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. 1–9, (2018 April.).
DOI: https://doi.org/10.17743/jaes.2018.0023

The Perception of Vocal Traits in Synthesized Voices: Age, Gender, and Human Likeness

ALICE BAIRD[1], STINA HASSE JØRGENSEN[2], EMILIA PARADA-CABALEIRO[1], NICHOLAS CUMMINS[1], SIMONE HANTKE[1,3], AND BJÖRN SCHULLER[1,4]

[1]*ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany*
[2]*Department of Arts & Cultural Studies, University of Copenhagen, Denmark*
[3]*MISP Group, MKK Technische Universität München, Germany*
[4]*GLAM, Imperial College London, UK*

alice.baird@informatik.uni-augsburg.de

The paralinguistics of the voice are the perceived states and traits that make that voice unique to the human body from which it resonates. In many cases the synthesized voice is produced by concatenated segments of recorded human speech, a complex process that can result in an arguably lifeless voice, which lacks the ability for free-expression among other human qualities. In recent years technology-based companies are developing their own synthesized voice identities, yet seemingly paying little attention to the stereotypical traits being heard. Do such synthetic voice traits differ from the human traits they are modelled on? To explore this, the presented perception study performed by 18 listeners evaluated the paralinguistic traits of gender, age, and human likeness in the IBM voice library. Results herein have shown a similar trend to a previous study by the authors with no voice achieving complete human likeness, no voice being perceived within a single age frequency band, and none tied solidly to their given binary gender—a novel finding as commercially available synthesized voices are typically developed to operate within binary identification structures.

1 INTRODUCTION

Human Computer Interaction (HCI) is one of the largest growing sub-fields within computer science [1]. Within HCI, the number of connected devices is predicted to increase three-fold by 2025[1], making integrated voice-based technologies, e.g., voice synthesis (as well as, Automatic Speech Recognition [2], and Emotion Recognition [3, 4]) a more prevalent part of daily-life. Voice-based interactions have previously been described as "frustrating and ineffective" [5], and in recent years large financial investments have been made by companies such as Amazon, Apple, Google, and IBM to improve both the design [6], and authenticity [7] of the synthesized voice.

With the bounty of synthetic voices now audible during daily interactions (e.g., via smart-devices, from navigation assistance, or bus stop announcers), and in the age of ubiquitous computing, voice user interfaces have been said to have the ability to complement and even replace the graphical user interface [5]. In this regard user perception of such synthesized voices can affect user interaction experience [8], and for some years research into user experience of synthetic voices has begun to grow [9–11].

Aesthetic design features of such voices obtain paralinguistic features and, like other societal constructs, we can prescribe categories that may guide our attitudes, thoughts, and behaviors [5]. The paralinguistics of speech are the non-linguistic effects, which indicate traits such as gender or personality [12] and may consequently bias our choice "of whom to like, whom to trust, and with whom to do business" [5].

Given this influence, and the increasing prevalence and quality of synthesized voices, there is a need to understand how the human states and traits being provided manifest post-processing and how they affect a user's perception of "artificially" designed voices. This is not only due to the need for human likeness in such voices, but additionally due to the possible affect that propagating binary identification (of in our case age and gender) stereotypes may have. In other realms of research, the process of

[1] A 2015 Statista report shows that from 2017–2025 (in billions), connected smart devices will increase from 20.35 to 75.44 globally.

identification is discussed more openly, presenting that identity is "performative" and can alter over time [13, 14]. Additionally, it is suggested that binary stereotypes, specifically as they pertain to gender identification, may alter societal power dimensions [15].

The authors' previous study [16], questioned how listeners identify synthetic voices, asking: are listeners attuned to the vocal characteristics related to gender, age, accent-origin, and human likeness? Other studies have explored synthesized voice perception, for example by comparing emotional recorded speech to emotional synthesized speech, for storytelling ability [11]. Recent engineering developments (such as Deep Neural Network (DNN) based systems [17]) have advanced the fidelity and (arguably) the *naturalness* and *expressiveness* of such voices.

1.1 Motivation and New Approach

As well as the motivating factors that have been mentioned previously—including prevalence, human-like imitation, and the impact of aesthetic design—there were some key results from our previous study [16] which evaluated the paralinguistic traits of accent-origin, age, gender, and human likeness on the 13 synthesized voices developed for the IBM Watson Text-to-Speech library [18]. In particular, it was previously observed that a parallel exists between human likeness and both gender and age, with "less human" voices being both younger and showing alternative gender traits. Additionally, and of most interest, all the voices used (which had been given binary gender values), were shown to have alternative gender traits, and in some cases up to 24.3%, divergence from the binary categorization provided by the company.

Offering a greater sense of freedom, this perception study is motivated by the [multi'vocal] collective;[2] a collective of artists, researchers, and developers working towards developing a more inclusive identity for synthesized voices through engaging users in the corpus construction process. This collective asks: is it possible to give listeners a creative role in designing semantic audio systems, and to what extent can participation be enabled in terms of gender identification in relation to the synthesized voice?

Assessing the response of 18 listeners from differing cultural backgrounds, this study evaluates the perception of the 13 IBM synthesized voices (cf., Table 1). This time excluding accent origin, we consider listener perception, specifically in connection to notions of gender identification, as well as age and human likeness. Now taking a closer look at these paralinguistic traits, we can consider the link that may exist between gender and human likeness, highlighting the androgynous nature of the machine. The testing environment not only brings light to the current "norms" we find in daily life, it also gives participants the opportunity to engage more deeply in the design of synthesized voices and allows them to reflect on their own relationship to the synthesized voice.

