10th ISCA Speech Synthesis Workshop
20-22 September 2019, Vienna, Austria

Speech Synthesis Evaluation — State-of-the-Art Assessment and Suggestion for a Novel Research Program

Petra Wagner[1,2], *Jonas Beskow*[3], *Simon Betz*[1,2], *Jens Edlund*[3], *Joakim Gustafson*[3], *Gustav Eje Henter*[3], *Sébastien Le Maguer*[4], *Zofia Malisz*[3], *Éva Székely*[3], *Christina Tånnander*[3], *Jana Voße*[1,2]

[1]Phonetics Workgroup, Faculty of Linguistics and Literary Studies, Bielefeld University
[2]CITEC, Bielefeld University
[3]Division of Speech, Music, and Hearing, KTH Royal Institute of Technology, Stockholm, Sweden
[4]Trinity College, Dublin

petra.wagner@uni-bielefeld.de

Abstract

Speech synthesis applications have become an ubiquity, in navigation systems, digital assistants or as screen or audio book readers. Despite their impact on the acceptability of the systems in which they are embedded, and despite the fact that different applications probably need different types of TTS voices, TTS evaluation is still largely treated as an isolated problem. Even though there is strong agreement among researchers that the mainstream approaches to Text-to-Speech (TTS) evaluation are often insufficient and may even be misleading, there exist few clear-cut suggestions as to (1) how TTS evaluations may be realistically improved on a large scale, and (2) how such improvements may lead to an informed feedback for system developers and, ultimately, better systems relying on TTS. This paper reviews the current state-of-the-art in TTS evaluation, and suggests a novel user-centered research program for this area.

1. Introduction

— Is that what people want?
— It's what we do.
(Tom Stoppard)

Synthetic speech is ubiquitous. We hear it in our daily lives as public transport announcements, when interacting with digital assistants or navigation systems, and synthetic voices have been made famous by personalities such as Stephen Hawking. Their perceptual quality has a strong impact on the acceptability of the systems in which they are embedded, and voice related quality issues are subject to much public discussion in online platforms, where journalists have even diagnosed ongoing "voice wars" [1]. Despite this, and despite the fact that speech synthesis technologies have undergone enormous technological developments in the past few years, TTS evaluation is approached in more or less the same way as in the late 1990s, when the International Telecommunication Union (ITU) contributed substantially towards evaluation standards [2]. However, these standards were not predominantly designed for TTS evaluation. Rather, they originated as general recommendations for assessing the output quality of speech transmission systems, where an undisturbed reference signal can be straightforwardly defined, and where the specific application and listening situation need not be taken into account. Rather, these context factors were treated as confounds that had to be controlled for in experimental settings.

The crucial problem with this underlying assumption is, that with respect to speech transmissions, there is no stable reference or gold standard that exists independently of a situation it is embedded in. This is easy to understand with the help of a thought experiment: Imagine a situation in which you consider the spoken delivery of an utterance as near perfect, e.g., when a highly skilled actor reads out a poem. Now imagine this exact style of delivery in a different social setting, e.g., a telephone-based inquiry, or by a person of a different gender, size, or personality. The result would most certainly not be perceived as "optimal" or "perfect", due to style mismatches between what is expected or situationally adequate, and what is perceived (cf. section 2).

In other words, just like clothes do not fit every person alike, and just like human speakers adapt their way of speaking to the situational needs and the audience they are addressing, the development of TTS is not an all-purpose or one-size-fits-all problem. Hence, the quality of a particular TTS will most likely not perceived in a stable fashion across various application contexts. This insight is mainstream for related domains such as the evaluation of dialogue systems, where a perceived system quality cannot be meaningfully assessed in a decontextualized fashion [3, 4]. First evidence supporting this claim also for the domain of TTS evaluation has been produced by [5], who show that the same TTS material is rated differently in a crowdsourced, non-interactive MOS rating, and an MOS rating following an interaction between a human and a virtual agent in a collaborative task. Despite these insights, a meta analysis [6] revealed that the vast majority of TTS evaluations remain to rely on decontextualized listening tests, where participants score the quality of isolated sentences rather than embedding them within realistic applications or meaningful interactions. Thus, our knowledge about the practical applicability of the various existing systems remains vague at best.

