Evaluation of speech synthesis

- Case study: the Blizzard Challenge

Evaluation case study: The Blizzard Challenge

- Annual evaluation of speech synthesis systems in which participating teams build a voice for their system using a common data set
- A large online listening test is used to evaluate the systems
- Goal:
  - understand and compare research techniques
- Method:
  - build voices on a **common dataset**
  - evaluate them in a **single listening test**
- The "hub" task is to take the released speech data, build synthetic voices, and synthesize a prescribed set of test sentences.
  - There are usually also several optional "spoke" tasks

Typical timeline

- Feb Databases released
- Mar Test sentences released
- Apr Deadline for submitting synthesized speech
- Apr Evaluation system goes live
- Jun End of Evaluation
- Jul Results distributed to teams
- Sep Presentation of results at a workshop

Benchmark systems can be used to compare across different evaluations

- **NATURAL** Natural speech from the same speaker as the corpus
- **FESTIVAL** The Festival unit-selection benchmark system
- **HTS** HMM-based benchmark system

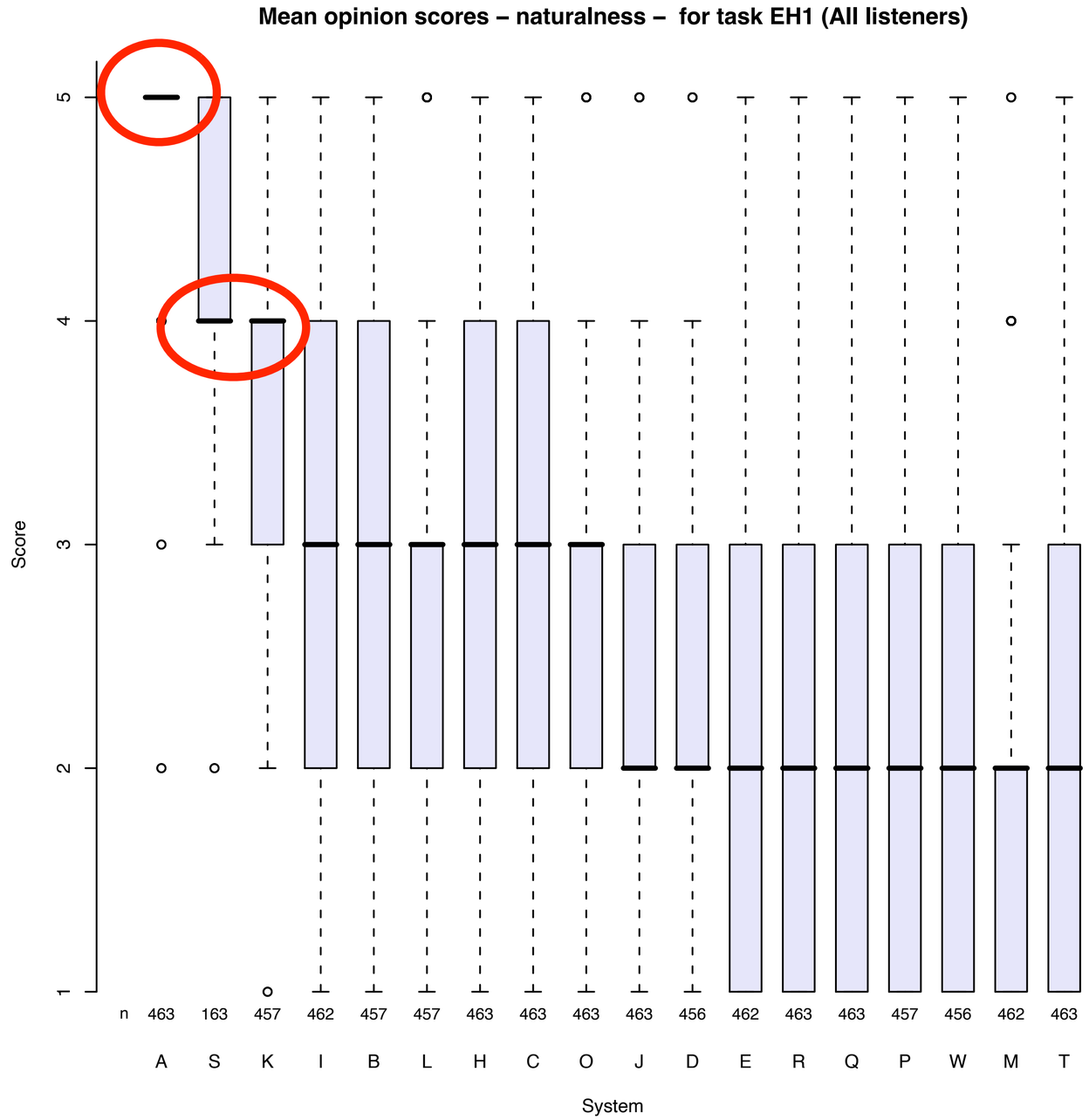
- Benchmark systems are intended to provide some comparability across listening tests
  - i.e., across years of the Challenge

