Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

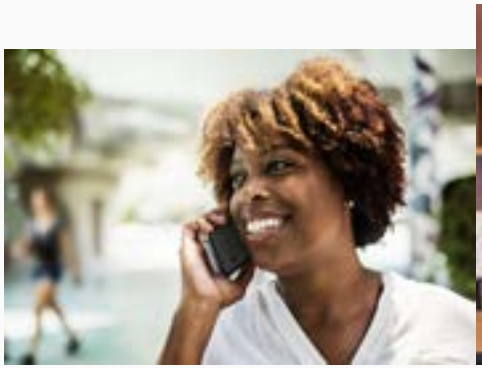
Speech Synthesis Evaluation

State-of-the-Art Assessment and Suggestion for a Novel Research Program

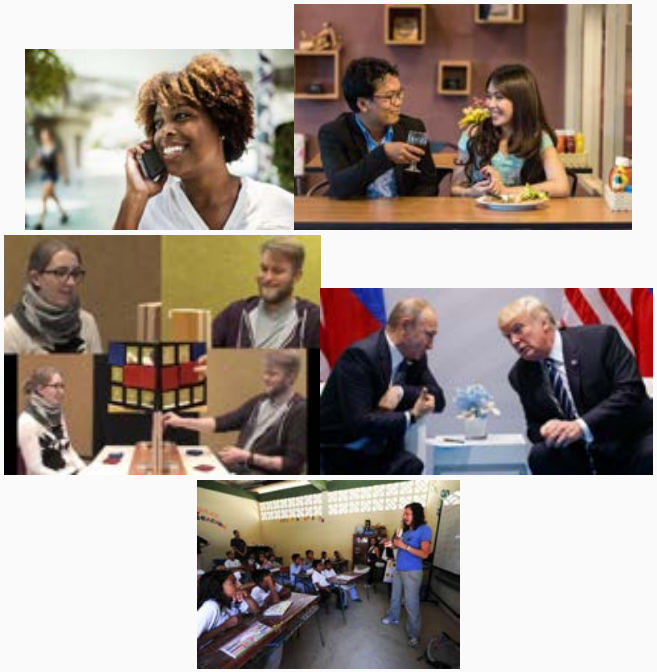
Petra Wagner[1,2], Jonas Beskow[3], Simon Betz[1,2], Jens Edlund[3], Joakim Gustafson[3], Gustav Eje Henter[3], Sébastien Le Maguer[4], Zofia Malisz[3], Éva Székely[3], Christina Tånnander[3], Jana Voße[1,2]

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

Speaking does not take place in the void



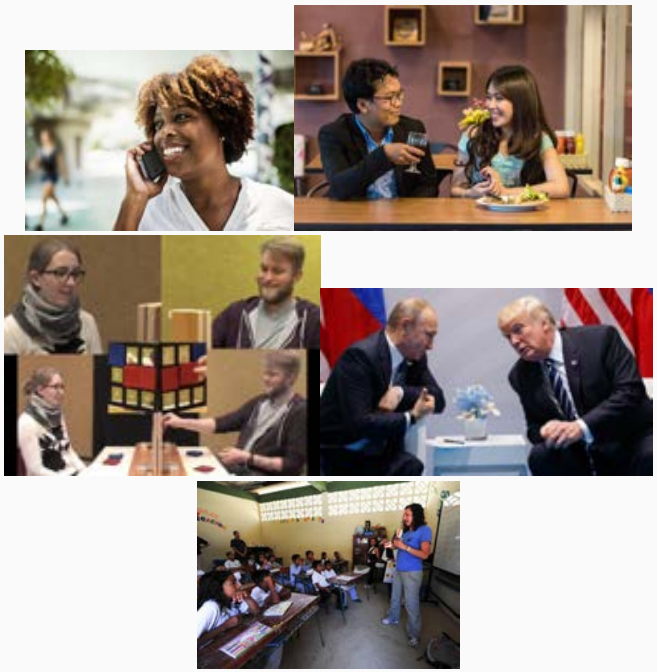
Speaking does not take place in the void



The way we speak depends to a large degree on...

