Interspeech 2018
2-6 September 2018, Hyderabad

The Perception and Analysis of the Likeability and Human Likeness of Synthesized Speech

Alice Baird[1], *Emilia Parada-Cabaleiro*[1], *Simone Hantke*[1,2], *Felix Burkhardt*[3], *Nicholas Cummins*[1], *Björn Schuller*[1,4]

[1]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, Univeristy of Augsburg, Germany
[2]Machine Intelligence and Signal Processing Group, Technische Universität München, Germany
[3] Deutsche Telekom, Berlin, Germany
[4]GLAM – Group on Language, Audio and Music, Imperial College London, UK.

alice.baird@informatik.uni-augsburg.de

Abstract

The synthesized voice has become an ever present aspect of daily life. Heard through our smart-devices and from public announcements, engineers continue in an endeavour to achieve naturalness in such voices. Yet, the degree to which these methods can produce likeable, human like voices, has not been fully evaluated. With recent advancements in synthetic speech technology suggesting that human like imitation is more obtainable, this study asked 25 listeners to evaluate both the likeability and human likeness of a corpus of 13 German male voices, produced via 5 synthesis approaches (from formant to hybrid unit selection, deep neural network systems), and 1 Human control. Results show that unlike visual artificially intelligent elements – as posed by the concept of the Uncanny Valley – likeability consistently improves along with human likeness for the synthesized voice, with recent methods achieving substantially closer results to human speech than older methods. A small scale acoustic analysis shows that the F0 of hybrid systems correlates less closely to human speech with a higher standard deviation for F0. This analysis suggests that limited variance in F0 is linked to a reduction in human likeness, resulting in lower likeability for conventional synthetic speech methods.

Index Terms: synthesized voices, human likeness, likeability.

1. Introduction

The anthropomorphisation of machines has been something of a curiosity for engineers throughout the 20th century [1, 2, 3], and is closely related to the advent of thought into artificially intelligent (AI) beings [4]. Giving a voice to such beings, has become a crucial consideration and necessary component for developers, as a means of achieving comfortable and 'social' Human Computer Interactions (HCI) [5]. As such voices are now embedded in our daily–life (from personal assistants [6] to humanoid robots [7]), this study explores the impact that human likeness may have on our perception of likeability.

From the recorded speech of repeated public announcements e. g., 'mind the gap', to the more complex (deep) machine learning Text-To-Speech (TTS) systems of today e. g., the IBM Watson System [8], or Deep Minds WaveNet [9], the methods for creating such voices have advanced substantially throughout the past decade [10, 8]. Now in the age of deep learning, it seems that true human likeness (or naturalness) of synthesized voices, is a more obtainable feature, with recent advancements focusing on specific speech features, including learning unknown pronunciations[11].

The perception of human speech is a well researched area [12, 13, 14]. However, considerably less efforts have been made towards the perception of synthesized speech. This is surprising as, synthesized voices are often a consideration in developing corporate identity (i. e., Apple's Siri, or Amazon's Alexa), and attractiveness (or likeability) is closely linked to commercial success [15]. Initial research has been made in relation to earlier synthesis methods, including the ability for such voices to portray personality traits [16], as well as the likeability of gendered synthesized voices [17]. Additionally, the effect of mixing human speech with synthetic speech as a means of improving likeability has also been evaluated [18].

Human likeness is a term commonly discussed in relation to AI through the concept of the *Uncanny Valley*. Coined by the Japanese robotisist Masahiro Mori in 1970 [19], this concept describes the 'almost but not quite' human likeness of visual features in humanoid robots, which may elicit an unfamiliar feeling or aversion to the AI 'being' [20]. The 'valley' is a point in the degree of human likeness when features are close (but not exact) replications of real-human features in which familiarity (and therefore in some way likeability) substantially decreases. The Uncanny Valley, has been extensively explored in relation to the visual attributes within AI [21, 22]. Yet, only briefly has the Uncanny Valley (in relation to robotic speech) explored [23], finding (as part of a multi-modal system) that human likeness is linearly related (in most cases) to the ability to portray emotion.In this way, [24] also evaluated the uncanny valley for text to speech (TTS), comparing two synthesis methods used within a *dialogue system*.

