13th edition of the Speech Synthesis Workshop
24-27 August 2025, Leeuwarden, The Netherlands

What is naturalness?

Ayushi Pandey[1], *Sébastien Le Maguer*[2], *Naomi Harte*[1]

[1]Sigmedia Lab, ADAPT Centre, School of Engineering, Trinity College Dublin, Ireland
[2]Univ. Helsinki, Finland

pandeya@tcd.ie, sebastien.lemaguer@helsinki.fi, nharte@tcd.ie

Abstract

Naturalness in Text-To-Speech (TTS) synthesizers is among the most widely evaluated aspect of TTS synthesizers. Despite the popularity, it has consistently been identified as a "nebulous" and "poorly defined concept, left to a listener's subjective interpretation of the term. Without a proper definition, researchers either continue to promote under-informative evaluation designs, or argue in favour of rendering the term obsolete.

As better methods of evaluation are being standardized, this paper presents a discussion around the definition of naturalness. Specifically, we describe naturalness as a multi-faceted perceptual attribute. While listener interpretation of the term naturalness has been covered in the previous literature, this paper serves to present a top-down approach. We enlist the perspectives on naturalness, as viewed by different practitioners of TTS or synthetic voices. First, we discuss why human-likeness is a desirable and important target in the development of speech synthesizers. We categorize the scope of naturalness within human-likeness along its use-cases. We next describe how a standalone understanding of human-likeness is not sufficient. We therefore provide an explanation of naturalness as appropriateness. The aim of this paper is to open a discussion around the meaning of naturalness, so that clear directions for its evaluations can be established.

Index Terms: naturalness, appropriateness, human-likeness, text-to-speech synthesis, evaluation of text-to-speech

1. Introduction

Among the statements that can be made with some certainty, we know that naturalness is the most widely tested perceptual attribute in TTS synthesizers. The question of naturalness is presented usually as "choose a score for how natural or unnatural the sentence sounded" in the Blizzard Challenge series [1]. While the question looks simple, the concept of naturalness itself has been considered "nebulous" [2], or "poorly defined" [1].

To address the ambiguity of the term, several research studies have argued in favour of suspending the term naturalness in favour of *appropriateness* [3]. The most recent edition of the Blizzard Challenge [4] has replaced naturalness with evaluating overall quality, as non-experts find it easier to understand. Despite this resistance, naturalness remains predominant in the literature as illustrated by a recent survey [5] showing that more than 50% of published TTS research continue to evaluate systems based on perceived naturalness. This can be explained by the fact that new TTS architectures continue to be released, and are expected to outperform previously established baselines on commonly evaluated dimensions. It is therefore critical to now question what the concept of "naturalness" entails. This can be achieved in two directions. The first direction, that we qualifies as "bottom-up", consists of exploring how listeners interpret naturalness. Shirali-Shareza et al.[6] have initiated investigating this direction by proposing an evaluation protocol, where *the listeners* are requested to provide what they understand by the naturalness of synthetic speech.

In this paper we propose to investigate the second direction which consists of exploring how researchers of different fields related to speech science and speech technology have understood naturalness. we present a top-down perspective on the concept of naturalness in synthetic speech. Specifically, we explain that **naturalness is a multi-faceted perceptual attribute**.

Human-likeness is a major component of TTS naturalness, and has diverse applications. This is discussed in Section 3. However, for many applications, human-likeness in TTS neither enough, nor suitable. In such cases, it is more important for the voice to be *appropriate* to the conversation, and match the overall expectations that a user has from the conversation. This is described in Section 4. Emerging from Sections 3 and 4 is a visual description of naturalness in Figure 1, that demonstrates its multi-faceted nature. Section 5 describes a set of general guidelines for diverse practitioners of synthetic voices. By no means is this review exhaustive, but we aim to *open* a discussion around the concept of naturalness.

2. Historical perspectives on naturalness

The use of the word "naturalness" first appears in Stewart's documentation [7] of an electrical speech synthesizer, referenced in [8]. In this documentation, naturalness is interpreted as both the quality and intelligibility of human speech. A semantic differentiation scaling analysis [9] revealed that "naturalness" and "intelligibility" were the two components of perceived quality of speech. This lead to dissociate the evaluation of naturalness and the evaluation of intelligibility. The category judgment method, or the Absolute Category Rating (ACR) method had already received C.C.I.T.T standardization for speech quality assessment in 1962. To the best of our knowledge, some of the first applications of a mean opinion score in speech synthesis, comes from Richards et al.'s [10] design of the opinion assessment score on listening effort.