[2] More information on [multi'vocal] can be found at the following URL: *http://multivocal.org*

Table 1. The voices used for this perception study and their associated traits as provided by IBM. (G)ender, (M)ale, (F)emale, (Am)erican. For the purpose of the study names have been abbreviated, e.g., `en-US_Lisa` = `US-F-1`.

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

2 THE PERCEPTION OF VOCAL TRAITS AND THE SYNTHESIZED VOICE

Paralinguistic traits of the voice are the properties that influence perception excluding the linguistic meaning. They are either short-term states (such as emotion or sleepiness), or long-term traits (such as age or gender) [12]. Such aspects of the voice, occasionally referred to as the "auditory face," may have an unconscious weighting on our perception of one another [19], which can be more prominent than visual features such as facial expression or personal style. Paraliguistics can also strongly affect human perception of essential interaction factors including "likeability" [20]. The engineers behind the synthesized voice technology have been attempting to replicate the "muscular vibrations" produced by the body for many years [21], and such human-like features could also transfer to the machines they embody, making understanding of such aspects an important factor for commercial success.

Listeners' perception of gender is said to be one of the first attributes people try to discern [22], yet gender perception can vary based on many factors including culture and social background [23]. Gender has been described as a "performative" [24] "identification" [25] of the self, which can be more flexible than the conventional binary classes of female and male [23]. The perception of gender in the human voice is a popular field of research [26–28], with some focus being given to voice synthesis [10, 11, 29]. In the case of human speech, the characteristics of maleness and femaleness in the voice have been studied for the effect that the fundamental frequency of the voice has on our perception of gender [26]. This study simulated the binary F0 (120 Hz male, and 240 Hz female) and found that there was more ambiguity perceived in the female tone placement on the gender spectrum. The gender of the synthesized voice has also been evaluated, but to a lesser extent, with some scholars showing mixed results. For example the results from [30] show that gendered voices are not categorical, whereas [9] argued that although not categorical the voices do still fall into binary stereotypes of gender.

The commercially available synthesized voices are predominately designed with binary identities[3]. The IBM library (using a Recurrent Neural Network driven acoustic concatenative synthesis system [31]) has been used for this study and is quite representative of the "binary" market, with all 13 voices being given binary gender labels (11 female and 2 male). The perception of such binary denominations of gender has been discussed, with psychological experiments [32] showing that men are more likely to trust the advice given by a female gendered machine, with the opposite being true for women.

3 METHODOLOGY

Many commercially designed synthesized voices are engineered in an endeavor to achieve natural speech [33] and to improve human interaction with the machines they embody [34]. Within humans these vocal traits can relate to personality and emotion, states which the field of artificial intelligence has been attempting to replicate through techniques such as fuzzy logic, genetic algorithms, and particle swarm optimization [35], and in the case of speech synthesis Deep Neural Network-based generative modelling [36]. The following section describes the parameters and testing environment put in place for listeners to evaluate the vocal traits of age, gender, and human likeness from a collection of synthesized voices.

3.1 The Voice Corpus

The corpus collected for this task was the IBM Watson Text-to-Speech (TTS) API, developed for the IBM Watson Developer Cloud [18][4]. The IBM Watson voices use an *expressive* TTS system [37], with *voice transformation* [18], allowing for the adaptation of many voice features including: glottal tension, speech rate, and pitch range, and the manipulation of the F0 (the fundamental frequency in the voice) for pitch rate and timbre. This has been implemented only on the American voices in the catalogue and so we have not applied it to our corpus. As well as this, post-segmentation the IBM system uses Pitch Synchronous Overlap and Add (PSOLA), a time-domain signal processing method, which modifies the pitch and speech duration between individual segments [31], giving natural flow between segments.

All 13 available voices in the IBM library, 11 female and 2 male, were selected for analysis (cf., Table 1 for futher details). As previously mentioned, we would like to note that this corpus has an unavoidable gender bias and has been chosen due to its binary attributes. Using the IBM API, 5 sentences were captured for each of the 13 available synthesized voices. Since linguistic content may bias the result (particularly for human likeness), we have chosen to only use the following "nonsense" sentences:

1. ‘ne kal ibam soud molen!’
2. ‘koun se mina lod belam’

These sentences are nonsense, and used within the GEneva Multimodal Emotion Portrayals (GEMEP) data set [38], a data set utilized for the popular 2013 Interspeech Computational Paralinguistics Challenge [39]. The average length of the audio files is 2.36 seconds. Files were originally captured in raw OGG and subsequently converted to stereo *wav* format (16-bit, 44.1 kHz) for platform compatibility.

