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

Hot topics in speech synthesis evaluation

Gérard Bailly[1], *Elisabeth André*[2], *Erica Cooper*[3], *Benjamin Cowan*[4], *Jens Edlund*[5], *Naomi Harte*[6], *Simon King*[7], *Esther Klabbers*[8], *Sébastien Le Maguer*[9], *Zofia Malisz*[2], *Roger K. Moore*[10], *Bernd Möbius*[11], *Sebastian Möller*[12], *Ayushi Pandey*[13], *Olivier Perrotin*[1], *Fritz Seebauer*[14], *Sofia Strömbergsson*[15], *David R. Traum*[16], *Christina Tånnander*[5,17], *Petra Wagner*[13], *Junichi Yamagishi*[18], *Yusuke Yasuda*[18]

[1]Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, France; [2]Univ. Augsburg, Germany; [3]NICT - Kyoto, Japan; [4]School of Information and Communication Studies, University College Dublin, Ireland; [5]Speech, Music & Hearing Department, KTH - Stockholm, Sweden; [6]Trinity College Dublin, Ireland; [7]Univ. Edinburgh, UK; [8]phAIstos Speech & Language Technology Services, USA; [9]Univ. Helsinki, Finland; [10]Univ. Sheffield, UK; [11]Language Science and Technology, Saarland Univ., Germany; [12]TU & DFKI - Berlin, Germany; [13]Karya, India; [14]Bielefeld Univ., Germany; [15]Karolinska Institutet, Stockholm, Sweden; [16]USC - Playa Vista, USA; [17]Swedish Agency for Accessible Media - Malmö, Sweden; [18]National Institute of Informatics - Tokyo, Japan

speech-and-synthesis-evaluation@googlegroups.com

Abstract

Speech synthesis is advancing rapidly, often reaching levels that challenge the distinction between synthetic and human speech. Its capabilities are increasingly diverse, and it can no longer be treated as a one-size-fits-all technology. Consequently, one-size-fits-all evaluations based solely on Mean Opinion Score (MOS) fail to reflect the specific requirements, conditions, and success criteria across the wide and ever-evolving landscape of applications and usage contexts. As evaluation necessarily becomes more task-oriented, determining what to evaluate and how to evaluate it must be an integral part of the evaluation process.

In this overview of the current speech synthesis evaluation landscape, we begin by revisiting a range of existing evaluation methodologies that are often overlooked in favour of MOS, despite offering valuable insights for specific tasks. We then highlight a set of emerging "hot topics" in speech synthesis, examining their unique demands and proposing directions for their evaluation.

The hot topics are structured around specific use cases, which serve as examples to highlight the speech synthesis capabilities that are critical in each context. The use case framing also facilitates a dual perspective, capturing both the evaluation of speech synthesis as an integrated part of an application and the assessment of its standalone capabilities.

Index Terms: Speech Synthesis, Evaluation, Incremental Processing, Interactive Systems

1. Introduction

Evaluation of speech synthesis has been largely driven over the past two decades by the Blizzard Challenge [1, 2], which proposes one systematic framework for large-scale text-to-speech synthesis evaluation. Its methods mostly comprise intelligibility assessment using transcription tasks [3], overall quality assessment using the Absolute Category Rating (ACR) paradigm – or MOS-test – described by ITU P.800 [4] and ITU-T P.800.1 [5], as well as adapted ACRs to evaluate the naturalness and the speaker similarity. While speech synthesis systems performance was more limited, a single quality scale may have sufficed to capture the dominant differences between systems. However, the results of the most recent Blizzard Challenge [6] add to the growing body of evidence that current Text-To-Speech (TTS) systems produce speech that, judged with a single ACR of isolated sentences, is perceptually indistinguishable from human recordings, while finer, local comparisons between the same generated sentences still succeeded to highlight weaknesses in the synthesis. Therefore, with ongoing improvements in quality and increasingly diverse capabilities, general-purpose evaluation for state-of-the-art speech synthesis has become untenable, and speech synthesis can no longer be treated as a one-size-fits-all technology. Consequently, one-size-fits-all evaluations necessarily fail to reflect the specific requirements, conditions, and success criteria across an ever-evolving landscape of applications and usage contexts.

At the same time, the consistency of the Blizzard Challenge –which has made it possible to compare TTS systems over the years [7]– has also anchored evaluation habits in the speech synthesis community, which sometimes uses experimental protocols systematically without linking them to a hypothesis and/or without precisely controlling experimental conditions. In that particular case, the limitations of standard evaluation metrics for speech synthesis—particularly the widespread use of Mean Opinion Scores (MOS)—have been widely acknowledged for decades: a substantial body of literature has documented the shortcomings of these approaches, ranging from statistical and validation issues to their lack of diagnostic value [8], insufficient protocol control [9], opaque definitions of the criteria to evaluate [10, 11], contextual sensitivity [12], and relevance to real-world tasks [13, 14, 7]. We do not aim to rehash these critiques here; rather, we take them as established. We thus step out of the frame of critiquing metrics, and turn instead to the broader question of what speech synthesis is, what it is for, and what its evaluation should tell us.

