



On the reliability of feature attribution methods for speech classification

Gaofei Shen¹, Hosein Mohebbi¹, Arianna Bisazza², Afra Alishahi¹, Grzegorz Chrupala¹

¹Tilburg University, The Netherlands

²University of Groningen, The Netherlands

{g.shen, h.mohebbi, a.alishahi}@tilburguniversity.edu,
a.bisazza@rug.nl, grzegorz@chrupala.me

Abstract

As the capabilities of large-scale pre-trained models evolve, understanding the determinants of their outputs becomes more important. Feature attribution aims to reveal which parts of the input elements contribute the most to model outputs. In speech processing, the unique characteristics of the input signal make the application of feature attribution methods challenging. We study how factors such as input type and aggregation and perturbation timespan impact the reliability of standard feature attribution methods, and how these factors interact with characteristics of each classification task. We find that standard approaches to feature attribution are generally unreliable when applied to the speech domain, with the exception of word-aligned perturbation methods when applied to word-based classification tasks.¹

Index Terms: speech processing, interpretability, feature attribution

1. Introduction

Large-scale self-supervised models such as wav2vec2 [1] and HuBERT [2] have shown impressive performance on various downstream speech processing tasks from automatic speech recognition to audio classification. As transformer models [3] have been increasingly adopted in speech processing, interpretability research for these models has also intensified.

An important research domain within interpretability is *feature attribution* which aims to quantify the contribution of different parts of the model input to its output. A variety of approaches to feature attribution been studied extensively for the domains of computer vision (CV) and natural language processing (NLP) models. A more limited body of work also exists for the domain of spoken language [4, 5, 6, 7, 8, 9, 10, 11].

A key challenge in research on feature attribution is the evaluation of the methods. Following the long-standing practice in research methods and psychometrics, we can distinguish two key concepts used to quantify the quality of a measurement method: *reliability* and *validity*. Validity refers to how well the method measures the quantity of interest. For feature attribution it largely overlaps with the concept of faithfulness: does the attribution highlight the features that in reality are the most important determinant features of the model’s output? On the other hand, reliability focuses on the consistency of a measurement, and answers the question: does the measurement give the same answer when repeated under similar conditions? For a measurement to be useful, it needs to score well on both of these dimensions. In this work we focus on evaluating the **reliability** of commonly used feature attribution methods as applied to

speech models. This aspect of evaluation is often neglected in prior research, but it is crucial to ensure that the whole endeavor of attributing model outputs to inputs rests on a solid foundation. Only when methods are shown to be reliable can we then focus on the question of their validity.

We believe this foundational work is especially needed in the speech domain. In contrast to the relatively intuitive saliency maps for CV or token-based attribution scores for NLP, the continuous and high resolution nature of the speech signal means that naive application of basic attribution methods leads to noisy and hard-to-interpret results. The choice of feature attribution methods for a particular task matters, as does the conditions under which the method is applied. In this paper, we focus on four attribution methods, and investigate two main aspects that might affect their reliability: the **input type** used for the feature attribution analysis, and the choice of the **attribution granularity**, or the timespan of aggregation or perturbation of input of a given attribution method.

Input type. Convolutional neural network (CNN) speech classification models can use either spectrograms and waveforms as input [5]. Meanwhile self-supervised transformer-based models like wav2vec2 [1] and HuBERT [2] operate in an end-to-end fashion and use waveform as input with a CNN block serving as the feature extractor. More traditional audio features such as log-Mel spectrogram are still being used by the popular Whisper model [12]. It is important to note that a spectrogram (time-frequency representation), and raw waveform are two different feature representations of the same *input signal*. We can convert a spectrogram to a waveform and vice versa. Thus to a large extent we can decouple input type from the specific target model. A more model-specific option is to attribute to the output of the CNN feature extractor block of models like wav2vec2 and HuBERT. Following the common practice in NLP, we call the output of the CNN feature extractor the *CNN embedding*. We can think of the CNN embedding as an even higher-level representation of the input signal than the typical time-frequency representation. Thus, for the most common models of interest we need to decide which input type is the most appropriate to use for feature attribution.