- Increasingly difficult to do in recent years due to rapidly-changing modelling paradigms

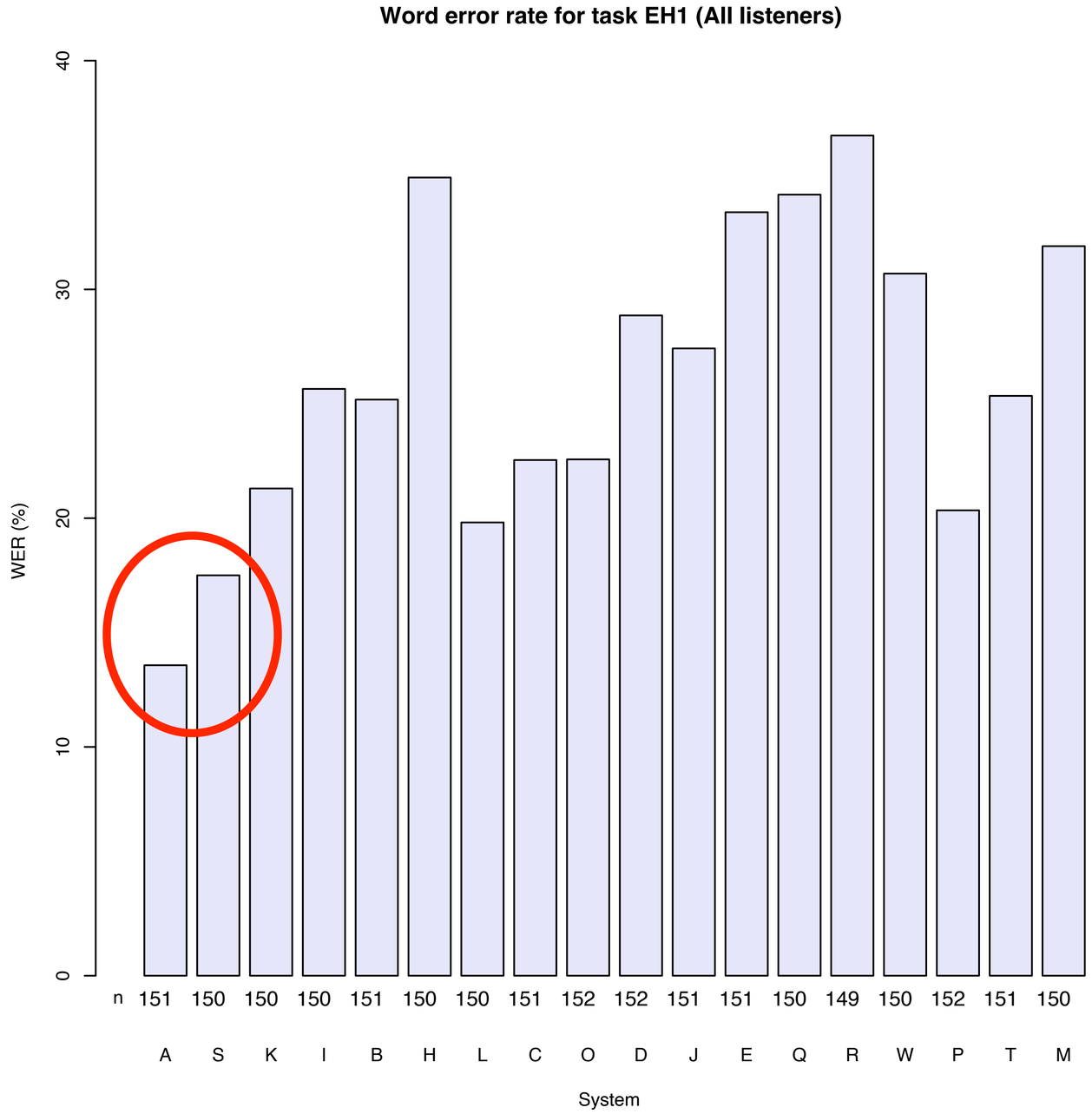
2008 systems

Natural speech	UPC
Festival benchmark	CMU
HTS benchmark	mXac
IIIT	I²R
INESC-ID	Nokia
CASIA	DFKI
VUB	TUD
AHOLAB	IBM
SUCLAST	NICT/ATR
USTC	Toshiba
CSTR/Cereproc	HTS

How the Blizzard Challenge presents results for MOS Naturalness



How the Blizzard Challenge presents results for Intelligibility as WER



Typical statements that can be made about the results

