Perception of paralinguistic traits in synthesized voices

Alice Baird, Stina Hasse Jørgensen, Emilia Parada-Cabaleiro, Simone Hantke, Nicholas Cummins, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Baird, Alice, Stina Hasse Jørgensen, Emilia Parada-Cabaleiro, Simone Hantke, Nicholas Cummins, and Björn Schuller. 2017. "Perception of paralinguistic traits in synthesized voices." In *Proceedings of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences - AM '17, London, United Kingdom, August 23 - 26, 2017*, edited by George Fazekas, Mathieu Barthet, and Tony Stockman, 17. New York, NY: ACM Press. https://doi.org/10.1145/3123514.3123528.

Perception of Paralinguistic Traits in Synthesized Voices

Alice Baird
Chair of Complex & Intelligent Systems,
Universität Passau, Germany
Alice.Baird@uni-passau.de

Stina Hasse Jørgensen
Department of Arts & Cultural Studies,
University of Copenhagen, Denmark
RHJ282@hum.ku.dk

Emilia Parada-Cabaleiro
Chair of Complex & Intelligent Systems,
Universität Passau, Germany
Emilia.ParadaCabaleiro@uni-passau.de

Simone Hantke
Chair of Complex & Intelligent Systems,
Universität Passau, Germany
MISP Group, MKK
Technische Universität München, Germany
Simone.Hantke@uni-passau.de

Nicholas Cummins
Chair of Complex & Intelligent Systems,
Universität Passau, Germany
Nicholas.Cummins@ieee.org

Björn Schuller
Chair of Complex & Intelligent Systems,
Universität Passau, Germany
Group on Language, Audio, and Music
Imperial College London, UK
Schuller@ieee.org

ABSTRACT

Along with the rise of artificial intelligence and the internet-of-things, synthesized voices are now common in daily–life, providing us with guidance, assistance, and even companionship. From formant to concatenative synthesis, the synthesized voice continues to be defined by the same traits we prescribe to ourselves. When the recorded voice is synthesized, does our perception of its new machine embodiment change, and can we consider an alternative, more inclusive form? To begin evaluating the impact of aesthetic design, this study presents a first–step perception test to explore the paralinguistic traits of the synthesized voice. Using a corpus of 13 synthesized voices, constructed from acoustic concatenative speech synthesis, we assessed the response of 23 listeners from differing cultural backgrounds. To evaluate if perception shifts from the defined traits, we asked listeners to assigned traits of age, gender, accent origin, and human–likeness. Results present a difference in perception for age and human–likeness across voices, and a general agreement across listeners for both gender and accent origin. Connections found between age, gender and human–likeness call for further exploration into a more participatory and inclusive synthesized vocal identity.

CCS CONCEPTS

• **Human-centered computing** → *User studies*; *Sound-based input / output*; *Personal digital assistants*;

KEYWORDS

Synthesized Voice, Humanisation of Synthesis, Human–Machine Interaction, Paralinguistic Traits, Personification Debate

AM '17, August 23–26, 2017, London, United Kingdom

ACM ISBN 978-1-4503-5373-1/17/08…$15.00
https://doi.org/10.1145/3123514.3123528\

ACM Reference Format:
Alice Baird, Stina Hasse Jørgensen, Emilia Parada-Cabaleiro, Simone Hantke, Nicholas Cummins, and Björn Schuller. 2017. Perception of Paralinguistic Traits in Synthesized Voices. In *Proceedings of AM '17, London, United Kingdom, August 23–26, 2017,* 5 pages.
https://doi.org/10.1145/3123514.3123528\

1 INTRODUCTION

The synthesized voice is an instrument of speech, which has evolved throughout the last century [8, 17, 18, 28, 30]. Often a disembodied assistant, it inhabits many mediums of our everyday–life, including: interactive speech and language training [10], guides for the blind on public transport [29], within smart-devices (i. e. Apple's®, Siri®), or smart–homes (i. e. Amazon's, Alexa), and educational devices, such as humanoid robots (i. e. Robokind's, Milo).

