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

Abstract

Spurred by the demand for transparency and interpretability in Artificial Intelligence (AI), the field of eXplainable AI (XAI) has experienced significant growth, marked by both theoretical reflections and technical advancements. While various XAI techniques, especially feature attribution methods, have been extensively explored across diverse tasks, their adaptation for the *speech* modality is comparatively lagging behind. We argue that a key challenge in feature attribution for speech processing lies in identifying informative acoustic features. In this paper, we discuss the key challenges in selecting the features for speech explanations. Also, in light of existing research, we highlight current gaps and propose future avenues to enhance the depth and informativeness of explanations for speech.

Keywords

Speech Models, Explainability, Feature Attribution

1. Introduction

> *Models are only as interpretable as their features.* [1]

Spoken language—as perhaps our most natural form of interaction—is the foundational element of many technologies we interact with in our daily lives [2], from virtual assistants to voice dictation [3, 4, 5]. More recently, the emergence of highly capable speech foundation models [6, 7, 8, 9] has also facilitated and expanded the adoption of speech technologies on an unprecedented multilingual scale. In light of this proliferation, a need arises to prioritize transparency and interpretability, qualities already demanded in the growing landscape of Machine Learning (ML).

As a response, the field of eXplainable AI (XAI) has risen prominently, with the aim of facilitating understanding of the rationale behind model decisions and fostering users' trust [10, 11, 12, 13]. XAI is also reinforced by the establishment of norms and legal frameworks, as seen in the European Union's General Data Protection Regulation, which enshrines the 'right to explanation', and the AI Act, which emphasizes transparency as a pivotal component of ML applications [14].

XAI encompasses various tasks and methods, such as identifying relevant model components for specific predictions, understanding the information processed by these components, and determining which input elements guide the model's predictions [15]. The latter task is the focus of *feature attribution* methods, which provide intuitive explanations by visualizing which input elements (e.g., pixels in an image or words in a sentence) have influenced the model's predictions. These methods assign a score to each input feature, quantifying its importance or contribution to the output: higher scores indicate greater importance of the corresponding input features for generating the output [16, 17, 18, 19]. They can help identify potential causes for errors and unexpected behaviors, as well as analyze the model's response to specific input properties. Overall, these explainability methods serve to present the reason why models make specific predictions by establishing a connection between input and output as a form of intuitive explanation for humans, thereby enhancing interpretability.[1]

Over time, ongoing efforts have aimed to refine feature attribution techniques and provide more effective explanations [22, 23]. However, it is essential to recognize that the effectiveness of feature attribution explanations relies not only on the techniques themselves but also on the informativeness of the input features used as explanatory variables. If an explanation highlights unintelligible or poorly informative features, it does little to enhance the understanding of the model's behavior

[1] Despite numerous efforts to differentiate the closely related concepts of explainability and interpretability, no consensus exists in the literature on their definitions [20]. In this paper, we adopt a perspective similar to that of Saeed and Omlin [21], where *explainability* refers to the process of extracting insights from a model's workings through specific techniques, while *interpretability* refers to the understanding process of those insights, crucial to make them actionable.

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

dfucci@fbk.eu (D. Fucci); bsavoldi@fbk.eu (B. Savoldi); mgaido@fbk.eu (M. Gaido); negri@fbk.eu (M. Negri); cettolo@fbk.eu (M. Cettolo); bentivo@fbk.eu (L. Bentivogli)

[1]. This can undermine key principles in XAI, such as *accuracy*—the property of correctly reflecting the factors that led the model to a specific decision including all relevant information—and *meaningfulness*—the property of offering explanations that are comprehensible to the user [24].[2]

In fields involving images or texts, feature representations are typically constrained to pixels and words, respectively. However, for speech, multiple input representations can be adopted, each emphasizing different acoustic aspects. Indeed, a sequence of speech elements does not only convey the meaning of what is said (like words in a text) but also bears a wealth of additional information useful for both human understanding and automatic processing (e.g. intonation, loudness, speaking rate). Consequently, when employing feature attribution methods, the resulting explanations can vary significantly in shape and focus on more or less informative characteristics depending on the type of speech representation used. To date, research on feature attribution for speech is notably limited to few applications—including classification [27, 28] and generative tasks [29, 30, 31, 32]—which offer a somewhat fragmented picture in the choice of speech representations, thus providing limited insights on the relation between the features considered and the explanations based upon them.

