Report from Dagstuhl Seminar 25032

Task and Situation-Aware Evaluation of Speech and Speech Synthesis

Jens Edlund[*1], Sébastien Le Maguer[*2], Christina Tånnander[*3], Petra Wagner[*4], and Fritz Michael Seebauer[†5]

1 **KTH Royal Institute of Technology – Stockholm, SE.** edlund@speech.kth.se
2 **University of Helsinki, FI.** sebastien.lemaguer@helsinki.fi
3 **Swedish Agency for Accessible Media – Malmö, SE.** christina.tannander@mtm.se
4 **Universität Bielefeld, DE.** petra.wagner@uni-bielefeld.de
5 **Universität Bielefeld, DE.** fritz.seebauer@uni-bielefeld.de

Abstract

Speech synthesis has now reached such human-likeness that its evaluation as a separate entity is no longer meaningful. In this Dagstuhl Seminar, we approach speech and speech synthesis evaluation from a multidisciplinary perspective. Our goal has been to establish a core network to reach all impacted research communities and to provide fundamental directions to develop the new standards of speech and speech synthesis evaluation.

Seminar January 12–15, 2025 – https://www.dagstuhl.de/25032
2012 ACM Subject Classification Human-centered computing → HCI design and evaluation methods; Computing methodologies → Natural language processing; Human-centered computing → Accessibility design and evaluation methods
Keywords and phrases evaluation, human-in-the-loop, speech technology, speech-to-text synthesis
Digital Object Identifier 10.4230/DagRep.15.1.84

1 Executive Summary

Jens Edlund
Sébastien Le Maguer
Christina Tånnander
Petra Wagner

This report documents the program and the outcomes of Dagstuhl Seminar "Task and Situation-Aware Evaluation of Speech and Speech Synthesis" (25032).

The recent advances in deep neural netowrks have pushed the boundaries for synthetic speech to the point where synthetic speech is, in some contexts, indistinguishable from human speech. Alongside a slew of well-known issues with deep fakes, this development raises fundamental questions concerning the evaluation of synthetic speech and its relation to the evaluation of human speech. Human speech and synthetic speech have traditionally been evaluated in different ways, with human speech often serving as an implicit or explicit gold standard for synthetic speech. At the same time, the technical distinction between synthetic and human speech is getting increasingly blurred: human speech is delivered through encoding/decoding processes that changes the signal fundamentally – most notably

[*] Editor / Organizer
[†] Editorial Assistant / Collector

in applications such as speech-to-speech translation and anonymisation – and voice cloning of recorded speech passes as speech synthesis. We hold that the fundamental question when evaluating speech these days is no longer "How similar to human performance is this?", but rather: "Is this "good" speech?"

The issue is made more complex still by the fact that what constitutes good speech, synthetic or human, is not a trivial question. Finally, standard evaluation methodologies fail to take into account the interdiciplinary nature of speech science and and speech technology through its assumption that one single evaluation metric should satisfy all requirements.

The goal of the Dagstuhl Seminar "Task and situation-aware evaluation of speech and speech synthesis" (25032) was to initiate shift in the different communities impacted by these evolutions. To do so, we gathered a total of 22 renowned researchers from various disciplines (among others: engineering, phonetics, user interface, computer science, speech pathology) to exchange about this fundamental issue. Exchange between groups was encouraged by organisating the seminar around working groups designed to explore the breadth of research fields and applications with stakes in speech and synthetic speech evaluation. This was also intended to encourage more active involvement of the participants both during and after the seminar. This hands-on approach came at the expense of formal talks and panel discussions, which were limited to two talks, both intended to get background "out of the way" and to allow the rest of the discussions to focus on the future.

In the present executive summary, our goal is twofold. Firstly, the presentation of the immediate outcomes of the dicussions that took place during the seminar. In addition, we are convinced that the manner in which the seminar was organized represents an contribution as it presents a process towards fruitful trans- and interdisciplinary exchange leading to a long term impact.

Each day of the three-day seminar was given a broad goal: The first day focussed on background, the second on innovation and solutions, and the third on consolidation and structuring. Each day was further divided into sessions, each with a specific result in mind.

Following a general introduction to the seminar, including Dagstuhl practicalities, the first day started out with three-minute participant presentations. As most of our participants have a long and broad set of experiences in the speech field, they were encouraged to focus their presentations on matters of direct relevance to the seminar, and they were also asked to specify their own interests in relation to the seminar description. Since this seminar was designed to gather and reconcile as much as possible of the collective experiences in speech and speech synthesis evaluation, with few presentations and much discussion and collaborative work, we include all personal statements in this report 4.1. They constitute a fair representation of the morning session. After a following discusssion and a brief introduction to the afternoon sessions, the morning session was concluded.

The afternoon session continued and concluded what we view as the background work in this seminar. First, two talks presented the limits of the current state-of-the-art in speech synthesis evaluation methodologies. A longer session dedicated to an extensive exchange about these methodologies and their limits followed the talks. Although these shortcomings were are well-known within the group of participants, experience has shown that discussion on improvements and evaluation innovation easily get stuck in repeated discussions about the shortcomings of current methods. Our goal was to get these discussions out of the way on day one, and then explicitly avoid going back in days two and three, to ensure that the remaining time of the seminar was dedicated to the exploration of new horizons.

At the end of day one, the organisers presented a set of speech and speech technology areas with suggested use cases, together with group assignments for all participants, in order to allow the participants to muse over these in preparation for the second day.