Development and implementation of synthesized voices is now a multi-million dollar industry, with some of the biggest technology companies including: Amazon, Apple, AT & T, Google, IBM, Nuance, and Yamaha successfully joining the field [1, 3, 12, 19, 26]. Such voices are now easily accessible and part of the 'norm', offering an instinctive interface between human and machine. Yet, arguably they lack the representation, of the diverse individuals who interact with them.

This paper (motivated by the [multi'vocal] collective [2]), is the start of a series of perception tests to explore the notions of identity, personhood and will that we ascribe to the synthesized voices talking to us. Presented herein is a perception test on a collection of synthesized voices produced by the IBM® Watson Text to Speech (TTS) system, an interactive synthesis system used to speak text input. This systems is able to adapt standard vocal features of the synthesized voice including; breathiness, and glottal tension. Offering control of prosody; pitch, and speaking rate, and expressiveness; 'Good News', and 'Uncertainty' [5].

Such human–like characteristics have been discussed in connection to the *personification debate* [4, 14]. This debate involves the need for interfaces to imitate human behaviour, asking if synthesized voices should adopt anthropomorphic traits, such as; gender, dialect, and portrayal of emotion [14]. However, the social and political implications involved in the design and aesthetic as discussed for other human–machine interactions, including; the internet [36]

Table 1: Traits provided by IBM®. Names will be abbreviated, e. g., `en-US_Lisa` = `US-F-1`.

IBM® Name	Study	Gender	Language
de-DE_Bridgit	DE-F-1	Female	German
de-DE_Dieter	DE-M-1	Male	German
en-GB_Kate	GB-F-1	Female	British English
en-US_Allison	US-F-1	Female	American English
en-US_Lisa	US-F-2	Female	American English
es-ES_Enrique	ES-M-1	Male	Castilian Spanish
es-ES_Laura	ES-F-1	Female	Castilian Spanish
es-LA_Sofia	LA-F-1	Female	Latin American Spanish
es-US_Sofia	US-F-3	Female	North American Spanish
fr-FR_Renee	FR-F-1	Female	French
it-IT_Fran.	IT-F-1	Female	Italian
ja-JP_Emi	JP-F-1	Female	Japanese
pt-BR_Isabela	BR-F-1	Female	Brazilian Portuguese

and humanoid–robots [27], have only recently begun to be discussed for the synthesized voice [23]. Although it is worth noting that the human voice, as a marker of identity, has been highlighted within cultural studies [15, 38].

Developing natural sounding synthesized voices requires collaboration from interdisciplinary fields, including the humanities (linguistics, philosophy, sociology, psychology), and engineering (machine learning, signal processing, interaction design). Predominately, focus lies with achieving naturalness, via concerns of corpus size, and processing capacity [14], with minimal consideration being given to the personification type, which may profoundly impact listeners.

Nevertheless, many designed synthesized voices are openly available, commonly through TTS systems, allowing for immediate user–interaction (i. e. Google Translate). State-of-the-art synthesized voices are popularly produced via concatenative sound synthesis, a method which automatically accesses minute segments within large speech corpora, to construct text input [32].

There have been many perception studies which mostly focus on improving, and better understanding naturalness of the synthesized voice. For example Gong et al. [11], explored the effect of human speech in combination with synthesized speech, showing that synthesized voices improved instruction understanding, better than the human–voice. Intelligibility, specifically for the blind was explored in [35], and [20] compared gendered synthesized voices for their authoritative influence. Nevertheless to the best of the authors knowledge, the paralinguistic (the study of speech, beyond the communicated message [34]) traits of the synthesized voice, have not yet been evaluated.