Some of the first perspectives specific to evaluating naturalness using the category judgement can be seen in the late 20th century. In the mid-80s, Pols et al. [11] describe their participation in the international ESPRIT SPIN project, which involved the development and testing of the office automation system using a speech interface. In this report, we note a few interesting points, which identify some important trends in speech synthesis. First, they make the assertion that quality and acceptability became the most important attribute of speech signals,

10.21437/SSW.2025-33

and that systems already could provide high levels of intelligibility. Next, they propose that a MOS-based evaluation of speech synthesizers' quality, can be expressed using a multidimensional format - including criteria such as naturalness, preference, acceptability. A reference to Nusbaum et al.'s [12] report is made, so it seems likely that *they* were the first to implement this. However, the exact document is not available through online/library searches. Therefore, naturalness evaluation through a Mean Opinion Score (MOS)-based listening test was first conducted in the mid-80s.

The subsequent years have seen active usage of MOS in assessing the quality of speech synthesizers; ranging from unit-selection synthesizers, Hidden Markov Model (HMM)-based synthesizers, and finally the most modern neural synthesizers. Their use has been extended to evaluation of synthetic speech in several languages [13], multiple contexts, such as interactive avatar or wizard-of-oz settings [14], crowd-sourced settings [15], as well as multi-modal speech synthesis [16].

As we have seen in this section, the conceptualisation of "naturalness" has evolved in parallel to the evolution of the quality of the speech produced by the different generation of synthesizers. Despite this continuous evolution, we have also seen that a clear distinction is now drawn between naturalness being considered as human likeness and what speech synthesis researchers recommend to focus on: appropriateness. While naturalness is seen as disjoint from appropriateness, we will show in the rest of our paper that naturalness encompasses both human-likeness and appropriateness. A visual representation of this is presented in fig. 1.

The next two sections are dedicated to explore each of these facets and

3. Naturalness as human-likeness

In a select set of applications, the concept of human-likeness is closely linked with naturalness. In covering the literature, we could identify three broad categories of reasons why human-likeness is a critical goal for synthetic voices.

3.1. Why should a synthesized voice sound like a human's?

First, attributing human-likeness, or anthropomorphism has historically been an important aspect of automata design [17]. Epley et al. [18] argue that human values are also attributed to non-human devices because of the compelling desire for human beings to form social bonds, especially in absence of human connections. This means that the human-likeness of automata is important for **social integration** of intelligent devices. Increasing the human-likeness of automata has been shown to increase trust [19] and error-tolerance [20] among human users. Specific to the voice of automata, an important finding comes from Schroeder et al. [21], where they specifically tested the effect of a human voice on assigning human authorship to written text. Participants were asked to determine whether a given text was written by a human or AI. Text stimuli were presented to them in 4 conditions: a) text (display only text), b) audio (listen to the human speaker reading out the text), c) subtitled video (watch an actor read the text with subtitles but no audio), and d) all modalities combined (audio, visual and text). While text was generated in all cases by a text generator, a participant was found most likely to determine that the text was composed by a human, if the participant listened to the human voice. Further, they showed that the human voice, especially one with rich intonational variation can "uniquely humanize" an interaction, even more effectively than the audiovisual medium. The authors argue that the voice communicates the presence of a creative mind more effectively, similar to biological movements indicate presence of living beings. This work highlights the importance of the human-likeness in synthetic voices, as a primary need for human communication.

Human-like voice serves their classic purpose for **advancing assistive technologies and healthcare**. Text-to-speech is a crucial application for learners with difficulties, especially to promote confidence and aid in destigmatizing their educational instruction. Dyslexic children have been shown to find the human voice more intelligible [22] and has improved their comprehension on quizzes [23]. Then, usability testing protocols for assistive devices developed for children with cerebral palsy also specify the requirement for a TTS to be "natural and human sounding" [24]. Developers of assistive and augmented technologies in Nepali identify that an "ideal" synthesizer must be able to "replicate speech as a human" [25]. Through these studies, we conclude that TTS can serve as a highly enabling assistive technology.

Finally, human-like voices are required for **advancing research in speech science**. Malisz et al. [26] argue that modern, neural synthetic speech can be a versatile tool for modelling human speech. Although many researchers insist on using naturally occurring corpora [27], the pre-processing of such corpora may be time-consuming as the data is often designed for other purposes. However, synthesizing human-like speech can either augment existing corpora, or facilitate the creation of suitable corpora. Finally, synthesizing speech can also aid in language preservation and documentation [28].

Through this discussion, we show that human-likeness is an important target for TTS synthesizers. For targeted applications of TTS, such as healthcare, social integration of voice-based assistants (VBA)s, naturalness can be synonymous with human-likeness. The next section describes the techniques applied by TTS practitioners to enhance and maximise naturalness.

3.2. What do synthesizers do to maximise human-likeness?

Developers of TTS synthesizers have long pursued the idea of naturalness as an important component of speech quality. A semantic differentiation scaling analysis [9] revealed that "naturalness", and "intelligibility" were the two components of perceived quality of speech. The first commercial synthesizers [29] were built on parallel arrangement of resonators whose inverse glottal filtering design [30] was reported to be perceptually "indistinguishable" from the human voice. Concurrently, concatenative synthesizers were maximising naturalness through mid-point concatenation of phonemes [31], joining together linear prediction coefficients into independent parameter sequences. Naturalness improved with unit-selection techniques [32], reducing selection and concatenation costs, while post-processing methods like PSOLA, pitch, and duration manipulation [33] further enhanced synthesizer quality.