The second day was dedicated to discover and explore new directions for speech and speech synthesis evaluation. This day was structured around four working group sessions interspersed with flash presentation plenary sessions. The goal of these plenary sessions was not only to fuel the discussion of the following sessions, but also to inform all the participants of the reflexion of each working group. The first three sessions were centred around high-level use cases known to be impacted by the recent evolution in speech synthesis.

The organizers defined the groups by taking into account the background and the interests of the participants as communicated in their personal statements. This strategy proved to be effective as no participants requested to change group. While the first two sessions aimed at developing the initial use cases and what falls under each use case, the third session focused on exploring potential methodologies. Note that the use cases, here, served primarily as a focal point for discussion, designed to capture specific TTS and speech characteristics and requirements as well as cover specific types of applications. Thus the goal was not to create fully fledged and ready-to-use protocols, but to explore what constraints and requirements future methodologies will have to meet.

Informed by these meetings, the organizers then defined a set of five methodology umbrellas which emerged from different use cases, and the last session was dedicated to exploring these umbrellas. For this session, the participants were left free to select the group to join. While this led to an uneven distribution, this also provided space for the participants to focus on some of their more immediate interests.

The last session of the day was a plenary general discussion of the day's events, aiming to allow participants to bring up points of criticism and complementary information.

The last day was dedicated to clean, collect, summarize the information produced previous day, with a clear aim at future work and practical ways to continue the work discussed in the seminar. The result was a less intensive and the organisers also decided to give more freedom for the participants to determine concrete activities they wanted to participate. Working group formed spontaneously to establish and engage in short term solutions to address the current flaws in speech evaluation.

The immediate outcome of the seminar is the establishment of a core network of renowned researchers from various disciplines dedicated to speech and speech synthesis evaluation. Due to its multidisciplinary and breadth, this community can reach a range of different research communities.

A major achievement in the wake of the seminar is a set of new guidelines for reviewing TTS papers with respect to evaluation. Members of the network are part of the organisation committees of Interspeech – the reference conference about speech science and speech technology – and the Speech Synthesis Workshop (SSW), and have promoted – for Interspeech – or enforced – for SSW the use of these guidelines. In addition, a discussion on edits and amendments to the ITU reports on TTS is currently undertaken with ITU.

Several papers are also underway as direct results of the seminar, including a position paper to SSW about the new directions in speech synthesis evaluation and a more substantial survey article to the journal Computer, Speech & Language (CSL).

To increase awareness, we are currently exploring other ways of dissemination such as designing tutorials to be presented at Interspeech 2026, as well as a repeating work shop series. We will also propose a special issue of the journal CSL dedicated to speech and speech synthesis evaluation.

In the longer term, the goal reaches far beyond documenting the current state of speech and speech synthesis evaluation, towards a dynamic process that avoids renewed fossilisation. We believe we have achieved this on three fronts. First, the seminar ensured the transmission

of information between generation of researchers. This is necessary to keep the field cohesive. Second, the seminar brought researchers both from academia and industry. This is critical to ensure that balanced is maintain between the different interests. Finally, the seminar provided the space to not only get new directions but also to have a core set of renowned researchers whose duty is now to impulse this change in their respective communities armed with the different resources provided by the activities of the seminar.

2 Table of Contents

3 Overview of Talks

3.1 Day 1, session 1b. MOS and its limitations and biases

Erica Cooper (NICT – Kyoto, JP)

Main reference Erica Cooper, Junichi Yamagishi: "Investigating Range-Equalizing Bias in Mean Opinion Score Ratings of Synthesized Speech", in Proc. of the 24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023, pp. 1104–1108, ISCA, 2023.
URL https://doi.org/10.21437/INTERSPEECH.2023-1076

This talk introduced the Mean Opinion Score (MOS) listening test protocol for evaluating synthesized speech, as well as some of the limitations and biases of this protocol. We described the collection of a large-scale MOS dataset covering synthesis systems from 2008-2022 and what we learned about listener preferences and the evolution of speech synthesis technology from this dataset. We presented the VoiceMOS Challenge, a shared task challenge for automatic evaluation of synthesized and processed speech, along with the lessons we learned from three years of running the challenge about which approaches work well for automatic prediction as well as what makes the task difficult. Lastly, we presented a case study on range-equalizing bias, which is the tendency of listeners to use the entire range of the rating scale regardless of the absolute quality of the samples presented to them, and demonstrated that MOS ratings for the same system can vary by over one point depending on the relative quality of the other samples included in the test.

3.2 Day 1, session 1b. The limits of the Mean Opinion Score for speech synthesis evaluation

Sébastien Le Maguer (University of Helsinki, FI)

Joint work of Sébastien Le Maguer, Naomi Harte, Simon King
Main reference Sébastien Le Maguer, Simon King, Naomi Harte: "The limits of the Mean Opinion Score for speech synthesis evaluation", Comput. Speech Lang., Vol. 84, p. 101577, 2024.
URL https://doi.org/10.1016/J.CSL.2023.101577

Speech synthesis has reached unprecedented quality, but its evaluation relies on methodologies developed more than two decades ago. Among these protocols, the Mean Opinion Score (MOS) test remains the overwhelming used standard. While not without controversy, and as contemporary synthesis systems produce speech remarkably close to human speech, it is now vital to determine how reliable this score is.

