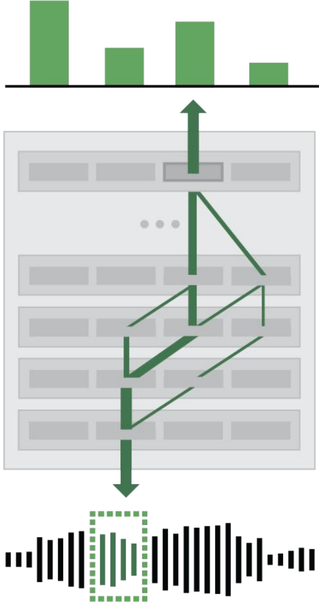
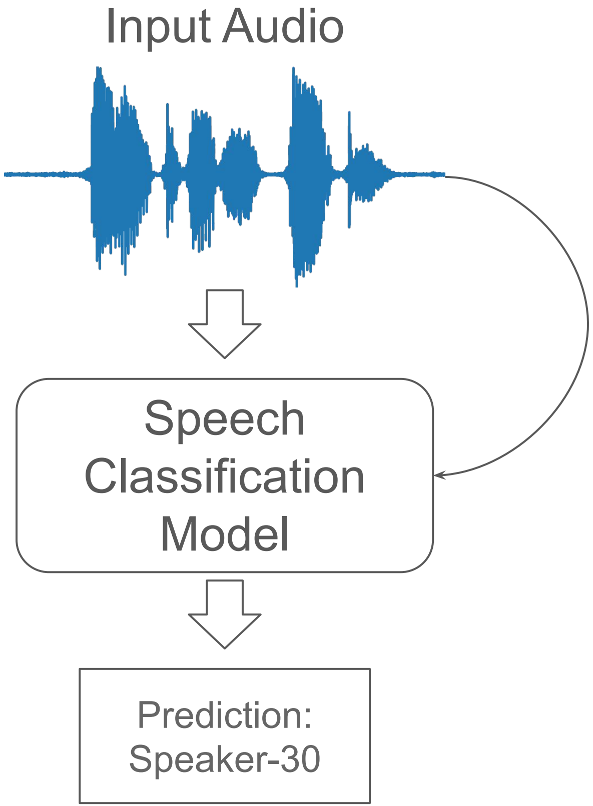
Feature attribution



Interpretability Techniques for Speech Models

Gaofei Shen, Tilburg University, The Netherlands

How does the model decide?



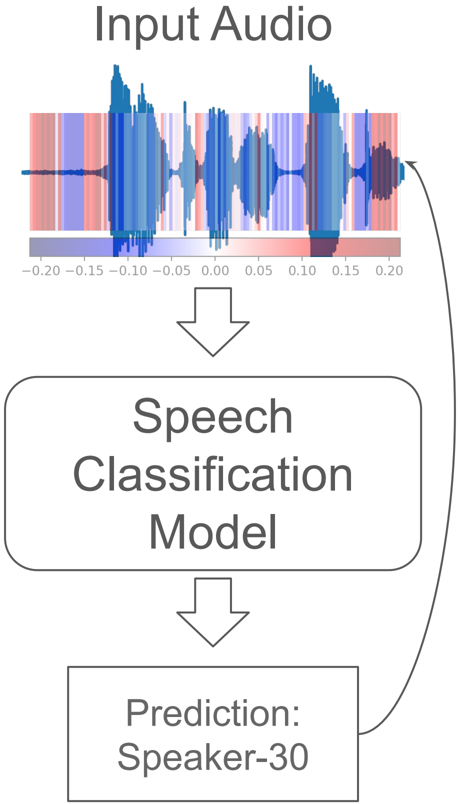
Input Audio

Speech Classification Model

Prediction: Speaker-30

Which part of the input audio enabled the model to correctly classify the speaker as Speaker-30?