In the present day, neural techniques claim to achieve naturalness of levels that have "never been reported". Where autoregressive architectures like WaveNet [34] initiated this shift, their non-autoregressive counterparts have achieved efficient parallelization in addition to naturalness. Recent advancements in diffusion-based and VAE-based TTS methods have further pushed the boundaries of naturalness and expressivity. Diffusion models iteratively refine noisy inputs, leading to higher fidelity, richer prosody, and natural cadence, making speech more human-like [35]. On the other hand, VAE-based approaches

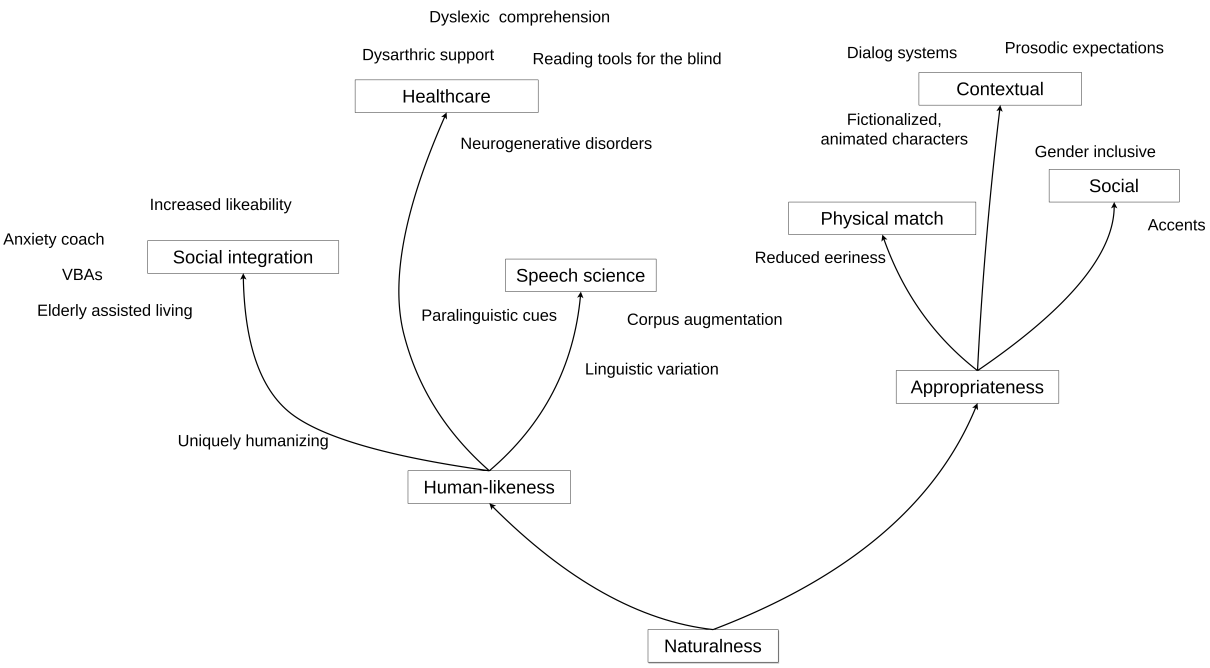

Figure 1: *What is naturalness? Naturalness of Text-to-Speech synthesizers a multi-component perceptual attribute, and can be viewed along both human-likeness and appropriateness. Terminal nodes show the broad applications that are more relevant to each sub-component, for example the importance of human-like speech is seen more in healthcare and assistive technologies. The surrounding text shows applications of each sub-component.*

provide better controllability and smooth latent representations, enabling finer manipulation of expressiveness while maintaining efficiency [36]. While diffusion models specialise in realistic waveforms, VAEs are better matched for prosody control and real-time applications.

Improvements in these architectures particularly revolve around enhanced expressivity, intonation patterns and fidelity, and latency - all of which are human-like characteristics in speech. This means, that on the engineering and development front, efforts are largely focused towards enhancing the human-like attributes of the synthetic voice. At this point, commercial use can be seen bootstrapping the standalone quality of human-likeness, but also situated awareness of the context.