3.2 Evaluation Parameters

As the traits of age, gender, and human likeness were shown to be of most interest in our previous study [16], the parameters being evaluated for this study have been modified to allow for a deeper analysis. In our previous study we also observed that listeners can, in most cases, identify the accent origin of the voices within this corpus and so have chosen not to re-evaluate this trait. Eighteen listeners of varying nationalities[5] evaluated the corpus. The listening task was completed in the iHEARu-PLAY online browser-based annotation platform [40], and traits were divided into three individual tasks:

- **Age:** The parameters for age have been reduced from a 10–point scale, covering ages from 0–100 years, to an 8–point scale covering from 15–55 years (each point corresponding to 5 years, i.e., 15–20, 21–25, etc.) with users restricted to a single choice. We chose this reduction as our previous study indicated that these voices were between the ages of 22–38 years old, and since we are reusing the same voice corpus, we do not expect any significant change to this result.
- **Gender:** For the task of gender identification, listeners are able to select a maximum of two options per speech instance. For each instance listeners can select a frequency bin of 20% denominations, for masculine and feminine, with the additional categorical option of Non Binary (NB) (the interface for the gender evaluation in iHEARu-PLAY is shown in Fig. 1). Rather than the culturally weighted categorical options either male or female, masculinity and femininity can be combined and relate to personality. Within this, NB can be described as a category that allows listeners to define voices as something that cannot be identified with in a binary structure [41]. The previous study considered nominal options for masculine, feminine, both or neither, and found that many of the voices were attributed values for

[3] Evaluating a selection of well-known commercially available synthesized voices, found the following to operate within a gender binary structure; Siri (Apple), Alexa (Amazon), Cortana (Microsoft), Google Assistant.

[4] Data was retrieved according to the terms of use set by IBM. All voices can be heard from the IBM service *https://www.text-to-speech-demo.mybluemix.net*.

[5] All listeners were fluent in English and the native language of listeners included 3 Danish, 4 English, 8 German, 1 Hebrew, 1 Hindi, and 1 Spanish.

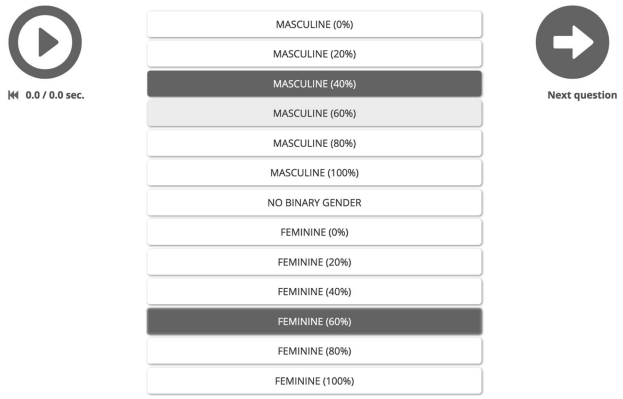

Fig. 1. The interface within the iHEARu-PLAY browser-based labelling platform for the evaluation of gender traits. Listeners could select a maximum of 2 options per instance.

gender which diverged from the binary traits provided by IBM. Due to this, for this study although binary-gender may be prominent, we give the option of identifying the synthetic voices heard outside of the binary system, and we can quantify this more accurately with the chosen paradigm.

- **Human Likeness:** As in our previous study [16], we use the term *Human-Like* from the *Uncanny Valley* [42] to describe how accurately the machine is able to mimic the human. For this study we broadened to a 6-point scale, 1 = Artificial (or Non-Human) and 6 = Human, allowing listeners to select a single value. The IBM Watson Text-To-Speech system uses a Deep Neural Network-based acoustic concatenation, accessing learned formations of text for each language and mapping this to the appropriate "chunk" (at frame level) in order to reorganize into speech. We expect that through the use of nonsense sentences in the current listening corpus, the speech may be less coherent, and listeners may be less able to rationalize the human nature of the voices.

3.3 The Listeners

The demographics of the listeners who participated was somewhat diverse (all were fluent in English and comprised of six different native tongues), their mean age was 30.6 years and of the 18 listeners 9 were male and 9 were female. As listeners were being asked to identify gender based on a *spectrum*, from 0% to 100%, we additionally asked our listeners to self-identify their own gender as a way of understanding better their own relationship to the subject of gender identification. In the same way as we laid out the gender task, listeners provided a maximum of two values for their gender identification. From this, we did find that our listeners as a group identified themselves as 44.2% masculine and 43.7% feminine. We also had one instance of 60% masculine "non-binary." Since these values may be heavily biased by cultural environment and social background [43], our listener number is not high enough to attempt to correlate this with the other part of the study. However, it does show that as a general group there is distance from the binary, which may have influence on the results and should be further analyzed on a more focused listener group.

4 EVALUATION OF TESTING PARAMETERS

Inspired by the approach made by [11] in which they asked listeners to rate binary-gender traits of synthesized voices, in this study we go beyond the binary of female/male with the inclusion of the option for non-binary gender traits. In the field of gender studies the use of these terms allows for broader identification possibilities and less constraint to rigid definitions of gender. For many years researchers have been debating the existing male and female terms, defining them as inadequate due to the cultural factors that may revolve around them, with their validity therefore becoming dependent on no cultural change [23]. For this study however, we realize that this attempt to open up the definitions of gender using feminine and masculine personality traits, may also cause yet another kind of gender stereotyping, as what is considered feminine or masculine is still culturally defined and could easily reproduce binary stereotypes.

However, we hope that by asking listeners to evaluate the voices heard with the ability to combine both gender categories as well as choose the non-binary option, we are promoting reflection over the gender categories being identified instead of promoting a reproduction of gender stereotypes. Additionally, the grading scale chosen of 20% steps could also be causing a bias, adjusting our grading scale to have the option 50 / 50 instead of only 40 / 60 may remove an unconscious weighting towards a specific gender.