Attribution granularity. While both gradient-based and perturbation-based methods have been tested on speech models, the interpretation of the attribution results differs. Gradient-based methods assign a score to every input value. The standard 16kHz sampling rate for wav2vec2 models means there are tens of thousands of attribution values for mere seconds of speech. These scores can then be aggregated over longer time-spans for ease of interpretation, but that happens as a post-processing step. Perturbation-based methods, on the other hand, can be applied in a top-down manner by directly perturbing larger chunks of the input signal and observing the change in the model’s out-

¹Code: <https://github.com/techsword/reliability-speech-feat-attr>

put. Thus, the choice of attribution method is coupled with the timespan of aggregation or perturbation.

In this work we investigate four feature attribution methods and quantify the impact of the choices regarding input type and granularity on their reliability. We apply these methods to speech classification models trained on three different tasks (one of which comprises three related subtasks). In order to quantify reliability, we use feature attribution agreement between pairs of separate training runs trained on the same data and applied to the same test input: we name this reliability score inter-(random)-seed-agreement (ISA). Our experiments show that even though the target classifier models learn the tasks and agree on the vast majority of the inputs, the attribution agreement is generally quite low, and that acceptable levels of reliability are only reached in very few specific conditions. Our findings highlight the inadequacy of standard approaches to feature attribution as applied to the speech domain, and underline the need for the development and careful evaluation of appropriate speech-specific attribution methods.

2. Related Work

Feature attribution has evolved along with advances in machine learning, with methods originally developed to visualize salient features in computer vision [13, 14], and adapted further for natural language processing [15, 16, 17]. Unlike in speech processing, the limitations of feature attribution methods have been carefully studied in vision and language models from both reliability and validity perspectives [18, 19, 20, 21, 22]. For example, gradient-based methods were demonstrated to be independent of parameters of later layers [21], while [22] argues that complete and linear attribution methods (such as Integrated Gradients) may perform no better than random guessing when identifying how models depend on features. By computing rank correlation, [20, 19] show that feature attribution methods (even those within the same family) often disagree in the explanation scores they produce. Unlike these approaches, we do not consider agreement between different attribution methods, but rather focus on the core issue of the reliability of a *single method* under *repeated measurement*: the consistency of a single attribution method applied to several randomly initialized fine-tuning runs of the same model architecture. Regarding the effect of different configuration details (e.g., aggregation level), our work is related to [23, 24] for textual data.

Applications and adaptations of attribution methods in the speech domain has been explored primarily with convolutional neural network (CNN) based models [4, 5, 25], given their architectural similarities with computer vision models. For example, [5] used Layerwise Relevance Propagation (LRP) to explain a CNN model trained on either waveform or spectrogram representations of audio signals. In contrast to their approach, where models trained on two different types of input data are compared, we examine the effect of different input types within a single model.

For speech classification, several studies have started to examine the validity of feature attribution methods. For example, [7] applied LIME [26] to a phoneme recognition task using the TIMIT [27] dataset, which provides manual labeling and segmentation at the phoneme level. They found that restricting input audio perturbations to a limited window around the phoneme of interest can improve the validity of LIME. Similarly, [8, 11] showed that discretizing attribution scores through phoneme- and word-level boundaries leads to more interpretable explanations for classification tasks. Despite these

Task	Accuracy	Overall Fleiss' κ	Error Fleiss' κ
Gender ID	0.999	0.999	0.356
Speaker ID	0.990	0.983	0.677
Intent Class.	0.998	–	–
Action	0.997	0.997	0.520
Object	0.999	0.999	0.602
Location	0.999	0.999	0.591

Table 1: *Agreement in model performance across different runs, measured using three metrics: Accuracy, Overall and Error Fleiss' κ on the test sets.*

studies on validity, the reliability of feature attribution for speech models remains largely unexplored.