3.3. How is human-likeness evaluated in TTS?

Evaluation techniques can be categorized in three broad approaches: subjective evaluations, objective evaluations and behavioural evaluations. Subjective evaluation is the method of obtaining feedback directly from the consumer of the voice. The most dominant method for collecting opinion is through a subjective listening test, and the opinion is captured most often through the MOS or the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) scores. The Blizzard Challenge series provides a comprehensive evaluation platform. Traditionally, the word “natural” has been implicitly used to refer to the human voice. For example, “..to find out how far our synthesizers are from natural speech.” [37], and “..*A* denoting natural speech” [38], where *A* referred to human speech. Next, objective evaluations are aimed at predicting user opinion so that the dependence on the time-consuming and expensive subjective evaluations can be diminished. Objective methods can be full-reference or non-intrusive, dependent on the application. In the full-reference method, synthetic speech is measured against human speech as a reference, using automatic, standardized measurements like the Perceptual Evaluation of Speech Quality (PESQ) [39], its successors [40] and competitors [41]. Conversely, the reference signal is discarded in non-intrusive methods [42]. Finally, behavioural metrics suspend user opinion, and collect feedback through physiological measures like pupil dilation, heart rate monitoring, neuronal activity detection [43].

Using synthetic speech as a research tool has attracted many phoneticians. For example, some researchers synthesize variation in speech, in terms of disfluencies, or pause locations [44] to understand perception of paralinguistic personality traits. Similarly, architectures have been designed to support fine-grained manipulation of low-level phonetic features [45]. Also, achieving prosodic control for speaker and style transfer [46], and cuing non-explicit pragmatic functions [47] is quite popular. In these studies, the evaluation of a TTS synthesizer is achieved rather indirectly, wherein its limits are questioned to generate nuanced speech phenomena.

By comparison, using techniques in phonetics directly for TTS evaluation has attracted modest attention. Specifically, Gutierrez et al. [48] incorporate the well-established Rapid Prosody Transcription paradigm [49] for obtaining locations of perceived prosodic errors. Additionally, Gessinger et al. [50] report differential effects of TTS techniques on phonetic entrainment patterns. Similarly, Pandey et al. [51] present a direct comparison between the human and synthetic voice, using fine-grained acoustic-phonetic features.

4. Naturalness as appropriateness

4.1. Why should synthesized speech be appropriate?

Although approaching human-likeness caters to several applications, some studies assert the importance of *appropriateness* in synthetic speech. Wagner et al. [2] assert that there is no "optimal" definition for quality. To address this, they recommend that application-specific and user-governed evaluation of synthesized speech must be preferred.

We categorize appropriateness into three aspects: contextual, physical and social, as shown in Figure 1. Physical appropriateness refers to a correspondence between the physical features and its voice of a robot or an agent in a multimodal conversational setting. Finally, social contexts address the expectations from the agent in specific social settings.

Contextual appropriateness encompasses those applications, where synthesized speech feels like a natural part of a larger discourse, or dialogue. First, read speech, which often serves as training data for several TTS designs is not considered appropriate in spontaneous conversational agents. For example, the MOS score for naturalness of utterances was found to vary with context-dependent instructions [52]. Similarly, ratings can fluctuate on the basis of length of the stimulus presented, where presenting longer paragraphs produces higher cognitive loads [3]. Other investigations have explored prosodic expectations of context on synthetic speech. For example, Gutierrez et al. [48] showed that sentential context (e.g. "Who ate the cake?", "MARY ate the cake.", as opposed to "Mary ate the CAKE") influences listener expectations, and that TTS voices trained on read speech alone cannot suffice. Similarly, participants in [53] used non-lexical, or discourse cues to determine those samples which prosodically matched their preceding context. Another example of contextual appropriateness comes from a corpus study on fictional characters in [54]. The authors show that voices of these characters revealed systematic differences from the human voice with respect to different personality types (e.g, good vs evil), and argue that the voice should fit the overall *narrative* context. In short, speech with artefacts was sometimes more important for some character personas and stories instead of human-likeness of the voice.

The next categorisation is made along **physical appropriateness**, where the synthesized voice must match the physical characters of the conversational agent. Particularly, [55] use a multimodal video analysis and show that a mismatch in physical features with voice can cause "eeriness". Similarly, McGinn et al. [56] found that the human voice was preferred for robots with human-like features (Nao, Stevie), while a synthetic voice was preferred for the more functional looking robots. Then, participants in Mara et al. [57] drew human-like facial features when they heard a human-like voice, and mechanical features when they heard a mechanical one. These observations appear robust across other languages, as demonstrated in Trovato et al. for Brazilian Portuguese [58]. They found that the human voice was not considered appropriate for a humanoid robot, especially when compared to a more anthropomorphic conversational agent. These findings suggest that a plain approach to human-likeness of synthetic speech is limited in an understanding of naturalness, especially in a multi-modal interaction setting.