The implications of synthesized ('robotic') speech have been evaluated in [25], and synthetic voice identity itself, has begun to be highlighted as problematic in recent literature [26]. In this regard, brief studies by researchers relating to human likeness in synthesis [27, 28] have begun to appear. This study aims to build on this, utilising the German Text-to-Speech Dataset [1] (GTTS). From GTTS, a selection of male voices (balance was not possible across genders) have been gathered, and this study asked 25 listeners from varied backgrounds to assess the level of human likeness, and likeability of a selection of 13 voices of varying synthesis methods (including formant, diphone, unit selection, hybrid unit selection, and state of the art hybrid deep neural network systems) as well as 1 human voice. In this way, we are evaluating if human likeness and likeability are linked for the synthesized voice, asking if there is an Uncanny Valley, and how close to human are current methods able to achieve?

[1]www.ttssamples.syntheticspeech.de

10.21437/Interspeech.2018-1093

2. Methodology

2.1. Corpus

The corpus used within this study is a subset of GTTS, consisting of 13 male voices, and 39 utterances. Through out the years of synthesis (details of voice years given in Table 1), it is observed that for a long period synthesis methods consisted of diphone (two-phones) concatenation from recorded speech. Formant synthesis was also a common signal processing method, associated with the 'robotic' somewhat monotonic synthetic voices. Within this corpora the following synthesis methods are included; *formant synthesis* [29], *concatenative diphone synthesis* [30] , *conventional non-uniform unit-selection* [31], and *hybrid non-uniform unit selection synthesis, using Liljencrants-Fant model* [32], and state-of-the-art *hybrid unit-selection synthesis* methods with *deep neural network* frameworks e. g., (IBM Watson (2016) [2], and Readspeak (2018) [3]). Additionally, as a control, one human male voice is included.

Each of the 13 voices speaks 3 sentences (a total of 39 utterances). Files were extract for the dataset and converted to mono *wav* format (16 bit, 44.1 kHz) for listening test. Sentences were designed to evaluate known problems for German natural language processing modules. The sentences include:

1: '*An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht. Dabei war eigentlich immer sehr schönes Wetter gewesen.*'

2: '*Dr. A. Smithe von der NATO (und nicht vom CIA) versorgt z. B. – meines Wissens nach – die Heroin seit dem 15.3.00 täglich mit 13,84 Gramm Heroin zu 1,04 DM das Gramm.*'

3: '*Die Manpowerdiskussion wird gecancelt, du kannst das File vom Server downloaden.*'

Within the corpus only male voices have been taken from the GTTS data set. Using only the male voices was a decision taken due to the imbalanced nature of gender across the dataset, and the year based selection of the voices. Seemingly, earlier years for German speaking TTS were dominated by male synthetic speech, and although there may be some effect on the listener perception due to the speaker gender [33], for this study we have chosen to prioritise balancing the dataset.

2.2. Evaluation Parameters

As a means of investigating the advancements in voice synthesis, the parameters of *likeability* and *human likeness* are evaluated. The listening task was completed in the iHEARu-PLAY online browser-based annotation platform [34], and traits were divided into 2 individual tasks. 25 listeners (ages ranging from 22–57 years) of varying nationalities[4], 14 male and 11 female, voluntarily evaluated the corpus of 13 voices. Taking into account the data quality management procedures as described in [35], high quality annotations were collected.

In previous studies by the authors, the effect of perception on native and non-native listeners [27] was evaluated, finding no substantial difference in results. Therefore, here, the multinational listener group is not expected to compromise the result.

Likeability: To evaluate likeability, the listeners were asked to 'Please listen to the voice and judge the level of Likability. i. e., How much do you like the voice speaking?'. For each utterance, listeners evaluated the Likeability on a 5–point Likert scale; 1= Not at all, 5=Extremely.

[2] www.console.bluemix.net/docs/services/text-to-speech/science

[3] www.readspeaker.com

[4] Native language of listeners included: 11 English, 10 German, 1 Spanish, 1 Japanese, 2 Chinese.