3. Methods

For speech classification, pre-trained models are typically paired with a lightweight feedforward neural network as a classification head and fine-tuned using labeled data. During fine-tuning, both the backbone model and classifier adjust their weights to emphasize the most relevant input features and learned representations, maximizing classification accuracy. We therefore assume that models starting from the same pre-trained checkpoint but fine-tuned with different random seeds identify and employ comparable relevant features for a given input, especially when they achieve consistently high accuracy. Accordingly, a reliable feature attribution method should consistently show the same pattern in highlighting the most important input features for a given utterance across such models. Our assumption aligns with that in [28].

3.1. Classification models

In our experiments, we use `wav2vec2-base`² model and fine-tune it with nine different seeds for three different speech classification tasks: gender and speaker identification (Gender ID and Speaker ID respectively), and intent classification (IC). We expect the IC task to rely on mostly on the presence of specific lexical items, while Gender ID and Speaker ID should rely mostly on lower-level acoustic features.

For Gender ID and Speaker ID, we use a subset of the Common Voice dataset [29]. We select 40 speakers with self-reported gender labels of masculine or feminine with 301 utterance for each speaker totalling 12,040 utterances. A stratified 80:20 train-test split was also applied before model fine-tuning. For intent classification we use the Fluent Speech Commands dataset [30] with the provided train-test split. We resample the waveforms in both datasets to 16kHz. The IC task comprises three related classification subtasks: Action, Object and Location. We use a separate classification head for each subtask.

During fine-tuning, we freeze the CNN feature extractor and the projection layers and only update weights of the transformer network and the final classification heads. This makes sure models with different seeds receive exactly the same input and allows us to have a fixed *CNN embedding* across different models investigated in this paper.

In order to lend further credibility to our assumption about the equivalency of models, we first evaluate the classification accuracy as well classification agreement of the different fine-tuning runs, shown in Table 1. If models' architecture as well as behavior is similar, we have more reason to believe that their internal computations are also equivalent.

²<https://huggingface.co/facebook/wav2vec2-base>

To measure the agreement between the model runs initialized with different random seeds, we report two separate versions of Fleiss’ κ . The overall Fleiss’ κ measures the agreement between all runs on the complete test set. The Error Fleiss’ κ measures the agreement between model decisions on the subset of the test data where at least one error was made. Fleiss’ $\kappa = 0$ if the agreement between runs is due to chance, and $\kappa = 1$ if the runs agree completely.

We can see that the classification accuracy is near perfect for all tasks and overall agreement is also very high. Agreement for the small subset of data points where an error was made is moderate. Given small percentage of overall errors, we believe this means the behavior of the models we investigate in this paper are sufficiently similar.

3.2. ISA reliability metric

To measure the reliability of feature attribution methods, we examine the consistency of their scores derived from models fine-tuned with different seeds. Specifically, we calculate the inter-seed agreement (ISA) metric based on a dynamic *top-p* metric.

$$\text{ISA} = \frac{1}{N} \sum_{n=1}^N \frac{|\text{top-p}(A_i)_n \cap \text{top-p}(A_j)_n|}{|\text{top-p}(A_i)_n|} \quad (1)$$

Here p is the percentage of the top indices of the attribution scores we are interested in. A_i and A_j are the attribution scores for the i -th and j -th model run. N is the number of samples in the dataset. The top- p function returns the top p percentage of indices of the attribution scores. The value of p is fixed at 20%. The intersection of the top- p indices is calculated for each sample in the dataset. The ISA score is the average of the pairwise percentage of shared indices of the top- p attribution scores for all combinations of random seed pairs. The higher the ISA score, the more the attribution scores of the models agree with each other. A baseline ISA can be obtained by randomly shuffling the attribution scores for each sample before calculating the ISA score.

We use the Captum library [31] to calculate feature attribution scores for the fine-tuned models on their respective datasets. We test two gradient-based methods: Saliency and Integrated Gradients (IG); and two perturbation-based methods: LIME and Feature Ablation (FA). For the perturbation-based methods, we use a feature mask to group waveform input into 10ms spans: this is done due to computational constraints as a highest tractable resolution for perturbing the waveform.

