cleanup for 260 data

2025-11-27

```
knitr::opts_chunk$set(echo = TRUE)
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
knitr::opts_chunk$set(warning = FALSE, message = FALSE)
install.packages("kableExtra")
```

```
## Installing package into '/usr/local/lib/R/site-library'
## (as 'lib' is unspecified)
```

```
library(openintro)
```

```
## Loading required package: airports
```

```
## Loading required package: cherryblossom
```

```
## Loading required package: usdata
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(tidyverse)
library(xtable)

lang_data <- read_csv("wcs_language_best_mse.csv")
speaker_data <- read_csv("wcs_per_speaker_mse.csv")
glimpse(lang_data)
```

```
## Rows: 110
## Columns: 5
## $ language_id       <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1~
## $ best_speaker_id   <dbl> 12, 14, 15, 3, 3, 6, 19, 25, 25, 11, 2, 17, 6, 1, 15~
## $ best_mse_cielab   <dbl> 0.14863657, 0.13044236, 0.10008133, 0.11998912, 0.13~
```

```
## $ median_mse_cielab <dbl> 0.1854640, 0.1645608, 0.1507974, 0.1580962, 0.153670~
## $ n_speakers        <dbl> 25, 24, 25, 35, 6, 27, 25, 25, 30, 25, 25, 25, 25, 2~
```

```
glimpse(speaker_data)
```

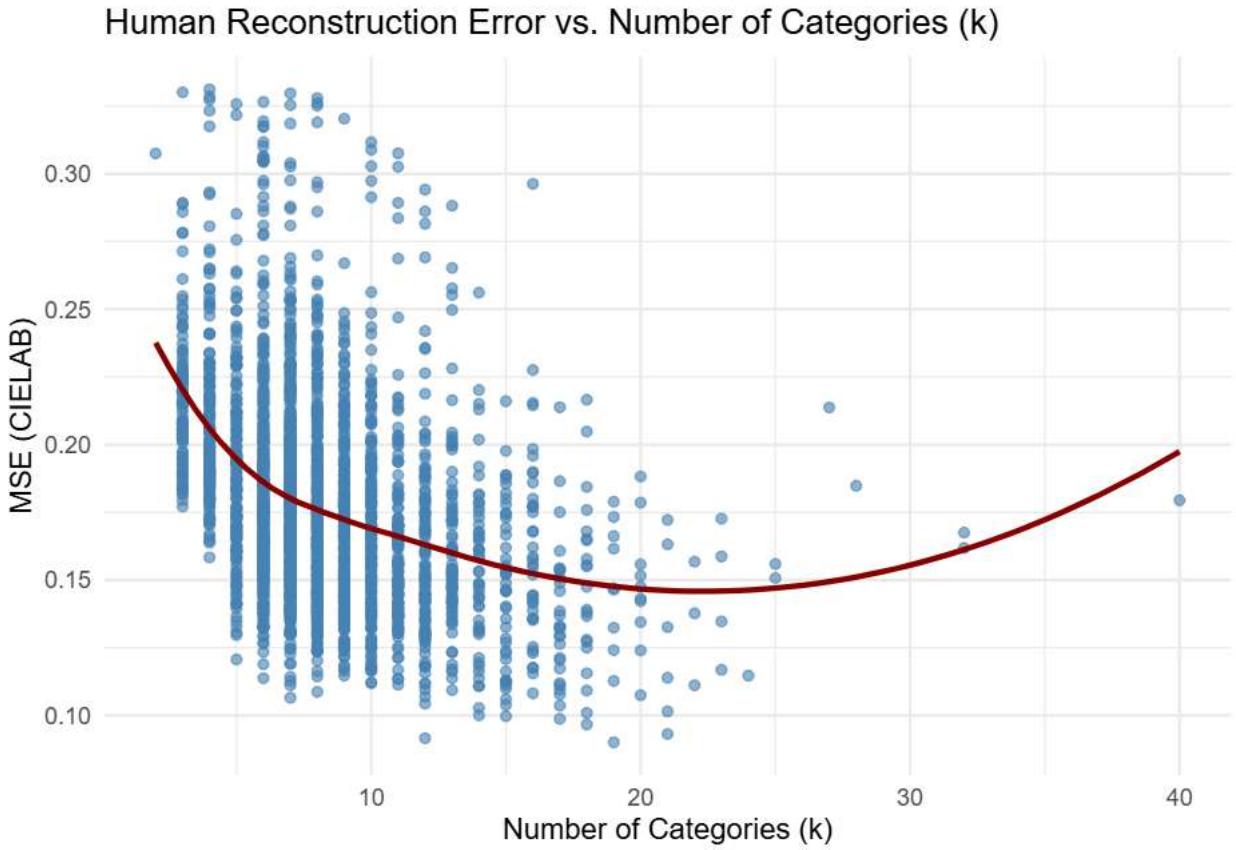
```
