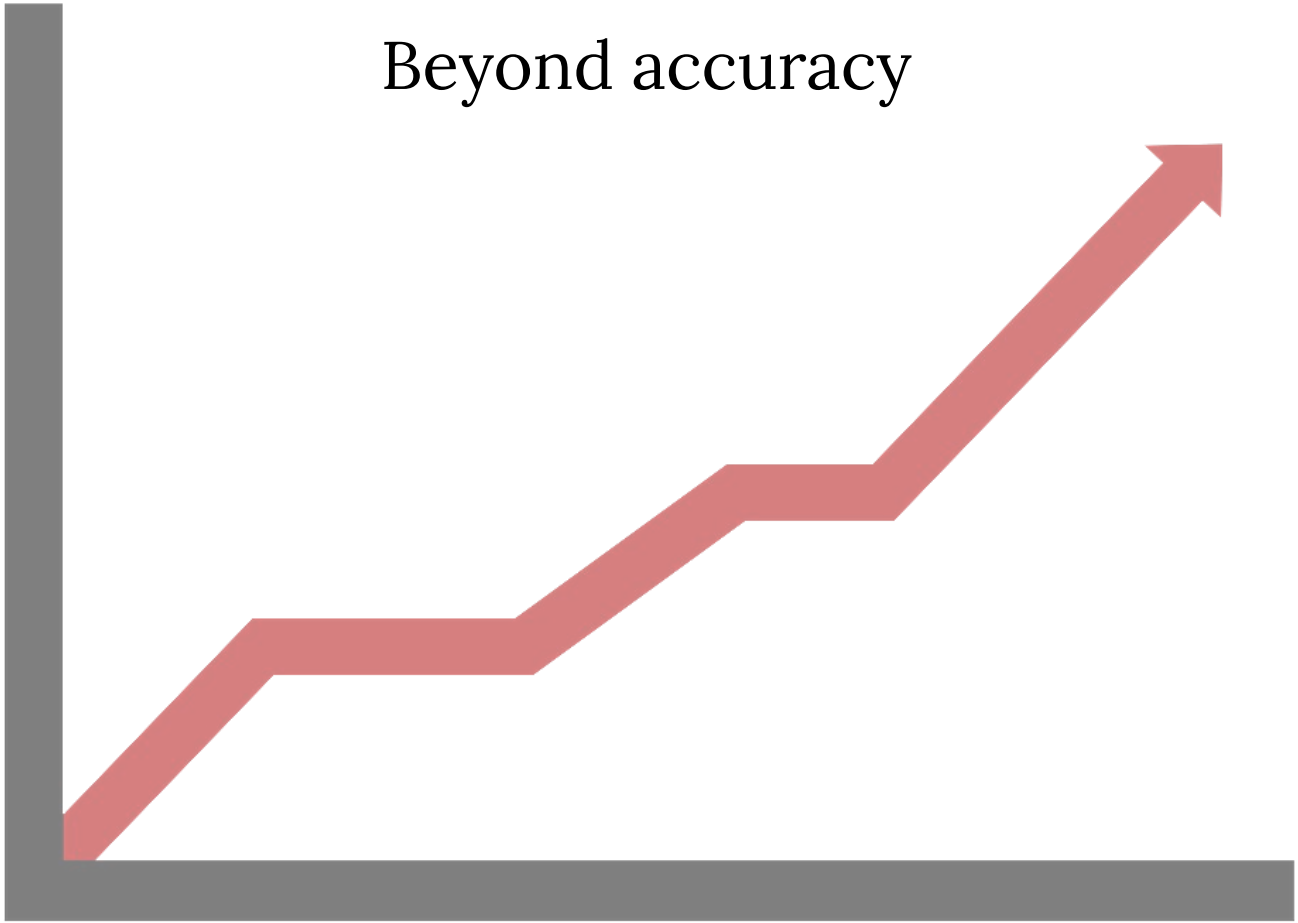
Challenges for interpretability of speech models

Grzegorz Chrupała

Department of Cognitive Science and AI
Tilburg University

The growing importance of real world speech speech applications.