There are some limitations to the listening test interface design. Due to these, we were not able to have multiple slider parameters for the gender task, and the design we finally chose is shown in Fig. 1. An important part of this task was that listeners could consider both gender traits, but we are aware this could have been complicated for listeners. In cases where listeners misunderstood the task, i.e., choosing more than 2 options totalling more than 100%, we were forced to disregard this annotator from the final results.

5 DISCUSSION OF RESULTS

The parameters evaluated by our 18 listeners—age, gender, and human likeness—have shown similarities to our previous study [16]; yet through this adapted paradigm we can see new trends as outlined here.

When evaluating the perception of age in this group of synthesized voices, the presented results confirm our previous impression. The voices are not showing binary results within frequency bins, and many voices varying in age quite substantially. For example, the youngest voice (`JP-F-1`), Mean age—M =25 years, and 95% Confidence Interval—CI from 23 to 28 years; and the oldest (`ES-M-1`) M =36 years, CI 34 − 39 years, both crossing 2 frequency bands. This distribution and other results are shown in Table 2.

For Age, when compared to our previous study, we made a more exhaustive evaluation. Focusing on the central age groups, i.e., from 15 to 55 years old in frequency bins of

Table 2. Results for the evaluated parameters of Age, Gender, and Human Likeness (HL). Age: Mean (*M*), (%) distribution across each 5-year frequency band. The values higher than 20% are highlighted in bold and a chromatic shading, mapped lowest-highest, 0–100%. Gender: as (m)asculinity, (f)emininity %, Non Binary (NB) separated. human likeness (HL): Mean (*M*) of 7-point scale, 0 = artificial, to 6 = human, converter to a %.

Name	Age									Gender			HL
%	M	15–20	21–25	26–30	31–35	36–40	41–45	46–50	51–55	**m**	**f**	**NB**	*m*
DE-F-1	36	00.0	06.3	03.1	**25.0**	**34.4**	15.6	12.5	03.1	10.5	89.5	0.00	43.2
DE-M-1	38	00.0	00.0	03.1	06.3	**43.8**	**31.3**	09.4	03.1	80.8	19.2	0.22	51.8
GB-F-1	37	00.0	00.0	06.3	18.8	**34.4**	**28.1**	06.3	06.3	20.9	79.1	0.22	34.4
US-F-1	33	00.0	06.3	09.4	**34.4**	**37.5**	06.3	06.3	00.0	10.8	89.2	0.45	42.00
US-F-2	30	03.2	19.4	**25.8**	19.4	16.1	12.9	00.0	03.2	00.9	99.1	0.00	57.14
ES-M-1	36	00.0	00.0	09.1	**31.3**	09.1	**36.4**	15.2	00.0	70.6	29.4	0.00	65.42
ES-F-1	32	06.3	00.0	**28.1**	**31.3**	12.5	09.4	06.3	03.1	20.4	79.6	0.22	46.00
LA-F-1	35	00.0	00.0	09.1	**36.4**	**30.3**	15.2	06.1	00.0	10.7	89.3	0.00	52.42
US-F-3	29	09.4	**28.1**	09.4	**21.9**	12.5	12.5	06.3	00.0	10.1	89.9	0.68	51.28
FR-F-1	26	03.1	**40.6**	**21.9**	**21.9**	09.4	03.1	00.0	00.0	10.4	89.6	0.00	64.28
IT-F-1	34	06.3	06.3	03.1	**25.0**	**34.3**	09.4	06.3	09.4	20.0	80.0	0.45	37.00
JP-F-1	25	**21.9**	18.8	**28.1**	09.4	18.8	00.0	03.1	00.0	10.0	90.0	1.36	21.00
BR-F-1	31	03.3	00.0	**20.0**	**40.0**	**30.0**	06.7	00.0	00.0	10.6	89.4	0.90	27.13

5 years, our results display that, as before, there was a tendency towards the middle age range (26 to 40 years), with other voices more clearly focused on a specific age range, e.g., FR-F-1 was identified by 40.6% of the listeners as 21−25 years old (M =26, CI 24−28 years). On the contrary, IT-F-1 presented a higher spread with most 34.3% marking age as 36−40 years, covering all frequencies, i.e., from frequency bin 15−20 years (06.3%), to frequency bin 51−55 years (09.4%) with M =34 and CI 31−37 years. This reveals that for the perceived age of these voices there is not such a clear agreement across listeners. Through the application of a chromatic scale on the age result (mapping lowest to higest result to a 0−100 color grading), cf., Table 2, we can see a clear clustering in the central age groups. Improving upon the perception of age task, we would in the future consider a continuous scale from 20−45 years, as this may offer deeper understanding on this clustered region.

Following our previous study we include the human likeness perception, i.e., the extent to which a synthesized voice is perceived as human or artificial. This time focusing on only nonsense utterances, we attempted to reduce the bias of linguistic meaning since it has been shown that native listeners are able to more accurately perceive vocal traits than non-natives [44], and in this case linguistic meaning could produce unconscious connections towards human related attributes. Our new evaluation confirms our previous results, yet displays even stronger tendencies, with ES-M-1 being identified as the most human (+5.42% from previous study) and JP-F-1 as the most synthetic (−15.00% from previous study). We hypothesize that this is a clear link to the use of nonsense utterances, particularly the ones we have chosen, "ne kal ibam sold molen" for example, which may be easier to pronounce in Spanish than any of the other languages in this library. To support this we begin to see that additionally other Romance languages in the library (excluding IT-F-1) also seem to be above 50% for human likeness. This assumption would benefit from a focused evaluation.