## Rows: 2,616
## Columns: 5
## $ language_id     <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ speaker_id      <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,~
## $ k_terms         <dbl> 6, 7, 6, 6, 7, 6, 7, 7, 7, 7, 8, 6, 8, 8, 6, 7, 7, 6, ~
## $ n_labeled_chips <dbl> 330, 330, 330, 330, 330, 330, 330, 330, 330, 330, 330,~
## $ mse_cielab      <dbl> 0.2107872, 0.1832399, 0.2055478, 0.1771522, 0.1925591,~
```

```
lang_data <- lang_data %>%
  rename(
    language_id = language_id,
    best_speaker_id = best_speaker_id,
    best_mse = best_mse_cielab,
    median_mse = median_mse_cielab,
    n_speakers = n_speakers
  )

speaker_data <- speaker_data %>%
  rename(
    language_id = language_id,
    speaker_id = speaker_id,
    k_terms = k_terms,
    n_chips = n_labeled_chips,
    mse = mse_cielab
  )

summary_table <- lang_data %>%
  arrange(best_mse) %>%
  slice(1:10)


print(xtable(summary_table), include.rownames = FALSE)
```

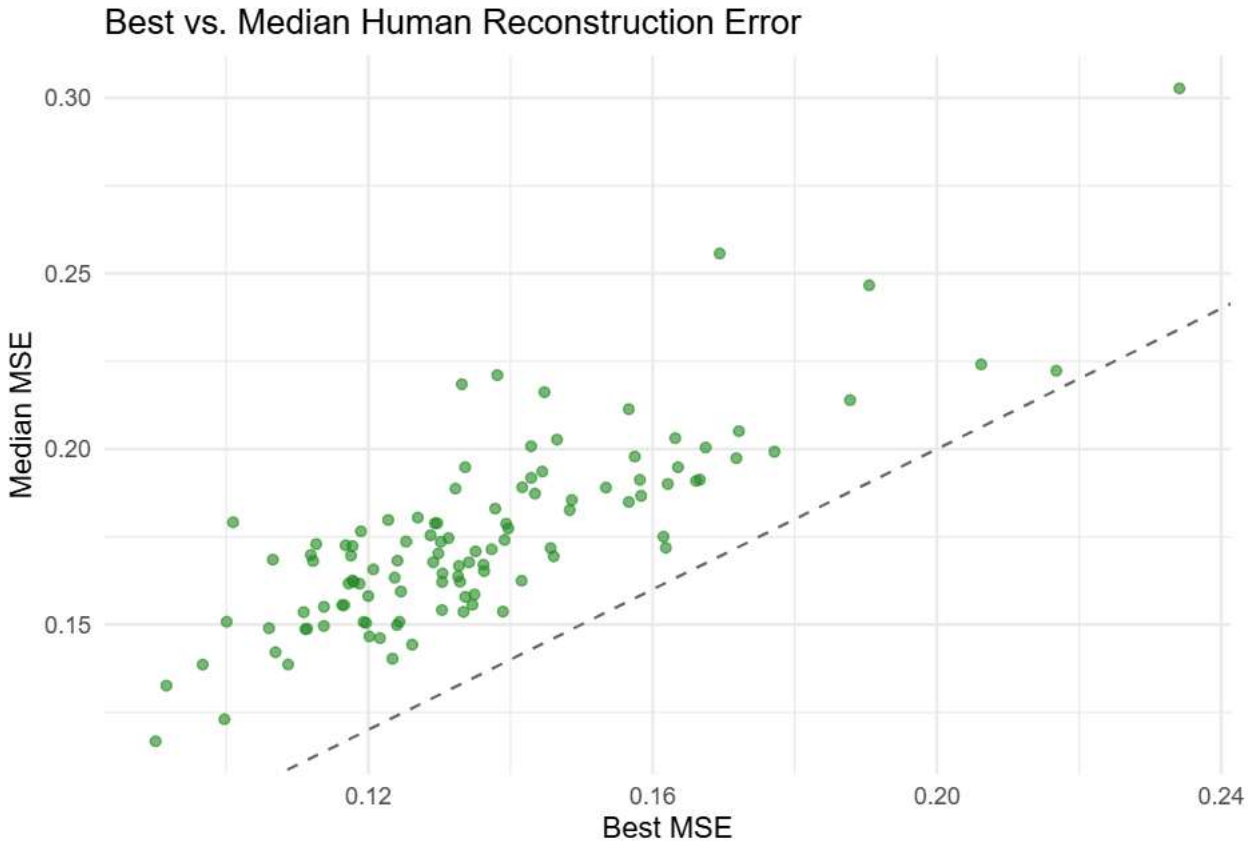
```
## % latex table generated in R 4.3.2 by xtable 1.8-4 package
## % Fri Nov 28 19:34:20 2025
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrrrr}