Table 1: *A summary voices used within this study. All voices are male, German speaking. Each voice speaks 3 sentences, with a total of 39 utterances. The human voice was recorded in 2017, and all synthesized voices are marked in the name.*

Name	Developer	Type
Human	-	30 years Native German
F_1974	Samples of a Audio-data Braille reader	Hardware Formant Synthesizer
D_1996	Technical University of Dresden	Diphone Synthesis
D_1997	BabelTech®	Diphone-Concatenation Synthesis
D_1998	Elan® now Acapella®	Diphone-Concatenation
D_2000	Voice Interconnet	Diphone Synthesis
D_2004	Atip®	Diphone-Concatenation
U_2007	Nuance®	Non-uniform unit-selection
F_2006	eSpeak	Formant Synthesis
F_2009	Meridian	Formant Synthesis
HU_2014	VoxyGen®	Hybrid non-uniform unit-selection
DN_2016	IBM Watson®	Non-Uniform Unit-Selection using DNNs
DN_2018	ReadSpeak®	Non-Uniform Unit-Selection using DNNs

Human likeness: As in our previous studies [27, 28], the term *Human likeness* has been used, coined from the concept of the *Uncanny Valley* [19] to describe how accurately the machine is able to imitate a human. For this task, listeners were asked to 'Please listen to the voice and judge the level of human likeness. i. e., How close to human would you rate the voice speaking?'. Again, using a 5–point Likert scale 1= Not at all, 5=Extremely.

3. Evaluation of Results

To evaluate how likeabiltiy is linked to improved human likeness, multiple comparisons between the listeners' perception of the evaluated classes are presented, across the voices (covering several years of synthesis). Mean results from the evaluation are included in Table 2. Reported p values, are significant under the conventional threshold of $p < .05$. With *eta square* (η^2) being used as a measure of effect size.

3.1. Analysis and Discussion

For both of the evaluated parameters, i. e., likeability, and human likeness, the null hypothesis that the samples of all the evaluated groups are perceived similarly have been rejected, given the significant difference in the medians shown by the Kruskal-Wallis test: $H(13) = 375.01$, $p < .001$ for likeability; $H(13) = 513.57$, $p < .001$ for human likeness. In order to evaluate between which specific voices are perceived as significantly different, a pairwise comparisons among the 13 groups (considering the post hoc test Dunn-Bonferroni), has been applied.

It is worth noting here that we are unable to perform a *one-way ANOVA*. During initial analysis, a *Shapiro-Wilk* test was used to check for normality and returned $p < .001$ for both perception evaluation parameters (likeability and human likeness). Furthermore, the results of a *Levene* test revealed that the variances between listeners' responses were also not equal, again at significant levels, $p < .001$, for both parameters. Therefore, using the results of both tests, we rejected the null hypothesis of the population variances being equal and the distribution normal, thus we switched to a non-parametric method of analysis.

From the pairwise comparisons, results display no signif-

Table 2: *The mean (m) and standard deviation (std) is shown for each voice, for results of both perception of human-likeness (H-L) and likeability (L). Results above* 2.5 *are highlighted.*

	H-L		**L**	
	m	*std*	*m*	*std*
F_1974	1.07	0.25	1.24	0.51
D_1996	1.94	1.08	2.13	1.02
D_1997	**2.65**	1.30	2.32	1.00
D_1998	1.92	0.98	1.82	0.85
D_2000	2.13	0.92	2.17	0.76
D_2004	1.86	0.87	1.74	0.81
U_2007	**2.66**	1.03	2.16	0.92
F_2006	1.22	0.49	1.32	0.57
F_2009	1.27	0.61	1.41	0.72
HU_2014	**3.95**	1.14	**3.27**	1.15
DN_2016	**3.91**	0.98	**3.29**	1.15
DN_2018	**3.88**	0.97	**3.52**	0.85
Human	4.76	0.68	3.69	1.19

icant difference in listener perception of likeability between Human and HU_2014, DN_2016, and DN_2018. This is observed through a comparisons of 3 voices, i. e., Human vs HU_2014, Human vs DN_2016, and Human vs DN_2018; (p = 1.00), and a small effect size η^2 = 0.03. For human likeness, a similar tendency is observed, not yielding significant results: Human vs HU_2014 (p = .792, η^2 = 0.21); Human vs DN_2016 (p = .560, η^2 = 0.27); Human vs DN_2018 (p = .381, η^2 = 0.31).

From this, it is observed that HU_2014, DN_2016, and DN_2018 (the most recent synthesis methods) are perceived to be as likeable and human like as the Human voice. Looking closely at the results of these voices shows that the hybrid synthesis of HU_2014 is as human like and likeable as DN_2016 and DN_2018; and to evaluate further the advancements due to DNN-base synthesis would require a more focused study.

Despite the similarity of HU_2014, DN_2016, and DN_2018, there is a continued increase in the perception of likeability and human likeness across the corpora, observed prominently within the plotted results in Figure 1 (A). This continued increase is not linear, and we would like to draw attention to the large fluctuation in results between the years of 2004 and 2009, is mostly due to the formant synthesis methods marked in the time-line.