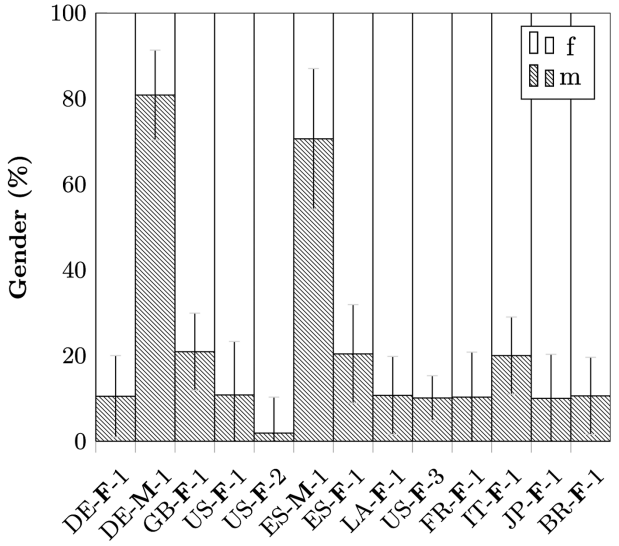

Fig. 2. Gender results in % for (m)asculinity and (f)emininity, and ± standard deviation for each voice. The given binary-gender is shown in the voice name, as either (F)emale or (M)ale.

In our previous evaluation there was a tendency towards the identification of synthetic voices as having both gender (feminine and masculine) traits. Nevertheless, our previous forced choice test could lead listeners into a binomial answer. For this reason, in our new test we chose to ask the listener a specific percentage for the traits of gender as well as providing the categorical option of non-binary. Interesting results have risen from this new evaluation, showing that even though this group of synthetic voices was originally given a binary-gender (i.e., male or female), the perception of listeners does not obviously mirror this intended attribute (cf., Fig. 2 for visual detail). On the contrary, our results display that with these revised testing parameters none of the voices can be concretely identified as having a binary gender, even those that have been given higher gender traits, i.e., US-F-2, still presents 0.94% masculinity,

and the most masculine DE-M-1 has 10.2% feminine traits. Furthermore ES-M-1, the most gender-ambiguously identified voice, has been perceived as 29.4% feminine, which could open new research questions of gender-ambiguity in relation to voice naturalness. Although given the unbalanced nature of our data set (2 of 13 male) it could be speculated that repeated listening to feminine voices introduces a bias towards pitch [45], and in future testing a more balanced data set would be optimal.

When asking the listeners about their perception of non-binary gender identities for the voices, i.e., if they perceive the voices as not belonging to a specific binary gender, the listeners have identified JP-F-1 as the most non-binary voice (1.36%); a voice that has also been identified as the least human (2.52%). The result could suggest that voices identified as non-binary by the listeners are those with a prominent artificial nature and evoke in the listener non-anthropomorphic attributes. As well as being those which technically struggle to produce the sentence structure required (as model training is based on corpora of native pronunciation), and as mentioned, JP-F-1 is not within the family of Romance languages, which are potentially a closer family to the nonsense sentence used here, meaning that the JP-F-1 available corpus may not have the ability to construct such a sentence.

We consider this result promising since it supports the idea that synthesized (artificial) voices much like the human voice [46] may not align with the conventional binary gender stereotypes, particularly those with lower human-like fidelity. Thus implying that newer techniques for synthesis (i.e., DNN-based), if improving the corpus size in consideration to diverse language pronunciation, may better retain their human-like attributes. One element of the machine voice that has been discussed previously is that it does not experience the same socialization as the human voice, e.g., socializing predominately with males and learning from others in the group [11]. On the other side, the variability of the identification of gender traits in these synthesized voices could more or less mirror the same tendencies that have been found in the human voice towards masculine or feminine identification [47].

When evaluating the results of this test we did not take the cultural background of the listeners into consideration. Our initial study [16] showed that listeners' cultural-background did not play a role when evaluating the paralinguistic traits of this group of synthesized voices. This aligns with previous research on cross-cultural perception of emotion in speech, of which the outcomes have shown that even though native speakers identify with more accurately emotional speech pronounced in their own languages than non-native speakers [44], the level of agreement between listeners from different cultures remains higher than chance [48].

6 CONCLUSIONS AND POSSIBLE DIRECTIONS

For this study we evaluated the responses of 18 listeners, asking them to consider the paralinguistic traits of age, gender, and human likeness from the 13 voices in the IBM Watson corpus. The results here have confirmed many of the assumptions that we had as a result of our previous study [16]. Both age and human likeness, for example, were consistent with our previous study—both the oldest (most-human) and youngest (least-human) voices were the same (ES-M-1, and JP-F-2, respectively). The given gender traits provided by IBM have shown to be too concrete and the voices cannot, in most cases, be allocated a dominant age. The addition of the non binary has shown some results that encourage further study, i.e., voices that have been given higher NB values, also showed to have lower human likeness values. This result seems as though listeners may have confused their understanding of NB, and from our testing paradigm may have incorrectly assumed that non-binary voices would also not be human. Since ultimately non binary voices are human, one possible future direction could include asking listeners to evaluate synthetic voices only for gender, or providing more alternative options in order to avoid creating this association.