In light of the above, this paper reflects on the impact of the chosen acoustic features in explaining the rationale behind speech models, aiming to gain a deeper understanding of the trade-offs associated with acoustic features. By first offering a gentle introduction to the rich and multidimensional nature of speech and its digital representation, we identify current gaps and potential avenues for effectively incorporating this multidimensionality into XAI for speech models. Our discussion will focus on two critical factors: i) the amount of information these features provide about the model's behavior, which influences the *richness* of the explanations, and ii) the level of detail of such information, which determines the *granularity* of the explanations. We will also explore how these aspects impact both the accuracy and meaningfulness of the explanations, ultimately shaping their overall interpretability.

2. The Correlates of Speech

To gain deeper insight into the complexities of defining informative features in speech, we explore key characteristics of speech and their implications for modeling.

Speech is a multifaceted phenomenon. It is grounded on the materiality of sound to convey linguistic content (i.e. *what is said*), which is modulated depending on several paralinguistic cues (i.e. *how is said*) entailing extensive variation—also for single individual speakers [33]. As such, it comprises several dimensions, which are hard to pin down individually, but collectively amount to what we intuitively and simply perceive as spoken language.

From a linguistic perspective, the spoken communication system consists of the combination of phonemes,[3] which are regarded as the smallest meaningful units of sounds [34, 35]. Physically, it involves the continuous flow of sounds shaped by the movements of our phonatory organs, transmitted as sound waves [36]. Perceptually, we process speech through three primary dimensions [37]: *i)* ***time***, or the sequential occurrence of sounds;[4] *ii)* ***intensity***, corresponding to the energy level of the wave due to the strength of molecular vibration, which we perceived as loudness; *iii)* ***frequency***, regarding the rate of vibrations produced by the vocal cords—interpreted as pitch—and whose modulation is responsible for shaping the type of speech sound.

These three elements, known as *acoustic correlates* [38], are specific to both speakers and phonemes. For example, speakers possess unique characteristics, including pitch and speaking rate [33], and also exhibit high variability stemming from various sociodemographic factors such as gender, age, and dialect [39]. In these cases, the speech content needs to be disentangled from the variability in its delivery. Conversely, language sounds exhibit variability in duration—e.g., /i/ in *ship* and *sheep*—and are distinguished by specific frequency ranges [36]. The frequency dimension also plays a vital role in shaping suprasegmental aspects of speech—broader phenomena that span multiple segments—such as intonation, obtained by varying pitch [40]. Pitch, for instance, has a distinctive function in tonal languages, where it is used to distinguish lexical or grammatical meaning [41]. But even in non-tonal languages, these prosodic elements are indispensable to delivering different meanings and intents, as the reader can perceive by reading out loud two contrastive sentences such as: "*You got the joke right*" and "*You got the joke, right?*", where pauses and prosody play pivotal roles.

All these factors add to the multidimensionality of speech, which feature engineering strives to encapsulate and that cannot be overlooked in the explanatory process.

3. Speech Representations

While various representations are used to encode speech in a digital format, three main types are commonly given

[2] The properties of *accuracy* and *meaningfulness* can be associated with those of *faithfulness* and *plausibility*, respectively [25, 26].

[3] Throughout the paper, we use the abstract category of *phoneme* to denote individual speech sounds. However, when discussing their actual realizations, it is more accurate to refer to them as *phones* [34].

[4] E.g. the order of sounds between /pɑt/ (*pot*) or /tɑp/ (*top*) differentiates two words.

as input to state-of-the-art speech models (for a review, see [42, 43]). Namely, waveforms, spectrograms, and mel-frequency cepstral coefficients (MFCCs), which are shown in Figure 1.

The ***waveform*** serves as the most fundamental representation of a signal, comprising sequences of samples (e.g., $16,000$ per second), each indicating the amplitude of the signal at a specific point in time—essentially, the fluctuations in air pressure over time. This type of representation is leveraged by models like Wav2vec [6].

The ***spectrogram*** results from feature engineering operations that decompose the speech signal into its frequencies, presenting a 2D visualization of frequency distributions over time. These representations are commonly depicted as heatmaps, where color intensity corresponds to the energy of a specific frequency at a given moment. The time unit in spectrograms is represented by a fixed-length window of a few milliseconds (e.g., 25), commonly referred to as a frame, whithin which a given number of waveform samples are encompassed. Notably, the articulation of sounds produces time-frequency patterns which are visible as darker regions [36]. Prominent examples of state-of-the-art models leveraging spectrograms are Whisper [9] and SeamlessM4T [44].