Finding the "Important" Parts: Feature Attribution



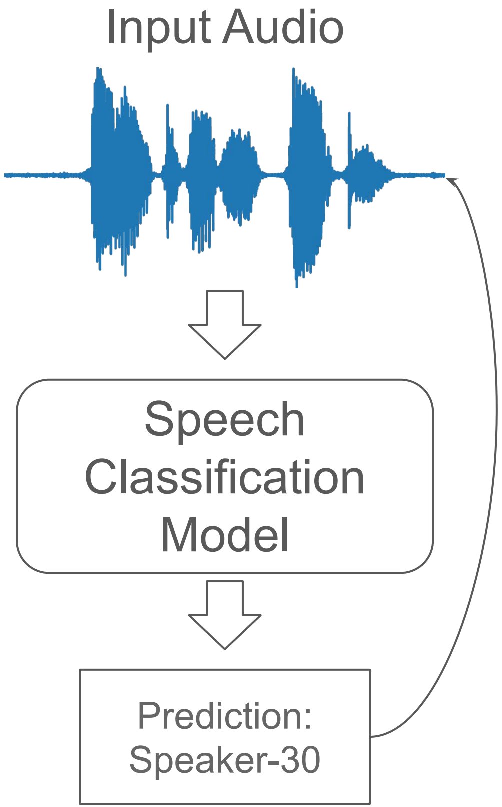
Input Audio

Speech Classification Model

Prediction: Speaker-30

Feature attribution is a family of techniques that help us assign **importance scores** to the ***input features***

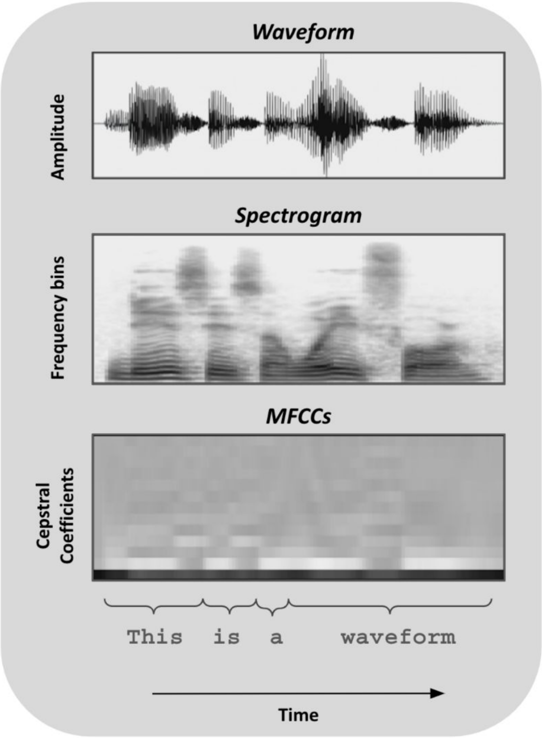
Help us answer the previous question!



Input Audio

Speech Classification Model

Prediction: Speaker-30



Fucci et al (2024)

