LITHUANIAN COMPUTER SOCIETY
VILNIUS UNIVERSITY, INSTITUTE OF DATA SCIENCE AND DIGITAL TECHNOLOGIES
LITHUANIAN ACADEMY OF SCIENCES

16th Conference on

DATA ANALYSIS METHODS
for Software Systems

November 27–29, 2025

Druskininkai, Lithuania, Hotel "Europa Royale"
https://www.mii.lt/DAMSS

VILNIUS UNIVERSITY PRESS
Vilnius, 2025

Contacts:

Dr. Jolita Bernatavičienė
jolita.bernataviciene@mif.vu.lt
Prof. Olga Kurasova
olga.kurasova@mif.vu.lt
Tel. (+370 5) 2109 315

https://doi.org/10.15388/DAMSS.16.2025
ISBN 978-609-07-1200-9 (digital PDF)

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

gerda.ana.melnik@gmail.com

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).

On each trial, two renditions of the same word were presented and listeners indicated whether A was more distorted, X was more distorted, or both sounded the same. Using three synthesis quality levels (LOW, MEDIUM, HIGH), we defined two conditions: LOW–HIGH (easier as larger quality gap) and LOW–MEDIUM (harder as small quality gap). Both groups performed better in the easier condition, yet they diverged as the quality gap narrowed: sighted listeners showed reduced accuracy, whereas blind listeners were comparatively stable. Crucially, difficulty was stimulus- and group-specific: the same lexical items shifted between "easy" and "hard" across groups, rather than increasing uniformly with nominal task complexity.

To probe the perceptual mechanisms used, we conducted item-level acoustic analyses contrasting cue differences between "easy" and "hard" items for each group. We performed segment-level (phoneme) annotation for each token, extracted acoustic measures per segment under each quality condition, and computed per-cue quality deltas (LOW–HIGH or LOW–MEDIUM) that were z-normalized. For each group, we ranked words by accuracy, took the top-10 "easy" and top-10 "hard," and plotted, for each listener group and cue, the difference between the mean z-normalized quality deltas of the top-10 easy and top-10 hard words. 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 isn't perceived the same way by all listeners. Listener background shapes how speech is judged, even when intelligibility is high. Accordingly, TTS evaluation should not collapse across heterogeneous listeners or across items. Protocols that report group-wise and item-level results, and inspect cue separability can expose asymmetric confusions that standard aggregate scores miss. Treating listener diversity as a first-order factor is necessary to avoid mischaracterizing model quality and to ensure that improvements generalize across populations.