When looking at Figure 1 (B), the cluster of the formant synthesis voices is observed in the lowest range, despite their 32 years age-gap. As well as this clustering, an observable gap between the diphone and unit-selection voices, and the Hybrid and DNN voices can be observed; shown by e. g., by the significant difference yielded between D_1997 vs DN_2018 for human likeness (p = .001, η^2 = 0.22). This observation could be related to the findings in [23], in which portrayal of emotion was found to be linked to the human likeness of synthesized voices. In relation to this, it could be observed that results of our study, are due to a more effective portrayal of emotion in the hybrid-DNN voices; however, this would need a detailed emotion specific analysis.

When looking at these results in relation to the concept of the Uncanny Valley [19], Figure 1 (B) sees likeability in place of familiarity. It is observed that unlike with visual elements, when likeability increases, so too does human likeness. However, in Figure 1 (B), it can also be observed that there is a tendency towards the 'valley' which warrants further exploration between the years of 1998 and 2014. Through the comparison between

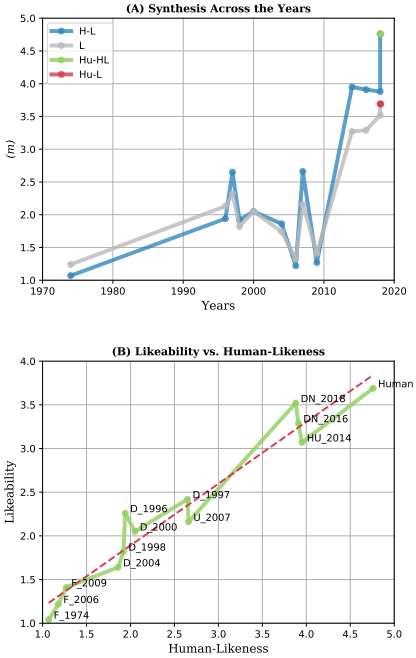

Figure 1: *Mean (m) results for human likeness (H-L) and likeability (L) for all the 13 voices evaluated in the corpora. (A) H-L and L of all voices, as well as Human-HL (Hu-HL) and Human-L (Hu-L) over time, (B) H-L against L for all voices.*

D_1996 vs D_2004, it is observed that these are perceived as highly similar, due to the small effect size of likeability and a none effect for human likeness (p = 1.00, η^2 = 0.03; and p = 1.00, η^2 = 0.00 respectively). On the contrary, D_2004 vs DN_2018 have been perceived as different, displayed by a medium effect size ($p > .001$ and η^2 = 0.54 for both likeability and human likeness), showing that voices above this year range (1998–2013, excluding the formant synthesis voice), are being perceived differently, to conventional synthesis methods as a whole, even those of lower human likeness. In this regard, despite this tendency towards the 'valley', it is observed consistently from the results that likeability does correspond to human likeness. Through computing a Pearson correlation of all human-likeness and likeability results from all voices, it is shown that the perception of these two is highly related, and there is a linear relationship between them (r = .489, $p < .001$). Therefore, a comparison to the conventional Uncanny Valley is not possible, as higher human likeness does not seem to match lower likeability at any point during the level of human likeness .

4. Acoustic Feature Analysis

As a means of evaluating the results from this study in more detail, focusing particularly on human likeness, a feature level analysis was made of 3 of the voices within the corpora. Figure 3. shows sentence 1, from 3 of the corpus voices (chosen due to their varied results, as well their polarising synthesis approaches): D_2004, DN_2018, and Human (as a control). The *smileF0* feature set from the open-source openSMILE feature extractor [36] was used, to automatically extract F0 at frame values (10ms) from the speech instances.

One observation from this selection of voices, is that the F0 of D_2004 has a much smoother trajectory (perhaps a more literal imitation); yet, this voice was perceived with less human

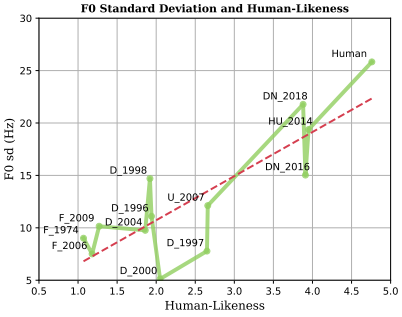

Figure 2: *The standard deviation (sd) result for the F0 of each voice against human likeness. We observe a relationship between the increase in human likeness and the value of standard deviation to the F0 (as shown from the dashed trend line).*

likeness and likeability (mean of 1.86 and 1.74 respectively, cf. Table 2). As well as this, the voice characteristics fit with the stereotype of monotonic synthesized speech, having an F0 mean (m) of 97.59 Hz and a smaller standard deviation (sd) of 9.7 Hz. This results is mentionable as – unlike visual elements – discussed through the Uncanny Valley – a somewhat precise imitation in the prosodic features is seen, which does not come through in perception of human likeness.