Beyond accuracy



What questions about models do we want to answer?

Modalities

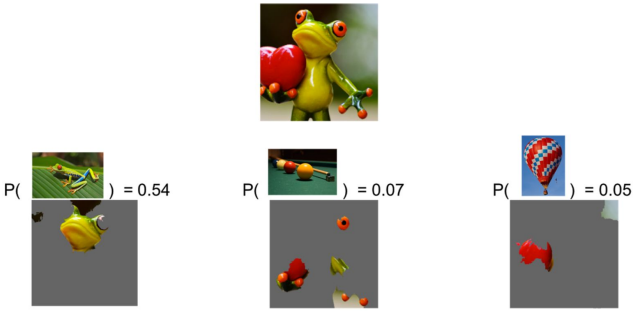
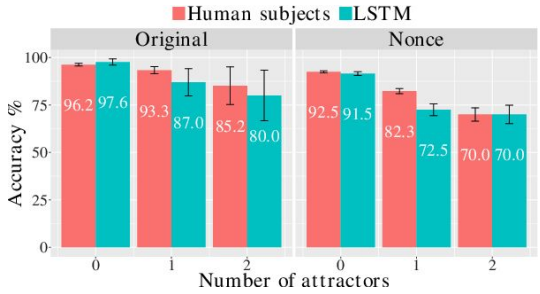
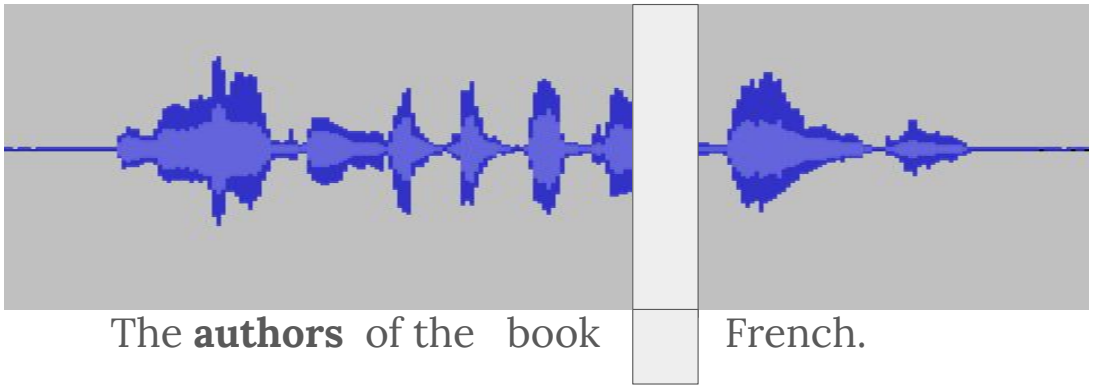
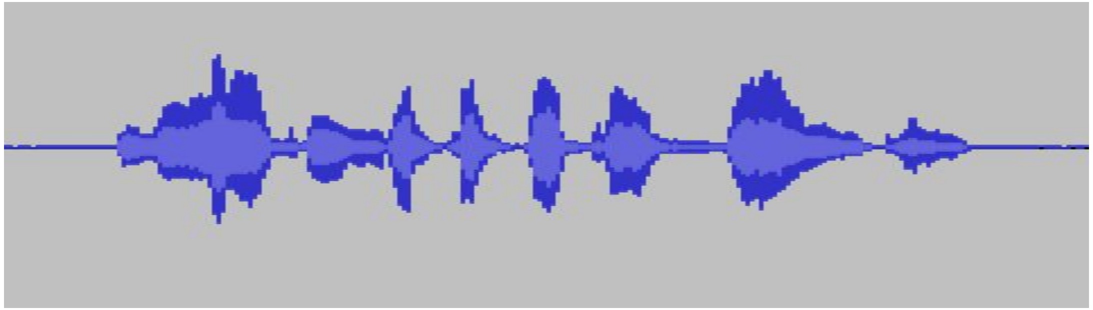


Figure 6. Explanation for a prediction from Inception. The top three predicted classes are "tree frog," "pool table," and "balloon." Sources: Marco Tulio Ribeiro, Pixabay (frog, billiards, hot air balloon).