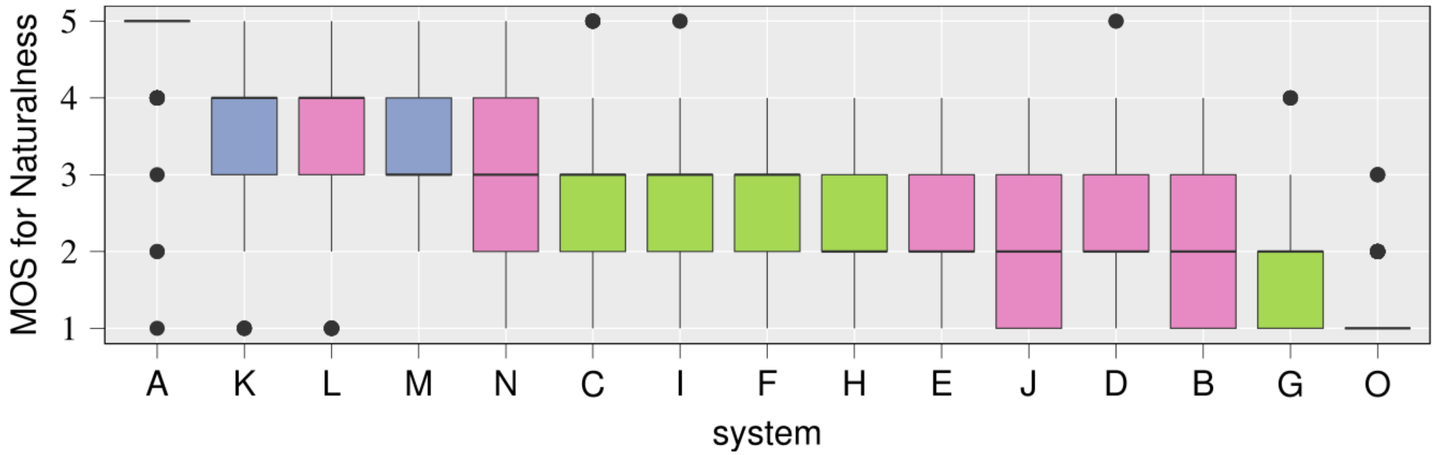
- Natural speech is **significantly** more natural and more similar to the original speaker than any synthesiser
- Systems S and K are both significantly more natural and more similar to the original speaker than all other synthesisers

- System S is **as intelligible as natural speech**
- But there is **no significant difference** in intelligibility between system S and a number of other systems (B,C,K,L,O,P)
  - so we cannot state that system S is more intelligible than other systems