3.3. Feature attribution conditions

Input feature types. We calculate the ISA metric for each of our three input feature types: waveform, spectrogram, and CNN embeddings. In order to enable attribution to the spectrogram for self-supervised models like wav2vec2 and HuBERT (which use raw waveform as their native input), we follow [10] by prepending an inverse short-term Fourier transformation (ISTFT) to the model. We use a hop length of 320 for the STFT and ISTFT transformations to keep the time resolution at 20ms to be consistent with the wav2vec2 model feature extractor.

For Integrated Gradients, LIME and Feature Ablation, a baseline input is needed. We use silence as our baseline: for the waveform input type we use the silence waveform directly, while for the other two input types we convert the silence waveform into the corresponding spectrogram or CNN embedding first. All of the attribution methods tested return both positive

and negative values: we do not do any additional processing to the attribution score before aggregation.

Granularity of aggregation. We test three granularities of aggregation: no aggregation, frame-level aggregation, and word-level aggregation. For frame-level aggregation, we sum the attribution scores for raw waveform input at 20ms intervals; spectrogram and CNN embedding input gets summed for every frame. For word-level aggregation, we aggregate the attribution scores at the word level using forced-alignment time stamps. We use Montreal Forced Aligner [32] to align our datasets with the provided transcriptions. To take varying word lengths into account, we mean-pool the attribution scores for each word. We also discard the non-word segments in the alignment.

Granularity of perturbation. For the perturbation methods only, the alternative to aggregation is to directly perturb specific timespans of the input. We test the effect of directly perturbing word-level segments of the input, based on the same forced-alignment time stamps as above.

4. Results

We organize the results into groups of comparisons. Within each group, we present the effects of the varied conditions in applying feature attribution methods. We visualize the central tendency (median) of the inter-seed agreement (ISA) scores as well as the spread around it via boxplots. An individual boxplot displays a set of 36 pairwise comparisons. We then visualize the median baseline score of randomly shuffled attribution scores for all pairwise comparisons via the dashed red lines.

4.1. Effects of input feature types

Figure 1 shows the ISA scores for attribution scores for the no-aggregation condition. For most combinations of method and input type the ISA scores are low to moderate (below 0.6). The exception is the task of Gender ID for the embedding input type and Integrated Gradients method, which shows a median around 0.7 but with a wide spread around it. We also note that the effect of attribution method is in general larger than that of the input type. Notably, in most cases LIME shows very low reliability, barely above the baseline. This indicates that LIME is not a suitable attribution method in a high-time resolution setting. At the same time, Integrated Gradients generally shows the highest reliability.

4.2. Effects of granularity of aggregation

To highlight the impact of granularity of aggregation, we plot the ISA scores of only the CNN embedding input type at various levels of aggregation in fig. 2. We can see that different aggregation levels do not alter the general reliability patterns we saw in fig. 1. As before, we generally see the best reliability for Integrated Gradients. For word-level attribution on the IC tasks we would expect a higher level of agreement, and yet we see that ISA scores do not surpass 0.6 for any of the condition combinations. This is surprising, as intuitively one would expect that aggregation should smooth out small variations in attribution scores and increase agreement.

4.3. Perturbations on the word level

Lastly, we also evaluate the reliability of perturbation methods applied to word-aligned input segments, in a setting analogous to that described in [11]. Figure 3 shows ISA scores for attribution scores generated with perturbation-based methods oper-

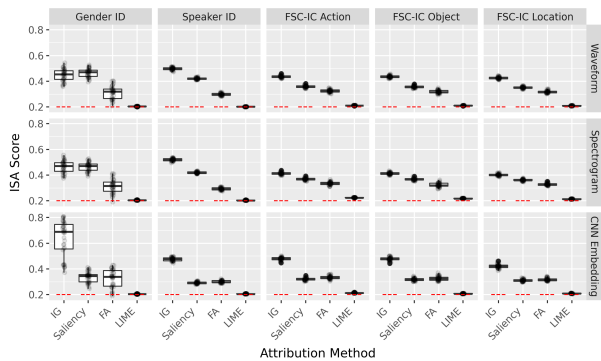


Figure 1: Distributions of ISA scores without aggregation. The rows indicate different input feature types, the columns are different tasks. Within each panel, each boxplot shows results from different attribution methods and the y-axis is the ISA score. The red dotted line indicates the randomly shuffled baseline. IG: Integrated Gradients, FA: Feature Ablation.