In a similar vein, recent times have seen an increasing number of papers criticizing traditional approaches to TTS evaluation [7, 5], or pointing out frequent methodological flaws such as the low validity of most TTS evaluations due to small participant numbers and a lack of diversity in the tested listener groups, especially in the light of vast individual differences between listeners [8, 9], which shows stronger for some traits (age, human-likeness) than others (gender, accent origin) [10]. Generally, these investigations point out the necessity for a better conceptual framing of the perception tasks, together with larger test populations and more careful statistical approaches.

Despite this repeatedly expressed scepticism of the way TTS evaluations are typically carried out, the majority of TTS evaluation appears to follow familiar, seemingly safe, paths. The likely reason for this is that alternative standards or at least clear-cut recommendations are still lacking.

10.21437/SSW.2019-19

This paper will take a first step towards suggesting an alternative program for synthesis evaluation, which is based on *contextual appropriateness* rather than an unrealistic notion of an existing gold standard (Section 2). We will then make a first suggestion for an alternative strategy towards speech synthesis evaluation, resting on an in-depth analysis of application centered user needs (Section 3), followed by an assessment of existing approaches towards synthesis quality measurement (Section 4). Finally (Section 5), we suggest that the design and standardization of suitable TTS evaluation schemes should be accepted as a necessary research area in its own right.

2. Contextual appropriateness as metric of speech quality?

We contend that human speech production is highly variable and comes in many different "styles", which are continuously adapted by speakers given dynamically changing social (tutoring, chatting, arguing, counseling...), individual (hearing problems, attitude, level of distraction, motivation, familiarity), linguistic (frequency, predictability, suprisal, importance) or environmental settings (external noise, mutual visibility, ...) [11, 12, 13, 14, 15, 16, 17, 18]. Due to this inherent contextual embedding, human speech production can never be "neutral" or "perfectly natural", and no speaking style therefore qualifies as a reference signal that a speech event of inherently less quality, e.g., a synthetic one, can be meaningfully compared to. Still, this remains an underlying assumption in much TTS evaluation research, where this reference or gold standard is often taken as being equivalent to "human read speech". Some researchers criticize this implicit assumption, and postulate an alternative reference such as "conversational speech" [19]. While such an approach may be useful for a particular application such as dialogue systems research, neither speaking style is inherently "neutral" or "natural": Read speech is entirely appropriate in certain contexts of human communication, e.g., when reading a story to a child, and conversational speech in others. Thus, while no style is inherently neutral, every style can be more or less appropriate for a given context, e.g., speaking loudly may be an optimal choice in a loud pub, but entirely inappropriate in more formal situations [18]. *Appropriateness* given a certain situation or application may be thus a better indicator of measuring the suitability of a certain speaking style over another. This is in line with the analysis by [20, 21], claiming that long-known problems of human machine interaction such as the *uncanny valley* can be modeled as a mismatch between a user's expectations and a machine's actual expression. In fact, attempts at defining suitable voices for robots have found that some human listeners prefer a robot to sound "robot-like", with the typical artifacts created by formant-based speech synthesis, even though these are often dispreferred in traditional listening tests [22]. In an evaluation dedicated to find a suitable synthetic voice for *Pepper* interacting with autistic children, [23] indeed find some support for the hypothesis of TTS quality to be predictable by a fit between what listeners expect a robot to sound like, and what it actually does sound like. Also, they confirm the hypothesis that human voices are not necessarily a suitable gold standard for TTS quality. Contrary to this, however, are the results by [24], finding human voices to be preferred in more complex tasks. It is unclear, though, whether this finding is really caused by the style of voice, or is an effect of the processing difficulties introduced by speech synthesis artifacts especially present in classic formant speech synthesis systems [25].

We therefore contend that even if the goal of the TTS evaluation is a "pure" system comparison, without an actual application in mind, some kind of conceptual framing may be advisable. Indeed, [19] report that simply asking listeners to imagine a particular interactive situation had an effect on listener's impressions. If no such framing is provided, listeners are forced to imagine *some* context in which they may listen to the TTS, and are prone to come up with a corresponding set of quality dimensions. Indeed, this factor is likely to be one of the causes for the strong variation found among participants of TTS evaluations [8, 9].