##   \hline
## language\_id & best\_speaker\_id & best\_mse & median\_mse & n\_speakers \\
##   \hline
## 17.00 & 2.00 & 0.09 & 0.12 & 30.00 \\
##   106.00 & 24.00 & 0.09 & 0.13 & 25.00 \\
##   27.00 & 10.00 & 0.10 & 0.14 & 25.00 \\
##   7.00 & 19.00 & 0.10 & 0.12 & 25.00 \\
##   3.00 & 15.00 & 0.10 & 0.15 & 25.00 \\
##   20.00 & 1.00 & 0.10 & 0.18 & 11.00 \\
##   84.00 & 11.00 & 0.11 & 0.15 & 25.00 \\
##   42.00 & 5.00 & 0.11 & 0.17 & 25.00 \\
##   24.00 & 18.00 & 0.11 & 0.14 & 25.00 \\
```

```
##   110.00 & 20.00 & 0.11 & 0.14 & 25.00 \\
##    \hline
## \end{tabular}
## \end{table}
```

```
ggplot(speaker_data, aes(x = k_terms, y = mse)) +
  geom_point(alpha = 0.6, color = "steelblue") +
  geom_smooth(method = "loess", se = FALSE, color = "darkred") +
  labs(
    title = "Human Reconstruction Error vs. Number of Categories (k)",
    x = "Number of Categories (k)",
    y = "MSE (CIELAB)"
  ) +
  theme_minimal()
```




```
ggplot(lang_data, aes(x = best_mse, y = median_mse)) +
  geom_point(alpha = 0.6, color = "forestgreen") +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "grey40") +
  labs(
    title = "Best vs. Median Human Reconstruction Error",
    x = "Best MSE",
    y = "Median MSE"
  ) +
  theme_minimal()
```


Best vs. Median Human Reconstruction Error
Median MSE
0.30
0.25
0.20
0.15
0.12
0.16
0.20
0.24
Best MSE