As a first-step for scrutinising the recent improvements of synthesized voice aesthetic design, this study assesses the response of 23 listeners from differing cultural backgrounds. Evaluating if the perceived vocal traits of age, gender, accent origin and human–likeness, can be assigned to each of the 13 synthesized voices in the IBM® Watson library (cf. Table 1), and if gender and accent origin differ from that which were previously attributed.

Through a more inclusive (broad in scale) trait annotation, this study hopes to begin understanding the personification assigned to synthesized voices. Given the success achieved in the field of computational paralinguistic's for vocal trait classification [21, 40], similar methodology for synthesized voices, may improve effectiveness, and also advance aesthetic understanding, to highlight the impact of synthesized voice personification.

2 SYNTHESIZED VOICE CONSTRUCTS

Since the *Voice Operating Demonstrator* was introduced in 1939 [17], speech synthesis has evolved, from circuitry replicating the vocal tract, to complex segmentation and automated reconstruction of recorded speech. Today's, synthesis methods include:*Formant, unit selection* (concatenation), *Hidden Markov Model* (HMM) (statistical parametric), and *coarticulation* synthesis [34].Predominately these techniques can be observed in two ways – rule–based manipulation of acoustic parameters, i. e. formant synthesis, and speech–based sample concatenation, i. e. unit selection [34].

Acoustic concatenation, is the method utilised by the IBM® Watson TTS system, and has become the popular for many others since the 1990's. This method takes a set of recorded sentences, and segments them into 10–100ms denominations, storing them in a corpora of varying sizes, which can be accessed to automatically reassemble the users text input. Depending on corpora size, concatenation synthesis can be very effective for TTS, and has the potential to pronounce all word combinations, achieving what some say is a natural human–like voice [14, 33].

Although it has been said that acoustic concatenation can offer greater naturalness, than previously developed methods, it does have some limited flexibility. Depending on the variation from the recorded speaker, long stretches of speech output can lack dynamic and often seem monotonous [16]. Some have suggested that naturalness can be increased through the addition of a synthetic fundamental frequency to the recorded segments [32], and more recently IBM® engineers are developing features including *Expressive Synthesis* and *Voice Transformation* [5] which allow for manipulation of human vocal feature such as F0, rate, and timbre. Additionally once optimal segments have been selected, the IBM® system uses Pitch Synchronous Overlap and Add (PSOLA), a time–domain signal processing method, which allows for modification of pitch and speech duration between individual segments [6].

Additionally, context–dependent HMM, is another method of synthesis implemented into HMM–based Speech Synthesis Systems (HTS). HMM synthesis generates sequences of minute time observations, to produce outputs based on their trained speech database [34, 39]. HMM–based synthesis has been shown to enhance human–like attributes, such as 'creaky voice' (vocal fry) [25]. More recently, engineers are exploring hybrid systems which use *Recurrent Neural Networks* (RNNs) and *Bidirectional Long Short Term Memory* (BLSTM) in combination with *Deep Neural Networks* (DNN)–based TTS systems. These have shown to outperform HMM synthesis, and produce 'smoother' speech [7], with research being pushed in this area across computational fields, it is likely that DNN-based systems will become the most successful for achieving natural human–like synthetic speech.

3 PERCEPTION TEST METHODS

To evaluate listeners perception of the considered traits (age, gender, accent origin and human–likeness) a corpus of 13 synthesized voices was rated by 23 listeners of varying nationalities[1]. All listeners were fluent in English, and some had knowledge of more than 3 other languages. The annotation task was undertaken in the iHEARu-PLAY online browser-based annotation platform [13], and traits were divided into individual tasks. Each task was created to allow for maximum inclusion, allowing for the diverse perception of personification for the voices.