Finally, **social appropriateness** is also important for a conversation to be natural. First, a conditional, task-specific preference for robotic voices has been seen when the task is more functional. Torre et al. [59] found that native speakers prefer an accent in robots which is already considered prestigious and trustworthy in human speech. Similarly, participants from New Zealand showed a selective preference for synthetic speech in their native accent, especially when the task involved a real-world interaction requiring trust [60]. Additionally, persuasive effects of native and non-native accents have been found to vary with different therapeutic approaches [61]. Other aspects of social appropriateness have been associated with expectations of gender. Through a comparative analysis of several results on voice and conversational agents, Seaborn et al. observe that gender influences the perception of a robotic voice "in line with stereotypes" [62]. For example, Nass et al. [63] report that despite explicitly removing all gender-specific information from the interaction, the perceptual attributes was stereotypical to gender roles. Gender was also reported to interact with personality features for specific applications, such as extroverted feminine voices were better matched for healthcare agents [64]. Even though the voice quality sounded non-human in older synthesizers, participants responded appropriately to gender cues [65]. Furthermore, methods of generating gender-neutral voices are proposed [66]. Although rated less natural than conventional voices [67], Danielescu et al. [66] argue that a gender-neutral voice caters to a broader spectrum of non-binary users.

4.2. How is appropriateness evaluated in TTS?

There is no standardized recommendation for evaluating appropriateness of task-specific or user-centered appropriateness in synthetic speech. Given the argument that there is no "one-size fits all" solution for the naturalness of synthesized voices, customized solutions have perhaps been preferred.

In the papers reviewed above, contextual appropriateness was evaluated using MOS ratings [52, 3]. However, research in human-robot interaction brings a wider variety of designs into the evaluation framework. The audience-response framework [68] collects instantaneous, real-time feedback from listeners. Researchers in psychoacoustics [69] observe the influence of length in correctly predicting the human-likeness of synthetic stimuli.

Commercial leaders of TTS voices also provide a firsthand user-centric approach. For example, Rime Labs incorporate phenomenon like infixation and non-verbal cues like laughter in their product. ElevenLabs aims to help content creators in reducing the dependence on recording studios and voice actors. For content to maximise its outreach, several companies also offer multilingual support in dubbing and voice cloning. A critical use-case seen across industries is to also ensure low-latency in conversational settings. This provides a seamless conversation between the human and AI partner.

5. General guidelines for evaluating naturalness

In the previous sections, we have seen the dimensions of naturalness and identified use-cases in each of them. We arrive at the definition that naturalness in TTS synthesizers is in effect a multi-faceted perceptual attribute. Since TTS is a fast-paced, dynamic field, we intend to open a discussion around naturalness rather than crystallize it. The following paragraphs will discuss how this definition can be a starting point, and how communities within TTS can motivate a better understanding of naturalness.

Naturalness has been a longstanding focus of technological development in the TTS development communities. Therefore, simple improvements in its evaluation practices can be both influential and highly generalizable. Promising directions are already visible in the Blizzard Challenge [4], where the use of the term naturalness is questioned for the first time. We recommend moving backwards from Figure 1. This means that a TTS practitioner or developer must *start* at the use-case. For example, if the end application is to build a TTS voice for aiding a neurogenerative disorder, then human-like traits - prosody and clarity - are required to be maintained in the TTS voice. For evaluating such a voice, TTS developers are strongly encouraged to use the term human-likeness. In contrast, if the focus is on performance in stylized or fictional contexts, such as gaming or storytelling, the goal shifts toward appropriateness—which demands architectural flexibility to model voice characteristics like accent, expressiveness, or emotional tone. This is also already being discussed in architectures like Parler-TTS [70].

Commercial leaders are well-positioned to contribute actively by designing and defining evaluation protocols for the TTS space. First, commercial, real-world experience identifies the most pressing use-cases of TTS voices. For example, automated voice agents requires reduced latency for enabling real-time conversations. Secondly, their user database spans millions of users worldwide, and aims to provide multilingual and multispeaker support. This can help researchers in evaluation draw robust generalisations across a variety of listener demographics. These investigations in collaboration with evaluation scientists can assert better understandings of naturalness, or natural conversations using synthetic voices.

Finally, there is a wealth of expertise in defining naturalness that phoneticians and speech scientists can contribute. Well before the breakthrough success of text-based conversational agents, computational linguists were designing evaluation paradigms for testing the functional competence of large language models. However, as discussed in Section 2.1, the wealth of phonetics techniques has not been explored enough in evaluating synthesizers. Fine-grained acoustic-phonetic analyses can pinpoint locations of distortion in synthetic voices. Especially, perceptual cues that connect with engagement, pleasantness and even trust in TTS voices can be evaluated.

6. Conclusion

In this paper, we define naturalness as a multi-faceted perceptual attribute in TTS synthesizers. We show that human-likeness of synthetic voices is an important and desirable target in TTS synthesizers. It caters to applications such as development of assistive devices, and creating stimuli for phonetics research. However, there are limitations to this approach. Thus, appropriateness of the voice to task and situation is also critical. Finally, we present a set of concrete guidelines for researchers and practitioners of Text-to-Speech voices.

Through this paper, we intend to open a discussion around the concept of naturalness. In designing evaluation designs, the readers are encouraged to bear the multi-faceted nature of naturalness in mind. Particularly, the use cases visually described should be incorporated in the design of their evaluation.

7. Acknowledgements

This work was conducted during the doctoral dissertation of the first author. It has the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224, the ADAPT Centre (Grant 13/RC/2106), and a Google Faculty Award.