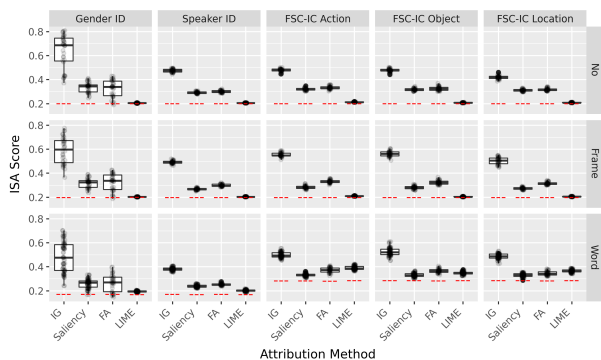


Figure 2: Distributions of ISA scores for the CNN embedding input type, at different levels of aggregation. The rows are levels of granularity of aggregation, the columns are different tasks. Within each panel, each boxplot shows results from different attribution methods and the y-axis is the ISA score. The red dotted line indicates the randomly shuffled baseline. IG: Integrated Gradients, FA: Feature Ablation.

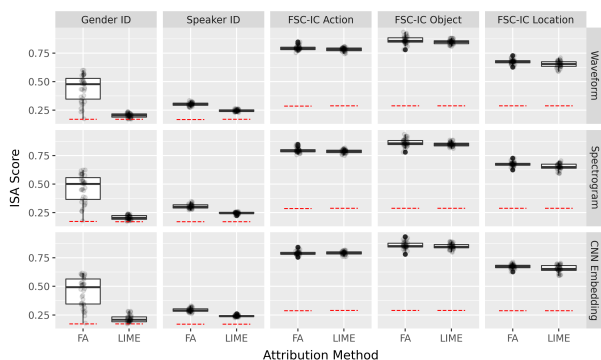


Figure 3: Distributions of ISA scores with perturbation operating directly on word-aligned segments. The rows indicate different input feature types, the columns are different tasks. Within each panel, each boxplot report results from different attribution methods and the y-axis is the ISA score. The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

ating directly on the word level. Here we see the Intent Classification tasks showing much higher ISA than in fig. 2. The reliability of perturbation-based attribution methods on word-based tasks is much higher when perturbing word-aligned than aggregating scores obtained from high-resolution features. We also observe that one of the IC tasks, Location, shows lower reliability than the other two. This perhaps reflects the degree to which these different subtasks can rely on redundant lexical cues.

Interestingly, we see less variations in the score pattern across different input feature types in fig. 3 than in fig. 1. This shows that perturbations done on the word-level granularity are less sensitive to the differences between different input feature types.

5. Discussion & Conclusion

Our findings show that the naive application of standard feature attribution methods to speech classification models generally leads to poor reliability. When attributing to high-resolution input, regardless of specific input types such as waveform, spectrogram or embeddings, even the most reliable of our methods, Integrated Gradients, does not surpass 50% inter-seed agreement for most tasks. Simply aggregating these scores does not improve reliability. The likely underlying issue is that the gradients or perturbation effect of such high resolution and highly correlated redundant features are very small and noisy.

Only in the case of directly perturbing word-aligned segments of the input, and only for the intent classification subtasks, do we see acceptable reliabilities. The likely explanation is that classification decisions for these tasks rely on specific words in the utterance, and that directly perturbing those specific words and only those words affects model output and thus attribution scores. There is thus little scope for models to disagree. On the other hand, tasks such as Gender ID and Speaker ID are unlikely to rely on specific words, and models may use redundant clues distributed over the whole utterance: thus for these tasks we do not find a consistently reliable combination of method and input-type to attribute to. Our findings suggest that in order for standard attribution methods to be applicable, the target speech classification task needs to be similar in nature to an equivalent text-based task, where token-based attributions are standard. Ideally, however, we would like feature attribution to be more wide applicable across varied speech tasks.