An embedding in a realistic application can also make interlocutors more sensitive for quality issues: in [5], it was found that an interactive setting increased listener's sensitivity for quality losses introduced by synthetic hesitations, even though the same hesitations increased their performance in a memory task.

For now, we believe that these conceptual framings can be carried out under controlled, laboratory conditions, as they are common practice in related fields such as Human Computer Interaction or Human Robot Interaction. In fact, some of these paradigms, e.g., preference tests, have already been successfully applied to the evaluation of prosodic styles [26].

Thus, our first contention is that TTS evaluation may profit from a change of perspectives, moving from the underlying assumption of a stable ideal baseline, to the perspective of choosing and tuning the parameters in such a way that they are most appropriate to a target application. Even if no such target application can be identified, it is advisable to provide some conceptual framing to participants in order to guide them to a set of speech quality dimensions that is comparable across participants and as general as possible, e.g., by instructing them "to imagine listening to a smartphone reading out a newspaper article". This type of framing is likely to affect the sentence material to be chosen for synthesis. Next, we need to specify the parameter space in which these applications are best evaluated.

Take Home Message 1:
There is no stable gold standard for optimal speech quality!

3. How to assess listener needs, expectations and preferences

A main problem with the paradigm sketched above is that we hitherto know very little about the individual and application centered needs and expectations of listeners with respect to TTS voices. Still, some approaches towards analyzing user preferences have been made: In an analysis of blind TTS users' preferences, [27] found out that participants often prefer formant synthesis over concatenative systems, as these perform better in ultrafast conditions. [28] conducted a large-scale survey on user preferences with respect to voices in car navigation systems. A more recent study directly used the intelligibility profiles of elderly listeners to fine-tune a TTS to their particular needs [29]. However, the general lack of information on user expectations poses a huge difficulty for TTS evaluations: if we want to come up with a diagnostic evaluation of our TTS voice that goes beyond a global assessment of quality, we need to ask precise questions, especially if questionnaires are being employed. Alternatively, we need to find diagnostics that point towards potential problems, without explicitly mentioning them. It is clearly the case that users may be unable to express an in-

formed opinion about their expectation of a TTS voice, other than, e.g., an opinion about the music or food they prefer.

Thus, while a first step towards better tailoring of TTS evaluations may lie in an in-depth analysis of needs, these needs probably arise only within a specific application context or interactive situation, and may evolve slowly over time and with increasing user experience. Our view on evaluation consequently changes from the perspective looking for a general-purpose synthesis to one that has much in common with an "audition scenario", where a highly skilled director or a team of experts cast several actors throughout a series of different scenes, until they have found the ideal person to perform a particular role.

Given the lack of available empirical data, we are currently confined to define the application-specific needs or relevant quality dimensions based on top-down assumptions, e.g., a TTS used in a noisy environment should be sufficiently clear, while a TTS used for leisure-time audio book reading should probably have some degree of expressivity. A first attempt at such a top-down analysis of user needs is given below in Table 1. Obviously, this table does not yet include an estimate for different user groups (elderly, children, non-native, distracted, visually impaired, ...), and will have to be fine-tuned to take into account different cognitive, physiological and personality traits and abilities.

Summing up, our second contention is that we need to intensify the analyses of listener's needs and expectations, to be able to develop suitably tailored evaluation settings. An additional strategy lies in exploring in developing useful diagnostic tools that point to potential issues during an ongoing interaction with a TTS.

Take Home Message 2:
We need to assess and take into account listeners' application-specific needs and expectations!