Evaluation Parameters

(1) **Age:** A 10–point scale from 1–100 (each point corresponding to 9 years i. e. 1–10, 11–20, etc.). The large scale was chosen for evaluation as this trait was undefined IBM®.
(2) **Gender:** Considering masculinity and femininity, and the additional options of, both or neither. Although binary–gender may be prominent, given the disembodiment of machine voices, the listener can evaluate if these voices show any androgynous or opposing gender qualities.
(3) **Accent Origin:** Using the nationality meta–data provided by iHEARu-PLAY, listeners could select from 249 nationality options. This test will evaluate if the listeners perceptions deviate from the IBM®assigned languages.
(4) **Human–Likeness:** Using a 5-point Likert scale, 1=Artificial (or Non-Human) and, 5=Human. Coining the term *Human–Like* from the *Uncanny Valley* [22], this test will evaluate if human qualities are lost through concantenation synthesis.

Voice Corpus

The corpus used for this annotation task, was collected from the IBM® Watson Text to Speech API, developed for IBM® Watson Developer Cloud [5][2]. Watson is an expressive TTS system [24], which allows for the alteration of an array of humanised parameters, including; glottal tension, speech rate, and pitch range. All 13 available voices, including 11 female and 2 male were selected for analysis (full details are shown in Table 1). Note there is an unavoidable gender bias which is representative of the synthesized voice market [3]. For this first–step evaluation, we chose to use the entire IBM® Watson corpus.

Using the IBM® service, five sentences were captured for each of the 13 available synthesized voices:

(1) 'ne kal ibam soud molen!' – Nonsense
(2) 'koun se mina lod belam' – Nonsense
(3) 'How are you?' – Sense, in defined language
(4) 'Thank you' – Sense, in defined language
(5) 'I love you' – Sense, in defined language.

[1]Native language of participating listeners included; 1 Arabic, 4 Danish, 3 English, 10 German, 2 Mandarin, 1 Spanish, 1 Teluga, and 1 Urdu.
[2]Data was retrieved according to the terms of use set by IBM®. The utilised data will not be made publicly available, all voices can be heard from the IBM® service 'https://text-to-speech-demo.mybluemix.net/'.
[3]Evaluating commercially available synthesized voice assistants, found the following to be female; Siri®(Apple®), Alexa (Amazon), Cortana (Microsoft), Google Assistant, and S-Voice (Samsung).

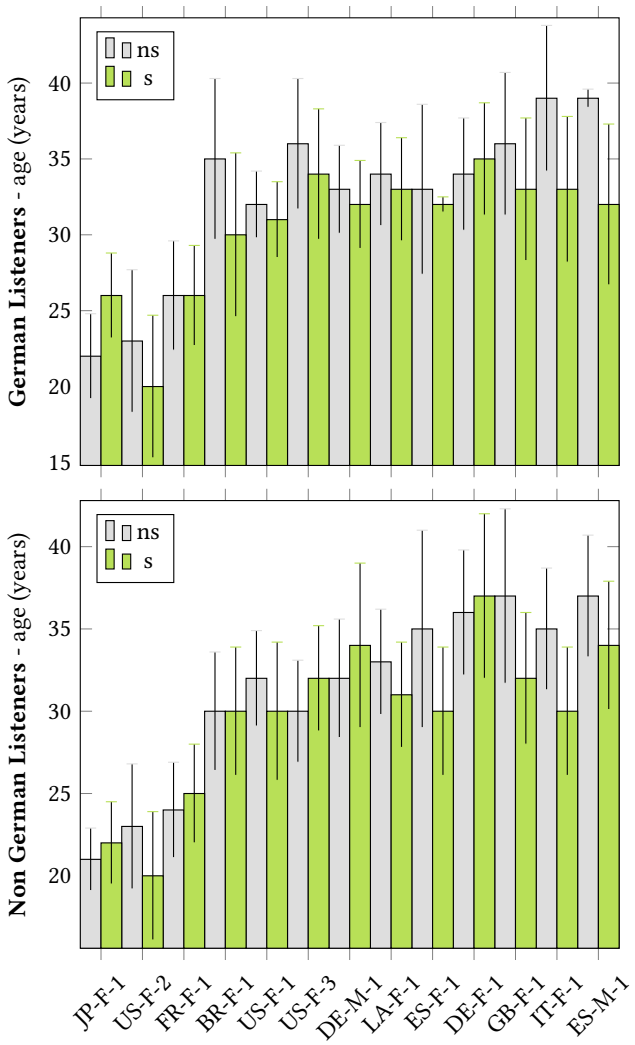

Figure 1: Mean age results for all voices, as perceived by German and Non German listeners. Showing nonsense (*ns*) and sense (*s*) utterances, and ± standard deviation for each.