- who we are
- who we are talking to
- the situation in which we are talking (transmission channel, communicative goal, task...)

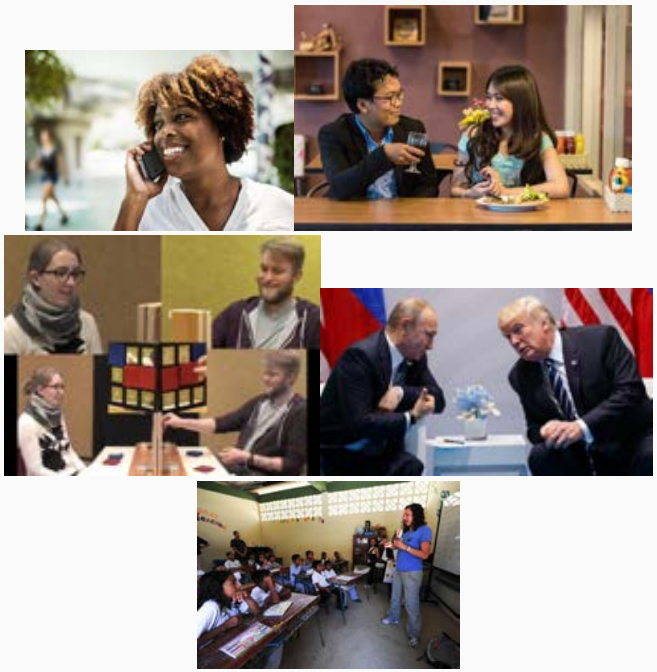
Speaking does not take place in the void



Felicitous (synthetic) speech needs to be

- adequately intelligible
- adequate in style with respect to the situation

Speaking does not take place in the void

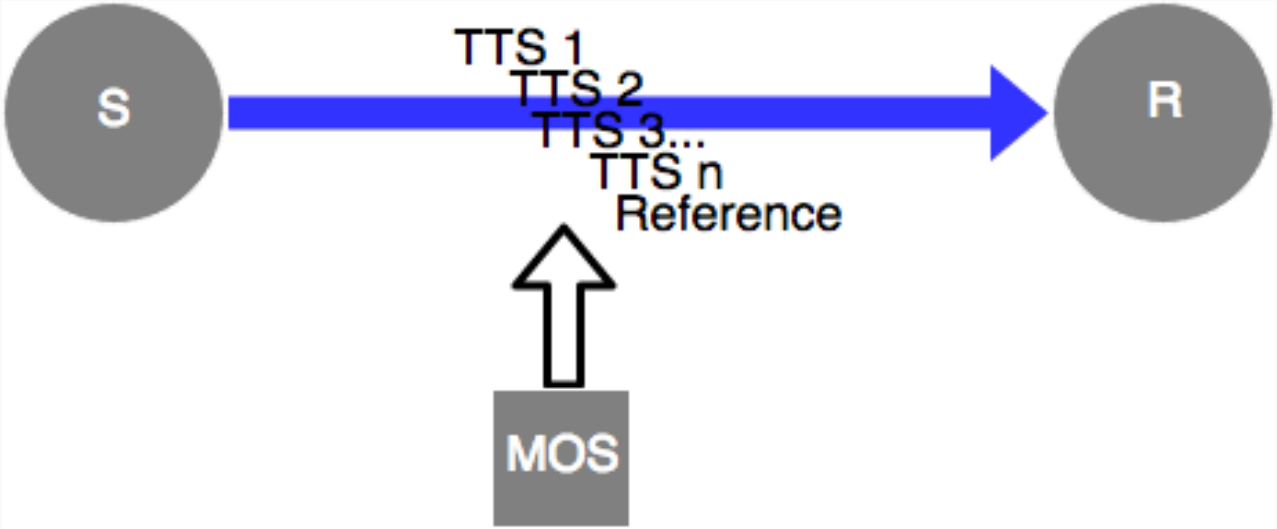


Felicitous synthetic speech needs to be

- adequately intelligible $\longrightarrow$ **solved**
- adequate in style with respect to the situation $\longrightarrow$??