In this talk, I will present a series of four experiments to question MOS. With these experiments we address the following questions: How stable is the MOS of a system across time? How do the scores of lower quality systems influence the MOS of higher quality systems? How does the introduction of modern technologies influence the scores of past systems? How does the MOS of modern technologies evolve in isolation?

The outcome of our experiments confirms the relative nature of MOS. It also suggest that we may have reached the end of a cul-de-sac with current evaluation methodologies and that we are in critical need to develop more suited protocols.

4 Working groups

4.1 Day 1, session 1a. The people – collected personal statements

Jens Edlund (KTH Royal Institute of Technology – Stockholm, SE), Elisabeth André (Universität Augsburg, DE), Gérard Bailly (University Grenoble Alpes, FR), Erica Cooper (NICT – Kyoto, JP), Benjamin Cowan (University College – Dublin, IE), Naomi Harte (Trinity College Dublin, IE), Simon King (University of Edinburgh, GB), Esther Klabbers (Beaverton, US), Sébastien Le Maguer (University of Helsinki, FI), Zofia Malisz (KTH Royal Institute of Technology – Stockholm, SE), Bernd Möbius (Universität des Saarlandes, DE), Sebastian Möller (TU Berlin, DE & DFKI Berlin, DE), Roger K. Moore (University of Sheffield, GB), Ayushi Pandey (Trinity College Dublin, IE), Olivier Perrotin (University Grenoble Alpes, FR), Fritz Michael Seebauer (Universität Bielefeld, DE), Sofia Strömbergsson (Karolinska Institute – Stockholm, SE), Christina Tånnander (Swedish Agency for Accessible Media – Malmö, SE), David R. Traum (USC – Playa Vista, US), Petra Wagner (Universität Bielefeld, DE), Junichi Yamagishi (National Institute of Informatics – Tokyo, JP), and Yusuke Yasuda (Nagoya University, JP)

Since this seminar was designed to gather and reconcile as much as possible of the collective experiences in speech and speech synthesis evaluation, with few presentations and much discussion and collaborative work, we include all personal statements in this report, as they give a picture of the diversity of the issues at hand.

4.2 Day 2, session 1. Use case: Speech variation and training

Naomi Harte (Trinity College Dublin, IE), Fritz Michael Seebauer (Universität Bielefeld, DE), and Sofia Strömbergsson (Karolinska Institute – Stockholm, SE)

A shared feature of the scenarios here can be summarized as the following challenge:

What is the difference/distance between a given exemplar (what a speaker/TTS system is currently able to produce), and a currently un-realized target exemplar (what a speaker/TTS system is supposed to achieve).

To complicate things, there are a **multitude of potential target-appropriate exemplars**; in other words, there is a range of variation that is acceptable among exemplars meeting the requirements of "having achieved the target".

Evaluation of speech comes into play in two distinct but closely related ways here. We want to use TTS to generate an appropriate set of exemplars for the given user and scenario. This requires that in the absence of the ideal reference, we can assess whether a generated exemplar meets the required standard. This could be how e.g. certain vowels should be realised in the low-resource language, or how the voice of the child with a specific speech order should progress. The second way evaluation occurs is when, given the exemplars,

we then compare those to what is currently produced by the speaker/system. Within an acceptable level of variation, we need to judge how well the speaker/TTS is approaching production levels in the exemplars.

Thus a shared feature of the scenarios is that the **evaluation requires some kind of expert knowledge**; for TTS in a low-resource language, the expert knowledge required is being a native (native-like?) speaker of the target language. The evaluation of whether a speaker with a speech disorder or a speaker learning an L2 has reached (or is approaching) the target, requires expert listeners like an SLP or a language teacher, respectively. To achieve automatic evaluation therefore requires that we somehow capture what the expert would look out for, and embed that into the evaluation protocol. We believe this part is a major challenge to get right, and also essential.

Presumably, **the evaluation criteria will rely on expert-relevant parameters** referring to phonetics, phonology, rhythm, intonation, etc.

The **evaluation task for the expert is highly context-dependent** – the expert evaluator might, for example, need to know the age of the speaker (and might have a more lenient criteria for what is a target-appropriate exemplar for a younger speaker/child voice than for an older speaker). This is needed because the target production of the use cases depends on the speakers current level of capability which needs to be assessed. Thus we have a situation of a **moving target, as our required output or expectation of what can be produced is changing with time**.

4.3 Day 2, session 1. Use case: Speech-to-speech

Simon King (University of Edinburgh, GB), Sébastien Le Maguer (University of Helsinki, FI), Sebastian Möller (TU Berlin, DE & DFKI Berlin, DE), and Junichi Yamagishi (National Institute of Informatics – Tokyo, JP)

In addition to the production of speech using a different input material (e.g., text), a various set of applications requires to produce a target speech signal given a source speech signal. Among these applications, we considered three use cases: video dubbing, voice privacy, audio-only translation and simultaneous interpretation. Video dubbing consist of introducing a new audio track to a video. The most well known example is the adaptation of a movie to a new language. At the opposite of the audio-only translation, which consists of transforming a speech signal from one language to another language (e.g., podcast, radio program), this also implies to ensure a consistency between the video and the new audio track. Closely related is the simultaneous interpretation which consists of creating a speech signal in real time. A concrete example of application is when large summits are hold in multiple languages such as the UN or the EU assemblies. Finally, voice privacy differs from the previous use-cases. It requires the exact content of the message carried by the source speech signal to be transferred to the target signal. This should be done by ensuring that the speaker identity of the source signal is stripped from the target signal. All these use cases, except voice privacy, do not require speech technology and can be achieved using human speakers (e.g., voice actors for dubbing).