Sentences (3), (4), and (5) were synthesised into the language corresponding to the assigned accent origin e. g., `DE-F-1` spoke sentence (3) as 'Wie geht es dir?'. Sentences (1) and (2), were designed as nonsense by GEneva Multimodal Emotion Portrayals (GEMEP) data set[37], and sentences (3), (4), and (5) were chosen due to being defined as the top 3 most Google translated sentences[9]. Audio files were originally captured in raw OGG and subsequently converted to *mp*3, 16 kHz, mono, 256 kb/s for compatibility with iHEARu-PLAY.

4 RESULTS AND DISCUSSION

To understand the perception of the evaluated traits, we first considered the listeners in two groups (German and non German) as there was unbalance across the listener groups. Our analysis (cf. Section 3) shows that there does not appear to be a relationship between listener nationality, and types of utterance. Results displayed

Table 2: Results for all evaluated traits. Age: Mean (m) and Standard Deviation (sd). Gender (%): Feminine (F), Masculine (M), Both (B), Neither (N). Accent Origin: 1st and 2nd most frequent. Human–Likeness: mean (m), mean difference (m_d) comparison, for 'least' (`JP-F-1`), and 'most' human (`ES-M-1`) vs. all. Values in bold, for gender and accent origin indicate results which correspond to the IBM® attribution and for human-likeness indicate $p < .05$ from a Tukey's post hoc test.

Name	Age		Gender (%)				Accent Origin		Human–Likeness (m, m_d)		
	m	sd	F	M	B	N	1st	2nd	m	JP-F-1	ES-M-1
BR-F-1	31	5.4	**90.8**	00.6	07.0	01.6	Spain	Italy	3.47	**0.91**	**0.73**
DE-F-1	35	6.6	**88.6**	00.0	08.7	02.7	**German**	UK	2.83	0.27	**1.37**
DE-M-1	33	11.6	24.3	**73.9**	00.9	00.9	**German**	Spain	2.93	0.34	**1.27**
ES-F-1	34	8.1	**93.7**	00.0	05.4	00.9	**Spain**	Italy	3.43	**0.87**	**0.77**
ES-M-1	38	10.2	06.9	**91.3**	00.9	00.9	**Spain**	Italy	4.20	**1.64**	−−
FR-F-1	25	10.0	**97.8**	00.0	02.2	00.0	**France**	Spain	3.74	**1.18**	0.46
GB-F-1	37	8.6	94.7	00.0	05.3	00.0	**UK**	USA	3.17	0.60	1.03
IT-F-1	37	4.7	87.8	0.8	08.7	02.7	**Italy**	Spain	3.01	0.45	**1.19**
JP-F-1	22	7.4	**94.7**	00.0	02.7	02.6	**Japan**	China	2.52	−−	**1.64**
LA-F-1	33	6.6	**93.9**	00.0	05.3	00.8	Spain	Italy	3.57	**1.00**	**0.63**
US-F-1	32	8.3	**84.4**	00.8	09.6	05.2	**USA**	UK	3.13	0.57	**1.07**
US-F-2	23	6.8	**94.7**	00.0	05.3	00.0	**USA**	UK	3.60	**1.04**	**0.60**
US-F-3	32	10.0	**96.5**	00.9	00.9	01.7	Spain	Italy	3.50	**0.93**	**0.70**

that listeners agree with the age grouping, for the 3 younger and 10 older voices: 20-26, 30-39 respectively.