4. Reviewing measures of TTS performance

Obviously, a straightforward way of finding out whether the estimated user needs are met by a system, is to simply ask or test listeners in a *subjective evaluation*. Another approach is to perform an *objective evaluation*, relying on an automated criterion that operationalizes an abstract quality dimension. Yet another, albeit less common strategy is to test whether the system allows listeners to perform an intended task better or worse, using a *behavioral evaluation*. Below, we give a short overview of the current state-of-the-art in objective, subjective, and behavioral TTS evaluation. More specifically, we will show that despite a current lack of informed quality dimensions, we already have a large repertoire of objective and subjective metrics at our disposal. In Table 1, we give examples for how a system's needs, or quality dimensions, can be operationalized in objective, subjective or behavioral evaluations. Some of these are not completely independent: comprehensibility may be regarded as a form of task success in an announcement system, and is likely to be a prerequisite for task success in most speech-based systems. Still, speech-based systems will often support tasks beyond the processing of speech-based information.

4.1. Objective assessment of TTS

Objective assessment generally consists of classifying system output to obtain a score. While the idea of scoring synthetic speech in an objective and automated manner is theoretically attractive, as it reduces the need for expensive, time-consuming, and noisy subjective evaluations, the truth is that our current objective metrics do not align well with human perception. This limits their use mostly to system tuning, while the final evaluation still must be based on a subjective listening test. Besides, not every trait that can be assessed subjectively has an objectively assessable counterpart. Furthermore, many of the more accurate objective measures require access to natural speech to compare against, or knowledge about the true noise signal in a speech-in-noise scenario, further limiting their applicability.

The most common speech aspects to score are intelligibility (especially in noisy or reverberant environments), but also segmental quality, and prosodic correlates such as pitch and voiced-unvoiced similarity to designated natural reference recordings. When trying to capture "naturalness", objective metrics tend to focus on spectral features, with prosody considered a secondary problem, an approach that seems to be based on a bias that is difficult to motivate from a phonetic point of view – besides the fact that "naturalness" is a nebulous concept in general (cf. section 2).

Speech quality assessment is mainly done using the mel-cepstral distortion (MCD) and the PESQ family of ITU standards [30], and use recorded natural speech as a reference against which the corresponding synthetic utterance is scored. MCD computation consists of time warping to align the two signals (in case the timings differ), computing the Euclidean distance between each aligned natural and synthetic mel-ceptral vector (frame), and averaging these distances over time.

There has been substantial effort to develop more advanced quality-assessment methods for synthetic speech based on machine learning, e.g., in Hinterleitner's PhD work [31] or Quality-Net [32], which learns to estimate PESQ scores without a natural reference. However in general, the correlation between system-level assessments might be passable, but stimulus-level correlations are low. More impressive results were reported by AutoMOS [33], but this system has only been trained and evaluated on a single speaker, and is not publicly available. However, with the advent of high-quality, probabilistic waveform-level synthesis models such as WaveNet [34], we finally have synthesizers capable of generating high-quality speech waveforms [25]. These models encode a lot of information about what a "natural"-sounding, or rather human-like, waveform may actually look and sound like. It is entirely possible that the likelihood that a trained waveform-level synthesizer assigns to a given speech waveform could be a useful indicator of whether or not that waveform is "human-like" or not, without actual access to a comparable utterance from a human speaker. However, this aspect has to our knowledge not yet been investigated. In any case, results need not transfer across speakers and might be sensitive to linear or nonlinear processing applied to the signals.

4.2. Subjective assessment of TTS

A popular approach to evaluate interaction quality employs questionnaires, explicitly asking users for their impression of various quality dimensions (e.g., likability, intelligibility, perceived intelligence). Given our lack of proper understanding of users' needs and expectations and quality dimensions, however, this method is risky, as it presupposes a good understanding of what a user actually misses or likes in the technical system. To overcome this problem, typical surveys employed in HCI or HRI tend to be very extensive [35], thereby trying to address all potential quality dimensions a user may have employed in

