Challenges for interpretability of speech models

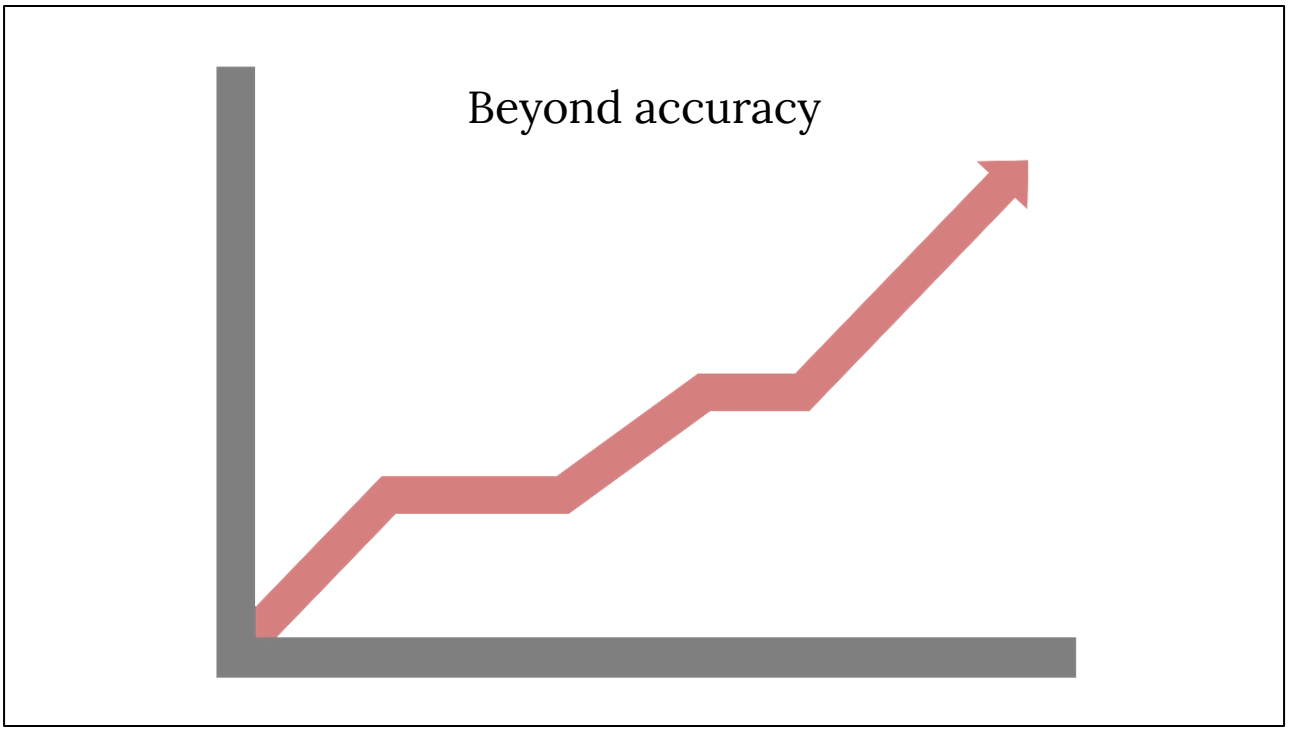
Grzegorz Chrupała

Department of Cognitive Science and AI
Tilburg University

What the special characteristics and challenges for the interpretability of models of spoken language?

The growing importance of real world speech speech applications.

- Spoken language translation
- Transription of
  - Medical records
  - Police interrogations
- Spoken dialog systems



We want AI models to be accurate.

What other properties are desirable?

- Calibration – model is accurate about its uncertainty.
- Robustness – perturbations do not affect predictions drastically.
- Safety – adversarial attacks do not break the model.
- Compliance with social and legal norms.
- Ability to reveal structure in data.
- Explainability – can justify its decisions.

What questions about models do we want to answer?

Which features of an input example led to a particular prediction?

Which input features or feature combinations are important for model decisions?

How does the model behave when faced with particular phenomena?

What information is encoded in activation patterns?

What algorithm does the model implement?