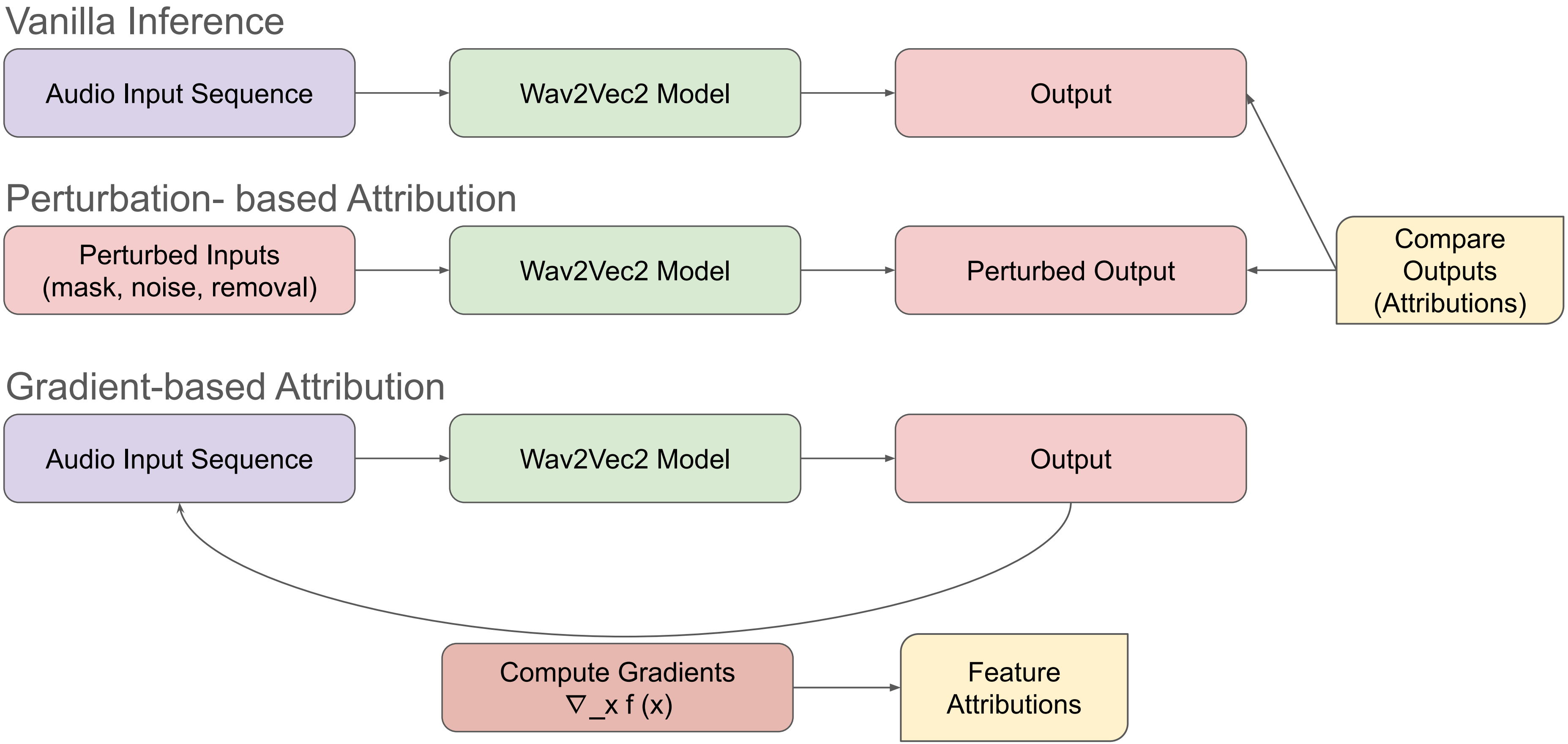
Flavors of Feature Attribution Methods

Gradient-based

- Derived from the gradient in model's internal computations
- Models have to be differentiable
- Fast (!)
- Examples:
  - Saliency
  - Integrated Gradients

Perturbation-based

- Modify input and measure changes in output
- Model agnostic
- Slow due to input perturbation
- Examples:
  - Occlusion



Vanilla Inference

Audio Input Sequence

Wav2Vec2 Model

Output

Perturbation- based Attribution

Perturbed Inputs
(mask, noise, removal)

Wav2Vec2 Model

Perturbed Output

Compare
Outputs
(Attributions)