This brings us to a deeper assumption that underlies much of speech synthesis evaluation: the idea of "human speech" as a gold standard. Claims that synthetic speech has reached human-level quality are based on comparisons with recordings

10.21437/SSW.2025-1

of human speech – either in direct side-by-side tests or through matched rating scales. The common practice of comparing synthetic speech to recorded human speech introduces a category error: it treats recorded human speech as a proxy for human speech, and then mistakes the proxy for the phenomenon it represents. However, playing back a recording of speech is, in effect, a form of speech synthesis; and indeed, many standard evaluation metrics were originally designed to assess the quality of recorded, transmitted and reproduced speech. Actual human speech, on the other hand, is not static or context-free. It is situated, emergent, and adaptive – shaped in real time by the speaker's intentions, the listeners' reactions, and the broader communicative context. It unfolds interactively, often in coordination with gesture, gaze, and other modalities, and is continuously modulated in response to what is happening as it is being produced. There is nothing in principle preventing us from developing speech synthesis that approximates actual human performance by exhibiting these same properties other than a lack of emphasis on them in system requirements and evaluation. However, as long as synthetic speech is evaluated primarily against static exemplars, we risk missing the qualities that matter most in many real-world applications, such as timing, adaptability, engagement, and communicative appropriateness.

What counts as "good" speech, be it human or synthetic, cannot be defined in isolation from how that speech is used. As speech synthesis systems are deployed across an increasingly wide range of applications – from audiobooks and assistive communication to real-time dialogue and expressive media – their functional requirements and communicative goals vary dramatically, and a growing body of work calls for more situated, task-sensitive evaluation approaches [13].

In this paper, we do not propose new general-purpose metrics to replace MOS or MUltiple Stimuli with Hidden Reference and Anchors (MUSHRA). Rather, we ask how evaluation can be made more responsive to the diversity of capabilities and use cases now emerging for modern TTS. To begin answering this, we take two steps. First, we highlight a set of "hot topics" in contemporary speech synthesis – a non-exhaustive but diverse set of use cases that exemplify new kinds of demands on synthesis systems and which requires additional methods and metrics to properly evaluate. Second, we revisit a set of existing evaluation methodologies that go beyond MOS, many of which are already available but remain underutilized, and discuss how these can be applied to the "hot topic" use cases.

2. The hot topics

We briefly present a set of use cases that represent some of the exciting applications for speech synthesis today and in the near future. Each application not only requires high-quality synthesis (as measured with the current standard metrics), but also have specific constraints not currently necessary for reading machines. These applications therefore impose additional demands which are not easily measurable using the previously mentioned standard metrics. While not a use case proper, we will also briefly discuss legal and ethical issues that arise with the high quality state-of-the-art synthesis, and how they should influence our TTS evaluations. The presentation of each case study presents *what* TTS features are worth modelling and evaluating, *when* and *why* these features would be necessary and *why* alternative/complementary evaluation methods are necessary.

2.1. Controllable speech synthesis

We consider controllable speech synthesis as synthesis in which low-level time-varying parameters are individually controlled or created at the frame level according to externally decided needs. Such modifications could be used by speech scientists, e.g., to create stimuli for classical perception studies such as sensorimotor calibration or categorical perception tasks [15, 16], to create illustration samples in phonetics pedagogy [17], in the augmented communication for digital arts [18], in the creation of suitable voices for non-human artefacts [19], or in voice substitution [20, 21]. Additional applications arise in speech pathology and language learning, where speakers need to adapt their production towards a target. Synthetic samples of acceptable targets could be provided for speakers to aid in acquiring their desired pronunciation. The evaluation of synthetic speech for such applications is hitherto rather uncharted territory and needs suitably adapted performance metrics. An important common characteristic within the above-mentioned use cases is their potential for making modifications, or creations of voices that potentially go beyond those features that are within the original speaker's physiological range (e.g., when performing voice modifications beyond the natural pitch range of an individual voice), or even beyond *any* human speaker's physiological range (e.g., when designing a suitable voice for a non-human artefact). In these cases, no reference may exist for comparison.

A suitable measure of quality therefore may not be the classical quality dimension of "human-likeness", but rather whether the synthetic signal is perceived or measured as suitably "congruent" with the original voice, or with the artefact or concept for which a voice is being designed. In the case of speech pathology or language learning, a framework that captures the intrinsic expert perceptual judgment of the therapist or teacher presents a major challenge. Further evaluation metrics could diagnose whether the intended modifications have indeed been successful, e.g., by performing acoustic-phonetic analyses [22], by letting the speakers themselves evaluate whether a synthesised *hypothetical* speech sample is an acceptable version of what they would have sounded like if they had in fact produced the sample [21], by employing eXplainable Artificial Intelligence (XAI) methodologies to analyse the modifications that led to the resulting speech signals [23], or by developing new metrics of 'appropriateness'.

2.2. Speech-to-speech conversion

Beyond the traditional speech generation from a different kind of input (e.g., text or articulatory movements), various applications require a target speech signal to be generated from a source speech signal. Among these applications are video dubbing, voice anonymization, audio-only translation and simultaneous interpretation. These applications not only have a diverse set of requirements, but they can also be polar opposites. For example, speech-to-speech translation would require the speaker identity to be preserved, while voice anonymization aims at masking this identity while preserving the semantic content.