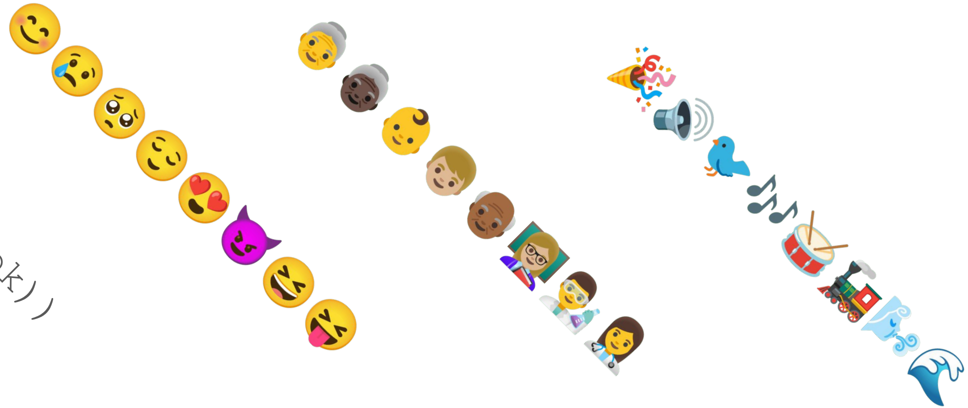
Yet the ratio of men who survive to the women and children who survive [is/are]

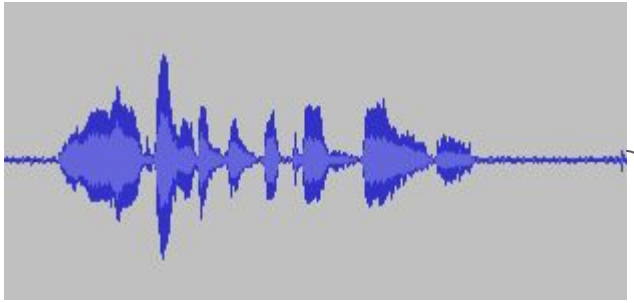


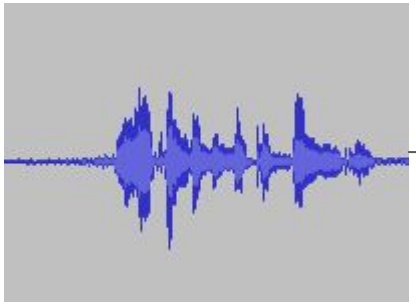


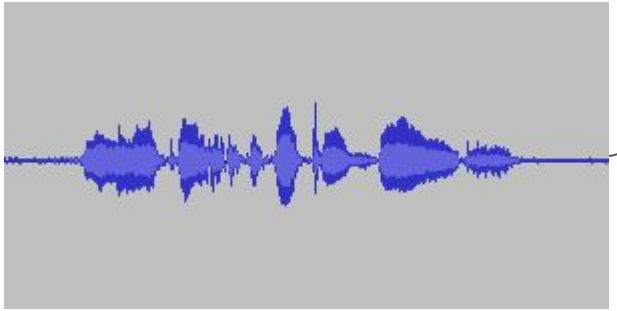
ðə ˈɔːθɚz əv ðə bʊk ɑɹ fɹɛntʃ

French (author (book))