Modalities

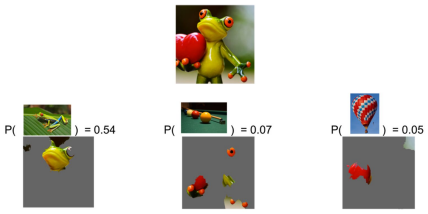
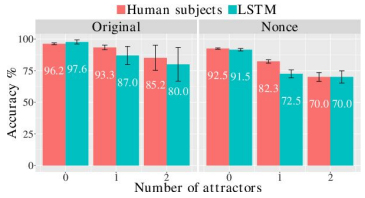
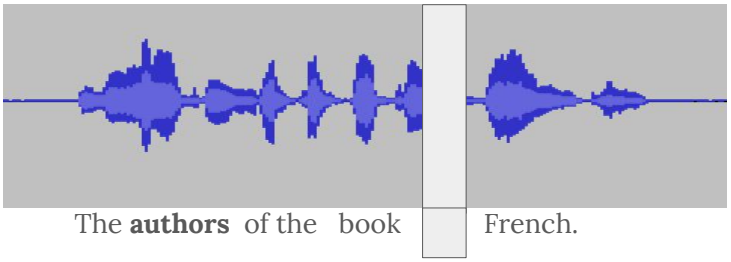


Figure 6. Explanation for a prediction from Inception. The top three predicted classes are "tree frog," "pool table," and "balloon." Sources: Marco Tulio Ribeiro, Pixabay (frog, billiards, hot air balloon).



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

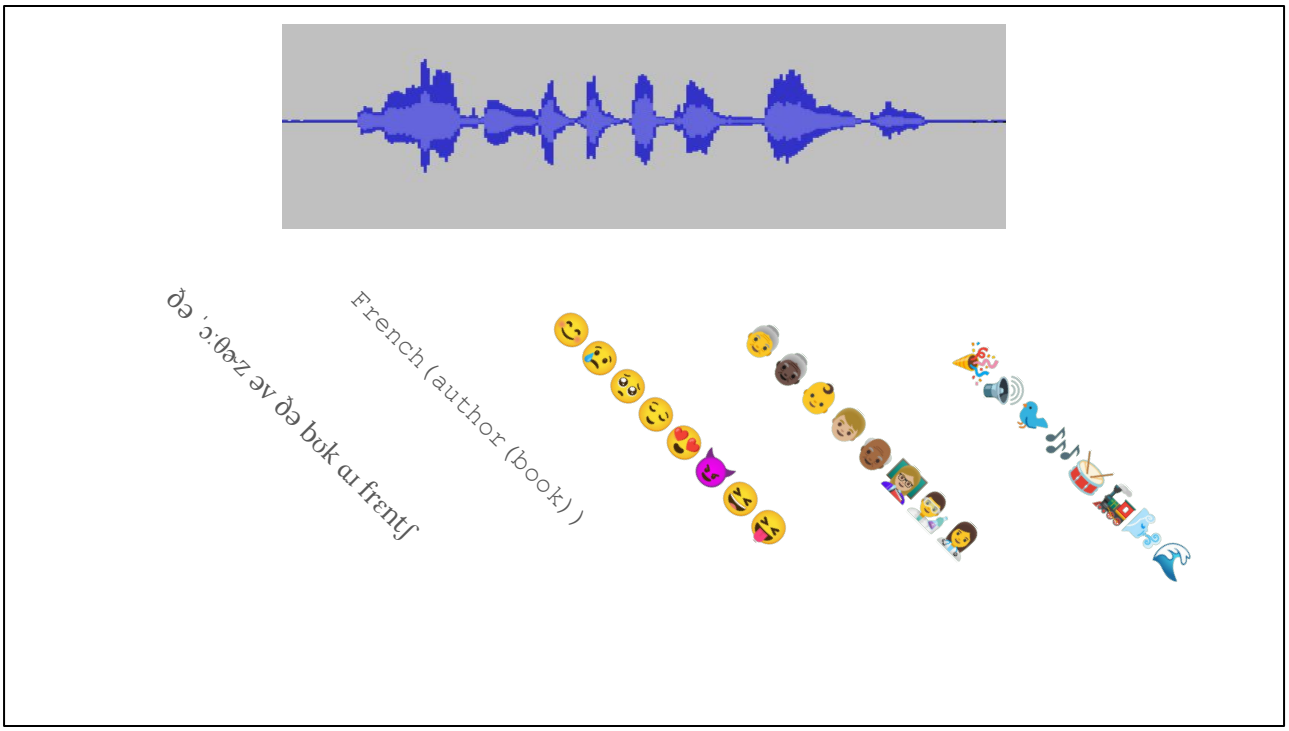


Much of initial work on interpretability was applied to computer vision and also to text processing models.

