Hot topics in speech synthesis evaluation

G. Bailly *(Univ. Grenoble Alpes, GIPSA-lab)*
E. André (Univ. Augsburg)
E. Cooper (NICT - Kyoto)
B. Cowan (University College Dublin)
J. Edlund (KTH)
N. Harte (Trinity College Dublin)
S. King (Univ. Edinburgh)
E. Klabbers (phAIstos Speech & Language Technology Services)
S. Le Maguer (Univ. Helsinki)
R. K. Moore (Univ. Sheffield)
B. Möbius (Saarland Univ.)
S. Möller (TU & DFKI – Berlin)
A. Pandey (Karya)
O. Perrotin (Univ. Grenoble Alpes, GIPSA-lab)
F. Seebauer (Bielefeld Univ.)
S. Strömbergsson (Karolinska Institutet, Stockholm)
D. R. Traum (USC – Playa Vista)
C. Tånnander (KTH, Swedish Agency for Accessible Media)
P. Wagner (Bielefeld Univ.)
J. Yamagishi (National Institute of Informatics – Tokyo)
Y. Yasuda (Nagoya University)

13th Speech Synthesis Workshop

Leeuwarden — The Netherlands — 24.08.2025

SotA of TTS technology and evaluation paradigms

- Current TTS produces speech perceptually indistinguishable from human recordings
  - When judged with a single ACR (e.g., MOS)
  - On isolated sentences & out-of-context

But …

➔ Increasing range of **interactive** applications
- Expressive audiobooks (i.e., long-form TTS), L1 & L2 training, assistive communication, speech-to-speech conversion, incremental interaction, etc.

➔ Claims for more **responsive evaluation paradigms**
- Multidimensional scales, online evaluation, task-specific performance measures, etc.

Dagstuhl Seminar 25032

Task and Situation-Aware Evaluation of Speech and Speech Synthesis

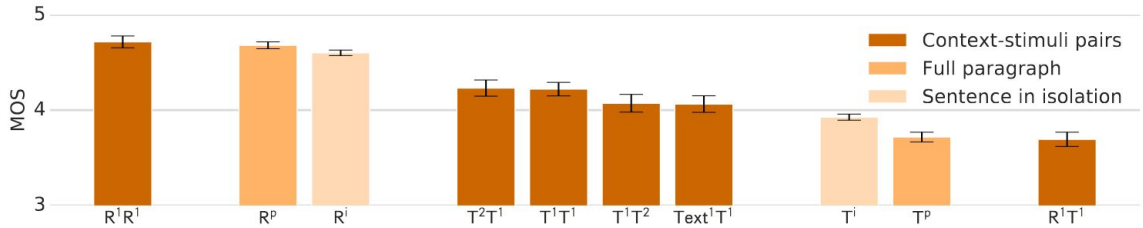
Jan 12 – 15, 2025


© SCHLOSS DAGSTUHL – LZI GMBH
licensed under Creative Commons License CC BY-NC-ND

1. The hot topics
2. A brief survey of methods and metrics

First step: long-form synthesis

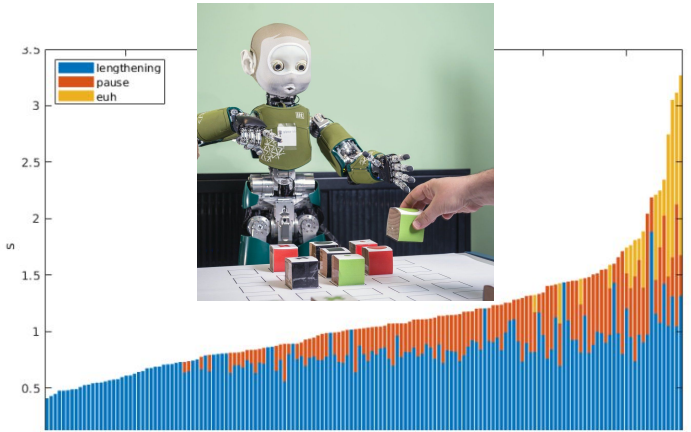
- Synthesis of paragraph, article or even an entire book
  - Styles (newspaper, advertising, exercising L1/L2 learners, storytelling…)
  - Character voices (narrators, dialogs, ...)
  - Structured information (images, maths, tables, …)
- Involves broader contextual information & understanding of the content
- Involves evaluation of
  - Consistency & variability: stable pace, style, pronunciation of names and terms while sustaining interest
  - Text contact: sense of genuine understanding of the content being read
  - Transitions, e.g. signal shifts between narration & dialogue, as well between characters in fictions



