Not All Listeners Are the Same: Acoustic Cue Use in Synthetic Speech Quality Judgments

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

Introduction

Standard TTS (text-to-speech) evaluation pipelines often assume a homogeneous "typical listener," risking misleading conclusions about perceived quality. We test this assumption by comparing congenitally blind and sighted listeners on a ternary AX discrimination task for assessing synthetic speech quality (using Lithuanian neural TTS).

Research questions

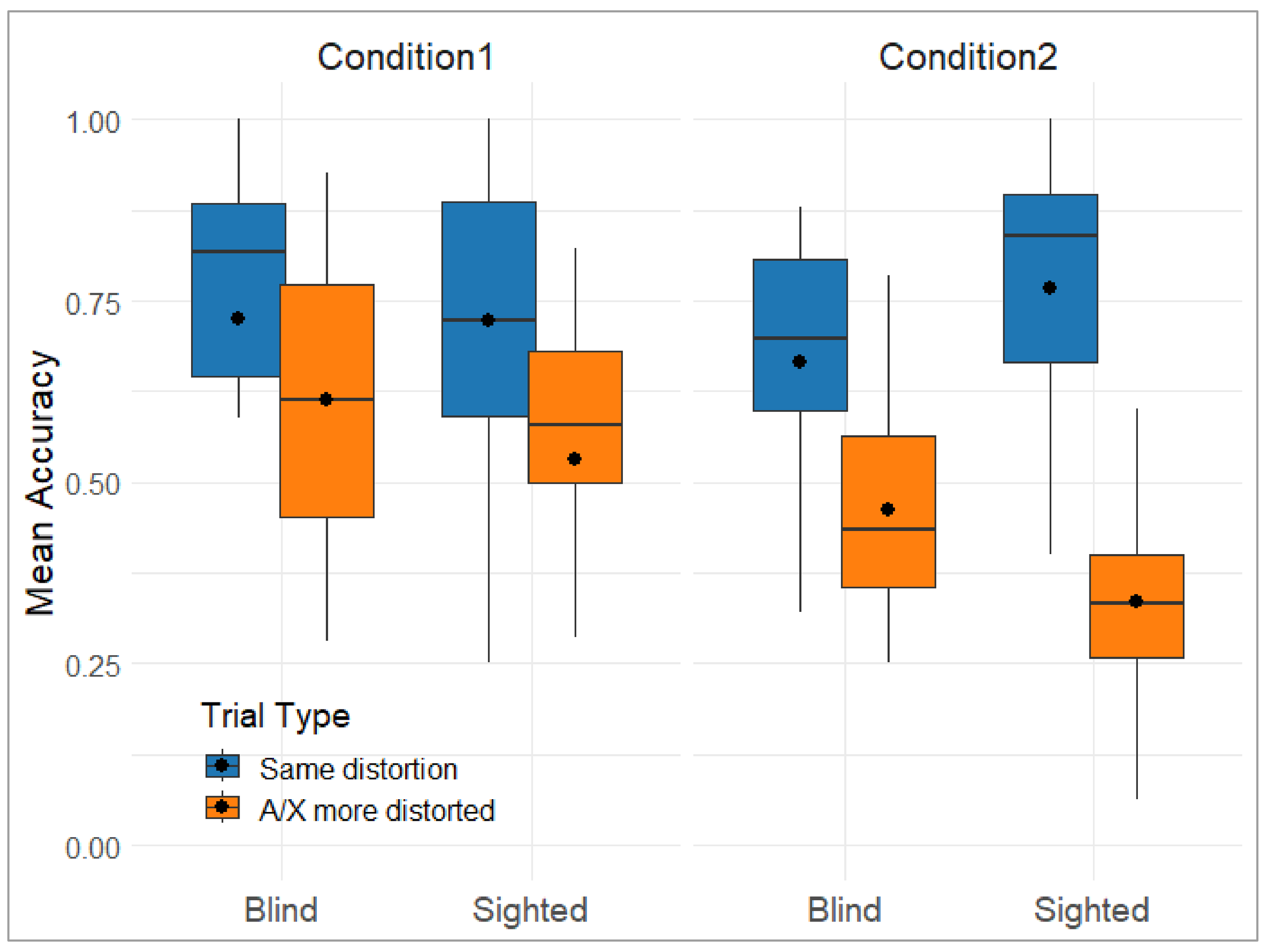
- **How do blind vs. sighted listeners perceive synthetic speech of different quality?** (replication of Melnik-Leroy & Navickas (2023) using a more sensitive task)
- **Do they rely on different perceptual strategies?**

Methods

- 2 groups x 2 conditions
- **Participants**: congenitally blind + sighted
- **Stimuli**: sequences of 1-3 words in **3 qualities** synthesized using Deep Neural Networks
- **Ternary Speeded AX** discrimination task: participants heard two versions of the same word, either in the same quality or in two different qualities.

	"Which item is more distorted: A, X, or are they the same?"
Condition 1 (easier)	A LOW quality ⟷ X HIGH quality
Condition 2 (harder)	A LOW quality ⟷ X MEDIUM quality

Results: Accuracy



Performance drops with difficulty more in sighted group.

Conclusions and Implications

The results showed that:
- The **same acoustic cues** that were **informative** for one group were uninformative, or even **confusing**, for the other, indicating distinct perceptual strategies. This tells us that synthetic speech is not perceived the same way by all listeners.