An observed difference when comparing `D_2004` and `DN_2018`, is that `DN_2018` has a noticeable *phrase-final lengthening* in as the prosodic boundary (highlighted by the red arrow), something typical to human speech [37], which is less prevalent in `D_2004`. The prosodic flow of `DN_2018` is dynamic, as is the `Human` voice (`Human` F0 m: 138.27 Hz, sd: 24.04 Hz; `DN_2018` F0 m: 108.61 Hz, sd: 21.89 Hz). In this regard (cf. Figure 2), we have calculated the sd of the mean F0, against human likeness for all voices, here observing a trend that voices with higher deviation from the mean F0 are also more human like. We speculate that `DN_2018` achieved both high human likeness, and likeability ratings as compared to `D_2004` (mean of 1.86 and 1.74 respectively, cf. Table 2), due to such prosodic behaviours.

5. Conclusions

Through this evaluation of the perception of human likeness and likeability across a corpora of 13 voices including 5 synthesis types, and 1 human control, it is clear that methodologies for voice synthesis generation have improved across the years for the evaluated parameters e. g., as shown by voice `DN_2018` against `D_2004`. However, state-of-the-art Hybrid-DNN voices have not yet made a substantial improvement over previous hybrid methods (`HU_2014`) methods (as no significance is shown between them). However, the results for the hybrid-DNN voices are promising for synthetic audio generation. With similar DNN methods being applied in multiple domains (e. g., music generation[9]), systems are seemly reaching a natural level of replication ability, which can only be positive for the computer audition community. In this regard, future analysis could include a variety of languages, as well as a focused synthesis selection in order to fully evaluate how DNNs are building upon the previous state of the art.

Of prominence, it was also observed that likeability and

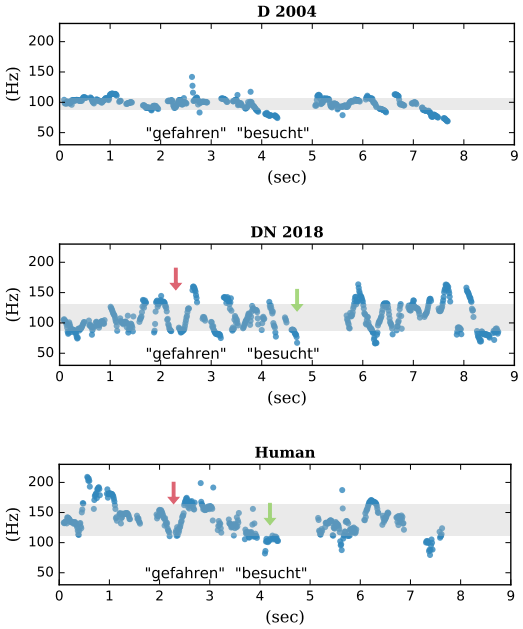

Figure 3: *F0 feature analysis for 3 voices within the corpora; `D_2004`, `DN_2018`, `Human`. The F0 standard deviation is plotted with the horizontal highlight, observing a larger variance in the prosodic pattern of `DN_2018`, and `Human` compared to that of `D_2004`. Labelled in blue (left arrow), a dynamic shift in the prosodic flow is observed. In green (right arrow), a human like pre-final lengthening of the phrase boundary is seen, both in `DN_2018` and `Human`.*

human likeness are correlated but unlike the concept for visual features discussed through the concept of the Uncanny Valley; there does not seem to be an aversion to the voices when they achieve more human like features. In reality, likeability is very much related to the synthesis methods, e. g., the clusterings mentions in Section 3. Despite this, it could be considered that there is not an Uncanny Valley for the synthesized voice, but rather an entire period, as it was shown that from 2014, voices synthesis methods achieve significantly higher results across both parameters.