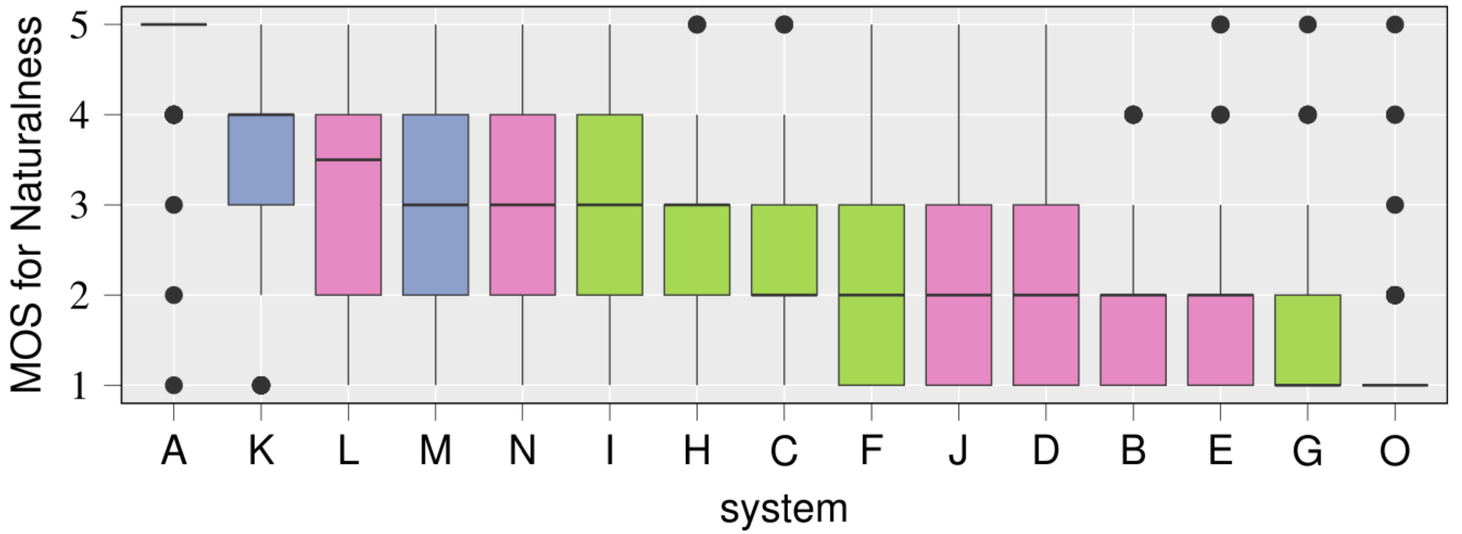
Le Maguer, King, Harte. "The limits of the Mean Opinion Score for speech synthesis evaluation." Computer Speech and Language 84, March 2024, DOI: 10.1016/j.csl.2023.101577