Evaluating read long-form TTS: Comparing the ratings of sentences and paragraphs (Clark et al, SSW 2019)

Further more: controllable speech synthesis

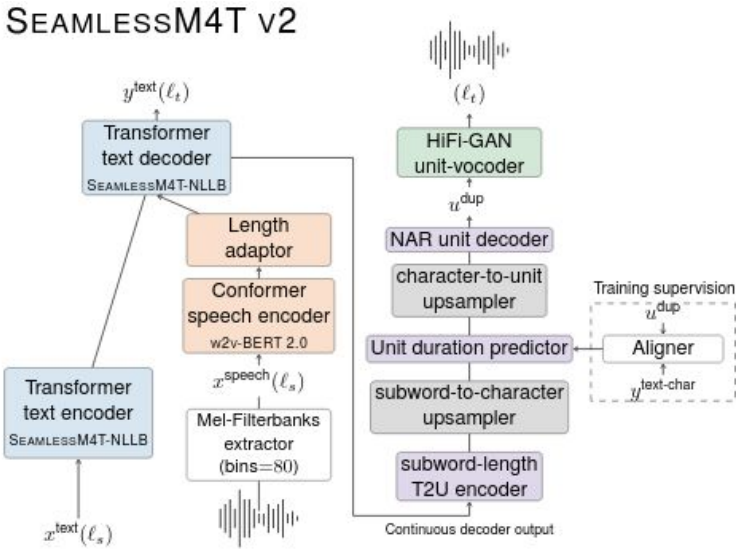
- Versatile speech synthesis
  - Use cases: art, digital therapy, phonetics, etc.
  - Involves user and expert judgment (e.g., artists, therapists, phoneticians), acoustic analyses, intrinsic evaluation of models
- Incremental TTS
  - Use cases: real-time spoken dialogue systems, assistive technologies, etc.
  - Synthesis with limited look-ahead
    - Not waiting for complete sentences or waiting for external events (gestures, gaze of interlocutors), waiting policy, …
    - GO/no GO decision on input pace, …
  - Involves assessment of timing and latency, timeliness and fluency
- Listening TTS: reacts on-the-fly to their interlocutors and environment
  - Involves objective and perceptual judgement of adaptation (e.g., Lombard effect, Chameleon, phonetic convergence, alignment of behavior, …)



Awaiting attention from the partner for a co-verbal deictic “That” gesture (Bailly & Elisei, NLG4HRI 2020)

Ultimately: interactive speech synthesis

- Speech-to-speech conversion/translation
  - Use cases: dubbing, anonymisation, etc.
  - Involves assessment of target speaker and content preservation, timing, real-time factor, expressivity, effect of non-speech information…
- Multimodal and embodied interaction
  - Involves assessment of complementarity between modalities, perceptual evaluation of speech and visual identity alignment, perceptual evaluation of the interaction task…



Barrault & al (2023). Seamless: Multilingual Expressive and Streaming Speech Translation. arXiv preprint arXiv:2312.05187

