1 Evaluation of speech synthesis

This document is to guide researchers at CSTR through subjective evaluation procedures of speech synthesis.

2 Standard tests

This section gives an overview of the various methods there are to evaluate synthetic speech. Some general points that need to be considered when thinking about designing your evaluation:

- Referenced vs non-referenced tests: In non-referenced tests subjects must assume some context for the stimuli prior to comparing them with their internal reference for that context; two sources of ambiguity: multiple contexts are possible for a given stimulus and different subjects may have different mental references for them.
- The question you ask influences the answer you will get [1]
- Some listeners are better than others; variability in subjects (not only materials)
- How many listeners do you need to get statistically reliable results? How do you calculate this?
- Can you use crowdsourcing? Are the results going to be reliable?
- The reliability, validity and sensitivity of the evaluation [2]

2.1 MOS

Participants have to judge stimuli on a scale from 1 to 5, where 1 is bad and 5 is excellent. Depending on what one wants to test, different questions should be asked. For instance: "Please listen to the sample and judge using a five point scale their ... quality / naturalness / similarity to a certain speaker

ITU 1994 recommends MOS with seven different questions ranging from "Listening Effort" to "Overall Quality", See also [2] for a description of an expanded MOS test in which questions are asked about the following: Listening Effort, Comprehension Problems, Speech Sound Articulation, Precision, Voice Pleasantness, Voice Naturalness, Humanlike Voice, Voice Quality, Emphasis, Rhythm, Intonation, Trust, Confidence, Enthusiasm, Persuasiveness.

Some considerations:

✗ What is the listener's internal model of speech perception? When asking a listener to rate naturalness or quality (or any of the above listed properties) what do they interpret these terms to mean? Are you measuring what you think you are?

✗ MOS test to measure similarity is problematic. How similar is one voice to the other? Often listeners are asked to do this task by comparing synthetic speech to a natural target sample, further complicating matters.

✓ Results are easier to analyse than preference tests' results

- Should natural speech be included?
- Use naturally produced sentences when possible
- A MOS test with X stimuli lasts around Y mins

2.1.1 DMOS

DMOS is a "degradation" or "differential" MOS test in which a reference sample is given and listeners are asked to rate the sample with respect to this reference.

2.2 MUSHRA

MUSHRA stands for MUltiple Stimuli with Hidden Reference and Anchor and is a methodology for subjective evaluation of audio quality. The main advantage over the Mean Opinion Score (MOS) methodology (which serves a similar purpose) is that it requires fewer participants to obtain statistically significant results. This is because all systems are presented at the same time, on the same samples (sentences), so that a paired t-test can be used for statistical analysis. Also, the 0-100 scale makes it possible to rate very small differences.

In MUSHRA, the listener is presented with the reference (labeled as such), a certain number of test samples, a hidden version of the reference and one or more anchors. The recommendation specifies that one anchor must be a 3.5 kHz low-pass version of the reference. The purpose of the anchor(s) is to make the scale be closer to an "absolute scale", making sure that minor artifacts are not rated as having very bad quality. (http://en.wikipedia.org/wiki/MUSHRA).

Some considerations:

- For TTS evaluation the choice for reference is natural speech but the choise for a anchor is not obvious. It should be a system which is consistenly, i.e. across sentences, always the worst.

✗ Can take a long time for participants to get through.

✓ It allows multiple comparisions to take place at the same time.

- A MUSHRA test of 20 screens with 12 sliders takes approximately 45mins

2.3 AB test, aka preference tests, paired or pairwise comparison tests

Listeners are presented with two speech samples and asked to indicate which one has more of a certain property. In a preference test, the question would be"which one do you prefer". But it could be: "which one sounds happier" (angrier/sadder etc.). The choices can be: "A","B", i.e., forced choice or "A", "B" and "no preference".

In AB tests for multidimensional scaling (MDS), the question is are the two samples the same or different. This can be in terms of naturalness, or similarity. Some considerations:

- Should choices be A-B or A-B-no pref? Interpretation of these results sometimes shaky. If more than 50% is "no pref" then surely the 2 systems are the same?

✓ in same/diff MDS can be use to visualise results

✗ if you want to compare many systems the test can become too big

✗ axes are undefined and sometimes difficult to interpret

- a preference test of X comparisons takes approximately Y mins

2.3.1 ABX

As in AB test, two samples are given but an extra reference sample is also given. Listener has to judge if A or B more like X. Again this can be in terms of naturalness, similarity, emotional quality.

2.4 Transcription

Transcriptions tests evaluate intelligibility of stimuli. Participants are asked to type what they could understand from each sentence, no matter if only a few words. Considerations:

✗ it can require a lot of post processing to account for homophones and misspellings

- sentences to be used: in noise Harvard sentences, in clean conditions SUS sentences to avoid ceiling effect. In noise matrix sentences are also available but there is a learning effect.
- SUS are odd, they can be a factor in the results
- a transcription test of 120 sentences takes around 45 minutes

3 Procedure

- Evaluation in the lab or using a crowd source system like AMT?
- How to advertise.
- How much to pay participants.
- Set up listening booths (set volume across booths).
- Create a consent form.
- Should I do a hearing screen before the test? How to?

4 Scripts

4.1 Listening tests

- Web scripts for MOS, preference and intelligibility tests can be found in.
- Matlab scripts for Hurricane style intelligiibility tests can be found in.
- MUSHRA test scripts.

4.2 Analysis

References

[1] Rasmus Dall, Junichi Yamagishi, and Simon King. Rating naturalness in speech synthesis: The effect of style and expectation. In *Proceedings of Speech Prosody*, Dublin, Ireland, 2014.

[2] Melanie D Polkosky and James R Lewis. Expanding the mos: Development and psychometric evaluation of the mos-r and mos-x. *International Journal of Speech Technology*, 6(2):161–182, 2003.