From the feature analysis it was also found that hybrid-DNN methods retain the human like variance (F0 sd) affecting the prosodic flow, which can be a strong indication for such high human likeness ratings – a finding which warrants further analysis across a selection of precise state-of-the-art voice synthesis methods, and in this regard the incorporation of an emotion-based evaluation may have insightful findings.

6. Acknowledgements

This work is funded by the Bavarian State Ministry of Education, Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B), and the European Union's Seventh Framework and Horizon 2020 Programmes under grant agreements No. 338164 (ERC StG iHEARu).

7. References

[1] S. W. Homer Dudley, R. Riesz, "A Synthetic Speaker," *Journal of The Franklin Institute*, vol. 227, no. 6, pp. 739–764, June 1939.

[2] G. Fant, "The Source Filter Concept in Voice Production," *Speech Transmission Laboratory-QPSR*, vol. 22, no. 1, pp. 21–37, 1981.

[3] R. Scha, "Virtual Voices," *Mediamatic Magazine*, vol. 7, no. 1, pp. 27–42, 1992.

[4] P. McCorduck, *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*. Natik, MA, USA: AK Peters Ltd, 2004.

[5] K. M. Lee and C. Nass, "Designing Social Presence of Social Actors in Human Computer Interaction," in *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, Ft. Lauderdale, Florida, USA, 2003, pp. 289–296.

[6] J. Sánchez and C. Oyarzún, "Mobile Audio Assistance in Bus Transportation for the Blind," *Journal of the National Institute of Child Health and Human Development in Israel*, vol. 10, no. 4, pp. 365–371, 2011.

[7] S. Shamsuddin, L. I. Ismail, H. Yussof, N. I. Zahari, S. Bahari, H. Hashim, and A. Jaffar, "Humanoid robot nao: Review of control and motion exploration," in *2011 IEEE International Conference on Control System, Computing and Engineering*, 2011, pp. 511–516.

[8] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, "Prosody Contour Prediction with Long Short-Term Memory, Bi-Directional, Deep Recurrent Neural Networks," in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, 2014, pp. 2268–2272.

[9] A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Driessche, E. Lockhart, L. C. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, and D. Hassabis, "Parallel wavenet: Fast high-fidelity speech synthesis," vol. abs/1711.10433, 2017.

[10] M. Schröder, "Approaches to Emotional Expressivity in Synthetic Speech," in *Emotions in the Human Voice*, ser. Culture and Perception. Oxfordshire, United Kingdom: Plural Publishing, 2009, vol. 3, ch. 19, pp. 307–323.

[11] K. Sawada, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Constructing text-to-speech systems for languages with unknown pronunciations," *Acoustical Science and Technology*, vol. 39, no. 2, pp. 119–129, 2018.

[12] M. Latinus and P. Belin, "Human voice perception," *Current Biology*, vol. 21, no. 4, pp. 143–145, 2011.

[13] D. Puts, S. Gaulin, and K. Verdolini, "Dominance and the evolution of sexual dimorphism in human voice pitch," *Evolution and Human Behavior*, vol. 27, no. 4, pp. 283–296, 2006.

[14] C. Tigue, D. Borak, J. J. O'Connor, C. Schandl, and D. Feinberg, "Voice pitch influences voting behavior," *Evolution and Human Behavior*, vol. 33, no. 3, pp. 210–216, 2012.

[15] B. D. Till and M. Busler, "The Match-Up Hypothesis: Physical Attractiveness, Expertise, and the Role of Fit on Brand Attitude, Purchase Intent and Brand Beliefs," *Journal of Advertising*, vol. 29, no. 3, pp. 1–13, 2000.

[16] C. Nass and K. Lee, "Does Computer-Synthesized Speech Manifest Personality? Experimental Tests of Recognition, Similarity-Attraction, and Consistency-Attraction." *Journal of Experimental Psychology*, vol. 7, no. 3, pp. 171–181, 2001.

[17] E. J. Lee, C. Nass, and S. Brave, "Can Computer-Generated Speech Have Gender?: An Experimental Test of Gender Stereotype," in *Proc. of CHI '00 Extended Abstracts on Human Factors in Computing Systems*, New York, NY, USA, 2000, pp. 289–290.

[18] L. Gong and J. Lai, "To Mix or Not to Mix Synthetic Speech and Human Speech? Contrasting Impact on Judge-Rated Task Performance versus Self-Rated Performance and Attitudinal Responses," *International Journal of Speech Technology*, vol. 6, pp. 123–131, 2003.