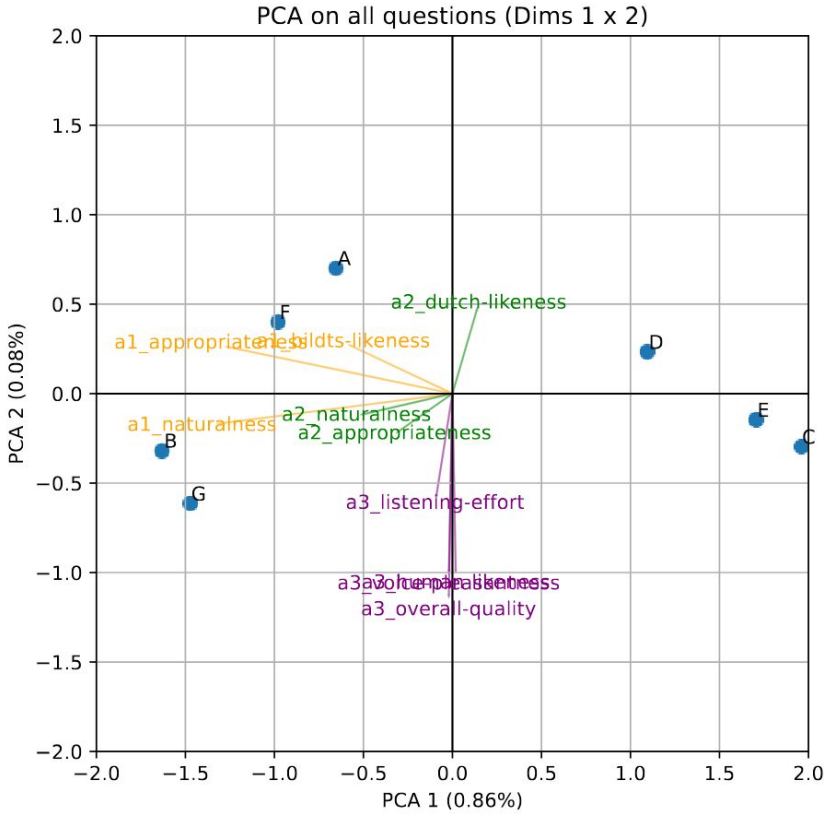
1. The hot topics
2. **A brief survey of methods and metrics**

Beyond MOS

- ***When*** to evaluate
  - **After** offline evaluation (multidimensional scales, verbal tagging, …)
  - **Before** gating experiments (evaluation of parts-of-speech, anticipatory behaviours, …)
  - **During** online evaluation (detection, continuous ratings, online monitoring, …)
- ***What*** to evaluate
  - **Language- and Task-specific** assessment (linguistic traps, performance, memorization…)
  - **Multimodality** (multimodal integration, complementarity, crossmodal binding, …)
  - **Intrinsic evaluation** of models (without synthesis: analysis of internal representations, …)

After | Offline evaluation

- Multidimensional scales
  - E.g., reading fluency assessment: accuracy, pace, phrasing & expressivity *(Razinski, PREL, 2004)*
  - Further reduction by PCA or MDS *(Mayo et al., Interspeech, 2005)*
- Free vs. semi-directed verbal tagging
  - Bag of words, text embeddings
- Transcriptions
  - E.g, Rapid Prosody Transcription paradigm *(Gutierrez et al., SSW, 2021)*
  - Memorization taks…



PCA of average MOS of Blizzard 2025 task MH1, highlighting the impact of audience on scores (a1 = native vs. a3 = non-native)

Before | Gating experiments

- Evaluating uncomplete verbal content
  - Word recognition process
    (Grosjean, Perception & Psychophysics, 1980)
  - Recognition of prosodic patterns
    (Aubergé et al., SSW, 1997)
- Assessment of anticipatory encoding
  - Key property for incremental & listening TTS
  - Anticipatory turn-taking, decoding intentions
  - Lower listeners' cognitive load