Images have explicit two-dimensional spatial structure, with a latent third dimension.

Text has an explicit one-dimensional sequential structure and an underlying hierarchical structure. It is symbolic rather than continuous in nature.

Spoken language shares some characteristics with these two modalities but it is also distinct in important ways.



Speech is a continuous one-dimensional signal which can be understood as a mixture of multiple channels, each of them carrying a distinct type of information.

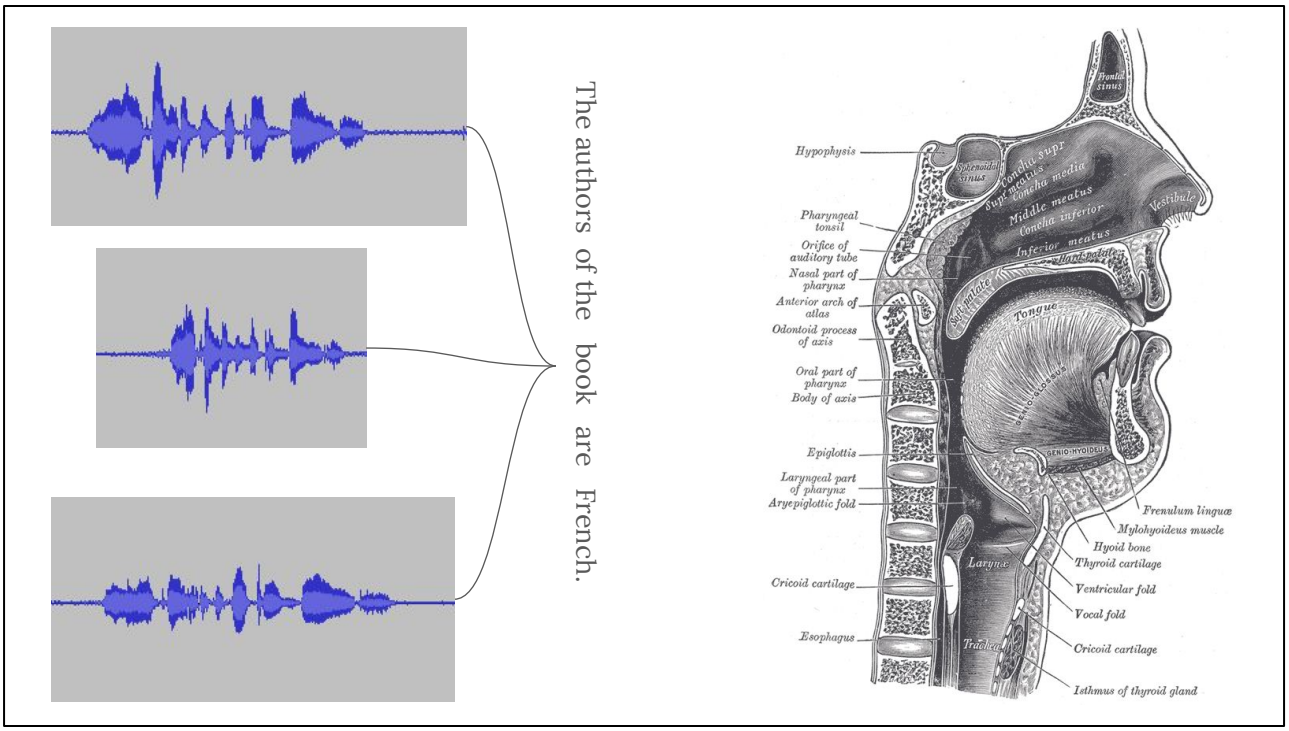
The main channel is the propositional content of spoken language, with an underlying discrete, hierarchical structure with layers corresponding to phonology, syntax, and semantics.

Other types of information are overlaid on this main channel, including paralinguistic features modulating the message, as well as speaker characteristics. These latter can be stable (such as speaker identity) or temporary (such as speaker emotional state).

Finally, speech signal usually also includes environmental sounds which can be a source of additional information as well as noise.

Due to these factors, models of speech tend to encode multiple

aspects of the speech signal, which may be hard to disentangle and may lead to confounds for interpretation.