[19] M. Mori, "Bukimi No Tani [The Uncanny Valley]," *ENERGY*, vol. 7, no. 4, pp. 33–35, 1970.

[20] F. E. Pollick, "In Search of the Uncanny Valley," in *Proc. User Centric Media*, Venice, Italy, 2009, pp. 69–78.

[21] M. Eaton, *Evolutionary Humanoid Robotics: Past, Present and Future*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 42–52.

[22] W. J. Mitchell, S. Kevin A Szerszen, A. S. Lu, P. W. Schermerhorn, M. Scheutz, and K. F. MacDorman, "A Mismatch in the Human Realism of Face and Voice Produces an Uncanny Valley," *i-Perception*, vol. 2, no. 1, pp. 10–12, 2011.

[23] T. J. Burleigh, J. R. Schoenherr, and G. L. Lacroix, "Does the Uncanny Valley Exist? An Empirical Test of the Relationship between Eeriness and the Human Likeness of Digitally Created Faces," *Computers in Human Behavior*, vol. 29, no. 3, pp. 759–771, 2013.

[24] J. Romportl, "Speech synthesis and uncanny valley," in *Text, Speech and Dialogue - 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings*, 2014, pp. 595–602.

[25] S. Wilson and R. K. M. , "Robot, Alien and Cartoon Voices: Implications for Speech-Enabled Systems," in *Proc. Vocal Interactivity in-and-between Humans, Animals and Robots*, SkÃűvde, Sweden, 2017, p. no. pagination.

[26] T. Phan, "The Materiality of the Digital and the Gendered Voice of Siri," *Transformations*, no. 29, pp. 23–33, 2017.

[27] A. Baird, S. H. Jørgensen, E. Parada-Cabaleiro, S. Hantke, N. Cummins, and B. Schuller, "Perception of Paralinguistic Traits in Synthesized Voices," in *Proc. of Audio Mostly Conference*, London, United Kingdom, 2017, pp. 1–5.

[28] A. Baird, S. H. Jørgensen, E. Parada-Cabaleiro, N. Cummins, S. Hantke, and B. Schuller, "The Perception of Vocal Traits in Synthesized Voices: Age, Gender, and Human Likeness," *J. Audio Eng. Soc*, vol. 66, no. 4, pp. 277–285, 2018.

[29] S. Awad and B. Guérin, "An Otimisation of Formant Synthesis Parameter Coding," *Speech Communication*, vol. 3, no. 4, pp. 335–346, 1984.

[30] M. Beutnagel, A. Conkie, and A. Syrdal, "Diphone synthesis using unit selection," in *Proc. of the Third ESCA/COCOSDA Workshop on Speech Synthesis*, Blue Mountains, Australia,, 1998, pp. 185–190.

[31] A. P. Breen and P. Jackson, "Non-Uniform Unit Selection and the Similarity Metric within BT's Laureate TTS System," in *Proc. of the Third ESCA/COCOSDA Workshop on Speech Synthesis*, Blue Mountains, Australia,, 1998, pp. 201–206.

[32] Y. Agiomyrgiannakis and O. Rosec, "Arx-lf-based source-filter methods for voice modification and transformation," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2009, pp. 3589–3592.

[33] C. J. Stevens, N. Lees, J. Vonwiller, and D. Burnham, "On-line experimental methods to evaluate text-to-speech (tts) synthesis: effects of voice gender and signal quality on intelligibility, naturalness and preference," *Computer Speech & Language*, vol. 19, pp. 129–146, 2005.

[34] S. Hantke, F. Eyben, T. Appel, and B. Schuller, "iHEARu-PLAY: Introducing a Game for Crowdsourced Data Collection for Affective Computing," *Proc. 1st International WASA 2015, ACII 2015*, pp. 891–897, 2015.

[35] S. Hantke, Z. Zhang, and B. Schuller, "Towards Intelligent Crowdsourcing for Audio Data Annotation: Integrating Active Learning in the Real World," in *Proceedings INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association*, ISCA. Stockholm, Sweden: ISCA, August 2017, pp. 3951–3955.

[36] F. Eyben, F. Weninger, F. Gross *et al.*, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proc. 21st ACM Int. Conf. Multimedia, MM 2013*. Barcelona, Spain: ACM, Oct 2013, pp. 835–838.

[37] S. Nooteboom, "The Prosody of Speech: Melody and Rhythm," *The Handbook of Phonetic Sciences*, pp. 640–673, 1997.