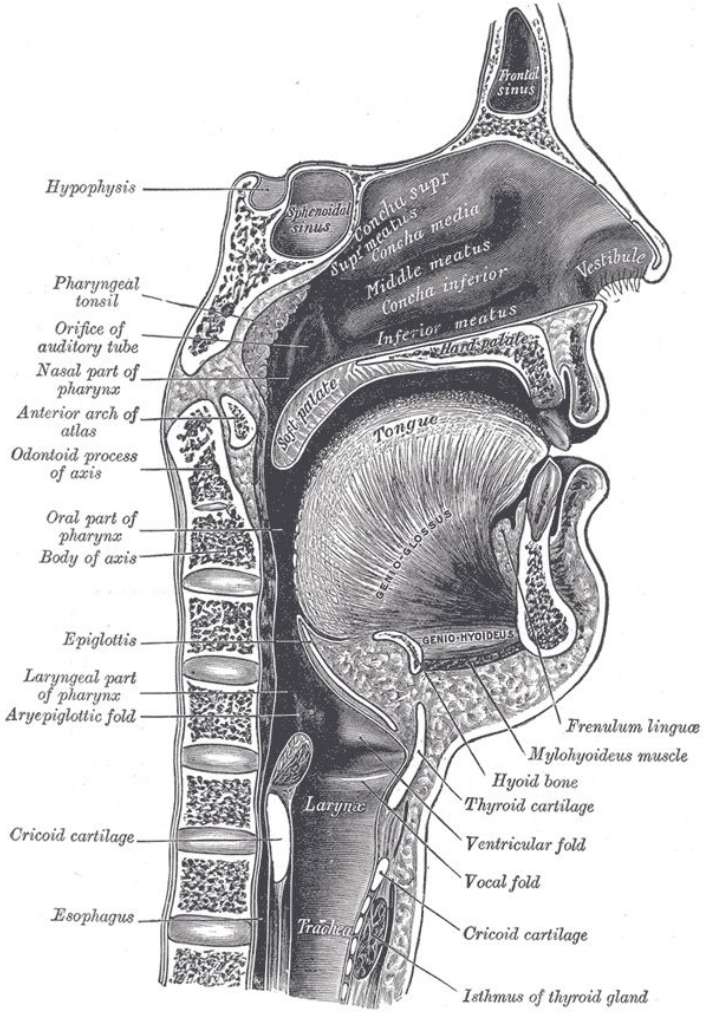








The authors of the book are French.



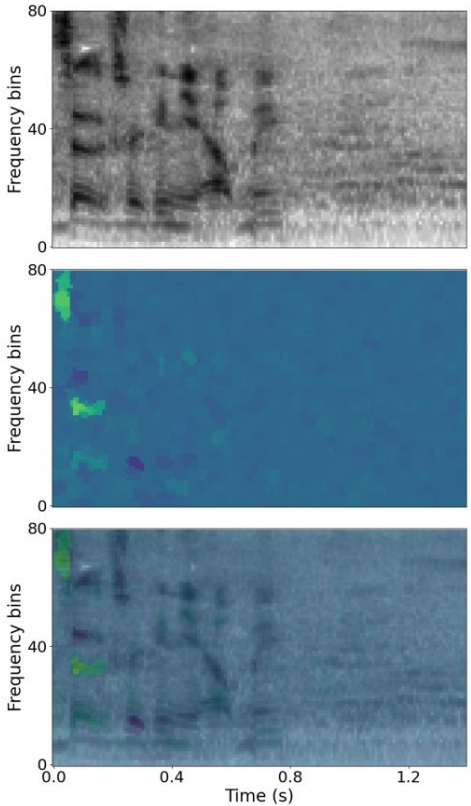


Figure 8: Example of saliency map (middle) for the token `so` (ASR), along with the corresponding spectrogram (top) and the map overlayed on the spectrogram (bottom).