Absolute Category Rating (ACR) - unfortunately *not* Absolute, but Relative

original results
from 2013

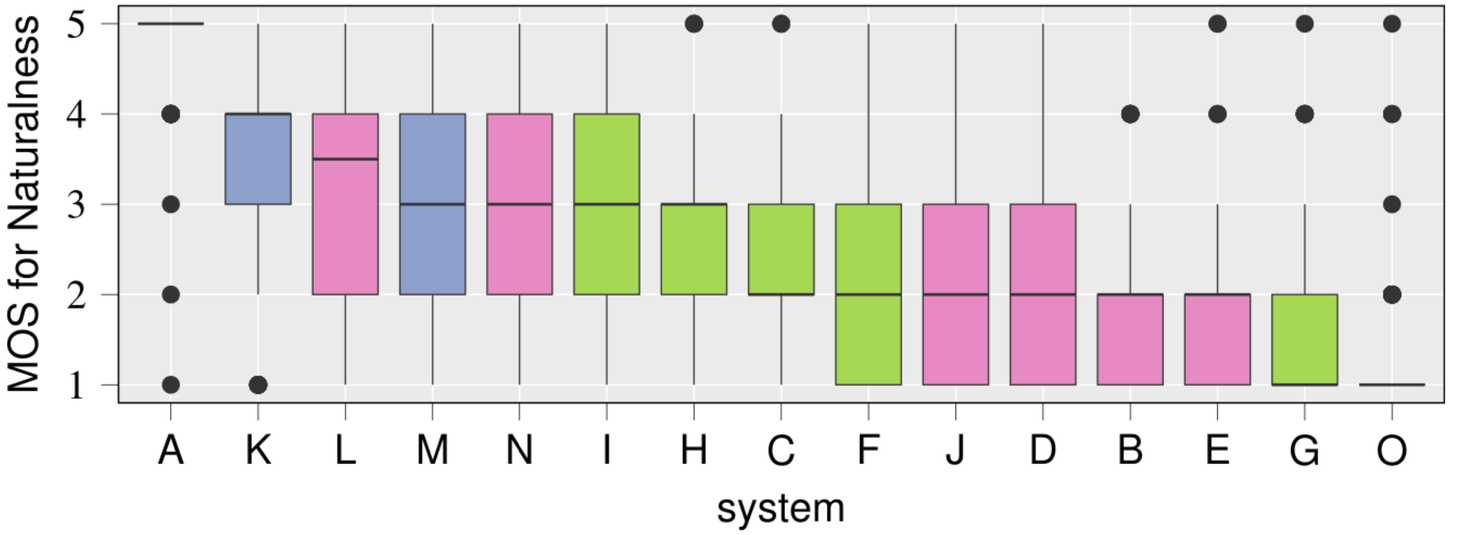


replication
performed in 2022

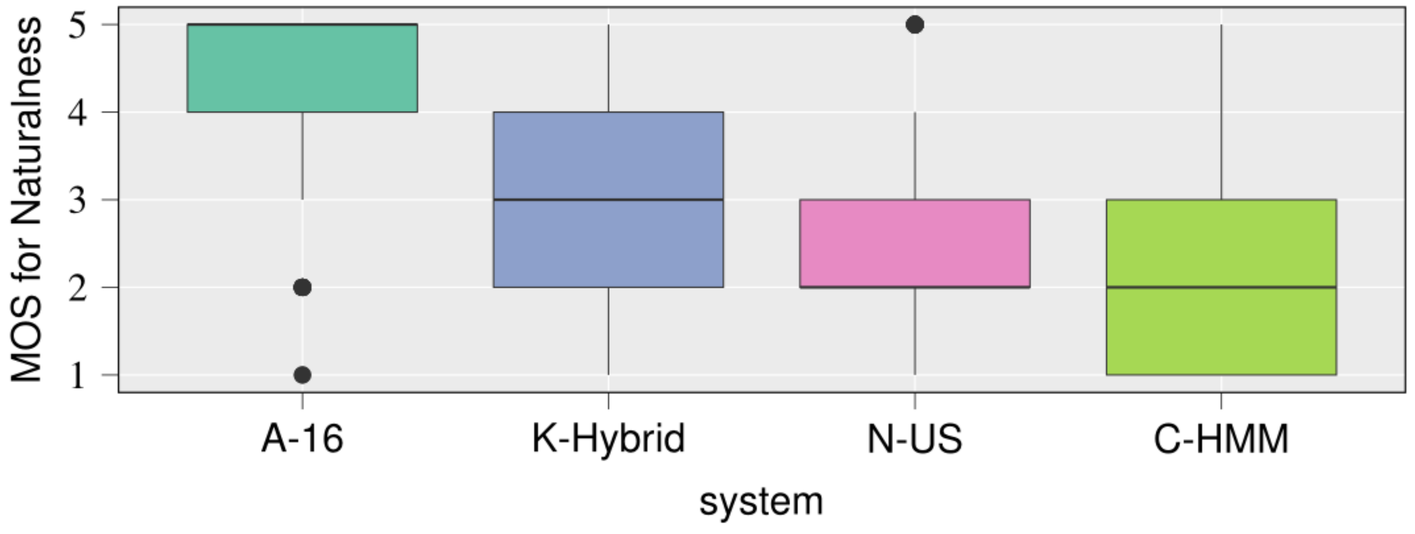


Absolute Category Rating (ACR) - unfortunately *not* Absolute, but Relative

replication
performed in 2022



testing *only* some of
the better systems



Evaluation of speech synthesis

- Discussion points

Discussion points - Blizzard Challenge

- is it possible to **cheat** on the Blizzard Challenge?
  - if so, what can you do to prevent cheating by participating teams
- can you design a challenge that **only** evaluates
  1. the front-end linguistic processor?
  2. a component of the front-end, e.g., LTS
  3. the waveform generator
- is the Challenge "**ecologically valid**"
  - discuss what that really means
  - can you think of improvements, to make it more valid
  - would your improvements change the outcomes / results / findings / conclusions ?