Application	Estimated needs	Possible evaluation
Virtual assistant	clear, pleasant voice	likability (s), intelligibility (o, s, b), comprehension (b), preference (b), voluntary interaction time (b), task success and efficiency (b)
Humanoid robot	humanoid (but not human-like) voice	perceived suitability (s), preference and interaction time (b), task success and efficiency (b)
Navigation	sufficiently loud, clear, timely	intelligibility (o, s, b), task success (b), comprehensibility (s, b)
Announcements	loud, clear	comprehension under noisy or distracted conditions (o, s, b)
Interactive travel guide	clear, pleasant	intelligibility (o, s, b), preference (b), voluntary interaction time (b), comprehensibility (s,b)
Screen reader	intelligible at high speed, informative prosody	intelligibility (o, s, b), comprehensibility (s, b), efficiency (b)
Audiobook (leisure)	slow, expressive	preference (b), voluntary interaction time (b)
Audiobook (educational)	optimized for online comprehension	comprehensibility (s, b), task success and efficiency (b)
Video game	convincing personality, expressive	preference and interaction time (b), personality fit (s), convincing (s) and easily identifiable (s, b) emotional display
Voice prosthesis	adaptable speaker identity, low latency	similarity to original voice (o, s), latency (o), long term user satisfaction (s)
Dialogue system	timely, incremental, suitable discourse markers	preference and voluntary interaction time (b), task success and efficiency (b), adaptive behavior (b)
Speech-to-speech translation	adaptable speaker identity	similarity to original voice (o, s), latency (o)

Table 1: *A first top-down sketch of listeners' demands on TTS for a variety of applications as well as ideas for their subjective (s), objective (o) or behavioral (b) measurement.*

her or his assessment. However, this poses a high risk of getting invalid responses, due to fatigue or boredom [36]. Also, the questionnaires do not normally address the amount of deviation from a user's expectations, which may considerably affect interaction quality. However, global subjective assessments of interaction quality remain a useful diagnostic.

Most metrics employed in questionnaires try to capture a global impression of signal quality such as mean opinion score (MOS) [37]. Alternatively, metrics target more fine-grained system diagnostics such as multiple stimuli with hidden reference and anchor (MUSHRA) [38], or pairwise comparison approaches that ultimately allow for a multidimensional scaling of systems, but rely on multiple assessments of comparable utterances across systems [39].

An alternative way of grasping TTS related problems during an ongoing interaction has been developed by [40]. In their auditory response system, they have third parties evaluate an interaction, and give a simple binary response in moments where "issues" arise. This method has the advantage of assessing subjective quality in course of an ongoing interaction. While behavioral and physiological metrics may provide alternative metrics for such real-time tracking of user experience, they may be overly sensitive and difficult to interpret. However, especially EEG and eye/mouse tracking may be suitable candidates for indicating mismatches between a user's expectation and the actual synthetic realization, and may therefore produce good estimates of subjectively experienced interaction quality.

4.3. Behavioral assessment of TTS

If a researcher is interested in less impressionistic measures of intelligibility, established measures are "semantically unpredictable sentences" (SUSs, [41]), together with word edit distance, word error rate estimates, or rhyme tests [42, 43]. With the advent of highly intelligible systems in recent years, the need for specific intelligibility measurements has become less of an issue. However, they may still play a role in more experimental systems such as articulatory synthesis.

Other than intelligibility, the measurement of *comprehensibility*, i.e., the degree to which a message's semantics and pragmatics has been understood, is largely under-researched and much less well understood. While some researchers postulate to assess it in content repetition tasks [44], [45] suggests it can only be assessed indirectly, e.g., by asking questions that allow for an inference about how well a listener has grasped a message's content.

Behavioral performance has been most meaningfully employed in evaluations of TTS embedded in interactive systems, e.g., by assessing the amount of retrieved information content (memory task efficiency) after an interaction between a listener and a dialogue system [5]. Related metrics are efficiency and effectiveness, which take into account task completion time or the duration of an interaction and often are employed when evaluating dialogue systems [4]. While a long interaction time typically is regarded an indicator of low interaction quality in assistance systems, a longer (voluntary) interaction time with a system intended to entertain, e.g., a game or an audiobook, may instead signify good system performance. Quality metrics are thus not independent from the application they are testing, and operationalizations need to be adjusted for each evaluation. Yet another form of behavioral analysis was chosen in [46], where participants' level of verbal adaptation to different interactive character displays was analyzed in a dialogue task. A high degree of adaptation to different characters (with individual voice profiles) was taken as evidence of a better user experience.

