.1338

5. Automatic prediction of warmth and competence

In this section, we present the automatic prediction of warmth and competence using the regression algorithms, linear regressor, and Support Vector Machine (SVM). For prediction of warmth, we have combined the subjective ratings of the scales, friendliness and likeability. For competence, we use the subjective ratings of skilfulness. The number of training examples we had were 40 for female and 36 for male voices. Hence, we perform a Leave-one-speaker-out cross validation. Table 7 shows the results of automatic prediction of warmth and competence. The first block consists of the prediction of warmth in male and female TTS voices. In the second block, we present the prediction performance for the characteristic, competence. The number of input features fed to the model in case of male and female voices and the characteristic predicted is presented. The performance of the models is evaluated with the metric, mean

Table 7: *Results of automatic prediction of warmth and competence from synthetic speech. AFs= number of acoustic features fed to the model, Ch. = characteristic, warmth (W) or competence (C) (2 attributes (likeability, friendliness) representing warmth and 1 attribute, skilfulness representing Competence) , LR = Linear Regression, SVR = Support Vector Regressor, MSE = mean squared error*

Model	Female			Male		
	AFs	Ch	MSE	AFs	Ch	MSE
LR	5	W	0.21	5	W	0.32
SVR	5	W	0.20	5	W	0.33
LR	3	C	0.47	2	C	0.35
SVR	3	C	0.45	2	C	0.34

squared error (MSE). We observe that the MSE score for female warmth is lower compared to that of MSE of male warmth with the same number of input features. In case of competence, eventhough the female input features are more than that of the male input features, the MSE scores of female are much higher than that of the male voices. The MSE scores of male warmth and competence display similar MSE scores with different number of input features.

6. Discussion

Table 8 presents the acoustic correlates of warmth in female voices. The vocal cues responsible for both the speaker attributes, friendliness and likeability are spectral Flux, first and the second formant (F1, F2). Additionally, third formant (F3) contributes to friendliness and Voiced slope contributes to likeability in female voices.

Table 8: *Warmth in female*

Friendliness	**Likeability**
Spectral Flux	*Spectral Flux*
F1 mean	*F1 mean*
F2 mean	*F2 mean*
F3 mean	SlopeV500-1500

Table 9 presents the acoustic correlates of warmth in male voices. The vocal cues responsible for both the speaker attributes, friendliness and likeability are loudness, first formant (F1) and unvoiced slope. Additionally, loudness rising slope contributes to likeability in male voices.

Table 9: *Warmth in male*

Friendliness	**Likeability**
F1 mean	*loudness*
SlopeUV500-1500	*F1 mean*
loudness	loudness rising slope
-	SlopeUV500-1500

Table 10 presents the acoustic correlates of competence in male and female voices. The vocal cues responsible for competence in male voices were fundamental frequency (F0) and voiced segment length. The acoustic correlates of competence in female voices were voiced slope, spectral flux and mfcc.

Table 10: *Competence in female and male voices*

Female	**Male**
Voiced Slope	F0
Spectral Flux	Voiced length
mfcc	-

From our analysis, we observed that the acoustic features intensity/loudness, spectral flux, fundamental frequency and its formants are the common acoustic features in both natural and synthetic voices contributing to various emotions and speaker characteristics [22, 25, 27]. We observe that the acoustic correlates of social speaker characteristics in synthetic speech can also be categorised into vocal quality (spectral parameters), timing (voiced segment length) and pitch (frequency parameters) as in natural speech [21, 22, 28].

The TTS voices, E (Google female voice) and Kendra (Amazon Polly female) display highest warmth among other TTS voices. The voices, Ivy (Amazon Female) and Kevin (Amazon male) display highest competence over the considered TTS voices.

The acoustic correlates predicted for each of the 3 attributes were obtained from the subjective evaluation conducted on a 15-item semantic differential scaling test. The subjective responses when requested for 3 scales (likeability, friendliness and skilfulness) alone might be different. Additionally, the models were trained on the 20 second long speech segments. Thus, we might have averaged the acoustic information present in the speech samples. Analysing the subjective ratings of individual speech samples could be interesting. Also, collection of the subjective ratings for a larger database and also different speech corpora (conversations, news reading, Semantically Unpredictable Sentences) is another future work. In the current work, the input dimensions (88) were higher than that of the number of training examples (40 for female, 36 for male) during automatic feature prediction. We have thus followed a recursive feature elimination approach for acoustic feature prediction. Therefore, as an extension to this work, we would perform an analysis with a larger dataset and unaveraged acoustic information.

7. 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.

8. References

[1] K. C. Raghavi, S. K. Rallabandi, S. Sitaram, and A. W. Black, "Speech synthesis for mixed-language navigation instructions," in *Proc. INTERSPEECH*, 2017.

[2] Y.-C. Huang and L.-C. Liao, "A study of text-to-speech (tts) in children's english learning," *Teaching English with Technology*, vol. 15, pp. 14–30, 01 2015.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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," *arXiv preprint:1904.04169*, 2019.

[8] 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.

[9] 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.

[10] 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.

[11] S. E. Asch, "Forming impressions of personality." *The Journal of Abnormal and Social Psychology*, vol. 41, no. 3, p. 258, 1946.

[12] R. F. Bales, "A set of categories for the analysis of small group interaction," *American Sociological Review*, vol. 15, no. 2, pp. 257–263, 1950.

[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] 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.

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

[19] 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.

[20] Sai Sirisha Rallabandi, Abhinav Bharadwaj, Babak Naderi, Sebastian Möller, "Perception of social speaker characteristics in synthetic speech," *in Proc. Interspeech*, 2021.

[21] K. R. Scherer, "Vocal affect expression: A review and a model for future research." *Psychological bulletin*, vol. 99, no. 2, p. 143, 1986.

[22] C. Gobl and A. N. Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Commun.*, vol. 40, pp. 189–212, 2003.

[23] C. Nass, U. Foehr, and M. Somoza, "The effects of emotion of voice in synthesized and recorded speech," 2001.

[24] J. Cahn, "The generation of affect in synthesized speech," *Journal of the American Voice I/O Society*, vol. 8, pp. 1–19, 1990.

[25] L. F. Gallardo and B. Weiss, "Perceived interpersonal speaker attributes and their acoustic features," *Proc. Phonetik & Phonologie*, 2017.

[26] P. Laukka, P. Juslin, and R. Bresin, "A dimensional approach to vocal expression of emotion," *Cognition and Emotion*, vol. 19(5), pp. 633–653, 08 2005.

[27] M. SCHROEDER, "Speech and emotion research : An overview of research frameworks and a dimensional approach to emotional speech synthesis," *Doctoral thesis, Phonus 7, Research Report of the Institute of Phonetics, Saarland University*, 2004.

[28] I. Murray and J. Arnott, "Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion." *The Journal of the Acoustical Society of America*, 1993.

[29] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*, 2013.

[30] B. Naderi and R. Cutler, "An Open source Implementation of ITU-T Recommendation P.808 with Validation," to appear in INTERSPEECH. ISCA, 2020.

[31] 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.

[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] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE transactions on affective computing*, 2015.