Measuring the Quality of Synthetic Speech

Gediminas Navickas, Gerda Ana Melnik-Leroy, Povilas Treigys
Institute of Data Science and Digital Technologies
Vilnius University

Introduction

In this study, we raise the question of synthetic speech evaluation, using a **controlled experimental paradigm**.

First, we evaluate experimentally the perceptual differences between objectively different synthesized speech qualities. Second, we compare the perception of two groups of listeners: **sighted and blind participants**.

Experimental Paradigm

The experimental paradigm is based on a **modified AX (same-different) discrimination task**, in which **participants hear two samples of synthesized speech in each trial and they have to answer whether they sound same or different**.

The advantage of this method is its simplicity and the possibility to record participants' perception without having to rely on subjective measures, such as the "naturalness" etc.

Description of the Experiment

Half of the trials ("same" trials) contained identical recordings (the same word of the same quality was played twice), half of them ("different" trials) were of different speech qualities (the same word was played in two different qualities). Both groups of participants completed two conditions (Table 1): in the Easy condition, the items in "different" trials were of low vs. high quality synthetic speech. Acoustic difference between those two qualities is large, the discrimination task was expected to be easier in this condition. Conversely, in the Difficult condition, participants heard medium vs. high quality synthetic speech stimuli, which differed in less acoustic distance.

	Blind	Sighted
Easy condition	low – high quality	low – high quality
Difficult condition	medium – high quality	medium – high quality

Table 1. The distribution of experimental conditions over participant groups

Experimental Setup

Stimuli for the tasks synthesized using: **Merlin – open source speech synthesis toolkit** for Deep Neural Network models adapted for Lithuanian language.

Three qualities of the synthetic stimuli: **low quality** (training data 400 sentences), **medium quality** (800 sentences), **high quality** (1600 sentences).

All stimuli for the experiment consisted of Lithuanian **words** or **short sequences of words** (1-3).

Results

1. Both user groups performed better on the Easy than on the Difficult condition. This suggests that they were sensitive to synthetic speech quality differences.

2. Blind participants performed better than sighted participants, especially in the Difficult condition. This suggests that blind listeners are much more sensitive to speech quality changes than the sighted ones.

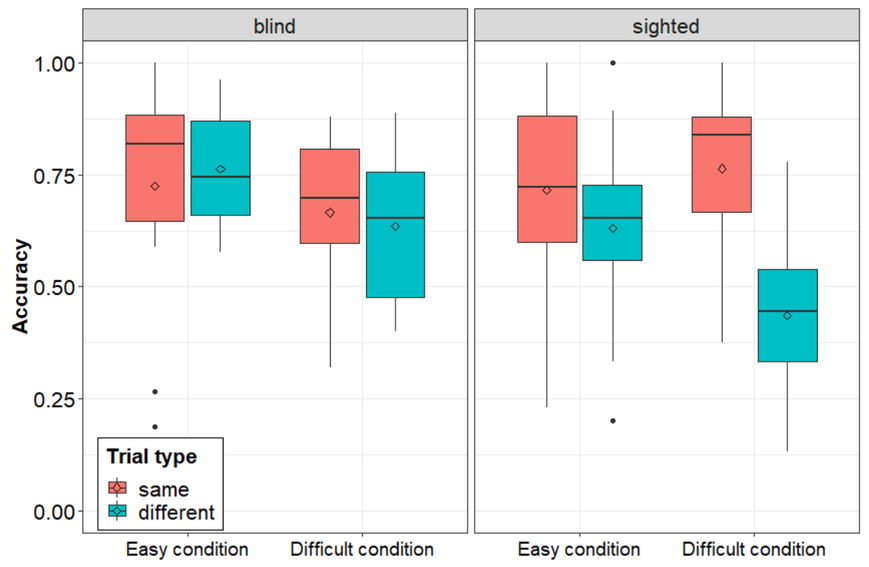

Fig. 1. Blind and sighted participants results in different conditions

Conclusions and Further Investigations

Blind participants had better discrimination abilities compared to sighted listeners, especially so in the difficult condition. This suggests that blind listeners might be much more disturbed by synthetic speech quality imperfections. Thus, speech quality should be adapted to this group.

The results show that this experimental paradigm can be used for the evaluation of synthetic speech quality and also for defining the level of perceptual accuracy for particular user groups. Importantly, unlike traditional methods which describe only the quality of the signal, the current method allows the evaluation of the perception of this signal.

In further research, other user groups (adults-children, native-non native speakers) could be investigated. The method could be used for evaluation of different speech synthesis systems.

Data Analysis Methods for Software Systems, 12th Conference Druskininkai, Lithuania, Hotel „Europa Royale", December 2 – 4, 2021

Contact: Gediminas Navickas, gediminas.navickas@mif.vu.lt