Each of these use cases has its own challenges. For example, it is imperative that part of the speaker identity is preserved for video dubbing, while for voiced privacy this identity should be stripped. Similarly, in the context of an audio only translation the speech-to-speech conversion can be achieved offline, while in the context of simultaneous interpretation the latency of the system will have an inpact on the comprehensibility of the interpretation. Even if we consider a system to be developped only for one use case, a "one fit all" evaluation protocol remains out of reach. For example, when producing a speech signal for dubbing, not only the speech signal has to be intelligibile but the synchronicity between the audio and the video should be preserved.

In order to design an adequate evaluation protocol, it will be important to determine which criteria have to be evaluated for each use case. During the working group, we identified six main criteria: the target speaker characteristics, the content preservation, the timing sensitivity, the expressivity of the speech, the real-time factor, and the impact of the non-speech acoustic information from the source signal. For each combination criterion/use case, different factors needs to be ensured. Table 1 summarizes the factors that we determined during the working group. This summary is non-exhaustive and should be seen as a starting point to determine what are the main points of evaluation.

There is a set of use cases where the source input is speech and the target output is speech. Source and target might be in the same or different languages. Some attributes, such as speaker identity, might be retained from source to target, or intentionally distinct. We considered dubbing videos, voice privacy, and audio-only translation. All of these can be achieved using natural speech, without technology. Applying speech technologies such as TTS or voice conversion was of course initially motivated by reducing cost. But technology has other advantages, including a lower barrier to entry (e.g., by reducing the skill level needed to produce a dub), increased speed/throughput of production, and a potentially wider portfolio of voices.

Table 1 Summary of factors to be evaluating when conducting an evaluation campaign dedicated to speech-to-speech use cases. All cells in bold correspond to critical point to be evaluated. If these are not assessed, the evaluation will be considered invalid.

CriterionUse Case	**Dubbing**	**Voice privacy**	**Translation**	**Interpretation**
Target speaker characteristics	Preserve Speaker ID (or) Fit to actor (or) **Consistency**	**Speaker ID masked**	Chosen by content producer	
Content	**Semantic fit** Artistic fit	Emotion recognition ER	**Semantic fit**	**Semantic fit**
Timing	**No overlap** Sync to video, lip sync			Ideal timing
Expressivity	Analogue to the source	Preserve within language **Mask speaker expressivity**	Analogue to the source	Neutrality
Non-Speech audio	Keeping other information Possibly diarization performance Noise robustness	Noise robustness	noise robustness	noise robustness
Real-Time Factor	offline	offline or online	offline	**online**

4.4 Day 2, session 1. Use case: Lengthy materials read aloud

Esther Klabbers (Beaverton, US), Erica Cooper (NICT – Kyoto, JP), Christina Tånnander (Swedish Agency for Accessible Media – Malmö, SE), and Yusuke Yasuda (Nagoya University, JP)

In Merriam-Webster, long-form is defined as "notably long in form in comparison to what is common or typical for works or content of a particular category". Since most TTS-generated outputs consists of single utterances or short sentences, "long-form" in a TTS context would include anything longer than a couple of sentences, for example a paragraph, news article or an entire book. Consequently, speech being evaluated need to have some degree of external validity to reflect a real-world reading/listening situation. Our discussions of what aspects of the speech to evaluate included the following topics:

Consistency and variability: The read speech should maintain consistency in audio quality, speaking rate and speech styles, as well as in the pronunciations of for example proper names and terms throughout the text. At the same time, the speech needs to be prosodic variation is crucial to ensure the speech remains comprehensibility and avoids causing listening fatigue.

Text contact: Whether delivered by a human narrator or TTS, the reader should exhibit what we here refer to as text contact: a sense of genuinely understanding the content being read.

Read dialogues: In fictional works, dialogue is a common feauture. It is essential to appropriately signal transitions between narrative text and dialogue, as well as between different characters within the dialogue.

4.5 Day 2, session 2. Methods

Sébastien Le Maguer (University of Helsinki, FI), Elisabeth André (Universität Augsburg, DE), Gérard Bailly (University Grenoble Alpes, FR), Erica Cooper (NICT – Kyoto, JP), Benjamin Cowan (University College – Dublin, IE), Jens Edlund (KTH Royal Institute of Technology – Stockholm, SE), Naomi Harte (Trinity College Dublin, IE), Simon King (University of Edinburgh, GB), Esther Klabbers (Beaverton, US), Zofia Malisz (KTH Royal Institute of Technology – Stockholm, SE), Bernd Möbius (Universität des Saarlandes, DE), Sebastian Möller (TU Berlin, DE & DFKI Berlin, DE), Roger K. Moore (University of Sheffield, GB), Ayushi Pandey (Trinity College Dublin, IE), Olivier Perrotin (University Grenoble Alpes, FR), Fritz Michael Seebauer (Universität Bielefeld, DE), Sofia Strömbergsson (Karolinska Institute – Stockholm, SE), Christina Tånnander (Swedish Agency for Accessible Media – Malmö, SE), David R. Traum (USC – Playa Vista, US), Petra Wagner (Universität Bielefeld, DE), Junichi Yamagishi (National Institute of Informatics – Tokyo, JP), and Yusuke Yasuda (Nagoya University, JP)