These results call for a much larger scale study, increasing listener numbers and broadening the corpus to improve the overall meaningful nature of the results herein. Additionally, as we have seen results that link human likeness with the corpus utilized for synthesis, adapting the testing data may also give fruitful results. Developing new prototypes for synthetic voices, e.g., by combining the alternative vocal features found to be both masculine and feminine, or by using comparable conventional methods for speech synthesis against the state-of-the-art (i.e., Hidden Markov Models [49], against WaveNet [36]), future studies could also go much deeper into the relationship of non-binary identification, and human likeness, evaluating if synthesized voices can be perceived as gender-ambiguous, i.e., if they are not so easily categorized as either male or female. In summary, through this evaluation, there are many points that can be extended upon and which clearly deserve further attention due to synthesized voice prevalence and the identity stereotypes we are currently being presented with. An additional and potentially novel finding from this perception study is that we see that the perception of synthesized voices is not a binary decision, much like the human identification process of which IBM has classified these voices. Future studies should focus on other commercially available voices, i.e., Apple's Siri and Amazon's Alexa, as they are also operating within a binary identification framework.

7 ACKNOWLEDGMENT

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 agreement No. 338164 (ERC StG iHEARu).

8 REFERENCES

[1] O. J. Øye, "HCI a Growing Field" (2016), http://bit.ly/2z8lZzg.

[2] S. J. Arora and R. P. Singh, "Automatic Speech Recognition: A Review," *Intl. J. Comp. App.*, vol. 60, no. 9, pp. 34–44 (2012 Dec.), [Online]. Available: 10.1007/1-4020-2673-0_3.

[3] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from the First Challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062–1087 (2011), [Online]. Available: 10.1016/j.specom.2011.01.011.

[4] K. R. Scherer, "Vocal Markers of Emotion: Comparing Induction and Acting Elicitation," *Computer Speech & Language.*, vol. 27, no. 1, pp. 40–58 (2013), [Online]. Available: 10.1016/j.csl.2011.11.003.

[5] C. Nass, S. Brave, *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*, 1st ed. (MIT Press, Cambridge, MA, USA, 2005).

[6] T. J. Seppala, "Amazon's Redesigned Echo Features Improved Sound, Alexa Smarts" (2017), URL http://engt.co/2ytQ73t.

[7] A. Robertson, "Google's DeepMind AI Fakes Some of the Most Realistic Human Voices Yet" (2016), URL http://bit.ly/2hDhsxf.

[8] R. Tamagawa, C. Watson, I. Han Kuo, E. MacDonald, and B. Broadbent, "The Effects of Synthesized Voice Accents on User Perceptions of Robots," *Intl. J. Social Robotics*, vol. 3, no. 3, pp. 253–262 (2011), [Online]. Available: 10.1007/s12369-011-0100-4.

[9] C. Nass, Y. Moon, and N. Green, "Are Machines Gender Neutral? Gender-Stereotypic Responses to Computers With Voices," *J. Applied Soc. Psych.*, vol. 27, no. 10, pp. 864–876 (1997), [Online]. Available: 0.1111/j.1559-1816.1997.tb00275.x.

[10] E. J. Lee, C. Nass, and S. Brave, "Can Computer-Generated Speech Have Gender? An Experimental Test of Gender Stereotype," *Conference on Human Factors in Computing Systems (CHI), The Hague, The Netherlands*, pp. 289–290 (2000), [Online]. Available: 10.1145/633292.633461.

[11] C. Nass, U. Foehr, S. Brave, and M. Somoza, "The Effects of Emotion of Voice in Synthesized and Recorded Speech," Association for the Advancement of Artificial Intelligence (AAAI) Technical Report (2001).

[12] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*, 1st ed. (Wiley, Chichester, UK, 2013).

[13] E. Sedgwick, *Tendencies* (Duke University Press, Durham, NC, USA, 1993).

[14] J. E. Muñoz, *Disidentifcations: Queers of Color and the Performance of Politics* (University of Minnesota Press, Minneapolis, MN, USA, 1998).

[15] J. Butler, *Bodies That Matter: On the Discursive Limits of "Sex"* (Roudedge, London, UK, 1993).

[16] A. Baird, S. H. Jørgensen, E. Parada-Cabaleiro, S. Hantke, N. Cummins, and B. Schuller, "Perception of Paralinguistic Traits in Synthesized Voices," *Proceedings of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences*, AM '17, pp. 1–5 (2017), [Online]. Available: 10.1145/3123514.3123528.

[17] H. Zen, A. Senior, and M. Schuster, "Statistical Parametric Speech Synthesis Using Deep Neural Networks," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 7962–7966 (2013), [Online]. Available: 10.1109/ICASSP.2013.6639215.

[18] IBM® Watson Developer Cloud, "Text to Speech" (2017), URL https://ibm.co/2sn8p7a.

[19] P. Belin, S. Fecteau, and C. Bedard, "Thinking the Voice: Neural Correlates of Voice Perception," *J. Trends in Cognitive Sciences*, vol. 8, no. 3, pp. 129–135 (2004), [Online]. Available: 10.1016/j.tics.2004.01.008.

[20] B. Schuller, S. Steidl, A. Batliner, E. Noth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "A Survey on Perceived Speaker Traits Personality, Likability, Pathology, and the First Challenge," *J. Computer Speech & Language*, vol. 29, no. 1, pp. 100–131 (2015), [Online]. Available: 10.1016/j.csl.2014.08.003.