The listeners evaluation displays a slight inclination towards the perception of voices pronouncing nonsense speech as younger, however our smaller corpus size, allows us only to interpret this as a tendency. We also consider that listeners' nationality and type of utterance do not effect listener perceptions; this could be due to the evaluated traits (age, gender, accent, and human–likeness) being related more to acoustic vocal characteristics, and have less relation to linguistic content. Since our results revealed this tendency across all evaluated traits, we will show results for the other parameters considering listener groups and utterances together, cf. Table 2.

The mean perceived ages for the voices; `JP-F-1`, `US-F-1`, and `FR-F-1` was younger (group 22–25) than for all the other voices, which were identified within the age group 31–38. The youngest mean age was 22 (standard deviation of 7.4 years) for `JP-F-1`, the voice defined by IBM® as a Japanese speaking female. The oldest age was given to `ES-M-1`, the Spanish male synthesized voice, shows a mean age of 38 years, and standard deviation of 10.2 years.

The listeners' perception of gender attributes for each synthesized voice is shown according to the four evaluated labels: femininity, masculinity, both, and neither. As can be seen, there is a clear relationship between gender associations and the IBM® given binary–gender. Despite this, `DE-M-1` (the German male voice), shows 73.9% masculine traits, and was evaluated by listeners to have 24.3% feminine traits. Acoustically comparing `DE-M-1` and ES-M-1 (IBM® Male, with 91.3% male attributes), we see that ES-M-1 has a formant structure simlar to that of human speech, whereas `DE-M-1` shows a broken formant structure which could be considered as more machine-like.

The first choice accent origin for all, except `BR-F-1` (Portuguese speaking Brazilian, identified as Spanish); were correctly identified in relation to the language defined by IBM®. Previous findings by [31], indicate that language 'family' can be correctly identified by a listener group which includes non-natives. Our results agree with this observation; with both first and second choices, falling into the previously defined language 'family' of each synthesized voice: *Neo-Latin* (Spanish, French, Italian, Portuguese), *Germanic* (German and English), and *East Asian* (Mandarin Chinese and Japanese). Nevertheless, listeners were not able to perceive the regional accent, e. g., all the Spanish speaking South American have been identified as Spanish.

Our analysis displays that the perception of human–likeness differs significantly across voices (cf. Table 2). The listeners consider `ES-M-1`, as the 'most human', rating this voice on average as 4.20 (84% − 1=0% Human and 5=100% Human), with an standard deviation of 1.00, and the 'least human' voice was `JP-F-1`, rated as 2.52 (39.6%), with standard deviation of 1.40.

The mean values and mean differences between the two 'least' and 'most' human voices are also given in Table 2. Values below the conventional threshold of $p < .05$, as obtained by Tukey's post hoc tests from a one way ANOVA, are highlighted in bold. The effect sizes d for the presented results vary from, 1.34 for (`JP-F-1` vs `ES-M-1`), to 0.51 for (`US-F-2` vs `ES-M-1`).

Evaluating our most androgynous result `DE-M-1` again, we see that this voice also received a much lower 'Human–like' result (2.93 = 48.8% Human), when compared to `ES-M-1` (4.20 = 80.2% Human). This leads us to assume a possible link between artificial voice, and perceived gender qualities. We also see a potential connection between age and human–likeness, as `JP-F-1`, our youngest voice (22), is also the least human (2.52 = 39.6%). However, `US-F-2` and `FR-F-1`, were also identified younger (25, 23), but received higher human–likeness.

5 CONCLUSION

In this study, we evaluated the responses of 23 listeners, who assigned paralinguistic vocal traits to 13 synthesized voices. Gender and accent origin, were compared to the IBM® given traits, and additionally age, and human–likeness were assigned to the voices. Of most prominence we have found that this selection of synthesized

voices can show a significant difference in the age result, which seems to parallel in most cases the result of human–likeness (i. e. the youngest voice was also perceived as the least human). Listeners did associate alternative gender traits to the voices, and in most cases were able to locate accent origin correctly as compared to the defined language locations previously attributed to them.