8. References

[1] S. King, "Measuring a decade of progress in text-to-speech," *Loquens*, vol. 1, no. 1, pp. e006–e006, 2014.

[2] P. Wagner, J. Beskow, S. Betz, J. Edlund, J. Gustafson, G. E. Henter, S. Le Maguer, Z. Malisz, É. Székely, C. Tånnander *et al.*, "Speech Synthesis Evaluation—State-of-the-Art Assessment and Suggestion for a Novel Research Program," in *Speech Synthesis Workshop (SSW)*, 2019.

[3] R. Clark, H. Silen, T. Kenter, and R. Leith, "Evaluating long-form text-to-speech: Comparing the ratings of sentences and paragraphs," 09 2019, pp. 99–104.

[4] O. Perrotin, B. Stephenson, S. Gerber, and G. Bailly, "The Blizzard Challenge 2023," in *Proc. 18th Blizzard Challenge Workshop*, 2023, pp. 1–27.

[5] A. Kirkland, S. Mehta, H. Lameris, G. E. Henter, E. Szekely, and J. Gustafson, "Stuck in the mos pit: A critical analysis of mos test methodology in tts evaluation," in *12th ISCA Speech Synthesis Workshop (SSW2023)*, 2023, pp. 41–47.

[6] S. Shirali-Shahreza and G. Penn, "Better replacement for tts naturalness evaluation," in *12th ISCA Speech Synthesis Workshop (SSW2023)*. ISCA, Aug. 2023, p. 197–203. [Online]. Available: http://dx.doi.org/10.21437/ssw.2023-31

[7] J. Q. Stewart, "An electrical analogue of the vocal organs," *Nature*, vol. 110, no. 2757, pp. 311–312, 1922.

[8] B. H. Story, "History of speech synthesis," in *The Routledge handbook of phonetics*. Routledge, 2019, pp. 9–33.

[9] V. E. McGee, "Semantic components of the quality of processed speech," *Journal of Speech and Hearing Research*, vol. 7, no. 4, pp. 310–323, 1964.

[10] D. Richards and J. Swaffield, "Assessment of speech communication links," *Proceedings of the IEE-Part B: Radio and Electronic Engineering*, vol. 106, no. 26, pp. 77–89, 1959.

[11] L. Pols and G. Boxelaar, "Comparative evaluation of the speech quality of speech coders and text-to-speech synthesizers," in *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 11. IEEE, 1986, pp. 901–904.

[12] H. Nusbaum, E. Schwab, and D. Pisoni, "Subjective evaluation of synthetic speech: Measuring preference, naturalness and acceptibility," in *Proceedings of the Institution of Electrical Engineers*, vol. 10. Speech Research Lab-Tech, Bloomington IN, 1984, pp. 391–407.

[13] H. A. Patil, T. B. Patel, N. J. Shah, H. B. Sailor, R. Krishnan, G. Kasthuri, T. Nagarajan, L. Christina, N. Kumar, V. Raghavendra *et al.*, "A syllable-based framework for unit selection synthesis in 13 indian languages," in *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*. IEEE, 2013, pp. 1–8.

[14] J. Mendelson and M. P. Aylett, "Beyond the listening test: An interactive approach to TTS evaluation." in *International Conference on Speech Communication and Technology (Interspeech)*, 2017, pp. 249–253.

[15] S. Betz, B. Carlmeyer, P. Wagner, and B. Wrede, "Interactive hesitation synthesis: modelling and evaluation," *Multimodal Technologies and Interaction*, vol. 2, no. 1, p. 9, 2018.

[16] W. Mattheyses and W. Verhelst, "Audiovisual speech synthesis: An overview of the state-of-the-art," *Speech Communication*, vol. 66, pp. 182–217, 2015.

[17] C. Fron and O. Korn, "A short history of the perception of robots and automata from antiquity to modern times," *Social robots: technological, societal and ethical aspects of human-robot interaction*, pp. 1–12, 2019.

[18] N. Epley, A. Waytz, and J. T. Cacioppo, "On seeing human: a three-factor theory of anthropomorphism." *Psychological review*, vol. 114, no. 4, p. 864, 2007.

[19] Q. Q. Chen and H. J. Park, “How anthropomorphism affects trust in intelligent personal assistants,” *Industrial Management & Data Systems*, vol. 121, no. 12, pp. 2722–2737, 2021.

[20] E. J. De Visser, S. S. Monfort, R. McKendrick, M. A. Smith, P. E. McKnight, F. Krueger, and R. Parasuraman, “Almost human: Anthropomorphism increases trust resilience in cognitive agents.” *Journal of Experimental Psychology: Applied*, vol. 22, no. 3, p. 331, 2016.

[21] J. Schroeder and N. Epley, “Mistaking minds and machines: How speech affects dehumanization and anthropomorphism.” *Journal of Experimental Psychology: General*, vol. 145, no. 11, p. 1427, 2016.