It becomes evident that speech synthesis evaluation will have to consider such conflicting goals and, while common metrics may be used, new recommendations will have to emphasize the condition of application of these metrics. In order to design an adequate evaluation protocol, it will be important to determine which criteria have to be evaluated for each use case. The following six main criteria can act as a starting point: (1) the target speaker characteristics, (2) the content preservation, (3) the

timing sensitivity, (4) the expressivity of the speech, (5) the real-time factor, and (6) the impact of the non-speech acoustic information from the source signal. Existing directions and protocols are already available. For example, Brannon et al. [24] have already proposed some directions for human dubbing. Detection methodologies such as the Audience Response System (ARS) [25] could be employed to identify local issues and coherency evaluation methodologies [26] can provide insights into the cross-modality requirements of the dubbing use-case.

2.3. Incremental speech synthesis

Most current TTS systems have access to the right-hand context (a full sentence or more) and produce speech based on sentence-level or longer textual input (e.g. long-form synthesis [27]). This speech generation process is modelled on a specific case of human speech: reading aloud. It is suitable when the full text is known in advance and consistency, coherence, and prosodic planning across long sequences are important. In contrast, most everyday human speech is not based on pre-written text. It is situated in a physical and social context, produced under time pressure, and shaped dynamically by interaction with others. This type of speech unfolds incrementally and often adaptively, responding to feedback, turn-taking cues, and environmental factors as it emerges.

Many speech-based applications require this kind of interactive, real-time speech behaviour. Spoken dialogue systems, speech-to-speech translation, and assistive technologies must begin generating speech with minimal delay, possibly before the full input is available. For example, in telephony, ITU-T Rec. G.114 recommends keeping one-way transmission delays below 400 ms to support natural conversational flow. This corresponds roughly to one or two words look-ahead if we do not allow any processing time at all for speech recognition, dialogue management, or translation, *inter alia*. Perhaps more importantly, near-instant responsiveness is a prerequisite for the implementation of many human interactive behaviours, such as convergence [28], Chameleon [29] or Lombard effects [30] as well as adjusting the speech in response to listener attention, hesitation, or backchannels.

Incremental Speech Synthesis (ISS) systems are designed to cope with these requirements. They generate speech from partial and continuously evolving input, operating on smaller units – such as frames, phonemes, syllables, or short text spans – rather than complete sentences. ISS is however challenging [31] and few proposals or systems exist [32, 33, 34].

ISS evaluation involves objective factors that are otherwise not part of speech synthesis evaluations: latency and timing. How soon does speech begin after input is received, and how smoothly does generation continue as more input arrives, how fast does the generation respond to changes? Equally important are the fluency and humanlikeness of the resulting speech, but the requirements vary with the task and the goals of the ISS. Perceptual evaluation criteria that reflect the variability and timing constraints are also added: is the response to some external event *timely*? A responsive system can conceivably be too fast or too slow, and too many repeated changes may impact the listener negatively.

These dimensions call for evaluation approaches that combine instrumentation (to measure timing and latency) with perceptual testing (to assess fluency and timeliness). Logging and aligning input-output timing can help quantify latency and stability, and, for example, Lombard-like adaptation to noise can be evaluated at least in part entirely without human involvement. But in order to confirm that Lombard-style adaptation actually increases comprehension in varying noise, listening tests would have to be conducted. For multimodal tasks, coordination accuracy between speech and for example visual or gestural events can also be evaluated, using either manual annotation or automatic alignment metrics.

2.4. Listening TTS

Complementary to incremental TTS, taking into account the constraints of the environment on-the-fly when producing a synthetic speech signal can enhance the user experience. A typical use-case concerns public announcement in a crowdy/noisy environment such as a train station or an airport. Selecting the speaker as well as the vocal force of the speech signal is particularly relevant in this case. Current TTS technology can already be more precisely controlled by numerous additional inputs such as style or speaker embeddings as well as so-called reference embeddings.

Taking into account the characteristics of the user (i.e., the listener) can also improve the interaction between the dialogue system and the user. Such well-documented phenomena as convergence mechanisms [28], Chameleon [29] or Lombard [30] effects can be modelled this way.

Pushing this reasoning further, TTS should be able to accept continuous driving signals that would monitor hyper/hypo-articulation or phonation as well as more global prosodic patterns in a closed loop.

2.5. (Humanlike) multimodal and embodied spoken interaction

The speech signal is just one component of speech-based interaction. As in humans, when interactive avatars, virtual characters and robots become the vehicle for the deployment of speech synthesis, multimodal cues may be used to augment communication. Facial displays, gaze, hand-gestures, head-nods and other non-verbal signals may all enter into the fray. How they are used will relate back to the use-case or user-story driving the overall system design, but as the role of speech within an overall interaction shifts, so must our perspective in designing the evaluation. To accommodate this dynamic nature, we encourage the definition of a gold standard process for selecting the evaluation practices and associated metrics rather than rigidly defining them. We identify three main goals for multimodal embodied systems: (1) make complementary and efficient use of all modalities involved in communication; (2) establish an identity; (3) serve as a dialogue partner in a social environment.