Gradient-based Attribution

Audio Input Sequence

Wav2Vec2 Model

Output

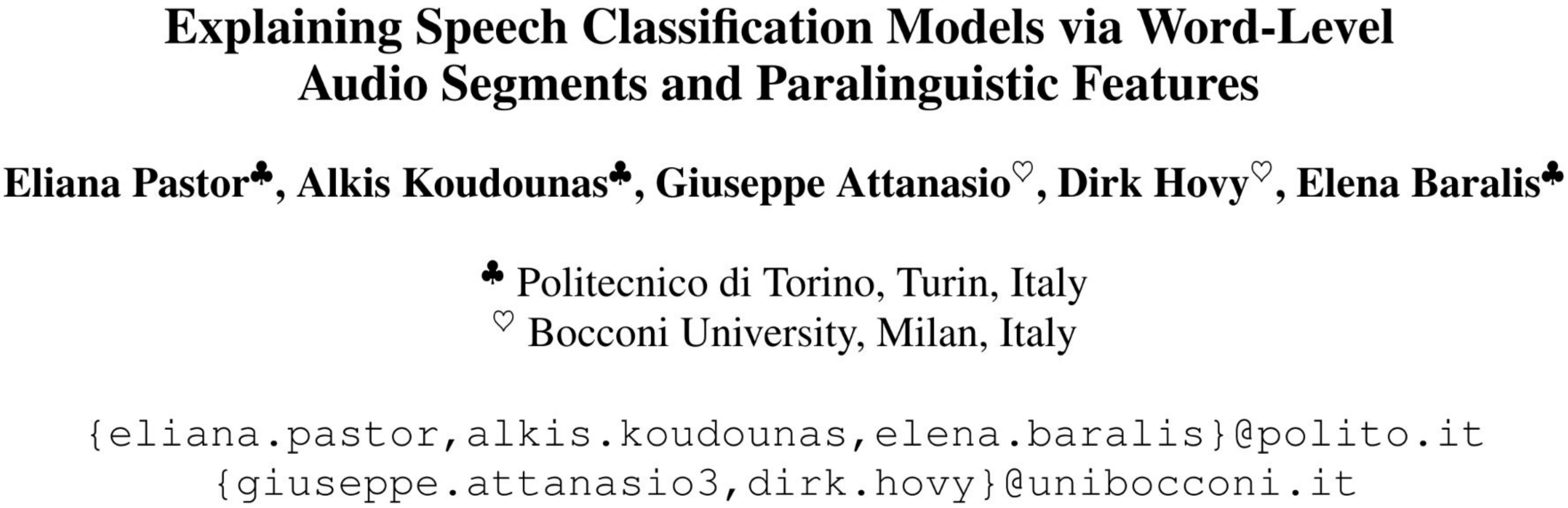
Compute Gradients
∇_x f (x)

Feature
Attributions

The Challenges in Speech

- Speech is not clean
    - Hard to find “baseline”
- There are no naturally occurring boundaries in speech
    - Difficult to “chunk” speech data in an intuitive way
- There are many ways of representing speech data
    - Waveform has time and amplitude
    - Spectrogram has time, amplitude AND frequency
- Information could be very spread out
    - E.g. Individual pitch data points doesn’t tell the complete story of the entire pitch contour

Case study 1



Explaining Speech Classification Models via Word-Level Audio Segments and Paralinguistic Features

Eliana Pastor[♣], Alkis Koudounas[♣], Giuseppe Attanasio[♡], Dirk Hovy[♡], Elena Baralis[♣]

[♣] Politecnico di Torino, Turin, Italy
[♡] Bocconi University, Milan, Italy

{eliana.pastor,alkis.koudounas,elena.baralis}@polito.it
{giuseppe.attanasio3,dirk.hovy}@unibocconi.it

Feature attribution for intent classification

- Insights from perturbing sections of waveform that correspond to word-level timestamps

- Relies heavily on annotation
  - (i.e. forced alignment)
- Departure from the continuous nature of speech signal encoding

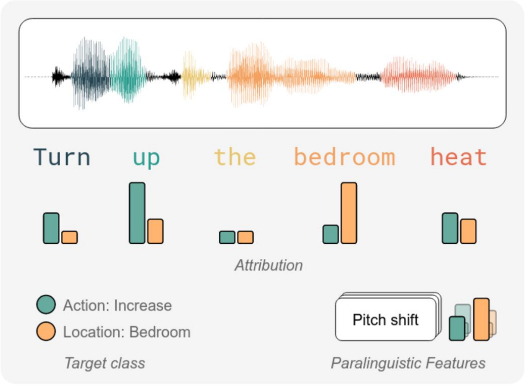
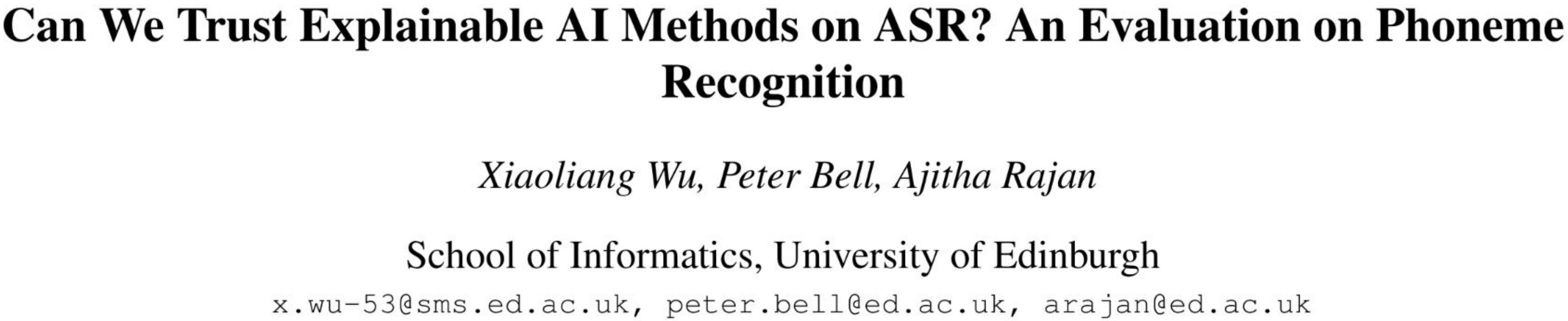