[22] V. Giannouli and M. Banou, “The intelligibility and comprehension of synthetic versus natural speech in dyslexic students,” *Disability and Rehabilitation: Assistive Technology*, vol. 15, no. 8, pp. 898–907, 2020.

[23] D. A. Brunow and T. A. Cullen, “Effect of text-to-speech and human reader on listening comprehension for students with learning disabilities,” *Computers in the Schools*, vol. 38, no. 3, pp. 214–231, 2021.

[24] C. Jreige, R. Patel, and H. T. Bunnell, “Vocalid: Personalizing text-to-speech synthesis for individuals with severe speech impairment,” in *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*, 2009, pp. 259–260.

[25] M. Basnet, N. Poudel, S. Dahal, and S. Subedi, “Aawaj: Augmentative communication support for the vocally impaired using nepali text-to-speech,” 2023.

[26] Z. Malisz, G. E. Henter, C. Valentini-Botinhao, O. Watts, J. Beskow, and J. Gustafson, “Modern speech synthesis for phonetic sciences: A discussion and an evaluation,” in *International Congress of Phonetic Sciencesnces ICPhS 2019 5-9 August 2019, Melbourne, Australia Melbourne Convention and Exhibition Centre*, 2019.

[27] M. Y. Liberman, “Corpus phonetics,” *Annual Review of Linguistics*, vol. 5, pp. 91–107, 2019.

[28] A. N. Chasaide, N. N. Chiaráin, H. Berthelsen, C. Wendler, and A. Murphy, “Speech technology as documentation for endangered language preservation: The case of irish.” in *ICPhS*, vol. 2015, 2015, p. 18th.

[29] D. H. Klatt, J. Tiao, and W. Tetschner, “Using dectalk as an aid for the handicapped,” *The Journal of the Acoustical Society of America*, vol. 75, no. S1, pp. S85–S85, 1984.

[30] J. Holmes, “The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer,” *IEEE transactions on Audio and Electroacoustics*, vol. 21, no. 3, pp. 298–305, 1973.

[31] G. E. Peterson, W. S.-Y. Wang, and E. Sivertsen, “Segmentation techniques in speech synthesis,” *The Journal of the Acoustical Society of America*, vol. 30, no. 8, pp. 739–742, 1958.

[32] A. J. Hunt and A. W. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 1. IEEE, 1996, pp. 373–376.

[33] T. Hain, P. C. Woodland, G. Evermann, M. J. Gales, X. Liu, G. L. Moore, D. Povey, and L. Wang, “Automatic transcription of conversational telephone speech,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 6, pp. 1173–1185, 2005.

[34] A. Van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “WaveNet: A generative model for raw audio,” *CoRR*, vol. abs/1609.03499, 2016. [Online]. Available: http://arxiv.org/abs/1609.03499

[35] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, “Diffwave: A versatile diffusion model for audio synthesis,” *International Conference on Learning Representations*, 2021.

[36] W.-N. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen *et al.*, “Hierarchical generative modeling for controllable speech synthesis,” *arXiv preprint arXiv:1810.07217*, 2018.

[37] A. W. Black and K. Tokuda, “The blizzard challenge-2005: Evaluating corpus-based speech synthesis on common datasets,” in *Ninth European Conference on Speech Communication and Technology*, 2005, pp. 77–80.

[38] S. King and V. Karaiskos, “The blizzard challenge 2013,” in *The Blizzard Challenge Workshop*, 2013, http://festvox.org/blizzard/bc2013/summary_Blizzard2013.pdf.

[39] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs,” in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[40] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl, “Perceptual objective listening quality assessment (polqa), the third generation itu-t standard for end-to-end speech quality measurement part i—temporal alignment,” *Journal of the Audio Engineering Society*, vol. 61, no. 6, pp. 366–384, 2013.

[41] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte, “Visqol: an objective speech quality model,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–18, 2015.

[42] D.-S. Kim, “Anique: An auditory model for single-ended speech quality estimation,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 821–831, 2005.

[43] J.-N. Antons, R. Schleicher, S. Arndt, S. Moller, A. K. Porbadnigk, and G. Curio, “Analyzing speech quality perception using electroencephalography,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 721–731, 2012.

[44] A. Kirkland, H. Lameris, E. Székely, and J. Gustafson, “Where’s the uh, hesitation? the interplay between filled pause location, speech rate and fundamental frequency in perception of confidence,” in *Proceedings of Interspeech*, 2022, pp. 18–22.

[45] P. Pérez Zarazaga, Z. Malisz, G. E. Henter, and L. Juvela, “Speaker-independent neural formant synthesis,” in *Proc. INTERSPEECH 2023*, 2023, pp. 5556–5560.

[46] J. Šimko, T. Törö, , M. Vainio, and A. Suni, “Prosody under control: A method for controlling prosodic characteristics in text-to-speech synthesis by adjustments in latent reference space,” in *Proceedings of 10th International Conference on Speech Prosody 2020, Tokyo, Japan*. ISCA, 2020.