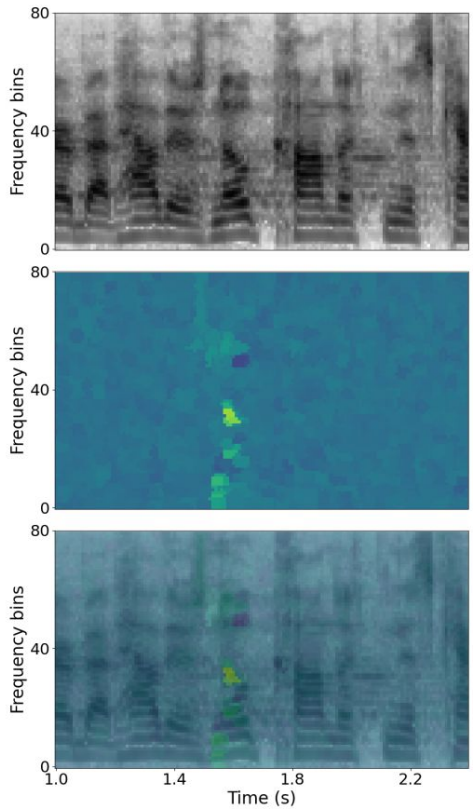


Figure 9: Example of saliency map (middle) for the token `no` (ASR), along with the corresponding spectrogram (top) and the map overlayed on the spectrogram (bottom).

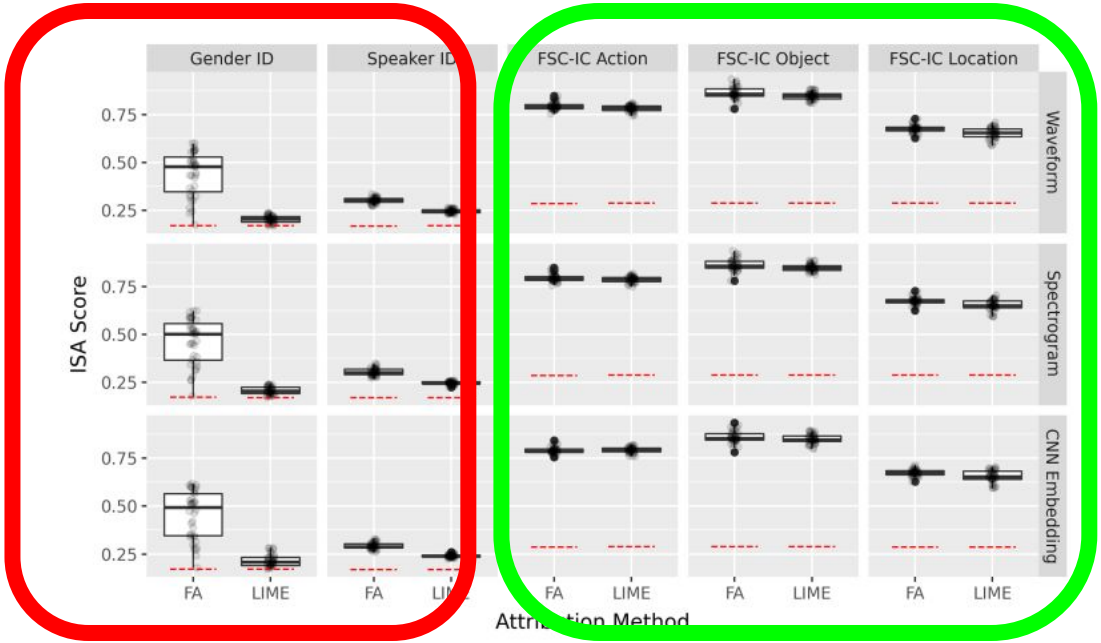


Figure 3: *Distributions of ISA scores with perturbation operating directly on word-aligned segments. The rows indicate different input feature types, the columns are different tasks. Within each panel, each boxplot report results from different attribution methods and the y-axis is the ISA score. The red dotted line indicates the randomly shuffled baseline. FA: Feature Ablation.*