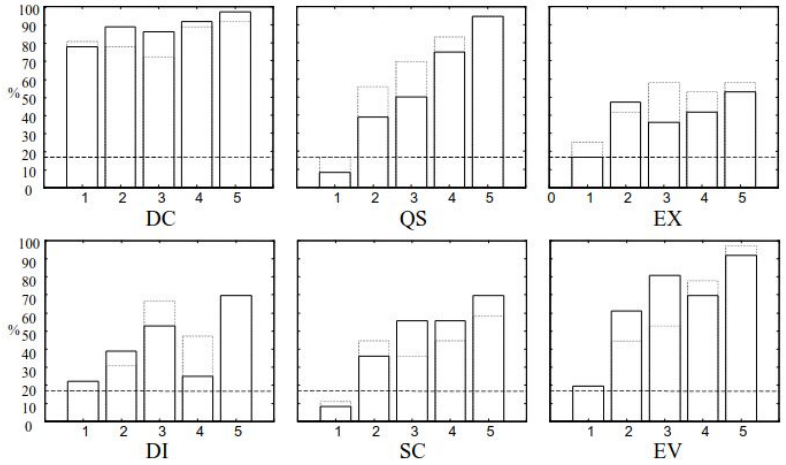
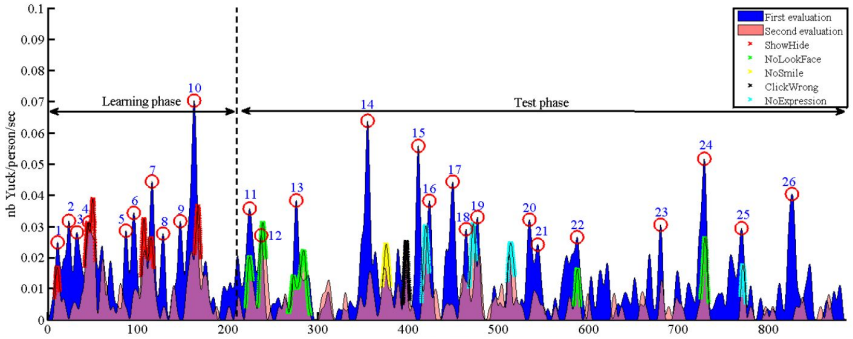


Figure 0.5. Identification rates for each gate with a normal order (solid lines) and a reverse order (dotted lines) for 5-syllable truncated utterances and the 6 attitudes.

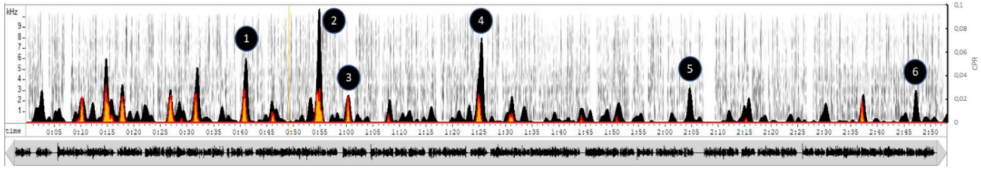
(Aubergé et al., SSW, 1997)

During | Online evaluation

- Raters' judgements as they experience the synthetic multimodal signals
  - Performed by end-users or third parties
  - Three Paradigms
    - Detection: Audience Response System (ARS) or Yuck responses (keypress)
    - Continuous ratings (sliders)
    - Online monitoring: close-shadowing, neurophysiological signals, etc.



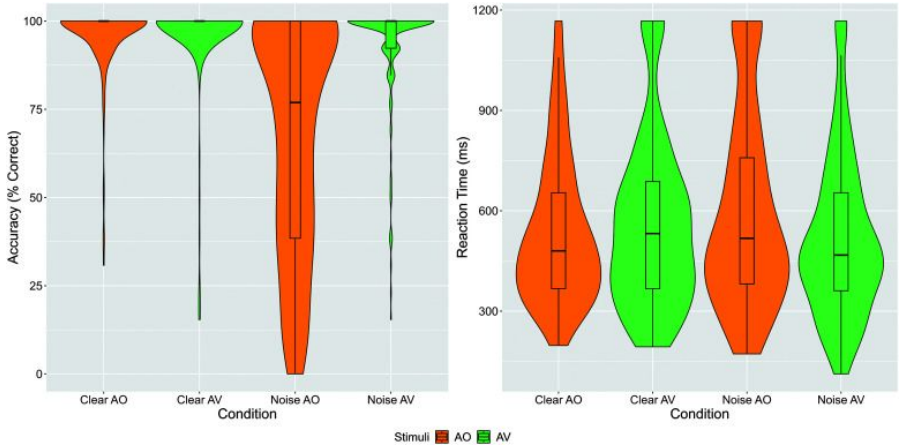
Density of Yuck responses for interactive robotic behaviours
(Nguyen et al, Gespin, 2017)



Density of Yuck responses for TTS. Event 2 is a fabricated error for sanity check!
(Tånnander et al., LREC, 2024).