Figure 1: Explanation with word-level and paralinguistic attributes for a sample in Fluent Speech Commands (Lugosch et al., 2019). Word-level audio-transcript alignment represented through color. Word-level attributions to explain the *Increase* (green, left boxes) and *Bedroom* (orange, right) target classes.

Case study 2



Can We Trust Explainable AI Methods on ASR? An Evaluation on Phoneme Recognition

Xiaoliang Wu, Peter Bell, Ajitha Rajan

School of Informatics, University of Edinburgh

x.wu-53@sms.ed.ac.uk, peter.bell@ed.ac.uk, arajan@ed.ac.uk

Different ways of segmenting speech

- LIME-WS
  - Uses TIMIT word-level segmentation
- LIME-TS
  - Uses a fixed 70ms window

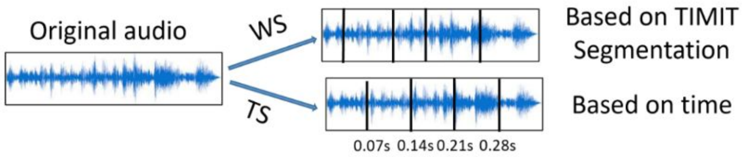
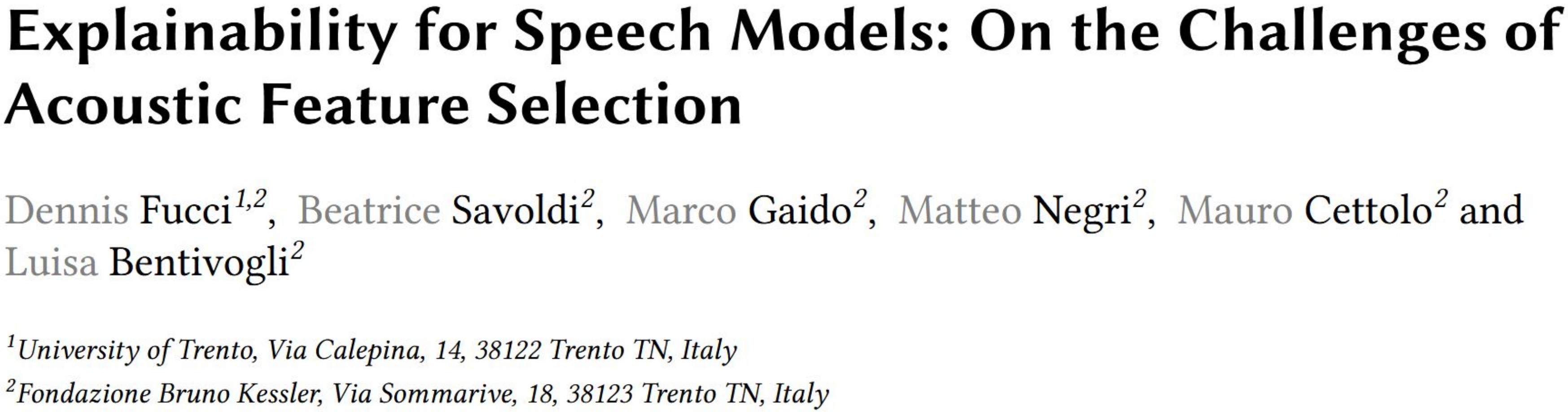