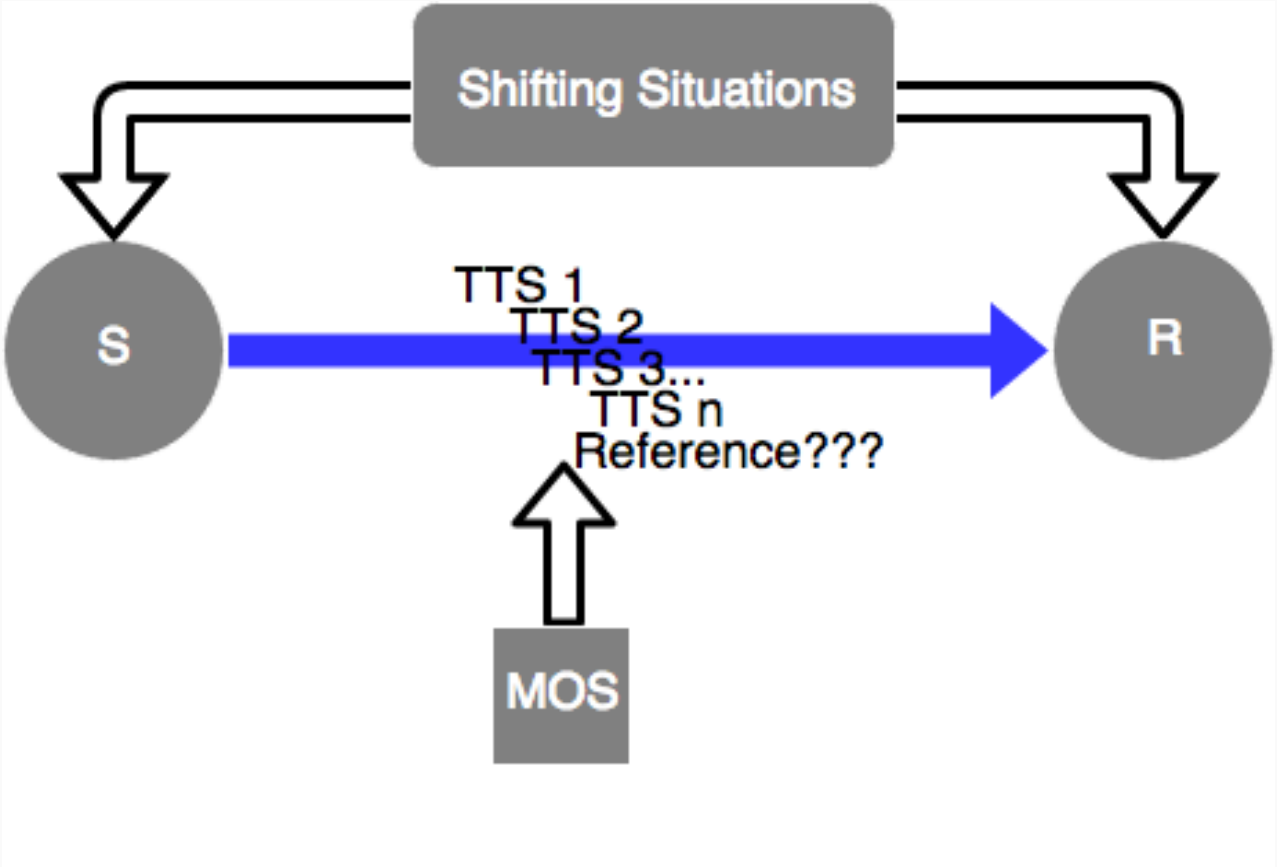
Testing speech quality

- Established protocols treat speech quality assessment as an estimate of the quality of the transmission, that a listener can compare to an internal or external reference (e.g., ITU-Rs for MOS, MUSHRA).



Testing speech quality

- Established protocols treat speech quality assessment as an estimate of the quality of the transmission, that a listener can compare to an internal or external reference (e.g., ITU-Rs for MOS, MUSHRA).
- With dynamically shifting "gold standards", we do not know whether results from such evaluations generalize across applications.



Problems are well known...