Like text, speech also has an underlying discrete hierarchical structure. However, text is a simple discrete encoding of syntax and semantics, while speech is a highly variable continuous signal with a much more complex relation to the underlying hierarchical structure.

Images or videos are variable continuous signals, like speech, but they are constrained by the physical reality. While speech reflects the constraints of biologically and culturally evolved human language.

We expect the representations and computations learned by models of spoken language to differ and to possibly require different approaches to their analysis and interpretation.

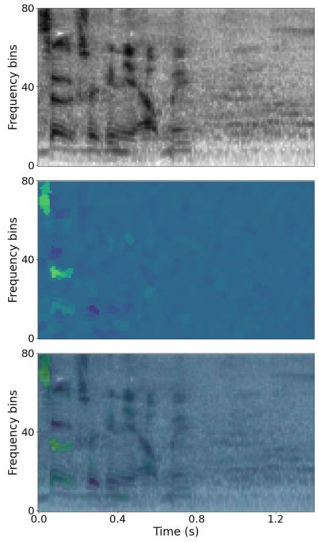


Figure 8: Example of saliency map (middle) for the token `so` (ASR), along with the corresponding spectrogram (top) and the map overlayed on the spectrogram (bottom).

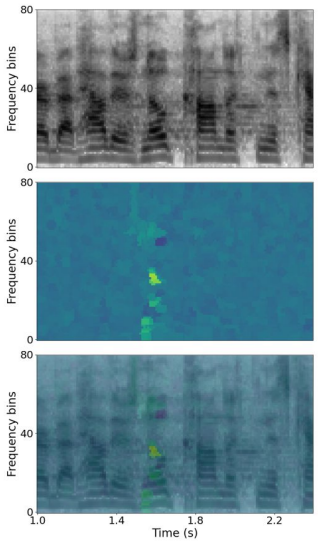


Figure 9: Example of saliency map (middle) for the token `no` (ASR), along with the corresponding spectrogram (top) and the map overlayed on the spectrogram (bottom).

Challenges with feature attribution and explainability

Naive humans have an intuitive grasp of features of text (such as words or word sequences) and of images (such as image segments).
They lack such understanding for most types of speech features, such as those that could be visualized via waveform or a spectrogram.

Example from Fucci, D., Gaido, M., Savoldi, B., Negri, M., Cettolo, M., & Bentivogli, L. (2024, November 3). *SPES: Spectrogram Perturbation for Explainable Speech-to-Text Generation.* arXiv. https://doi.org/10.48550/arXiv.2411.01710

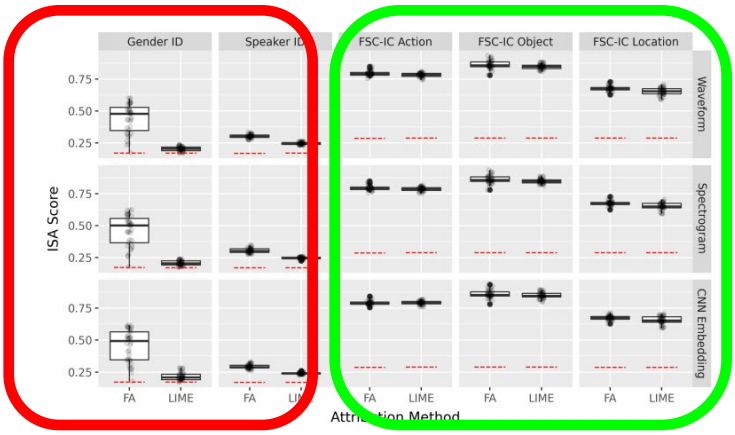


Figure 3: *Distributions of ISA scores with perturbation operating directly on word-aligned segments. The rows indicate different input feature types, the columns are different tasks. Within each panel, each boxplot report results from different attribution methods and the y-axis is the ISA score. The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

Beyond ease of interpretation, recent work shows that feature attribution for speech classification suffers from low reliability. Speech features in a spectrogram or waveform are high resolution and highly correlated and redundant, and simply aggregating them does not fully resolve the problem caused by these characteristics.

Example from Shen, G., Mohebbi, H., Bisazza, A., Alishahi, A., & Chrupała, G. (2025). On the reliability of feature attribution methods for speech classification. In Proceedings of Interspeech 2025.