As an example of complementarity, in the speech-only scenario developers may design a complex speech-based solution to ascertain misunderstanding in a user. If the system can see the user, visual cues in the face may be a better signal to use than the original speech solution. Likewise, in conveying directions, the availability of gesture to point, may greatly enhance the message (e.g., see [35]). Thus, some aspects that were vital in speech-only synthesis evaluation may disappear and other new considerations will arise. Evaluation must capture this in order to adequately assess the speech in this new context.

As for identity, human likeness which is highly prized in most TTS applications is not necessarily a requirement in this use case (also see 2.1). By contrast, consistency between voice and body is crucial to allow distinctiveness and recognition of agents as individual entities (having an agent that has one body associated with multiple voices or the same voice associated with different bodies may be confusing and off-putting). Also,

identification of voice might be part of a brand identity. Plausibility/appropriateness is also a criterion, i.e., the voice matches the embodiment. Perception studies have been conducted to assess whether a voice aligns with a given agent, presented by an image of a human face [36], or a robot [37]. In addition, coherence between voice and visual appearance can be evaluated through open-ended labeling tasks [37], in which separate groups of participants independently describe the auditory and visual stimuli using freely chosen attributes.

Two complementary methods are relevant for evaluating embodied voices: (1) using explicit detailed questionnaires; (2) obtaining implicit global preference about interaction experiences with several voices. In the former case, borrowing metrics from voice coaching would be an interesting direction to investigate. In the latter case, the interlocutor plays the role of a feature extractor, which is highly dependent on his/her social background and environment. These approaches can be carried out off-line (post experiment), or online. In that case, instead of an explicit "yuck button" [38], we can use a "yuck" test (recognizing laughter) to naturally measure participants' recognition of unexpected situations. The choice of evaluators is also highly important, as evaluation from the participants involved in the interaction and of external viewers should be quite complementary. In the latter case, the question of realistic immersion in the interaction can have a great impact on the assessment. Performing these evaluations longitudinally is a means of assessing familiarisation.

2.6. Long-form TTS

Long-form text can include anything from a few sentences to a paragraph, an article, or an entire book. The content may be fictional, necessitating the synthesis of dialogue and character voices, or non-fictional, such as a textbook containing structured information, mathematical formulas, and tables. Compared to shorter texts, the preprocessing and synthesis of long-form content must account for broader contextual information, both linguistically and structurally (see e.g., [27, 39]).

In this context, it becomes necessary to shift evaluation from a simple statistical aggregation of listeners' overall opinions to more diagnostic approaches to identify specific problematic or anomalous pronunciations, prosodic realizations, or parts that were difficult to understand. The following examples illustrate aspects of speech that are especially important to evaluate for the use case of long-form synthesis: (1) *consistency and variability*, referring to the need for a stable speaking rate, style, and pronunciation of names or terms throughout the text, while still allowing for sufficient prosodic variation to sustain interest and correctly communicate the structure of the text; (2) *text contact*, the sense of genuine understanding of the content being read; and (3) *dialogue transitions*, the ability to appropriately signal shifts between narrative and dialogue, as well as between different characters in fictional texts.

While humans are the consumers of synthesized speech and consequently essential for final evaluation, listening tests are costly, especially for long-form synthesis. Therefore, automatic methods can be effectively used for intermediate and diagnostic evaluation, for example, ASR to identify less intelligible sections, LLMs to identify anomalies, and silence detection to identify pauses that are too long. ARS, where listeners indicate perceived shifts, is effective for assessing dialogue transitions. Finally, selecting challenging test passages is crucial to thoroughly evaluate system performance.

2.7. Ethical and Legal Concerns

While not a use case in itself, as speech synthesis goes mainstream, several issues arise regarding not only whether certain capabilities can be created and how effective they are, but also whether it is appropriate or allowable to deploy them in specific contexts. These include cases where similar human speech would be problematic (e.g., incitement to lawless actions, threats, obscenities, defamation, fraud, hate-speech), but also concerns specific to technology, including whether the voice inappropriately makes use of unauthorized training data, audio deepfakes that mimic specific humans without authorization, voices that allow a machine to represent itself as human, or re-use of a familiar voice for new illicit purposes (e.g. a scambot that sounds like a trusted Siri or Alexa). New evaluation methods may be required to determine whether a voice use case is allowable. This may require more transparency or additional augmentations (digital watermarks) that facilitate legal and ethical evaluations while maintaining usability and effectiveness.

3. A brief survey of methods and metrics for speech synthesis evaluation

Given the heterogenous and sometimes directly contradictory aims of the topics and use cases presented in the previous section, it is clear that a single evaluation metric like MOS or MUSHRA is insufficient to effectively evaluate all of them. In this section we review a range of evaluation methodologies and metrics – some long-standing, others newer – that offer more task-sensitive, diagnostic, or contextual insights. We organize these discussions as addressing three fundamental questions regarding the evaluation: when to evaluate, what to evaluate, and within each section, how to evaluate. Table 1 summarizes the relationship between the hot topics introduced in the previous section and the proposed evaluation methods described here.