The ***MFCCs*** offer another 2D representation where each coefficient captures important details about how the frequency content of the signal changes over time. Like spectrograms, MFCCs offer information about both frequency and time, but in a more compact form. MFCCs are commonly used in the implementation of ASR models within popular toolkits like Kaldi[5] [45] and Mozilla DeepSpeech[6].

Overall, though different in nature, these three types of representations are all effectively exploited by current speech models.[7] For human understanding, however, they actually vary in terms of informativeness with respect to the acoustic correlates discussed in §2. Indeed, although both intensity and frequency are somewhat discernible in waveforms, qualitative distinctions of patterns specific to pitch or phoneme frequencies are rarely feasible [36]. Comparatively, spectrograms and MFCCs are richer and more descriptive, because they capture the multiple dimensions of time, frequency, and intensity with finer detail. Still, spectrograms are more conducive to phonetic analyses, given the established knowledge in analyzing frequency patterns over time within this representation [36] In contrast, MFCCs are rarely used for phonetic analysis [46].

Overall, while weighting the informativeness and selection of speech representations requires a certain expertise in speech processing, being aware of the trade-offs they intrinsically entail is crucial for carefully conducting XAI examination in speech. Indeed, it is precisely upon such input features—and their trade-offs—that explanations are built.

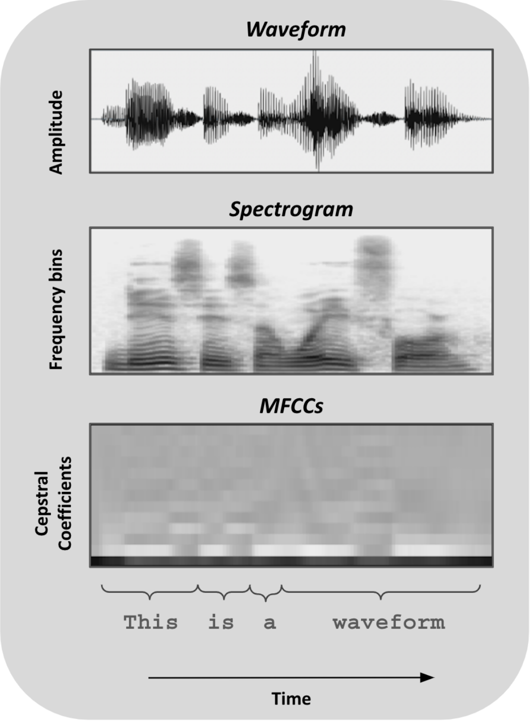

Figure 1: Schematic illustration of the primary speech representations used by state-of-the-art speech models for the utterance “This is a waveform”. The features were computed using Librosa 0.10.1 [47].

4. Richness of Explanations

Considering the foregoing, there is a causal relationship wherein explanatory possibilities in speech XAI are inherently limited by the richness of the audio features used, specifically the dimensions they encapsulate. This limitation directly correlates with the richness of the resulting explanations. Also, owing to the compatibility of current models with various representation types, the explanations generated are inevitably confined by the specific input features provided to the model. To exemplify, if models process audio as waverfoms—which poorly represent the frequency dimension for human understanding—explanations accounting for such a correlate will be out of reach. In fact, previous works by Wu et al. [31] and Wu et al. [32], based on waveforms solely focus on the temporal dimension to explain ASR.

[5] See https://kaldi-asr.org/doc/feat.html.

[6] See https://deepspeech.readthedocs.io/en/master/DeepSpeech.html.

[7] We are not aware of any recent study attributing higher systems performance depending on the used representation.

In these cases, to avoid limiting the understanding of the models' behavior to one single dimension it would be advisable to ***explore alternative techniques that offer deeper insights into how models process other acoustic correlates.*** For instance, Pastor et al. [28] integrated counterfactual explanations to specifically investigate whether selected paralinguistic features such as pitch, speaking rate, and background noise were influent for the model's prediction. Additionally, various techniques exist to analyze how models extract relevant patterns from waveforms through convolutions [48, 49, 50].

When the selected input features represent multiple dimensions, as in the case of spectrograms or MFCCs, the decision to only account for one of these dimensions becomes arbitrary. For example, two models tested by Wu et al. [31], namely, DeepSpeech [51] and Sphinx [52], are fed with spectrograms and MFCCs, respectively. However, explanations based on raw waveforms are provided for these models. This inconsistency between the features used in explanations and those used by the models inevitably offers only a partial overview of the models' behavior and limits the exploration of important acoustic aspects. This, in turn, can impact the accuracy of the explanations, which ideally should encompass all relevant information.