While measures related to task performance may be indicative of listening effort or a system's comprehensibility, they typ-

ically fail to unveil *why* and *where* the problems occurred during the interaction. To tackle this issue requires methods that continuously monitor the interaction. Here, both behavioral and physiological metrics of speech synthesis have been explored: [47] combined eye tracking in a visual world paradigm with subjective judgments to explore a facilitating effect of a TTS for listening comprehension. [48] looked at response times and task performance durations in a simple GUI-based interactive game, where listeners had to move around geometric shapes according to a synthetic voice's instructions. Also, some first attempts of using *physiological* rather than behavioral metrics such as pupil dilation or EEG exist [49, 39].

Generally, the behavioral (or physiological) assessment methods described here have the advantage that they do not expect listeners to have an informed opinion about their preferences or expectations. Such an informed opinion is unlikely unless participants have prior experience of TTS-based systems (cf. section 3). However, it is still unclear to what extent behavioral metrics correlate with subjectively experienced quality.

Summing up, although a wide range of metrics have been explored, and contextualizations are possible, TTS evaluations still predominantly rely on global quality estimates using MOS-based tests based on randomly chosen individual utterances. At least some of the approaches sketched above can be easily set-up, and could be carried out resource-efficiently, e.g., using web-based interfaces allowing for crowdsourcing approaches, and have the potential to provide alternatives, or at least supplements, to traditional evaluation procedures.

Take Home Message 3:
Suitable alternatives to traditional decontextualized TTS evaluation procedures exist!

5. Conclusion

To conclude, it seems to be mostly a lack of alternative recommendation standards that prevent current TTS evaluations from being more insightful and less mono-cultured. We therefore end this paper with a proposition, namely that the development of a set of best practice recommendations (rather than a standardization) is a profitable research area in its own right.

Our proposition parallels similar suggestions within the HCI community, striving to enhance the technology-centered concept of "Quality of Experience" with the more user-centered concept of "User Experience" [50]. To initiate research in this area, a few guiding questions could be:

1. Are there cases in which global impressions of subjective quality actually generalize across applications, thus rendering more complex evaluations unnecessary?
2. How can we improve our estimates of user needs (and corresponding quality dimensions)?
3. Do mismatches between user expectations and synthetic styles predict interaction quality in a reliable fashion?
4. Do behavioral (e.g., eye gaze) or subjective (e.g., audience responses) online measures of TTS quality reliably point to local issues that affect global interaction quality?
5. Which dimensions of subjective quality do the other metrics (objective, physiological, behavioral) actually assess?
6. How can novel machine learning and high quality synthesis such as WaveNet be put to use in TTS evaluation?
7. How can we meaningfully generalize from our short-time evaluations to long-time user experience?

Take Home Message 4:
The development of a set of best practice recommendations for TTS evaluation should be a research area in its own right!

6. References

[1] M. Wollerton, "Voice wars: Siri vs. Alexa vs. Google Assistant three voice assistants are fighting for space in your smart home – is there a clear winner?" 2018. [Online]. Available: https://www.cnet.com/news/voice-wars-siri-vs-alexa-vs-google-assistant/

[2] "Methods for subjective determination of transmission quality," International Telecommunication Union, ITU-R Recommendation P.800, 1996.

[3] S. Möller, *Quality Engineering: Qualität kommunikationstechnischer Systeme*. Berlin, Heidelberg: Springer-Verlag, 2010.

[4] "Subjective quality evaluation of telephone services based on spoken dialogue systems," International Telecommunication Union, ITU-R Recommendation ITU-P.851, 2003.

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

[6] P. Wagner and S. Betz, "Speech synthesis evaluation – realizing a social turn," in *Tagungsband Elektronische Sprachsignalverarbeitung (ESSV)*, 2017, pp. 167–172.

[7] J. Mendelson and M. Aylett, "Beyond the listening test: An interactive approach to TTS evaluation," in *Proceedings of Interspeech*, 2017, pp. 249–253.

[8] M. Wester, C. Valentini-Botinhao, and G. E. Henter, "Are we using enough listeners? No! — an empirically-supported critique of Interspeech 2014 TTS evaluations," in *Proceedings of Interspeech*, 2015, pp. 3476–3480.