3.1. When to evaluate

3.1.1. After: Off-line evaluation

Currently, most evaluations (including MOS-tests) happen off-line, after the context in which the speech was used, or out of the context of use. We can distinguish three types of evaluations: (1) **Multidimensional scales**, where listeners are tasked to rate multiple specific dimensions of the speech signal. Examples of dimensions can be *accuracy, pace, phrasing* and *expressivity* [40] for reading fluency assessment. The dimensions to query are usually determined by analysing correlations in ratings of multiple scales by the intended target audience. This is accomplished either via multi-dimensional scaling [41] or factor analysis [42]. (2) **Free vs. semi-directed verbal tagging**, where listener impressions are collected via free text input [43] or semi-directed selection in large pre-defined set of labels [44]. Such tagging is often used to explore subjective impressions for properties or functions of the speech signals with large dimensions (e.g., 412 emotional labels for audiovisual emotions in [45]). (3) **Transcribing**, where listeners act as an annotator by identifying and marking problematic regions of the synthesized speech signal [46].

3.1.2. Before: Gating experiments

Off-line evaluations are usually performed using "complete" verbal content (e.g. sentences, paragraphs, etc). Gating verbal content gives a snapshot of the evaluation of incomplete stim-

Sect.	What	Why	How to evaluate
2.1	Controllable SS	Speech science, digital arts & voice substitution	Task-specific performance measures
2.2	Speech-To-Speech	Dubbing, speech translation, interpretation, voice privacy	Task-specific evaluation
2.3	Incremental TTS	Dialogue systems, speech-to-speech translation, speaking aids	On-line evaluation
2.4	Listening TTS	Sport commentaries, Lombard & adaptive conversation	On-line comprehension
2.5	Embodied interaction	Conversational agents, multiparty interactions	Task-specific evaluation
2.6	Long-form TTS	Audiobooks, synthesis of structured documents	On-line comprehension

Table 1: *Case studies that are worth exploring in the near future, that elicit speech synthesis evaluation methods reviewed in this paper.*

uli and its redundant properties if any. As an example, [47] has shown that the initial portion of a specific acoustic signal is sometimes sufficient to predict properties of the complete signal with good accuracy – phone or word identity, length of a syntagm, etc. – of the whole signal. Similarly [48] have shown that listeners can perceive attitudes before the end of sentences. Such anticipatory encoding of verbal information in the signal is a key property for incremental and listening TTS (see Sections 2.3 and 2.4) as well as anticipating turn-taking, decoding intentions and lowering the listeners' cognitive load. Checking such properties in synthetic signals may therefore also be of relevance.

3.1.3. During: On-line evaluation

Several evaluation methods aim at soliciting raters' judgements as they experience the synthetic multimodal signals (i.e., on-line). This can be performed by end users or third parties – possibly by crowd-sourced workers. Three paradigms deserve attention: (1) **Detection**: the Audience Response System (ARS) [25] or Yuck-responses [38, 49]: Here, raters are asked to press a key each time they perceive a specific event (e.g. phone detection, speaker change, ...) or misbehaviour (e.g. inappropriate intonation, pause length ...) (2) **Continuous ratings**: In time-continuous annotations [50], raters are asked to use sliders to continuously score a specific dimension of the speech signal, such as arousal or valence for the evaluation of emotional speech. Note that these methods require a post-hoc alignment between ratings and recordings, to estimate the reaction time of the raters and analyse signal-dependent contributions [51]. (3) **On-line monitoring**, where behaviours of listeners are monitored when hearing natural vs. synthetic speech. This encompasses close-shadowing delays [52], a comparison between neurophysiological signals (e.g., pupillometry [53]) and monitoring listener behaviors such as facial expressions and laughs ("yucks").

3.2. What to evaluate?

We can consider both black box (looking only at the output speech - Section 3.2.1) and glass box (looking at the internals of models - Section 3.2.3) approaches. We can also consider how speech is evaluated as one of several modes of communication (Section 3.2.2).

3.2.1. Language- and Task-specific evaluation

These measures concern the question of "how well speech synthesis does its job" for the task it is involved in. We should first not forget that the impressive results obtained on major languages do not always extend to low-resourced languages or languages with very different phonological and/or morphosyntactic structures. Recently, several multilingual TTS covering several thousands of languages [54, 55] have been proposed. Here, the devil is in the details: it remains to be seen if language-specific "traps" are really captured by such systems. Another issue concerns the use of speech synthesis in less controlled environments that laboratories. To address this, task-specific measures such as put forward in Human-Computer Interaction (HCI) deserve attention, in particular: **Objective assessment** Performance: Time-to-complete, task-specific errors, learnability, memorization **Subjective assessment** Longitudinal studies should investigate the impact of familiarity and adaptation to individual voices. **Control conditions** Assessment should be performed comparing cohorts of subjects using speech vs. non speech interfaces.

3.2.2. Multimodality

In virtual agents and robots, speech is embedded in a broader communicative context that includes facial expressions, gaze, gestures, and head movements - all of which shape the interaction. Crucially, it is not only the presence of these modalities, but the quality of their coordination with speech that determines communicative effectiveness. This interplay can be evaluated using various methods: (1) **Impoverishment of modalities** particularly by adding noise to speech [56] (2) **Tests of multimodal integration and crossmodal binding** using coherent [26] or incoherent modalities (e.g. McGurk effect in [57]).