What | Evaluation of multimodal synthesis

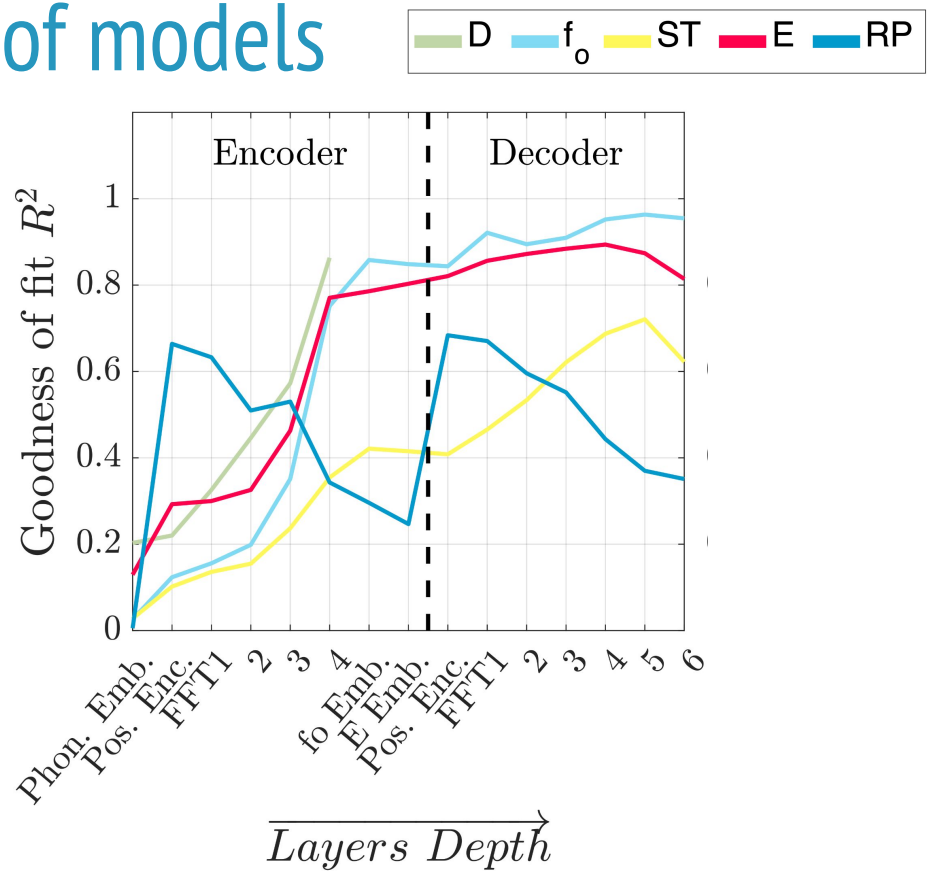
- Quality of coordination, redundancy & complementarity
- Methods:
  - Impoverishment of modalities (e.g., AV speech: speech in noise vs. display of face profile)
  - Multimodal integration & crossmodal binding using coherent or incoherent modalities (e.g. McGurk effect)



Reassessing the Benefits of Audiovisual Integration to Speech Perception and Intelligibility
(O'Hanlon et al., JLSHR, 2025)

What | Intrinsic assessment of models

- Objective evaluation of internal representations built by models or components
  - Ablation studies
  - Causal regression of acoustic parameters
  - Insights into phonetic encoding, long-term dependencies, covariations between modalities, etc.



Internal encoding of prosodic features in Fastspeech2 (Lenglet et al., CSL (submitted), 2025)

Conclusions & further challenges

- SotA TTS technology has achieved remarkable progress but performance saturates on simple text reading tasks

➡ Let's move beyond our comfort zone!

- We are now equipped to challenge new situations with much more diverse evaluation criteria
  - Focus on particular properties of proposed architectures & applications
- Feed speech science with new insights in speech communication
  - Technology for science and world comprehension: explainability issues
  - New paradigms for probing speech, language & brains
- Ethical and legal concerns
  - Ease legal and ethical evaluation of technology while maintaining usability & effectiveness

Further discussion on dedicated Google Group

speech-and-synthesis-evaluation@googlegroups.com