[21] H. Dudley, "Fundamentals of Speech Synthesis," *Journal of Audio Engineering Society*, vol. 3, no. 4, pp. 170–185 (1955 Oct.).

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

[23] K. Franck and E. Rosen, "A Projective Test of Masculinity-Femininity," *J. Consulting Psychology*, vol. 13, no. 4, pp. 247–256 (1949).

[24] J. Butler, *Gender Trouble* (Routledge, Oxon, UK, 1990).

[25] A. Jones, *Seeing Differently* (Routledge, Oxon, UK, 2012).

[26] R. O. Coleman, "A Comparison of the Contributions of Two Voice Quality Characteristics to the Perception of Maleness and Femaleness in the Voice," *J. Speech and Hearing Res.*, vol. 19, no. 11, pp. 168–180 (1976).

[27] M. Latinus and P. Belin, "Human Voice Perception," *Current Biology*, vol. 21, no. 4, pp. 143–145 (2011), [Online]. Available: 10.1016/j.cub.2010.12.033.

[28] K. Johnson, E. A. Strand, and M. D'Imperio, "Auditory Visual Integration of Talker Gender in Vowel Perception," *J. Phonetics*, vol. 27, no. 4, pp. 359–384 (1999), [Online]. Available: 10.1006/jpho.1999.010.

[29] T. Phan, "The Materiality of the Digital and the Gendered Voice of Siri," *Transformations*, vol. 1, no. 29, pp. 23–33 (2017).

[30] J. W. Mullennix, K. A. Johnson, M. T. Durgun, and L. M. Farnsworth, "The Perceptual Representation of Voice Gender," *J. Acoust. Soc. Amer.*, vol. 98, no. 6, pp. 3080–3095 (1995).

[31] IBM Watson Developer Cloud, "The Science Behind the Service" (2017), URL https://ibm.co/2jCyEDE.

[32] E. Lee, "Effects of 'Gender' of the Computer on Informational Social Influence: The Moderating Role of Task Type," *Intl. J. Human-Computer Studies*, vol. 58, no. 4, pp. 347–362 (2003), [Online]. Available: 10.1016/S1071-5819(03)00009-0.

[33] P. Alku, H. Tiitinen, and R. Natanen, "A Method for Generating Natural-Sounding Speech Stimuli for Cognitive Brain Research," *Clinical Neurophysiology*, vol. 110, no. 8, pp. 1329–1333 (1999), [Online]. Available: 10.1016/S1388-2457(99)00088-7.

[34] H. F. Olson, H. Belar, and E. S. Rogers, "Research Towards a High Efficiency Voice Communication System," *Journal of Audio Engineering Society*, vol. 14, no. 3, pp. 233–239 (1966 July).

[35] T. Sutikno, M. Facta, and G. A. Markadeh, "Progress in Artificial Intelligence Techniques: From Brain to Emotion," *J. Telecomm. Computing Elec. and Control*, vol. 9, no. 2, pp. 201–202 (2010).

[36] A. V. D. 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*, p. 1609:03499 (2016).

[37] J. F. Pitrelli, R. Bakis, E. M. Eide, R. Fernandez, W. Hamza, and M. A. Picheny, "The IBM Expressive Text-to-Speech Synthesis System for American English," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1099–1108 (2006 July), [Online]. Available: 10.1109/TASL.2006.876123.

[38] T. Bänziger, H. Pirker, and K. Scherer, "GEMEP-GEneva Multimodal Emotion Portrayals: A Corpus for the Study of Multimodal Emotional Expressions," *Proceedings Language Resources and Evaluation*, pp. 15–19 (2006).

[39] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valence, and S. Kim, "The Interspeech 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," *Proceeding of the International Speech Communication Association Conference (INTERSPEECH)*, pp. 148–152 (2013).

[40] S. Hantke, F. Eyben, T. Appel, and B. Schuller, "iHEARu-PLAY: Introducing a Game for Crowdsourced Data Collection for Affective Computing," *Proceedings of the 1st International Wireless Algorithms, Systems, and Applications (WASA) 2015*, pp. 891–897 (2015), [Online]. Available: 10.1109/ACII.2015.7344680.

[41] M. Sycamore, *Nobody Passes: Rejecting the Rules of Gender and Conformity* (Seal Press, Emeryville, CA, USA, 2006).

[42] M. Mori, "Bukimi no Tani [The Uncanny Valley]," *Energy*, vol. 7, no. 4, pp. 33–35 (1970).

[43] Y. Kashima, S. Yamaguchi, U. Kim, S. Choi, M. Gelfand, M. Yuki, "Culture, Gender, and Self: A Perspective from Individualism-Collectivism Research," *J. Personality and Social Psychology*, pp. 925–937 (1995). [Online]. Available: 10.1037/0022-3514.69.5.925.

[44] P. Laukka, *Vocal Expression of Emotion: Discrete Emotions and Dimensional Accounts*, Ph.D. thesis, University of Uppsala (2004).

[45] K. T. Hill and L. M. Miller, "Auditory Attentional Control and Selection during Cocktail Party Listening," *Cerebral Cortex*, vol. 20, no. 3, pp. 583–590 (2010), [Online]. Available: 10.1093/cercor/bhp124.