[9] A. Rosenberg and B. Ramabhadran, "Bias and statistical significance in evaluating speech synthesis with mean opinion scores," in *Proceedings of Interspeech*, 2017, pp. 3976–3980.

[10] A. Baird, S. H. Jørgensen, E. Parada-Cabaleiro, S. Hantke, N. Cummins, and B. Schuller, "Perception of paralinguistic traits in synthesized voices," in *Proceedings of Audio Mostly*, 2017.

[11] M. Aylett and A. Turk, "The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech," *Language and Speech*, vol. 47, no. 1, pp. 31–56, 2004.

[12] B. Lindblom, *Explaining Phonetic Variation: A Sketch of the H&H Theory*. Kluwer Academic Publishers, 1990, pp. 403–439.

[13] D. Watson, J. Arnold, and M. K. Tanenhaus, "Tic tac TOE: Effects of predictability and importance on acoustic prominence in language production," *Cognition*, vol. 106, no. 3, pp. 1548–1557, 2008.

[14] É. Lombard, "Le signe de l'élévation de la voix," *Annales des Maladies de l'Oreille et du Larynx*, vol. XXXVII, no. 2, pp. 101–109, 1911.

[15] J. Pierrehumbert and J. Hirschberg, "The meaning of intonational contours in the interpretation of discourse," in *Intentions in Communication*, P. Cohen, J. Morgan, and M. Pollack, Eds. Cambridge, MA: MIT Press, 1990, pp. 271–311.

[16] D. Jurafsky, A. Bell, M. Gregory, and W. D. Raymond, "Probabilistic relations between words: Evidence from reduction in lexical production," *Typological Studies in Language*, vol. 45, pp. 229–254, 2001.

[17] Z. Malisz, E. Brandt, B. Möbius, Y. M. Oh, and B. Andreeva, "Dimensions of segmental variability: Interaction of prosody and surprisal in six languages," *Frontiers in Communication*, vol. 3, pp. 25:1–18, 2018.

[18] P. Wagner, J. Trouvain, and F. Zimmerer, "In defense of stylistic diversity in speech research," *Journal of Phonetics*, vol. 48, pp. 1–12, 2015.

[19] R. Dall, J. Yamagishi, and S. King, "Rating naturalness in speech synthesis: the effect of style and expectation," in *Proceedings of Speech Prosody*, 2014.

[20] R. K. Moore, "A Bayesian explanation of the 'Uncanny Valley' effect and related psychological phenomena," *Scientific Reports*, vol. 56, no. 2, p. 864, 2012.

[21] ——, "Appropriate voices for artefacts: Some key insights," in *Proceedings of the 1st International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots*, 2017.

[22] A. Hönemann and P. Wagner, "Adaptive speech synthesis in a cognitive robotic service apartment: An overview and first steps towards voice selection," in *Tagungsband Elektronische Sprachsignalverarbeitung*, 2015, pp. 135–142.

[23] F. Burkhardt, M. Saponja, J. Sessner, and B. Weiss, "How should Pepper sound – preliminary investigations on robot vocalizations," in *Tagungsband Elektronische Sprachsignalverarbeitung*, 2019, pp. 103–110.

[24] E. Rodero, "Effectiveness, attention, and recall of human and artificial voices in an advertising story. Prosody influence and functions of voices," *Computers in Human Behavior*, vol. 77, pp. 336–346, 2017.

[25] 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 *Proceedings of ICPhS*, 2019.

[26] H.-L. Cao, L. C. Jensen, X. N. Nghiem, H. Vu, A. De Beir, P. G. Esteban, G. Van de Perre, D. Lefeber, and B. Vanderborght, "DualKeepon: a human-robot interaction testbed to study linguistic features of speech," *Intelligent Service Robotics*, vol. 12, no. 1, pp. 45–54, 2019.

[27] D. Moers, P. Wagner, and S. Breuer, "Assessing the adequate treatment of fast speech in unit selection speech synthesis systems for the visually impaired," in *Proceedings of the 6th Speech Synthesis Workshop*, 2007, pp. 282–287.