4.5.1 Best Practices

This group was interested in coming up with meta-guidelines to help researchers choose and develop evaluation methods for speech synthesis applications. One immediate challenge, lies in the fact that evaluation needs often, if not always, be tailored towards the use case it is supposed to measure. Following this, the guidelines outlined here need to be rather general, to still be applicable for all possible use case evaluations. We structure this abstracts in two different parts: First there is a non-exhaustive list of potential recommendations and second is a list of suggestions how to provide resources that aid researchers in implementing them. In listing the recommendations, we do not restrict ourselves by catering towards implementation problems of disseminating them in specific communities, but rather aim to provide a topline. The first guideline refers to the statistical validity of proposed methods. This mirrors the provided guidelines in adjacent fields like psychology. Examples for criteria to consider include adequate sample sizes for the number of investigated factors. The next point also contains scientific standards as a whole. For example the order randomisation of stimuli, the choice of adequate baseline references and control conditions, or the general fit of experiment design. We also consider that a plea towards open science would be sensible, suggesting the open sourcing of evaluation material and analysis code. In a less general recommendations we conclude that if possible, the use case of a given system should be considered when choosing or designing an evaluation protocol. On that same note, language considerations play into adapting already existing scales. These are findings borrowed from the field of psychometric research, that concepts might not be translateable between cultures and languages, and thusly scales would have to be re-developed or adapted if used in other contexts. This extends to older scales, that were designed to be valid for specific distributions of stimuli and might either be adapted for newer systems, or statistically corrected in the analysis phase. Most of these recommendations are geared towards individuals or institutions that have the time and resources to concern themselves with the pitfalls of designing proper evaluation protocols. Admitting that this is not always the reality of scientific research our last suggestion would be to consider collaboration or outsourcing the evaluation to ensure a certain quality standard is being kept. Moving into the topic of how to provide resources for

aspiring TTS researchers to use, when deciding to evaluate their systems. We consider using a linear, or branched, web-form template which instructs the user with guided questions akin to a decision tree. A similar resource could be a database which contains examples of different evaluation kinds and reference papers of confirmed quality, that could be queried by researchers looking for guidance. In terms of community work, we propose to somehow identify and showcase particularly well executed examples of evaluation, for example in form of an award. A more prescriptive approach would see the requirement of a standard in an evaluation section that should fulfill specific criteria, analogous to the well established standards of having a background in an introduction. Finally we consider what the specific meta-criteria might be to determine these "good" examples. It might be the thoroughness of the reporting done on all aspects of the decision making process, or the inventiveness with regards to already established methods.

4.5.2 Meta-structure

Starting from the use cases and the evaluation proposals. How can these be characterised, categorised?

Can the proposed evaluation be used for other use cases with no or little changes? Which ones? What do they have in common?

How do other evaluations fall within the same system?

Since I'm the only person turning up for this Group, I'm taking the liberty of promoting the evaluation "meta" taxonomy I've used in the past and which underpinned the EAGLES standards and resources activities coupled with more recent characterisation of the speech technology field ...

In particular, I suggest that is useful to partition use-cases into four broad categories according to whether a given scenario involves interactivity and/or requires real-time processing. For example, in a two-dimensional plot, voice-enabled artefacts would fall in the interactive/real-time quadrant, automated announcements would fall in the non-interactive/real-time quadrant, film dubbing would fall in the non-interactive/non-real-time quadrant, and long-form reading would fall in the interactive/non-real-time quadrant. (Note: I have a LaTeX template suitable for this diagram – slide 4 in my presentation.)

For real-time/interactive use cases, I also recommend distinguishing between three behavioural domains:

- the physical domain of objects and actions (for mechanical support),
- the information domain of knowledge and data (for cognitive support),
- the social domain of agents and relations (for emotional support).

Interaction in the physical and information domains typically involves formulaic speech acts – "command-and-control" or "question-and-answer" respectively – which usually conform to a strict "turn-taking" protocol for dialogue. Interaction in the social domain involves more fluid conversational behaviour with considerable overlap between interlocutors. These domains are not mutually-exclusive, hence any particular use-case will have a balance of requirements across all three domains. (Note: I have a diagram that captures this – slide 5 in my presentation.)

Of course, there are many factors which influence the ultimate performance of spoken language systems. This means that, not only is it is necessary to distinguish between "capabilities" and "requirements" of the components technologies (such as speech synthesis), but it also important to emphasise that the purpose of introducing spoken language technology into an application is to achieve the appropriate operational benefits. However, successful

implementation of spoken language systems depends only indirectly on the technical features of the system components. The anticipated application benefits need to be expressed in terms of application requirements which in turn need to be expressed as technical requirements. On the other side of the coin, the features of the technology need to be expressed in terms of technical capabilities which in turn need to be expressed as application capabilities. Both the technical and application capabilities/requirements are multi-dimensional in nature and thus require assessment of "goodness" across a set of relevant application and technical factors. (Note: I have a diagram that captures this – slide 6 in my presentation.)

Finally, it is important to acknowledge that a key "goodness" parameter (alongside obvious dimensions such as intelligibility) is the "appropriateness" of a particular synthetic voice to a given communicative context. In this regard it is useful to distinguish the (dynamic) situational context from the (static) embodied context. The latter would likely be motivated by suitable design principles/priors, and thus evaluation (and therefore, optimisation) could be performed off-line. The former implies synthesis that is reactive to changing conditions, and thus evaluation (and therefore, optimisation) would require on-line evaluation.