5.1. Limitations and future directions

While the scope of this paper is limited to assessing reliability, the methods found to be reliable will also need to be evaluated for validity. While previous works have used cross-method agreement as a proxy for validity, we believe that a more direct measure of alignment with the target model will be needed. Our work focused on reliability for attribution to features along the time dimension. The other important axis for audio data is the frequency domain: for certain tasks it may be more useful to attribute to frequency-based features. Additionally, for audio data generally, and for speech specifically, certain high-level features such as loudness, pitch-contours, or specific aspects of timbre may also be interesting targets for attribution: for these cases standard attribution techniques such as those evaluated here are not directly applicable. Ultimately, spoken language may be sufficiently different from images or text data that only feature attribution techniques tailored to the speech domain will prove reliable enough to be useful.

6. Acknowledgements

This publication is part of the project *InDeep: Interpreting Deep Learning Models for Text and Sound* (with project number NWA.1292.19.399) of the National Research Agenda (NWA-ORC) program.

7. References

- [1] A. Baeovski, H. Zhou, A. Mohamed, and M. Auli, “Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” 2020.
- [2] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhota, R. Salakhutdinov, and A. Mohamed, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units,” Jun. 2021.
- [3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention Is All You Need,” Dec. 2017.
- [4] A. Prasad and P. Jyothi, “How Accents Confound: Probing for Accent Information in End-to-End Speech Recognition Systems,” in *ACL 2020*, D. Jurafsky, J. Chai, N. Schlueter, and J. Tetreault, Eds. Online: ACL, Jul. 2020, pp. 3739–3753.
- [5] S. Becker, J. Vielhaben, M. Ackermann, K.-R. Müller, S. Lapschkin, and W. Samek, “AudioMNIST: Exploring Explainable Artificial Intelligence for audio analysis on a simple benchmark,” *Journal of the Franklin Institute*, vol. 361, no. 1, pp. 418–428, Jan. 2024.
- [6] X. Wu, P. Bell, and A. Rajan, “Explanations for Automatic Speech Recognition,” in *ICASSP 2023*. Rhodes Island, Greece: IEEE, Jun. 2023, pp. 1–5.
- [7] —, “Can We Trust Explainable AI Methods on ASR? An Evaluation on Phoneme Recognition,” in *ICASSP 2024*, Apr. 2024, pp. 10 296–10 300.
- [8] S. Gupta, M. Ravanelli, P. Germain, and C. Subakan, “Phoneme Discretized Saliency Maps for Explainable Detection of AI-Generated Voice,” Sep. 2024.
- [9] D. Fucci, M. Gaido, B. Savoldi, M. Negri, M. Cettolo, and L. Bentivogli, “SPES: Spectrogram Perturbation for Explainable Speech-to-Text Generation,” Nov. 2024.
- [10] E. Mancini, F. Paissan, P. Torrioni, M. Ravanelli, and C. Subakan, “Investigating the Effectiveness of Explainability Methods in Parkinson’s Detection from Speech,” Nov. 2024.
- [11] E. Pastor, A. Koudounas, G. Attanasio, D. Hovy, and E. Baralis, “Explaining speech classification models via word-level audio segments and paralinguistic features,” in *EACL 2024*, Y. Graham and M. Purver, Eds., Mar. 2024.
- [12] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *Proceedings of the 40th ICML*. PMLR, 2023. [Online]. Available: <https://proceedings.mlr.press/v202/radford23a.html>
- [13] M. D. Zeiler and R. Fergus, “Visualizing and Understanding Convolutional Networks,” in *Computer Vision – ECCV 2014*. Cham: Springer International Publishing, 2014.
- [14] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” in *ICLR 2014, Workshop Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2014. [Online]. Available: <http://arxiv.org/abs/1312.6034>