From this first–step study, our future work will include further human–likeness perception testing for a wider range of synthesized voices, created via differing synthesis methods. As well as evaluating closely the relationship of human–likeness and age, and accuracy of accent origin, via consideration to a more focused listener group. We also hope to additionally consider short–term emotional states, as a way to advance understanding of how the aesthetic design of synthesized voices can impact those in ear–shot.

ACKNOWLEDGMENTS

This work was supported by the European Union's Seventh Framework and Horizon 2020 Programmes under grant agreements No. 338164 (ERC StG iHEARu) and No. 688835 (RIA DE-ENIGMA).

REFERENCES

[1] Amazon. 2017. The Alexa Fund. (2017). http://amzn.to/2fD1COc/

[2] A. Baird, F. Tollund Juutilainen, S. Hasse Jorgensen, and M. Steensig Pelt. 2017. [multi'vocal], Exploring Representation, Identity and Aesthetics of Synthesized Voices. (2017). http://www.multivocal.org/

[3] M. Beaulieu. 2002. *Wireless Internet Applications and Architecture: Building Professional Wireless Applications Worldwide.* Pearson Education, Boston, MA, USA.

[4] C. Yen C. Nass. 2010. *The Man Who Lied to His Laptop: What We Can Learn About Ourselves from Our Machines.* Penguin Group, New York, NY, USA.

[5] IBM® Watson Developer Cloud. 2017. Text to speech. (2017). https://ibm.co/2vLOhNE

[6] IBM® Watson Developer Cloud. 2017. The Science Behind the Service. (2017). https://ibm.co/2vtyDnu

[7] Y. Fan, Y. Quan, F. Xie, and F. Soong. 2014. HMM-based synthesis of creaky voice. *In Proc. Interspeech* (2014), 964–1968.

[8] G. Fant. 1981. The Source Filter Concept in Voice Production. *STL-QPSR* 22, 1 (1981), 21–37.

[9] L. Ferlazzo. 2015. The Most Translated Words Using Google Translate Are. (2015). http://bit.ly/2wArIZI

[10] J. Ferrell. 1999. System and Method for Multimodal Interactive Speech and Language Training. (23. 03. 1999).

[11] L. Gong and J. Lai. 2003. 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* 6 (2003), 123–131.

[12] Yamaha Group. 2014. Designing the New Sound. Annual report 2014. (2014). http://bit.ly/2vsTIOR

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

[14] R. A. Harris. 2005. *Voice Interaction Design: Crafting the New Conversational Speech Systems.* Morgan Kaufmann Publishers /Elsevier, San Francisco, CA, USA.

[15] S. Hasse. 2016. Stemmernes Politik I Samtidskunsten. *TerrÃęn: Dansk Samtidskunst, Aarhus Universitetsforlag* (2016), no pagination.

[16] J. Hirschberge. 2006. Speech Synthesis: Prosody. *In Encyclopedia of Language & Linguistics* 7 (2006), 49–55.

[17] S. Watkins Homer Dudley, R. Riesz. 1939. A Synthetic Speaker. *Journal of The Franklin Institute* 227, 6 (June 1939), 739–764.

[18] U. Jekosch. 2005. *Voice and Speech Quality Perception: Assessment and Evaluation.* Springer-Verlag Berlin Heidelberg, Heidelberg, Germany.

[19] A. Kharpal. 2017. Amazon Voice Assistant Alexa could be a Billion Dollar Mega-Hit by 2020. (2017). http://cnb.cx/2vWx8QX

[20] E. Ju Lee, C. Nass, and S. Brave. 2000. Can Computer-generated Speech Have Gender?: An Experimental Test of Gender Stereotype. In *CHI '00 Extended Abstracts on Human Factors in Computing Systems (CHI EA '00).* ACM, New York, NY, USA, 289–290.