4.5.3 Users and user expectations

An important consideration for the evaluation of speech synthesis systems is the selection of a set of human raters that is representative for the population of end users of the system. Options range from drawing random samples from the general population if the target application of the system is a general-purpose speech synthesis system, to a system that serves the need of users with specific properties or needs. Examples for the latter are people with physiological (e.g., vision, hearing, reading) or neurological impairments (e.g., aphasia, personality disorders, people who would benefit from Easy Language). For these more specific user groups, coverage by a randomly drawn set of human evaluators does not appear to be feasible. A perfect match between evaluators and target users is required for end users with cognitive or neurological impairments. Furthermore, developers of synthesis systems are interested in diagnostic information about deficiencies in the performance of their systems. While the developers themselves should be excluded from independent evaluation efforts involving their own systems, some degree of expertise in speech science and technology is required to identify errors in the synthesis output. That said, the paradigm known as "yuck detection" has recently been applied with some success: evaluators hit a button whenever they perceive a flaw in the synthesized signal, which allows system developers to inspect locations with accumulated indications of flaws. End users may also have different expectations of the synthesis quality and the system's capabilities, which may introduce judgment biases. A synthesis system that achieves near-human-likeness in overall quality and intelligibility may be expected to be also capable of maintaining a dialog, manage turn-taking, and perhaps even be omniscient, even though such capabilities are beyond the scope of the speech synthesizer itself. Moreover, it is presently unclear how users cope with inconsistent, apparently stochastic synthesis output resulting from the statistical nature of state-of-the-art speech generation methods. In summary, depending on the aim of the evaluation and the application domain of speech synthesis, designing subjective evaluation setups needs to make judicious decisions about the composition of the pool of human synthesis evaluators.

4.5.4 Pitfalls and problems

There are some common problems and pitfalls that apply to all evaluation protocols and for all use cases. These can be in the design of the protocol, during its implementation and execution, or when interpreting the results.

A typical design pitfall is failing to adequately simulate the intended use case and thus create an ecologically-invalid design. The two principal errors made in implementation are to use poorly-chosen materials or unsuitable participants (or objective measures).

Examples of poor design: evaluating speech as audio-only when the use case is a talking robot; ...

Examples of poorly-chosen materials: including incomplete or ungrammatical sentences in the test set; using sentences from out-of-copyright novels when the intended use case is spoken dialogue.

Examples of unsuitable participants: listeners who are unfamiliar with the accent of the synthetic speech and so are unable to accurately judge speaker similarity; ...

Examples of incorrect results interpretation: ...

Design

- Use Case 1:
  - side effects also need to be assessed
- Use Case 2:
  - how to determine the right evaluator
  - is the listener qualified to evaluate the sample
  - resource issue
- Use Case 3:
  - over sensitivity of audio only evaluation in case of multimodal (some artefacts may not matter)
- Use Case 4:
  - user expectation/familiarity needs to be taken into account
  - difference of meaning/implications between same terminology (e.g., what is "fair"?)
- Use Case 5:
  - assuming traditional way of doing the evaluation is the gold standard ⇒ metrics should not become the target
  - over sensitivity of audio only evaluation in case of multimodal (some artefacts may not matter)

Execution

- Use Case 1:
  -
- Use Case 2:
  - different experience between observer and user (⇒ overthinking from the observer perspective, influence of the recovery)
  - user vs participant
- Use Case 3:
  - what if the system if not fast enough yet ⇒ offline, how to deal with fine grain
- Use Case 4:
  - input processing
  - no access to the final use ⇒ defining the delta
  - "inappropriate" material deployed during the test
  - proficiency of the participant in the language
  - cognitive load (dual language – evaluation/evaluated language)

- Use Case 6:
  - what about participants who don't usually listen to audiobooks?
  - expectations from participant how have already read the book?
  - duration tolerance (5min is ok but 2h) + long term consistency

Interpretation

- Use Case 1:
  - what if basic test works only for formant synthesis?
- Use Case 3:
  - over interpretation of the results

4.6 Notes on proposed methods and method requirements

4.6.1 Suggested evaluation methods for some use cases 1

Use case	Needs of end users	Metric
Designing for an artefact	aligned multimodal interaction affordances	positioning of persona (within space of possible voices)
Targeted manipulation in a high-quality voice	disentanglement of effects within an experimental setting	exact reproduction of intended contrast
	preservation of original "voice"	objective or subjective analysis of signal degradation and/or speaker ID

4.6.2 Suggested evaluation methods for some use cases 2

1. Virtual human coach- visual and tactile; social context- weird aspects of imitation; derogatory aspects of voice mimicry?
2. Augmented Accessibility- Identity uniqueness; creating voice that satisfies constraints; unique characteristics of the conditions; age appropriateness; speech loss; representation and social relationship
3. Agent/Robot mediator

Metrics/Observations

1. Appropriateness; social acceptability; trust; representation matching
2. Appropriateness; social acceptability; social and dialogue cues; temporal aspects of interaction; qualitative identification of dimensions of voice key to the user (interviews); communication behaviours (turns etc); identity matching/augmentation; consistency; configurability, satisfaction with result

Experiment Design

1. Third party vs user/creator evaluation; Explicit- asking them; Implicit – Engaging and interacting Comparators? control conditions?