[47] H. Lameris, S. Mehta, G. E. Henter, J. Gustafson, and É. Székely, “Prosody-controllable spontaneous tts with neural hmms,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[48] E. Gutierrez, P. O. Gallegos, and C. Lai, “Location, location: Enhancing the evaluation of text-to-speech synthesis using the rapid prosody transcription paradigm,” *ArXiv*, vol. abs/2107.02527, 2021.

[49] J. Cole and S. Shattuck-Hufnagel, “New methods for prosodic transcription: Capturing variability as a source of information,” *Laboratory Phonology*, vol. 7, no. 1, 2016.

[50] I. Gessinger, E. Raveh, I. Steiner, and B. Möbius, “Phonetic accommodation to natural and synthetic voices: Behavior of groups and individuals in speech shadowing,” *Speech Communication*, vol. 127, pp. 43–63, 2021.

[51] A. Pandey, S. Le Maguer, J. Carson-Berndsen, and N. Harte, “Production characteristics of obstruents in WaveNet and older TTS systems,” in *Proc. Interspeech 2022*, 2022, pp. 2373–2377.

[52] R. Dall, J. Yamagishi, and S. King, “Rating naturalness in speech synthesis: The effect of style and expectation,” in *Proceedings of Speech Prosody*. Citeseer, 2014.

[53] S. Wallbridge, P. Bell, and C. Lai, "It's Not What You Said, it's How You Said it: Discriminative Perception of Speech as a Multichannel Communication System," in *Proc. Interspeech 2021*, 2021, pp. 2386–2390.

[54] S. Wilson and R. K. Moore, "Robot, alien and cartoon voices: implications for speech-enabled systems," in *1st Int. Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR-2017)*, 2017, pp. 40–44.

[55] W. J. Mitchell, K. A. Szerszen Sr, 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.

[56] C. McGinn and I. Torre, "Can you tell the robot by the voice? an exploratory study on the role of voice in the perception of robots," in *2019 14th ACM/IEEE international Conference on human-robot interaction (HRI)*. IEEE, 2019, pp. 211–221.

[57] M. Mara, S. Schreibelmayr, and F. Berger, "Hearing a nose? user expectations of robot appearance induced by different robot voices," in *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 355–356.

[58] G. Trovato, J. Ramos, H. Azevedo, A. Moroni, S. Magossi, H. Ishii, R. Simmons, and A. Takanishi, "Designing a receptionist robot: Effect of voice and appearance on anthropomorphism," in *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2015, pp. 235–240.

[59] I. Torre and S. Le Maguer, "Should robots have accents?" in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2020, pp. 208–214.

[60] R. Tamagawa, C. I. Watson, I. H. Kuo, B. A. MacDonald, and E. Broadbent, "The effects of synthesized voice accents on user perceptions of robots," *International Journal of Social Robotics*, vol. 3, pp. 253–262, 2011.

[61] S. Alam, B. Johnston, J. Vitale, and M.-A. Williams, "Would you trust a robot with your mental health? the interaction of emotion and logic in persuasive backfiring," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 2021, pp. 384–391.

[62] K. Seaborn, N. P. Miyake, P. Pennefather, and M. Otake-Matsuura, "Voice in human–agent interaction: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 4, pp. 1–43, 2021.

[63] C. Nass, Y. Moon, and N. Green, "Are machines gender neutral? gender-stereotypic responses to computers with voices," *Journal of applied social psychology*, vol. 27, no. 10, pp. 864–876, 1997.

[64] B. Tay, Y. Jung, and T. Park, "When stereotypes meet robots: the double-edge sword of robot gender and personality in human–robot interaction," *Computers in Human Behavior*, vol. 38, pp. 75–84, 2014.

[65] E. J. Lee, C. Nass, and S. Brave, "Can computer-generated speech have gender? an experimental test of gender stereotype," in *CHI'00 extended abstracts on Human factors in computing systems*, 2000, pp. 289–290.

[66] A. Danielescu, "Eschewing gender stereotypes in voice assistants to promote inclusion," in *Proceedings of the 2nd conference on conversational user interfaces*, 2020, pp. 1–3.

[67] C. Yu, C. Fu, R. Chen, and A. Tapus, "First attempt of gender-free speech style transfer for genderless robot," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2022, pp. 1110–1113.

[68] J. Edlund, C. Tånnander, and J. Gustafson, "Audience response system-based assessment for analysis-by-synthesis." in *ICPhS*, 2015.

[69] A. Pandey, J. Edlund, S. Le Maguer, and N. Harte, "Listener sensitivity to deviating obstruents in WaveNet," in *Proc. INTERSPEECH 2023*, 2023, pp. 1080–1084.

[70] D. Lyth and S. King, "Natural language guidance of high-fidelity text-to-speech with synthetic annotations," 2024.