- Betz et al. (2018)
- Mendelson & Aylett (2017)
- Rosenberg et al. (2017)
- Wester et al. (2017)
- Wester et al. (2015)
- King (2015)
- ITU-R P.800 (2004)

Problems are well known...

- Betz et al. (2018) $\longrightarrow$ MOS-scores do not generalize!
- Mendelson & Aylett (2017)
- Rosenberg et al. (2017)
- Wester et al. (2017)
- Wester et al. (2015)
- King (2015)
- ITU-R P.800 (2004)

But little is happening...

We lack clear-cut guidelines for alternative evaluation procedures. These could be developed within a novel research program centering on listeners' context specific needs and expectations.

From “all purpose” TTS to “appropriate TTS”

- Blind users often prefer formant-based systems over unit selection (Moers et al., 2007)
- TTS quality of robots partially predictable by fit between expected and realized “robot-like” voice quality (Burkhardt et al., 2019), in line with “uncanny valley” effect (Moore, 2012, 2017)
- Conceptual framing useful to normalize for users’ imagined system application (Dall et al., 2014)
- Embedding an evaluation in a realistic dialogue task increases users’ sensitivity for synthesis artefacts (Betz et al., 2018)

A first attempt of a expectations/needs assessment

Application	**Estimated needs**
Virtual assistant	clear, pleasant voice
Humanoid robot	humanoid (but not human-like) voice
Navigation	sufficiently loud, clear, timely
Announcements	loud, clear
Interactive travel guide	clear, pleasant
Screen reader	intelligible at high speed, informative prosody
Audiobook (leisure)	slow, expressive
Audiobook (educational)	optimized for online comprehension
Video game	convincing personality, expressive
Voice prosthesis	adaptable speaker identity, low latency
Dialogue system	timely, incremental, suitable discourse markers
Speech-to-speech translation	adaptable speaker identity

Do we need new metrics altogether?

Established objective Metrics

- So far, objective metrics do not align well with listening tests
- Approaches often rely on problematic "natural baseline" or "gold standard" as reference (but cf. Hinterleitner, 2017; Fu et al., 2018)
- Approaches focus on spectral features, and mostly ignore prosodic aspects
- **Idea: likelihood of waveform-level synthesizer as indicator of "human-likeness"**

Do we need new metrics altogether?

Subjective Metrics

- Mean Opinion Scores (MOS) for global impression of quality (ITU-R P.800)
- MUSHRA for pairwise comparisons useful for multidimensional scaling; needs multiple assessments of comparable utterances across systems (ITU-R.B.S.1534)
- Questionnaire-based subjective scores based on (multidimensional) questionnaires; problem of fatigue or boredom; between-subjects (many participants needed) (Bartneck et al., 2009)
- Alternative approach: online tracking of listening quality using Audience Response Systems (Edlund et al., 2015)
- **Idea: Further develop methods of relating online quality tracking with global impressions**

Do we need new metrics altogether?

Behavioral and physiological metrics

- Intelligibility metrics (SUSs, word edit distance...) established but less important
- Measurement of comprehension much less well understood.
- Rarely used: task success, task efficiency, interaction time
- **Idea: combining global behavioral, subjective metrics (e.g., task success) with metrics monitoring cognitive load in the ongoing interaction, e.g., eye tracking, response time** (Rajakrishnan et al., 2010; Betz et al., 2017; Govender and King, 2018

Combining needs and established metrics

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

So why am I giving this talk?

So why am I giving this talk?

Because this table has been generated without knowing whether it really helps predicting users' "quality of experience"...

A very preliminary first set of guidelines

1. Move away from "all purpose TTS" to "context-appropriate synthesis development/evaluation" – or see how widely applicable a system is.
2. Even if no application context is defined, provide suitable conceptual framing.
3. Conduct user need analysis to determine speech quality space.
4. Go beyond subjective and into behavioral metrics (e.g., global task performance).
5. Develop online estimates of speech quality to pinpoint problems (and combine them with global quality assessments).

Some questions for a novel research program

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

Questions and Comments!! (And who's on board with us?)

Questions and Comments!!

Suggestion for evaluation procedure