4.6.3 Suggested evaluation methods for some use cases 3

Criterion/Use case	**Dubbing (Audio/Visual)**	**Voice privacy/Anonymization**	**Translation (audio only, e.g podcast)**	**Simultaneous interpretation (e.g., EU parliament, meeting)**
Target speaker characteristics	Speaker ID Speaker embedding Fit to actor Consistency	Speaker ID Speaker embedding	Chosen by content producer	Don't care
Content	AVSR/ASR WER Artistic and semantic fit	ASR WER Emotion recognition ER	BERT score Semantic fit (e.g., journalistic fit)	BERT score Semantic fit (e.g., journalistic fit)
Timing	Must have :: No overlap Nice to have :: Sync to video, lip sync			When to speak? (turn taking, cognitive load, latency)
Prosody	Expressivity target being the "same" as the source	Expressivity same (within language, without speaker)	Expressivity target being "same" as the source	Neutrality
Non-speaker related information in the source (channel, music ...)	Keeping other information Possibly diarization performance Noise robustness	Noise robustness		
Offline / Online	offline (real time factor)	offline / online	offline	oneline

Evaluation methods. Long-form speech should primarily be evaluated by humans, although automatic methods can be used in some cases.

We identified two distinct evaluation scenarios: (1) general evaluation of a TTS voice developed for long-form content; and (2) evaluation of the final product, such as an audiobook.

In the second scenario, automatic methods such as **ASR** (Automatic Speech Recognition) can compare its output with the input to the TTS; **LLMs** (Large Language Models) can analyse linguistic complexity to identify anomalies; and a silence detector can identify pauses that are too long, which might occur when a human narrator loses focus.

In our selected use case, the shifts between narrative text and dialogue and between speakers, we propose using **ARS** (Audience Response System), where the respondents push a button whenever they perceive a shift. This procedure is followed up with questions about perceived difficulty. Similar methods could evaluate whether the listeners can identify the speaker in multi-party dialogues, for instance by clicking a picture of the speaker.

Finally, we agreed that the test data selection is crucial and must involve challenging text passages to thoroughly assess the system's capabilities.

4.7 Day 2, session 1. Use case: Simulation and stimuli generation (for speech science)

Zofia Malisz (KTH Royal Institute of Technology – Stockholm, SE), Roger K. Moore (University of Sheffield, GB), and Petra Wagner (Universität Bielefeld, DE)

This group discussed quality dimensions and evaluation metrics in which synthesis (or the analysis of synthetic speech) is used with the expressed reason to conduct scientific research. Perhaps the most typical case for this will be the application of synthesis within the area of speech science, phonetics or related fields, but it may also include other disciplines in which the vocalizations and their quality are concerned. Within the group, two concrete use cases were discussed and further operationalized:

Use Case 1: Manipulation of voice alongside a certain dimension of scientific interest (f0, prominence, age, manner or place of articulation etc.)

Use Case 2: Designing a voice for an artefact (that is not normally considered to have a voice)

For Use Case 1, we identified two core dimensions of quality, which are known challenges for state-of-the-art synthesis systems:

(1) the successful disentanglement of effects within a speech science experimental setting. This could be operationalized by measuring the exact reproduction of the intended contrast encoded by this dimension (e.g., pitch modification). However, a truly successful manipulation would necessarily also check that the manipulation does not result in an unintended modification of those features that characterize the original "voice", or speaker identity, e.g., by not only changing pitch, but also the perceived gender.

(2) Thus, the "preservation of original voice" is our second quality dimension. This could be operationalized in several ways, but example metrics would be an objective analysis of signal degradation and the preservation of the speaker ID. Another metric would be subjective listening tests to perform these analyses.

For Use Case 2, we identified two core dimensions of quality:

(1) aligned multimodal interaction affordances, which could be evaluated with evaluation methods that have been established in HCI, including the alignment with the user, ventriloquist effect, or general usability of the artefact. A second dimension of quality would be (2) the adequate "positioning of the persona" established by an artefact's voice. This could be measured by perceived appropriateness or the artefact's voice, or rather, the voice's congruency with its perceived embodiment. An example for this would be that a "speaking briefcase" would probably be expected to have a voice that is somewhat muffled (due to the fabric it consists of), and has a voice that is aligned with its size.

During our discussions it emerged that our two use cases share one quality dimension, which is the "appropriateness, or congruency" of a generated voice with the persona it belongs to. If extended, this quality dimension could also define the boundary between human-like and "super-humanlike" voices, or between different speaker groups or even species.

4.8 Day 2, session 1. Use case: Incremental TTS; incremental speech (i.e. live speech production)

Bernd Möbius (Universität des Saarlandes, DE), Gérard Bailly (University Grenoble Alpes, FR), Jens Edlund (KTH Royal Institute of Technology – Stockholm, SE), and Ayushi Pandey (Trinity College Dublin, IE)

We're defining these systems loosely as signal generation systems (i.e. a TTS engines) meeting the following criteria:

Condition the signal on a stream of tokens that is smaller than what is common.

If "sentence" or "utterance" is the typical unit size for TTS; we may use lemma/word, grapheme, or phoneme.

1. Leave open a continuous input channel representing external phenomena (e.g. surrounding noise)
2. Leave open a continuous input channel representing phenomena that are conditioned or created by self (e.g. own audio, reactions of others).
3. Be able to take decisions, midstream, based on these extra input channels. It may pay off to view the standard conditioning stream as intention, and the other streams as controls.