4.2. Room for improvement

4.2.1. What to evaluate

Naturalness and intelligibility remain the main evaluation criteria for speech synthesis, with judgements being elicited from listeners on a Lickert scale (Likert, 1932). Naturalness remains poorly defined, although listeners do seem to have a clear idea of what is being asked of them given the consistency of their judgements. Intelligibility is measured, as noted in Section 4.2.2, in a particularly unrealistic, or 'ecologically invalid', way.

Blizzard also adds an evaluation of speaker similarity to the mix. This was introduced initially only as a check that participants were using the provided recordings and not entering pre-built systems. With the advent of speaker-adaptive approaches, and for unit selection entires employing voice conversion, speaker similarity became a useful dimension of the evaluation in its own right.

Loquens, 1(1), January 2014, e006. eISSN 2386-2637 doi: http://dx.doi.org/10.3989/loquens.2014.006

4.3.1. Whole system vs. component-level evaluations

As we mentioned in Section 2, Blizzard only attempts end-to-end system evaluations. Moreover, it also bundles in the data preparation stages such as alignment with the text and optional hand-corrections performed by some participants. In other words, it evaluates the totality of the *systems components* and the *engineering skill and effort* needed to make it work well on a new database. Conclusions about which *method* is "best" are therefore inevitably filtered through the level of expertise and available resources of the team implementing that method. This may be a partial explanation of the "failure" of some entries: the idea had merit, but the implementation was flawed. The availability of resources for checking and correcting the data varies widely between participants. To quantify the effect this has on overall quality, one year's Challenge did release hand-checked alignments but this was found to be of limited use because it does not guarantee consistency across systems, since some may use a different phonetic inventory or pronunciation dictionary. Some participants have themselves investigated the benefits of manual annotations (Chu et al., 2006).

Providing linguistic specifications may appear to be one way to isolate the waveform generation component, but it would not be possible for some participants to modify their systems to use an externally-provided linguistic specification.

Loquens, 1(1), January 2014, e006. eISSN 2386-2637 doi: http://dx.doi.org/10.3989/loquens.2014.006

Evaluation of speech synthesis

- Group activity: design a listening test

Group activity: design a listening test

- Step 1
  - define the hypotheses more precisely
  - what aspects to evaluate; what task(s) for the listeners
- Step 2
  - materials
  - listeners: type, recruitment, vetting,…
- Step 3
  - interface design
  - should you show the text to the listener? will you play examples of natural speech?
- Step 4
  - sanity checking results, detecting listeners who cheat, removing outliers
  - mock-up how the results will be presented in your paper

start in class

finish on your own,
with help in lab sessions

Systems to be evaluated

- I tried varying the contents of the database, and found it had a strong effect on the synthetic speech

- Step 1
  - write down at least two clear hypotheses that could be tested
  - what aspects of the speech would need to be evaluated, to test those hypotheses?
  - what task(s) are you going to ask your listeners to do?
  - what systems will you need to build?

Summary of today's class: what we learned about evaluation

- **Why did you evaluate your system? What do you need it to deliver?**
  - Information that helps you decide how to improve it
  - Confirmation that it works for your use case
- **Don't overlook the obvious**
  - Intelligibility (suitably measured, perhaps WER, perhaps something else)
  - Naturalness (can you define it? do you need to?)
- **But that is not enough**
  - High-level, whole system "performance"
  - Low-level, specific aspects / system components (e.g., pronunciation)
- **Does your evaluation deliver what you need?**

Summary of today's class: what we learned about evaluation

- **Always start with a clear hypothesis**
  - including an explanation of *why* you believe it to be true

- Design an experiment that tests the hypothesis (and nothing else)
  - Keep it simple and direct

A look forward to neural approaches

- methods for synthesis have rapidly advanced
- yet approaches to evaluation have **barely changed**

- but evaluation is even more relevant than before
- so we'll need to revisit this topic later in the course (whilst reading recent papers)

Natural language guidance of high-fidelity text-to-speech with synthetic annotations

Dan Lyth[1], *Simon King*[2]

[1]Stability AI
[2]University of Edinburgh, UK

Table 2: *Naturalness and relevance results (with 95% confidence intervals)*

Model	**MOS**	**REL**
Ground truth	3.67 ± 0.09	3.62 ± 0.06
Ours	3.92 ± 0.07	3.88 ± 0.06
Audiobox	2.79 ± 0.09	3.19 ± 0.06

arXiv:2402.01912