[28] B. Aschenberner and P. Wagner, "A diagnostic evaluation of the speech input and output devices in GPS navigation systems," *Sprache und Datenverarbeitung*, vol. 2, pp. 135–146, 2005.

[29] R. Nishimura, T. Nagao, A. Ichimanda, and N. Kitaoka, "Study on editing method to improve speech intelligibility based on speech perception characteristics of elderly people," *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics*, vol. 30, no. 6, pp. 351–389, 2018.

[30] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," International Telecommunication Union, ITU-R Recommendation P.862, 2003.

[31] F. Hinterleitner, *Quality of Synthetic Speech*, ser. T-Labs Series in Telecommunication Services. Berlin, Germany: Springer, 2017.

[32] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Quality-Net: An end-to-end non-intrusive speech quality assessment model based on BLSTM," in *Proceedings of Interspeech*, 2018, pp. 1873–1877.

[33] B. Patton, Y. Agiomyrgiannakis, M. Terry, K. W. Wilson, R. A. Saurous, and D. Sculley, "AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech," in *Proceedings of the NIPS 2016 End-to-end Learning for Speech and Audio Processing Workshop*, 2016.

[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," *arXiv preprint 1609.03499*, 2016.

[35] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International Journal of Social Robotics*, vol. 1, pp. 71–81, 2009.

[36] P. J. Lavrakas, *Encyclopedia of Survey Research Methods*. Thousand Oaks, CA: SAGE Publications, Inc., 2008, ch. Respondent Fatigue, p. 743.

[37] "Mean opinion score terminology," International Telecommunication Union, ITU-R Recommendation P.800.1, 2016.

[38] "Method for the subjective assessment of intermediate quality level of audio systems," International Telecommunication Union, ITU-R Recommendation ITU-R.BS.1534-3, 2015.

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

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

[41] C. BenoÎt, M. Grice, and V. Hazan, "The SUS test: a method for the assessment of text-to-speech synthesis intelligibility," *Speech Communication*, vol. 18, no. 4, p. 381–392, 1993.

[42] W. Voiers, A. Sharpley, and C. Hehmsoth, "Research on diagnostic evaluation of speech intelligibility," Air Force Cambridge Research Laboratories, Bedford, MA, Tech. Rep. AFCRL-72-0694, 1975.

[43] U. Jekosch, "The cluster identification test (CLID)," in *Proceedings of the International Conference on Spoken Language Processing*, 1992, pp. 205–208.

[44] K. Fellbaum, "Anmerkungen zu den begriffen "Verständlichkeit" und "Verstehbarkeit" bei der Sprachqualitätsmessung," in *Tagungsband Elektronische Sprachsignalverarbeitung*, 2014, pp. 240–247.

[45] S. A. Duffy and D. Pisoni, "Comprehension of synthetic speech produced by rule: a review and theoretical interpretation," *Language and Speech*, vol. 35, pp. 351–389, 1992.

[46] J. Gustafson, J. Boye, M. Fredriksson, L. Johanneson, and J. Königsmann, "Providing computer game characters with conversational abilities," in *Proceedings of the International Conference on Intelligent Virtual Agents*, 2005, pp. 37–51.

[47] R. Rajakrishnan, M. White, S. R. Speer, and K. Ito, "Evaluating prosody in synthetic speech with online (eye tracking) and offline (rating) methods," in *Proceedings of the 7th Speech Synthesis Workshop*, 2010, pp. 276–281.

[48] S. Betz, S. Zarrieß, and P. Wagner, "Synthesized lengthening of function words – the fuzzy boundary between fluency and disfluency," in *Proceedings of the 8th Workshop on Disfluency in Spontaneous Speech*, 2017.

[49] A. Govender and S. King, "Using pupillometry to measure the cognitive load of synthetic speech," in *Proceedings of Interspeech*, 2018, pp. 2838–2842.

[50] I. Wechsung and K. De Moor, *Quality of Experience Versus User Experience*, ser. T-Labs Series in Telecommunication Services. Cham, Switzerland: Springer, 2014.