Human Assessment of AI Soundscape Descriptions

Instructions for Respondents

Thank you for agreeing to contribute to the *AI Soundscape Description* Project.

To avoid noise interference, please wear headphones to participate in this experiment.

Along with this information sheet, you were sent 2 Excel files (P*n*_excel with local links, P*n*_google with Google Drive links*)*. The excel sheet includes a list of the recordings you have been assigned to evaluate, **some from synthetic datasets and others from natural datasets**, and their captions. Some of these captions were created by a Large Language Model (LLM) and some were created by a human describer. When you have completed your assessment, please email the completed Excel file (*either file you used*) to Yuanbo.Hou@UGent.be .

Assessment

We introduce the Transparent Human Benchmark for Soundscapes (THumBS)[1] rubric.

1) THumBS uses 2 main scores each on a scale from 1 to 5: **Precision** and **Recall**.

Precision (P) $\epsilon[1, 5]$, measures how precise the caption is given the soundscape.

Recall (R) $\epsilon[1, 5]$, measures how much of the salient information (e.g. objects, attributes, and relations) from the soundscape is covered by the caption.

The overall score is computed by averaging precision and recall and deducting penalty points.

2) There are 3 possible penalty points: **Fluency, Conciseness,** and **Irrelevance.**

Each of these penalties should range from 0 to -2.

Fluency (F) $\epsilon[-2, 0]$, refers to the quality of captions as English text regardless of the given image. If a fluency problem is expected to be easily corrected by a text postprocessing algorithm (e.g. grammatical error correction, the penalty should be -0.1. 0.5+ points should be subtracted for more severe problems such as duplication, ambiguity, and broken sentences.

Conciseness (C) $\epsilon[-2, 0]$, penalty is applied for repetitive descriptions. These do not need to be direct repetitions of text and should be applied if the same idea or object are repeated in the description.

Irrelevance (I)[2] $\epsilon[-2, 0]$, penalty is applied for descriptions which introduce irrelevant details or sound sources do not present in the soundscape.

Layout

A column is available for your comments on the description to expand upon or add context to your scores. Example assessments are given in the **'Example' sheet** of the Excel file. ***Please see the example first.*** Please complete your assessments in the **'Assessment' sheet**.

Authors: *Yuanbo Hou (UGent), Qiaoqiao Ren (UGent), Andrew Mitchell (UCL), Wenwu Wang (UoS), Tony Belpaeme (UGent), Jian Kang (UCL), Dick Botteldooren (UGent).*

[1] Based on THumB for image captioning proposed by Kasai et. al. (2022).

[2] Adapted from the Irrelevance metric of Kreiss et al. (2022).