1 Group members

Cantao Su, Weihao Jiang, Yinqiu Wang

2 Lab Book S6

For this labbook you will design a listening test to evaluate a hypothetical TTS system. This assignment can be done in a group of 2-3 people, but you are also free to make up a "group of one" if you want to work on this assignment solo. In this lab book you don't have to conduct the experiment and you don't have to synthesize speech. For illustration purposes you can use dummy set of recordings.

Your hypothetical TTS system is made for a specific usage domain, for example:

1. TTS system for a minority language (you specify which minority language you would focus on).
2. TTS module for a specialized dialogue system (you specify the domain for which your dialogue system would apply).
3. TTS system for a pathological speech (you specify pathology, whether it is oral cancer, ALS or anything else).
4. Or your own challenging use case!

2.1 Assignment 1

[10pts] *Design a listening test*:

Task:

Choose your hypothetical TTS scenario (see the examples above) and specify where and how you would use such a system. Design an evaluation protocol that would include the description of the following:

1. Which aspects of synthetic speech you evaluate.
2. What you include in the listening task.
3. Your test design.
4. The materials (stimuli) you use.
5. The number and categories of listeners you recruit.
6. How you present the stimuli and collect the listeners' responses (to illustrate your user interface you can use https://rug.eu.qualtrics.com/ or any other platform for response collection).

Motivate your design choices.

Submission

Submission of this group project is a LaTeX file containing the structured description of your evaluation protocol with link(s) to your dummy experiment setting. Feel free to submit any additional files that you deem necessary. Don't forget to include the description of authors' contributions.

Answer

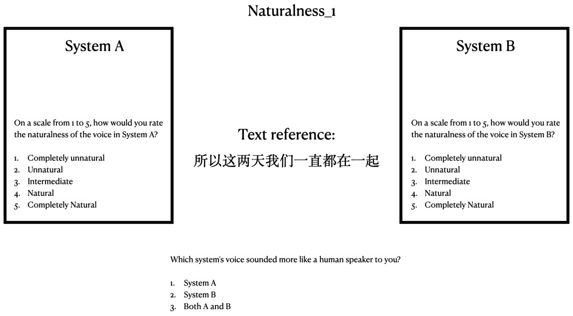
1. Aspects of our evaluation
   When evaluating synthetic speech, we mainly consider the following aspects: naturalness, intelligibility, and emotional expression, accent and intonation perception.

2. What we include in the listening task
   To evaluate the synthetic speech, we designed these listening tasks below:

   - Naturalness: The listener is presented with two audio samples, one from each system, with a shared text reference for context.
     The listener rates the naturalness of each system on a 5-point scale and indicates which system sounds more human-like.
   - Intelligibility: The listener is asked whether a specific word was clear in the sentence from each TTS system.
     After listening to both samples, the listener selects which system produced clearer speech overall.
   - Emotional Expression: The listener rates how well the intended emotion was conveyed in each TTS system's speech on a 5-point scale.
     The listener chooses which system better conveyed the emotional tone.
   - Accent and Intonation perception: The listener evaluates whether the intonation of each TTS system is appropriate for standard Mandarin.
     The listener decides which system has an accent closer to standard Mandarin pronunciation.

3. Test Design

   - Judgments: Relative Judgments
     We will employ relative judgments to directly compare the performance of the two TTS systems. This approach enables listeners to make assessments based on a side-by-side comparison, which can be more intuitive for detecting subtle differences in quality between the two systems.
   - Interface:
     The interface, as depicted in the image, presents two audio clips from Systems A and B side by side. For each set, listeners will be provided with a text reference to ensure that their evaluations are based on the audio quality rather than the content. Listeners will rate the naturalness of each system on a 5-point Likert scale and then choose which system's voice sounds more human-like to them.



Naturalness_1

System A

On a scale from 1 to 5, how would you rate the naturalness of the voice in System A?

1. Completely unnatural
2. Unnatural
3. Intermediate
4. Natural
5. Completely Natural

Text reference:
所以这两天我们一直都在一起

System B

On a scale from 1 to 5, how would you rate the naturalness of the voice in System B?

1. Completely unnatural
2. Unnatural
3. Intermediate
4. Natural
5. Completely Natural

Which system's voice sounded more like a human speaker to you?

1. System A
2. System B
3. Both A and B

Figure 1: Demo of Interface

   - Test Size:
     The test will consist of 40 audio samples in total. This will be divided into 5 sets for each aspect of evaluation (naturalness, intelligibility, emotion recognition, accent, and intonation), with each set containing 2 audio samples from each TTS system for direct comparison.
   - Within-Subject Design:
     All listeners will hear exactly the same stimuli to ensure that each system is evaluated under the same conditions. The order of stimuli will be randomized for each listener to control for order effects and fatigue bias, ensuring that the position in the sequence does not unduly influence the listener's judgment.

- Rational:
  The chosen design aims to balance thoroughness with listener fatigue management. Forty audio samples are enough to cover a wide range of scenarios without overwhelming the listeners. Randomizing the order of presentation mitigates the risk of the listeners' responses being influenced by the sequence rather than the content. The side-by-side comparison facilitates a direct and clear understanding of the relative merits of each system.

4. The materials (stimuli) we use.
   We use "normal" materials, which are the segments of Mandarin from Chinese movies. At the same time, these segments exhibit significant emotional fluctuations.

5. The number and categories of listeners we consider recruiting.
   For this task, we plan to recruit 500 listeners to ensure the reliability and generalizability of the results. Our listeners will primarily consist of Cantonese native speakers from the areas near Hong Kong and Guangzhou, who are also required to understand Mandarin. Within this broad category, we will further divide the listeners into two groups. The first is the general listeners who do not have a professional background in voice processing and can provide a general public perception of the TTS voice. The second group is the professional listeners, who have some understanding of voice processing and can provide a more professional and in-depth analysis.

6. presentation:
   You can access the demonstration through the following link Survey Link. However, since Qualtrics doesn't support audio insertion, we have prepared a PowerPoint for further demonstration, which you can find in the attachment.

7. Speech Dataset Credit to:
   Chenye Cui, Yi Ren, Jinglin Liu, Feiyang Chen, Rongjie Huang, Ming Lei, Zhou Zhao from Zhejiang University and AlibabaGroup on their website and their thesis provide valuable insights, which you can find via the link