3.2.3. Intrinsic evaluation of models

Over the past few years, the adoption of rapidly emerging architectures for each component of a speech synthesis system (encoder, decoder, duration model, speaker and style embeddings, *inter alia*) has led to a remarkable diversity of models. However, evaluating the speech output alone –unless conducted across a large combination of TTS modules or through large-scale ablation studies– offers limited insight into why some modules perform better than others, according to any criteria listed in the article. Conversely, methods that directly evaluate internal representations built by these models provide a means to understand how speech and language are modelled by the implemented architectures. Such evaluations offer valuable insights into how information is encoded, for instance in terms of phonetic encoding [58], controllability of acoustic features [23, 59, 60, 37], long-term dependencies (incremental or long-form TTS) [33], or co-variations between modalities.

4. Conclusions

TTS technology has achieved remarkable progress over the last two decades. The current review of evaluation methods and case studies shows that we are now equipped to challenge new situations with much more diverse evaluation criteria. It is up to us, as a research community, to move beyond what has been our comfort zone and to properly address these new situations.

5. Acknowledgements

This document is one of the outcomes of Dagstuhl seminar 25032, "Task and Situation Aware Evaluation of Speech and Speech Synthesis", which was organised by Jens Edlund, Sébastien Le Maguer, Christina Tånnander and Petra Wagner in January 2025. The document is a joint effort from all participants, who are listed in alphabetical order in the author list, with the exception of the first author position, reserved for the presenter of the work (an homage to the original hot topics paper that inspired our title and structure).

6. References

[1] A. W. Black and K. Tokuda, "The Blizzard Challenge - 2005: Evaluating corpus-based speech synthesis on common datasets," in *Interspeech*, Lisbon, Portugal, September 4-8 2005, pp. 77–80.

[2] S. King and V. Karaiskos, "The Blizzard Challenge 2013," in *Blizzard Challenge Workshop*, Barcelona, Spain, September 3 2013. [Online]. Available: http://www.festvox.org/blizzard/bc2013/summary_Blizzard2013.pdf

[3] C. Benoît, M. Grice, and V. Hazan, "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences," *Speech Communication*, vol. 18, no. 4, pp. 381–392, 1996. [Online]. Available: http://www.sciencedirect.com/science/article/pii/016763939600026X

[4] I. T. Union, "Methods for objective and subjective assessment of quality," International Telecommunication Union, Tech. Rep. ITU-T P.800, August 1996. [Online]. Available: https://www.itu.int/rec/T-REC-P.800-199608-I/en

[5] ——, "Methods for objective and subjective assessment of speech and video quality," International Telecommunication Union, Tech. Rep. ITU-T P.800.1, July 2016. [Online]. Available: https://handle.itu.int/11.1002/1000/12972

[6] O. Perrotin, B. Stephenson, S. Gerber, G. Bailly, and S. King, "Refining the evaluation of speech synthesis: A summary of the blizzard challenge 2023," *Computer, Speech & Language*, vol. 90, p. 101747, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S088523082400130X

[7] S. Le Maguer, S. King, and N. Harte, "The limits of the mean opinion score for speech synthesis evaluation," *Computer, Speech & Language*, vol. 84, p. 101577, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230823000967

[8] S. Zielinski, F. Rumsey, and S. Bech, "On some biases encountered in modern audio quality listening tests-a review," *J. Audio Eng. Soc*, vol. 56, no. 6, pp. 427–451, 2008. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=14393

[9] 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 *Interspeech*. Dresden, Germany: ISCA, September 6-10 2015, pp. 3476–3480.

[10] 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 *ISCA Speech Synthesis Workshop*. Grenoble, France: ISCA, August 26-28 2023, pp. 41–47. [Online]. Available: https://www.isca-speech.org/archive/ssw_2023/kirkland23_ssw.html

[11] F. Seebauer, M. Kuhlmann, R. Haeb-Umbach, and P. Wagner, "Re-examining the quality dimensions of synthetic speech," in *ISCA Speech Synthesis Workshop*. Grenoble, France: ISCA, August 26-28 2023, pp. 34–40.

[12] J. O'Mahony, P. Oplustil-Gallegos, C. Lai, and S. King, "Factors Affecting the Evaluation of Synthetic Speech in Context," in *ISCA Speech Synthesis Workshop*. Budapest, Hungary: ISCA, August 26-28 2021, pp. 148–153. [Online]. Available: https://www.isca-speech.org/archive/ssw_2021/omahony21_ssw.html

[13] 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 *ISCA Speech Synthesis Workshop*, 2019, pp. 2019–19.

[14] E. Cooper, W.-C. Huang, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "A review on subjective and objective evaluation of synthetic speech," *Acoust. Sci. Technol.*, vol. 45, no. 4, pp. 161–183, Jul. 2024.

[15] I. Raharjo, H. Kothare, S. S. Nagarajan, and J. F. Houde, "Speech compensation responses and sensorimotor adaptation to formant feedback perturbations," *The Journal of the Acoustical Society of America*, vol. 149, no. 2, pp. 1147–1161, 2021.