[21] E. Marchi, F. Eyben, G. Hagerer, and B. W. Schuller. 2016. Real-time Tracking of Speakers' Emotions, States, and Traits on Mobile Platforms. In *Proc. Interspeech 2016.* ISCA, ISCA, San Francisco, CA, 1182–1183.

[22] M. Mori. 1970. Bukimi No Tani [The Uncanny Valley]. *ENERGY* 7, 4 (1970), 33–35.

[23] T. Phan. 2017. The Materiality of the Digital and the Gendered Voice of Siri. *Transformations* 29 (2017), 23–33.

[24] J. F. Pitrelli, R. Bakis, E. M. Eide, R. Fernandez, W. Hamza, and M. A. Picheny. 2006. The IBM Expressive Text-to-Speech Synthesis System for American English. *IEEE Transactions on Audio, Speech, and Language Processing* 14, 4 (2006), 1099–1108.

[25] T. Raitio, J. Kane, T. Drugman, and Gobl C. 2013. HMM-based Synthesis of Creaky Voice. *In Proc. Interspeech* (2013), 2316–2320.

[26] B. B. Read. 2011. IVR: Nuance Acquires PerSay to Bring Voice Biometrics to Market. (2011). http://bit.ly/2uv4YNr

[27] J. Robin. 2008. 'Robo-Diva R&B' Aesthetics, Politics, and Black Female Robots in Contemporary Popular Music. *Journal of Popular Music Studies* 20, 4 (2008), 402–423.

[28] M. R.Schroeder. 2004. *Computer Speech: Recognition, Compression, Synthesis.* Springer-Verlag, Heidelberg, Germany.

[29] J. Sánchez and C. Oyarzún. 2011. Mobile audio assistance in bus transportation for the blind. *Official journal of the the National Institute of Child Health and Human Development in Israel* 10, 4 (2011), 365–371.

[30] R. Scha. 1992. Virtual Voices. *Mediamatic Magazine* 7, 1 (1992), 27–42.

[31] K. Scherer, R. Banse, and H. Wallbott. 2001. Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures. *Journal of Cross Cultural Psychology* 32, 1 (2001), 76–92.

[32] M. Schröder. 2001. Emotional Speech Synthesis: A Review. *In Proc. Interspeech* (2001), 964–1968.

[33] M. Schröder. 2009. Approaches to Emotional Expressivity in Synthetic Speech. In *Emotions in the Human Voice*, Krzysztof Izdebski (Ed.). Culture and Perception, Vol. 3. Plural Publishing, United Kingdom, Chapter 19, 307–323.

[34] B. Schuller and A. Batliner. 2013. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing.* Wiley, Hoboken, NJ, USA.

[35] A. Stent, A. Syrdal, and T. Mishra. 2011. On the Intelligibility of Fast Synthesized Speech for Individuals with Early-onset Blindness. In *Proc. ACM SIGACCESS (ASSETS 2011).* ACM, New York, NY, USA, 211–218.

[36] T. Streeter. 2003. The Romantic Self and the Politics of Internet Commercialization. *Cultural Studies* 17, 5 (2003), 648–668.

[37] K. Scherer T. Bänziger, H. Pirker. 2006. GEMEP-GEneva Multimodal Emotion Portrayals: A corpus for the study of multimodal emotional expressions. In *In Proc. Language Resources and Evaluation.* 15–19.

[38] A. Weidman. 2014. Anthropology and Voice. *Annual Review of Anthropology* 43 (October 2014), 37–51.

[39] J. Yamagishi. 2006. An Introduction to HMM-Based Speech Synthesis. Technical report, Technical report. *Tokyo Institute of Technology* (2006).

[40] Y. Zhang and B. Schuller. 2016. Towards Human-Like Holisitc Machine Perception of Speaker States and Traits. In *Proc. of the Human-Like Computing Machine Intelligence Workshop, MI20-HLC.* Springer, Windsor, U. K. 'no pagination'.