[46] D. H. Klatt and L. C. Klatt, "Analysis, Synthesis, and Perception of Voice Quality Variations Among Female and Male Talkers," *J Acoust. Soc. Amer.*, vol. 87, no. 2, pp. 820–857 (1990), [Online]. Available: 10.1121/1.398894.

[47] S. Lattner, M. Meyer, and A. Friederici, "Voice Perception: Sex, Pitch, and the Right Hemisphere," *J. Human Brain Mapping*, vol. 24, pp. 11–20 (2004), [Online]. Available: 10.1002/hbm.20065.

[48] K. R. Scherer, R. Banse, and H. G. Wallbott, "Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures," *J. Cross-Cultural Psychology*, vol. 32, pp. 76–92 (2001), [Online]. Available: 10.1177/0022022101032001009.

[49] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech Synthesis Based on Hidden Markov Models," *J. Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252 (2013), [Online]. Available: 10.1109/JPROC.2013.2251852.

THE AUTHORS


Alice Baird


Stina Hasse Jørgensen


Emilia Parada-Cabaleiro


Nicholas Cummins


Simone Hantke


Björn Schuller

Alice is a research assistant for the ZD.B Chair of Embedded Intelligence for Healthcare and Wellbeing, University of Augsburg, Germany, where she is involved with the Horizon 2020 project DE-ENIGMA, analysis of vocal and linguistic cues. Alice as recently been awarded a ZD.B Ph.D Fellowship (2018–2021), in which she will research speech monitoring and synthetic soundscape generation. Alice has an MFA in Sound Arts from Columbia Universities, Computer Music Center, and a BA in Music Technology from London Metropolitan University. Alice works across an array of disciplines, predominately in the realm of paralinguistic speech and intelligent audio analysis. Focusing her research efforts towards applications of computing for health and wellbeing, with consideration to novel "in the wild" data collection methodologies.

•

Stina Hasse Jørgensen is a Ph.D. student at the Department of Arts and Cultural Studies, University of Copenhagen, where she is focusing on the politics and aesthetics of synthesized voices through practice-based research. Her post-doctoral research at the IT University of Copenhagen will begin late spring 2018. Stina has published articles on art, design, technology, and sound in journals such as *Digital Creativity*, *Transformations Journal*, *Body, Space & Technology Journa*l, and *Cultural Analysis Journal*, and has presented work at conferences such as ISACS, TERRAEN, ISEA, Aesthetics, Ethics and Biopolitics of the Posthuman, Global Lives Project - UC Berkeley, Re-New IMAC, and NORDIK.

•

Emilia Parada-Cabaleiro is a post-doctoral researcher at the University of Augsburg (Germany), where is involved in the EU-FP7 ERC project iHEARu. Her main research interests lay in the intersection between, psychology, music, and technology, mainly involved in the study of human perception of emotions in music and speech. She received a Ph.D. in history, science and techniques of music from the University of Rome Tor Vergata (2016). She has a Masters degree in musicology from the Higher Conservatory of Vigo (Spain), a Bachelor degree in music education (University of Vigo), a Master degree in music management (St. Cecilia Conservatory), a Post-Graduate in sonic arts (Tor Vergata University), and a Professional diploma in piano performance and in music therapy. She has working experience in music archive-keeping data management, musical pedagogy, and has been involved in projects related to music and emotions in clinical environments.

•

Dr. Nicholas Cummins received his Ph.D. in electrical engineering from UNSW Australia in February 2016. He did his undergraduate BE degree at UNSW, graduating with first class honors in 2011. Currently, he is a postdoctoral researcher at the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing at the University of Augsburg, Germany, where he is involved in the Horizon 2020 projects DE-ENIGMA, RADAR-CNS, and TAPAS. His current research interests include areas of affective computing and computer audition with a particular focus on the understanding and analysis of different health states. He has (co)authored over 40 conference and journal papers (374 citations, h-index 9). Dr. Cummins is a reviewer for IEEE, ACM, and ISCA journals and conferences as well as serving on program and organizational committees. He is a member of ISCA, IEEE, and the IET.

•

Simone Hantke received her Diploma in media technology in 2011 from the Technische Hochschule Deggendorf and her Master of Science in 2014 from the Technische Universität München, one of three of Germany's Excellence Universities. She currently is a PhD Candidate at the Institute for Human-Machine Communication at the Technische Universität München and working in the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing at University Augsburg. She is working on her doctoral thesis in the field of affective computing and speech recognition, focusing her research on data collection and new machine learning approaches for robust automatic speech recognition and speaker characterization. Her main area of involvement has been with the EU-FP7 ERC project iHEARu. In the scope of this project she leads the development of crowdsourcing data collection and annotation for speech processing and is the lead author of iHEARu-PLAY.

•

Björn Schuller received his diploma in 1999, his doctoral degree for his study on automatic speech and emotion recognition in 2006, and his habilitation (fakultas docendi) and was entitled Adjunct Teaching Professor (venia legendi) in the subject area of signal processing and machine intelligence for his work on intelligent audio analysis in 2012 all in electrical engineering and information technology from TUM in Munich, Germany. Since 2017, he is Full Professor and ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing at Augsburg University, Germany, and Centre Digitisation.Bavaria (ZD.B) in Garching, Germany. At the same time, he is a Reader (Associate Professor) in machine learning in the Department of Computing at Imperial College London, UK, since 2015 where he heads the Group on Language Audio & Music (GLAM), previously being a Senior Lecturer since 2013.