Since this starts with a capability description, in quite technical terms, it's perfectly feasible to have a feature assessment – checkboxes pretty much – that must be passed in order for a system to count. That narrows down our concept space neatly.

Also, this can be seen as a component. In which case API descriptions can be used as capability descriptors, and a clear eval step is to see if the effect of using the API is as follows. So claimed capabilities and capabilities delivered.

As with embodied/virtual agents, more of a technology than a use case (and there are of course relevant pure technology evaluations). Use cases include anything interactional that isn't heavily turn based, as well as fast and responsive transmission. Incremental rendering is also at least in principle less resource (memory) consuming, And a prerequisite for any kind of situational real-time adaptation of the speech, such as the Lombard effect.

- Interruptible systems
- Realtime adaptation (e.g. Lombard)
- Responsive speech
- Listening speakers
- Feedback-sensitive speech
- Self-correcting speech

Good use cases may include: listening speakers (interruptible, feedback responsive), self-corrective speech, Lombard speech. And possibly streaming speech.

4.9 Day 2, session 1. Use case: (Human-like) embodied spoken dialogue

Petra Wagner (Universität Bielefeld, DE), Elisabeth André (Universität Augsburg, DE), Benjamin Cowan (University College – Dublin, IE), and Olivier Perrotin (University Grenoble Alpes, FR)

Our group conducted a deep discussion on the use and evaluation of synthesis when applied embodied speech systems, i.e., where a speech synthesiser is embodied either within an external agent (virtual or robot) or a human user (in the case of Assistive Augmented Communication (AAC) interactions). Overall, we could not identify gold standards for the evaluation of such systems, which go largely beyond clarity and intelligibility. Evaluation is extremely task-specific and so are the associated metrics. Instead, we encourage the definition of gold standard process for selecting the evaluation practices and associated metrics rather than to define them. We identified two main goals of such embodied system:

1. establish an identity;
2. being a dialogue partner in a social environment.

As for identity, human likeness which is highly prized in most TTS applications is not necessarily a requirement in this use case, for both agent and patient voice embodiment. By contrast, consistency being voices and their body is crucial (as opposed to have several agents with similar voices or vice-versa) to allow distinctiveness and recognition of agents/patients as individual entities. Also, identification of voice might be part of a brand identity. Plausibility is also a criteria, i.e. the voice matches the embodiment, to favour expectation matching from interlocutor. Having identified those criteria, two complementary evaluation methods are relevant. On the one hand, the use of explicit detailed questionnaires addressed to both participants in interaction with the system under evaluation, and external participants watching the interaction. On the other hand, the getting of implicit global preference about interaction experiences with several voices. Finally, performing these evaluations longitudinally is a mean to integrate the familiarisation of the participants with the new voice to evaluate.

The second direction is the assessment of the interaction, mainly being whether the message is conveyed properly (either via linguistic and/or paralinguistic information), and whether the voice impact user language and dialogue. A variety of paradigms have been listed for evaluation, transversally of the task-specificity of interactions. First, as for identity assessment, the evaluation can be carried out either by extracting explicit metrics from the speech signals or questionnaires, or by inferring implicit metrics from the interlocutor reaction to the speech interaction. In the former case, borrowing metrics from voice coaching would be an interesting direction to investigate. In the latter case, the interlocutor has the role of feature extractor, which is highly dependant on his/her social background and environment. These approaches can be carried out offline (post experiment), or online. In that case, in a similar fashion that the yuck test, the laugh test naturally measures unexpected situations. The choice of evaluators is also of high importance, as evaluation from the participants involved in the interaction and the one of external viewers should be quite complementary. In the latter case, the question of realistic immersion in the interaction can have great impact on the assessment. Finally, the quantifying of user-agent interplay such as entrainment/adaptation to voice/linguistic markers alignment could be one aspect of assessing interaction, but is highly task- and environment-dependent to assume whether alignment is an expected or unexpected phenomena to happen in a statisfying interaction.

Participants

- Elisabeth André
Universität Augsburg, DE
- Gérard Bailly
University Grenoble Alpes, FR
- Erica Cooper
NICT – Kyoto, JP
- Benjamin Cowan
University College – Dublin, IE
- Jens Edlund
KTH Royal Institute of Technology – Stockholm, SE
- Naomi Harte
Trinity College Dublin, IE
- Simon King
University of Edinburgh, GB
- Esther Klabbers
Beaverton, US
- Sébastien Le Maguer
University of Helsinki, FI
- Zofia Malisz
KTH Royal Institute of Technology – Stockholm, SE
- Bernd Möbius
Universität des Saarlandes – Saarbrücken, DE
- Sebastian Möller
TU Berlin, DE & DFKI Berlin, DE
- Roger K. Moore
University of Sheffield, GB
- Ayushi Pandey
Trinity College Dublin, IE
- Olivier Perrotin
University Grenoble Alpes, FR
- Fritz Michael Seebauer
Universität Bielefeld, DE
- Sofia Strömbergsson
Karolinska Institute – Stockholm, SE
- Christina Tånnander
Swedish Agency for Accessible Media – Malmö, SE
- David R. Traum
USC – Playa Vista, US
- Petra Wagner
Universität Bielefeld, DE
- Junichi Yamagishi
National Institute of Informatics – Tokyo, JP
- Yusuke Yasuda
Nagoya University, JP