- [15] I. Covert, S. M. Lundberg, and S.-I. Lee, “Explaining by removing: A unified framework for model explanation,” *J. Mach. Learn. Res.*, vol. 22, pp. 209:1–209:90, 2020.
- [16] J. Bastings and K. Filippova, “The elephant in the interpretability room: Why use attention as explanation when we have saliency methods?” in *BlackboxNLP 2020*, A. Alishahi, Y. Belinkov, G. Chrupała, D. Hupkes, Y. Pinter, and H. Sajjad, Eds. Online: ACL, Nov. 2020, pp. 149–155.
- [17] H. Mohebbi, A. Modarressi, and M. T. Pilehvar, “Exploring the role of BERT token representations to explain sentence probing results,” in *EMNLP 2021*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and Punta Cana, Dominican Republic: ACL, Nov. 2021, pp. 792–806.
- [18] D. Pruthi, R. Bansal, B. Dhingra, L. Baldini Soares, M. Collins, Z. C. Lipton, G. Neubig, and W. W. Cohen, “Evaluating explanations: How much do explanations from the teacher aid students?” *TACL*, vol. 10, 2022.
- [19] S. Krishna, T. Han, A. Gu, S. Wu, S. Jabbari, and H. Lakkaraju, “The disagreement problem in explainable machine learning: A practitioner’s perspective,” *TMLR*, 2024.
- [20] M. Neely, S. F. Schouten, M. J. R. Bleeker, and A. Lucic, “A song of (dis)agreement: Evaluating the evaluation of explainable artificial intelligence in natural language processing,” in *HAAI*, 2022.
- [21] L. Sixt, M. Granz, and T. Landgraf, “When explanations lie: Why many modified bp attributions fail,” in *ICML*, 2019.
- [22] B. Bilodeau, N. Jaques, P. W. Koh, and B. Kim, “Impossibility theorems for feature attribution,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 121, 2022.
- [23] J. Bastings, S. Ebert, P. Zablotskaia, A. Sandholm, and K. Filippova, ““Will you find these shortcuts?” a protocol for evaluating the faithfulness of input saliency methods for text classification,” in *EMNLP 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. ACL, Dec. 2022.
- [24] Y. Chen, K. Zhang, F. Hu, X. Wang, R. Li, and Q. Liu, “Dynamic Multi-granularity Attribution Network for Aspect-based Sentiment Analysis,” in *EMNLP*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: ACL, 2024.
- [25] H. Muckenhirn, V. Abrol, M. Magimai-Doss, and S. Marcel, “Understanding and Visualizing Raw Waveform-Based CNNs,” in *Interspeech 2019*. ISCA, Sep. 2019.
- [26] M. T. Ribeiro, S. Singh, and C. Guestrin, ““Why Should I Trust You?”: Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD ICKDD*, 2016.
- [27] Garofolo, John S., Lamel, Lori F., Fisher, William M., Pallett, David S., Dahlgren, Nancy L., Zue, Victor, and Fiscus, Jonathan G., “TIMIT Acoustic-Phonetic Continuous Speech Corpus,” 1993.
- [28] P. Atanasova, J. G. Simonsen, C. Lioma, and I. Augenstein, “A diagnostic study of explainability techniques for text classification,” in *EMNLP*. Online: ACL, Nov. 2020, pp. 3256–3274. [Online]. Available: <https://aclanthology.org/2020.emnlp-main.263/>
- [29] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, “Common voice: A massively-multilingual speech corpus,” in *Proceedings of the 12th LREC*. ELRA, May 2020. [Online]. Available: <https://aclanthology.org/2020.lrec-1.520/>
- [30] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, “Speech Model Pre-training for End-to-End Spoken Language Understanding,” Jul. 2019.
- [31] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson, “Captum: A unified and generic model interpretability library for PyTorch,” 2020.
- [32] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, “Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi,” in *Interspeech 2017*. ISCA, Aug. 2017, pp. 498–502.