[16] Z. Malisz, G. E. Henter, C. V. Botinhao, O. Watts, J. Beskow, and J. Gustafson, "Modern speech synthesis for phonetic sciences: A discussion and an evaluation," in *International Congress of Phonetic Sciences (ICPhS)*. Australian Speech Science & Technology Association Inc, 2019, pp. 487–491.

[17] F. Rautenberg, M. Kuhlmann, F. Seebauer, J. Wiechmann, P. Wagner, and R. Haeb-Umbach, "Speech synthesis along perceptual voice quality dimensions," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[18] L. Feugère, C. d'Alessandro, B. Doval, and O. Perrotin, "Cantor digitalis: chironomic parametric synthesis of singing," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2017, pp. 1–19, 2017.

[19] R. K. Moore, "Is spoken language all-or-nothing? implications for future speech-based human-machine interaction," *Dialogues with social robots: enablements, analyses, and evaluation*, pp. 281–291, 2017.

[20] T. Mills, H. T. Bunnell, and R. Patel, "Towards personalized speech synthesis for augmentative and alternative communication," *Augmentative and Alternative Communication*, vol. 30, no. 3, pp. 226–236, 2014.

[21] S. Strömbergsson, Å. Wengelin, and D. House, "Children's perception of their synthetically corrected speech production," *Clinical Linguistics & Phonetics*, vol. 28, no. 6, p. 373–395, Jan. 2014. [Online]. Available: http://dx.doi.org/10.3109/02699206.2013.868928

[22] A. Pandey, J. Edlund, S. Le Maguer, and N. Harte, "Listener sensitivity to deviating obstruents in wavenet," in *Interspeech*, Dublin, Ireland, 2023, pp. 1080–1084.

[23] M. Lenglet, O. Perrotin, and G. Bailly, "A closer look at internal representations of end-to-end Text-to-Speech models: How is phonetic and acoustic information encoded?" 2025. [Online]. Available: http://dx.doi.org/10.2139/ssrn.5217280

[24] W. Brannon, Y. Virkar, and B. Thompson, "Dubbing in practice: A large scale study of human localization with insights for automatic dubbing," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 419–435, 05 2023. [Online]. Available: https://doi.org/10.1162/tacl_a_00551

[25] C. Tånnander, J. Edlund, and J. Gustafson, "Revisiting three text-to-speech synthesis experiments with a web-based audience response system," in *Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, 2024, pp. 14 111–14 121.

[26] A. Vatakis and C. Spence, "Crossmodal binding: Evaluating the "unity assumption" using audiovisual speech stimuli," *Perception & psychophysics*, vol. 69, pp. 744–756, 2007.

[27] W. Zhang, C.-C. Yeh, W. Beckman, T. Raitio, R. Rasipuram, L. Golipour, and D. Winarsky, "Audiobook synthesis with long-form neural text-to-speech," in *ISCA Speech Synthesis Workshop*, 2023.

[28] J. H. Manson, G. A. Bryant, M. M. Gervais, and M. A. Kline, "Convergence of speech rate in conversation predicts cooperation," *Evolution and Human Behavior*, vol. 34, no. 6, pp. 419–426, 2013.

[29] T. L. Chartrand and J. A. Bargh, "The chameleon effect: The perception–behavior link and social interaction." *Journal of personality and social psychology*, vol. 76, no. 6, p. 893, 1999.

[30] S. A. Zollinger and H. Brumm, "The lombard effect," *Current Biology*, vol. 21, no. 16, pp. 614–615, 2011.

[31] A. Köhn, "Incremental natural language processing: Challenges, strategies, and evaluation," *arXiv preprint arXiv:1805.12518*, 2018.

[32] H. Buschmeier, T. Baumann, B. Dosch, S. Kopp, and D. Schlangen, "Combining incremental language generation and incremental speech synthesis for adaptive information presentation," in *Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2012, pp. 295–303.

[33] B. Stephenson, L. Besacier, L. Girin, and T. Hueber, "What the Future Brings: Investigating the Impact of Lookahead for Incremental Neural TTS," in *Interspeech*, Shanghai, China, October 25–29 2020, pp. 215–219.

[34] T. Saeki, S. Takamichi, and H. Saruwatari, "Incremental text-to-speech synthesis using pseudo look-ahead with large pretrained language model," *IEEE Signal Processing Letters*, vol. 28, pp. 857–861, 2021.

[35] R. Younes, F. Elisei, D. Pellier, and G. Bailly, "Impact of verbal instructions and deictic gestures of a cobot on the performance of human coworkers," in *2024 IEEE-RAS 23rd International Conference on Humanoid Robots (Humanoids)*. IEEE, 2024, pp. 1040–1047.