Implications for TTS evaluation and optimization:
- **Evaluate by group**: TTS quality should be reported separately for different listener groups (e.g., blind vs sighted), not just as a single global score.
- **Evaluate by item**: Item-level analyses should be standard, to detect words that systematically fail for specific groups.
- **Design for cue robustness**: TTS optimization should prioritize acoustic cues that work consistently across groups, or explicitly flag group-specific trade-offs.
- **Accessibility testing**: Systems used in assistive technologies (e.g., screen readers) should undergo dedicated evaluations with their target user populations.

Results: Strategies

- The top hardest and easiest items – differ for blind and sighted groups

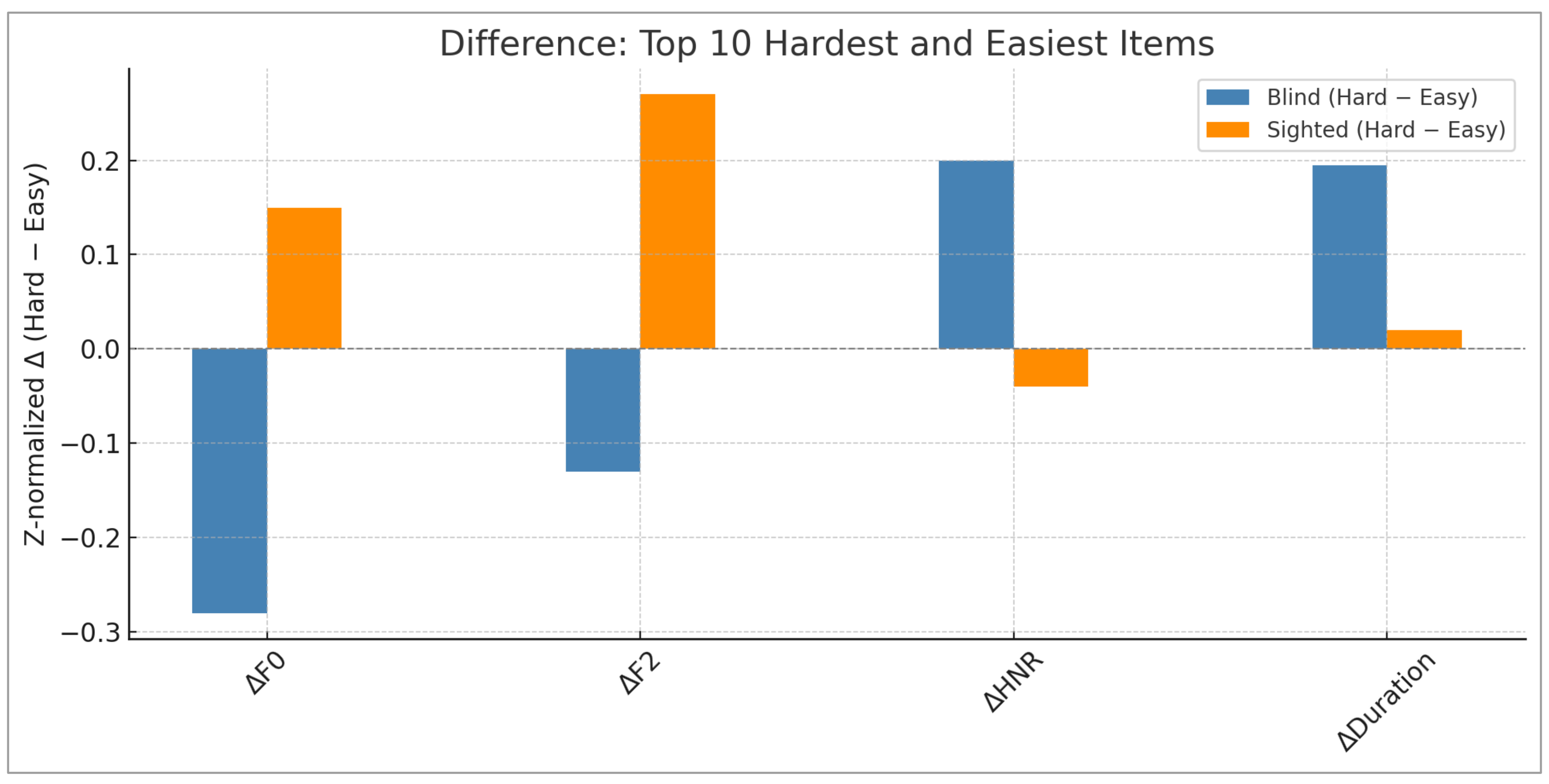
=> Do blind and sighted listeners rely on different acoustic cues?

Approach:
- Identified top 10 easiest and hardest items per group
- comparison between LOW and MEDIUM quality TTS versions

Preprocessing:
- Phoneme-level acoustic annotation
- Focus on key acoustic cues: F0-F3, Duration, HNR, Intensity, Jitter, Shimmer

Cue Informativeness: for each cue and phoneme:
- Calculated ΔCue = value in LOW – value in MEDIUM TTS
- A larger |ΔCue| means that LOW and MEDIUM are more acoustically separated on that cue.

- Compared ΔCue between easy and hard items
  Negative bar:
  cue difference larger in easy items → likely *perceptually helpful*
  Positive bar:
  cue difference larger in hard items → likely *perceptually confusing or unused*



Results:
- **Blind** listeners benefited from F0 and F2 differences
- For **Sighted** F0 and F2 uninformative or confusing
- **Suggests asymmetric cue weighting and group-specific perceptual strategies**.

Contact: Gerda Ana Melnik-Leroy, gerda.melnik@mif.vu.lt