Figure 2: *Different segmentation used by (LIME-WS,LIME) and LIME-TS.*

Special shoutout



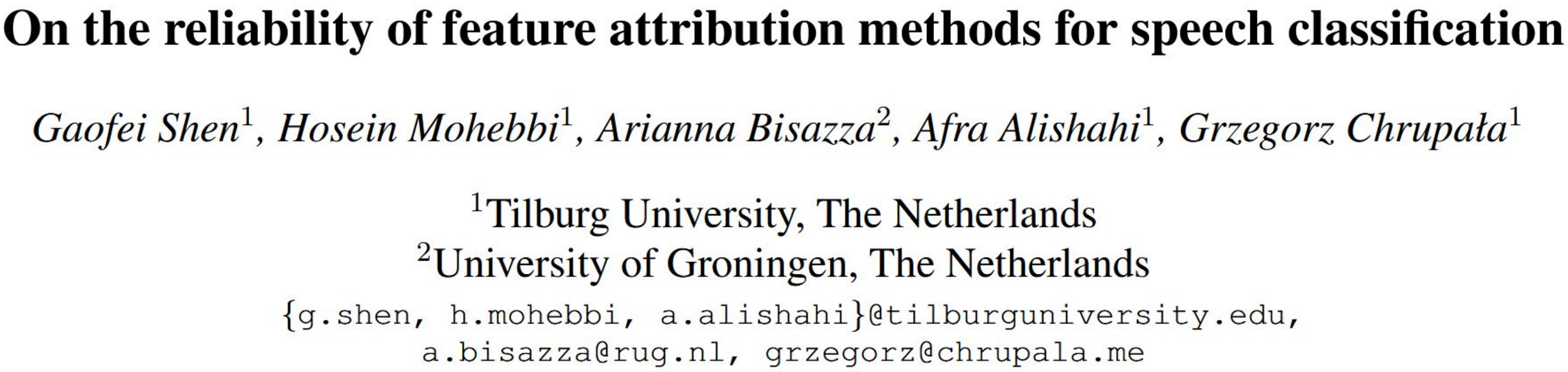
Explainability for Speech Models: On the Challenges of Acoustic Feature Selection

Dennis Fucci[1,2], Beatrice Savoldi[2], Marco Gaido[2], Matteo Negri[2], Mauro Cettolo[2] and Luisa Bentivogli[2]

[1] *University of Trento, Via Calepina, 14, 38122 Trento TN, Italy*

[2] *Fondazione Bruno Kessler, Via Sommarive, 18, 38123 Trento TN, Italy*

Self plug



On the reliability of feature attribution methods for speech classification

Gaofei Shen[1], *Hosein Mohebbi*[1], *Arianna Bisazza*[2], *Afra Alishahi*[1], *Grzegorz Chrupała*[1]

[1]Tilburg University, The Netherlands
[2]University of Groningen, The Netherlands
{g.shen, h.mohebbi, a.alishahi}@tilburguniversity.edu,
a.bisazza@rug.nl, grzegorz@chrupala.me

References

Pastor, E., Koudounas, A., Attanasio, G., Hovy, D., & Baralis, E. (2024). Explaining Speech Classification Models via Word-Level Audio Segments and Paralinguistic Features. In Y. Graham & M. Purver (Eds.), Proceedings of the 18th EACL. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.eacl-long.136

Wu, Xiaoliang, Peter Bell, and Ajitha Rajan. “Can We Trust Explainable AI Methods on ASR? An Evaluation on Phoneme Recognition.” *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2024, 10296–300. https://doi.org/10.1109/ICASSP48485.2024.10445989.

Fucci, D., Savoldi, B., Gaido, M., Negri, M., Cettolo, M., & Bentivogli, L. (2024). Explainability for Speech Models: On the Challenges of Acoustic Feature Selection. In F. Dell’Orletta, A. Lenci, S. Montemagni, & R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024) (pp. 373–381). CEUR Workshop Proceedings. https://aclanthology.org/2024.clicit-1.45/

Shen, Gaofei, Hosein Mohebbi, Arianna Bisazza, Afra Alishahi, and Grzegorz Chrupala. “On the Reliability of Feature Attribution Methods for Speech Classification.” 2025, 266–70. https://doi.org/10.21437/Interspeech.2025-1911.

Thank you!

https://interpretingdl.github.io/speech-interpretability-tutorial/

Feature attribution:

edu.nl/rk8vb

Notebook

Context-Mixing:

Notebook