P5 - Attribution-based explainability methods for speech classification models

Explainable and Trustworthy AI Course

Politecnico di Torino - 2023/2024

Reference teachers: Eliana Pastor, Eleonora Poeta (+ Alkis Koudounas)

Project. Given the increasing adoption of speech classification models, recent works in explainable AI address the issue of their lack of interpretability. This project focuses on identifying research gaps in feature attribution methods for speech classification models and proposes novel solutions to improve speech model interpretability.

Overview.

Speech classification models are widely adopted for various tasks, such as intent classification, keyword spotting, and emotion recognition. Speech models are increasingly adopted end-to-end, i.e., directly processing the audio signal without transcribing and operating on the text data. However, these models often lack interpretability, not revealing the reason behind individual prediction. To address this limitation, recent works address this issue by proposing feature attribution explainability methods. Some works operate on spectrograms [2, 1] and highlight parts of the spectrogram impacting predictions close to saliency map visualization in image classification tasks. Other works operate on the audio signal itself [3, 4, 6, 5], providing which audio segments impact the prediction, including equal-width audio [3] or phoneme-level [5] or word-level audio segments [4]. Moreover, recent works also focus on the paralinguistic aspects [4].

Goal.

The task of the project is first to review existing explanation methods for speech-based classification systematically. Then, the project aims to identify existing research gaps in the explainable AI literature for speech classification, focusing on feature-attribution explainability methods. Examples of research gaps include the combination of word-level explanations with paralinguistic aspects and assessing the quality of explanations.

Required analysis, implementation, and evaluation.

- **Literature Review.** Conduct a systematic review of existing explanation methods for explaining the behavior of speech classification models.
- **Identification of Research Gaps.** Identify key research gaps to address for the context of explainability methods for speech classification.
- **Implementation.** Select a specific research gap to address. Propose and implement a methodology to address the identified research gap. This may involve proposing a novel approach or adapting existing explainability methods to suit the context of speech classification.
- **Evaluation**. Assess the effectiveness and applicability of the newly implemented approach.

References

[1] Sören Becker et al. *AudioMNIST: Exploring Explainable Artificial Intelligence for Audio Analysis on a Simple Benchmark.* 2023. arXiv: `1807.03418 [cs.SD]`.

[2] Marco Colussi and Stavros Ntalampiras. *Interpreting deep urban sound classification using Layer-wise Relevance Propagation.* 2021. arXiv: `2111.10235 [cs.SD]`.

[3] Saumitra Mishra, Bob L Sturm, and Simon Dixon. "Local interpretable model-agnostic explanations for music content analysis." In: *ISMIR.* Vol. 53. 2017, pp. 537–543.

[4] Eliana Pastor et al. "Explaining Speech Classification Models via Word-Level Audio Segments and Paralinguistic Features". In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers).* Ed. by Yvette Graham and Matthew Purver. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 2221–2238. URL: `https://aclanthology.org/2024.eacl-long.136`.

[5] Xiaoliang Wu, Peter Bell, and Ajitha Rajan. "Can we trust explainable ai methods on asr? an evaluation on phoneme recognition". In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE. 2024, pp. 10296–10300.

[6] Xiaoliang Wu, Peter Bell, and Ajitha Rajan. "Explanations for Automatic Speech Recognition". In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 2023, pp. 1–5. DOI: `10.1109/ICASSP49357.2023.10094635`.