[36] P. van Rijn, S. Mertes, D. Schiller, P. Dura, H. Siuzdak, P. M. C. Harrison, E. André, and N. Jacoby, "Voiceme: Personalized voice generation in TTS," in *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 2588–2592. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-10855

[37] P. van Rijn, S. Mertes, K. Janowski, K. Weitz, N. Jacoby, and E. André, "Giving robots a voice: Human-in-the-loop voice creation and open-ended labeling," in *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024*, F. F. Mueller, P. Kyburz, J. R. Williamson, C. Sas, M. L. Wilson, P. O. T. Dugas, and I. Shklovski, Eds. ACM, 2024, pp. 584:1–584:34. [Online]. Available: https://doi.org/10.1145/3613904.3642038

[38] R. Poppe, K. P. Truong, and D. Heylen, "Backchannels: Quantity, type and timing matters," in *International workshop on intelligent virtual agents (IVA)*. Springer, 2011, pp. 228–239.

[39] R. Clark, H. Silen, T. Kenter, and R. Leith, "Evaluating Long-form Text-to-Speech: Comparing the Ratings of Sentences and Paragraphs," in *10th ISCA Workshop on Speech Synthesis (SSW 10)*. ISCA, 2019, pp. 99–104.

[40] T. V. Rasinski, "Assessing reading fluency." *Pacific Resources for Education and Learning (PREL)*, 2004.

[41] C. Mayo, R. A. J. Clark, and S. King, "Multidimensional scaling of listener responses to synthetic speech," in *Interspeech*, 2005, pp. 1725–1728.

[42] F. Hinterleitner, G. Neitzel, S. Möller, and C. Norrenbrock, "An evaluation protocol for the subjective assessment of text-to-speech in audiobook reading tasks," *Blizzard Challenge Workshop*, 2011.

[43] Y. Greenberg, N. Shibuya, M. Tsuzaki, H. Kato, and Y. Sagisaka, "Analysis on paralinguistic prosody control in perceptual impression space using multiple dimensional scaling," *Speech Communication*, vol. 51, no. 7, pp. 585–593, 2009.

[44] G. Bailly, R. Legrand, M. Lenglet, F. Elisei, M. Garnier, and O. Perrotin, "Emotags: Computer-assisted verbal labelling of expressive audiovisual utterances for expressive multimodal tts," in *Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, 2024, pp. 5689–5695.

[45] W. Junek, "Mind reading: The interactive guide to emotions," *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, vol. 16, no. 4, p. 182, 2007.

[46] E. Gutierrez, P. Oplustil-Gallegos, and C. Lai, "Location, Location: Enhancing the Evaluation of Text-to-Speech synthesis using the Rapid Prosody Transcription Paradigm," in *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 25–30.

[47] F. Grosjean, "Spoken word recognition processes and the gating paradigm," *Perception & psychophysics*, vol. 28, no. 4, pp. 267–283, 1980.

[48] V. Aubergé, T. Grepillat, and A. Rilliard, "Can we perceive attitudes before the end of sentences? the gating paradigm for prosodic contours," in *Eurospeech*, 1997, pp. 871–874.

[49] I. A. de Kok, "Listening heads," Ph.D. dissertation, University of Twente, Twente, The Netherlands, 2013. [Online]. Available: https://ris.utwente.nl/ws/portalfiles/portal/6032366/thesis_I_de_Kok.pdf

[50] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, 2011.

[51] S. Mariooryad and C. Busso, "Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations," in *International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2013, pp. 85–90.

[52] G. Bailly, "Close shadowing natural versus synthetic speech," *International Journal of Speech Technology*, vol. 6, pp. 11–19, 2003.

[53] A. Govender and S. King, "Using pupillometry to measure the cognitive load of synthetic speech," *System*, vol. 50, no. 100, pp. 2018–1174, 2018.

[54] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi *et al.*, "Scaling speech technology to 1,000+ languages," *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.

[55] F. Lux, S. Meyer, L. Behringer, F. Zalkow, P. Do, M. Coler, E. A. P. Habets, and N. T. Vu, "Meta learning text-to-speech synthesis in over 7000 languages," in *Interspeech*, 2024, pp. 4958–4962.

[56] C. Benoît and B. Le Goff, "Audio-visual speech synthesis from french text: Eight years of models, designs and evaluation at the icp," *Speech Communication*, vol. 26, no. 1, pp. 117–129, 1998. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639398000454

[57] D. Cosker, S. Paddock, D. Marshall, P. L. Rosin, and S. Rushton, "Toward perceptually realistic talking heads: Models, methods, and mcgurk," *ACM Transactions on Applied Perception (TAP)*, vol. 2, no. 3, pp. 270–285, 2005.

[58] A. R. Vaidya, S. Jain, and A. Huth, "Self-supervised models of audio effectively explain human cortical responses to speech," in *International Conference on Machine Learning (ICML)*, vol. 162, July 17-23 2022, pp. 21 927–21 944. [Online]. Available: https://proceedings.mlr.press/v162/vaidya22a.html

[59] S. Sadok, S. Leglaive, L. Girin, X. Alameda-Pineda, and R. Séguier, "Learning and controlling the source-filter representation of speech with a variational autoencoder," *Speech Communication*, vol. 148, pp. 53–65, 2023.

[60] M. Jacquelin, M. Garnier, L. Girin, R. Vincent, and O. Perrotin, "Exploring the multidimensional representation of unidimensional speech acoustic parameters extracted by deep unsupervised models," in *IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. Seoul, Korea: IEEE, April 15 2024.