To prioritize explanation accuracy and conduct analyses considering the crucial role of acoustic correlates such as frequency, it is advisable to ***take into account all dimensions embedded in the speech representation***. This approach is exemplified by the works of Markert et al. [30], who provide explanations that account for the most influential elements in MFCCs, as well as Trinh and Mandel [29] and Becker et al. [27], who base the explanations on spectrograms. In the work by Markert et al. [30], however, it is challenging to connect the results with specific acoustic parameters due to the complexity of analyzing MFCCs (see §3), which significantly undermines the meaningfulness of the explanations. In contrast, explanations using spectrograms offer valuable insights into how machines process speech, producing both accurate and meaningful results. For instance, Trinh and Mandel [29] demonstrated that neural ASR models focus on high-energy time-frequency regions for transcription, while Becker et al. [27] found that lower frequency ranges, typically associated with pitch, exhibit higher attribution scores in speaker gender classification tasks [27], showing some alignment with human speech processing. However, interpreting these insights requires specialized expertise, which can reduce the meaningfulness of explanations for non-experts. This highlights that, even in speech, the balance between accuracy and meaningfulness can vary depending on the context [24].

5. Granularity of Explanations

Another critical factor concerning the informativeness of input features is the level of granularity at which the features are considered during the explanatory process. This decision affects the level of detail in the resulting explanations and, consequently, accuracy—as more detailed explanations may more accurately reflect the model's behavior—and their meaningfulness—as detailed and comprehensive explanations can be more difficult to interpret [12, 24].

In the time domain, for example, input features are highly fine-grained. As discussed in §2, spectrograms typically contain frames spanning tens of milliseconds, capturing detailed frequency content within each frame, whereas waveforms are composed of samples taken at much shorter time intervals—for instance, as mentioned in §2, there can be $16,000$ samples in just one second. This level of detail poses great challenges for (human) comprehension, particularly for a broader audience, since mapping groups of frames/samples in an explanation to recognizable speech units is highly time-consuming and requires specialized expertise.

Accordingly, to address the issue and make explanations for speech more broadly accessible, previous works have leveraged textual transcripts within the explanation process. More specifically, Wu et al. [32] and Pastor et al. [28] resort to the alignment of audio to text, either for individual phonemes or words, respectively, and apply explainability techniques to such units. While this approach helps decipher the contribution of input features based on more intuitive linguistic units, it diverges from how current models process speech features in small frames and samples [43]. This divergence risks overlooking the model's behavior and compromises the accuracy and effectiveness of the explanations. For instance, whether ASR systems rely on shorter or longer time intervals than individual words remains unclear [29]. Therefore, analyzing this aspect requires a more granular approach at the time level.

In light of the above, ***explanations should be obtained with low-level units*** to avoid biasing explanations towards human understanding. The use of audio-transcript alignment to aid analysis of explanations can be very useful but should occur downstream of the explanation process, not upstream. In this way, we can maximize the use of all available units to generate detailed and accurate explanations, and then aggregate scores from individual frames or samples to create more compact representations at the level of phonemes or words, ensuring flexibility in the meaningfulness of the explanations according to specific needs. This bottom-up approach mirrors practices in the text domain, providing adaptability in defining attribution units that can range from subwords to words or phrases [53, 54].

6. Conclusion

This paper has examined the role of acoustic features and their selection for explaining speech models. More specifically, we considered a specific subfield of XAI, namely, *feature attribution*, which connects input features to outputs as a form of explanation. Previous research has not explicitly addressed how to incorporate features into the explanation process within the speech domain, where input is encoded in more varied ways compared to other fields, such as text. This has led to diverse approaches, each with different implications for what can and cannot be explained about model behavior, and with the risk of not fully or accurately representing the model's functioning.

By discussing the key characteristics of speech and the properties of the most adopted acoustic features, we argue that explanations should ideally encompass all available dimensions, particularly time and frequency, as both are essential for a comprehensive understanding of the models' rationale. We have also discussed challenges associated with aligning explanations at high granularity with human understanding, emphasizing solutions that provide flexibility in the analysis, allowing for adjustments between more or less detail as needed.

Building on these insights, our ongoing research focuses on developing feature attribution techniques that operate on spectrograms at the finest possible unit level, integrating both time and frequency dimensions. Our aim is to generate explanations that are accurate and meaningful for experts, as well as adaptable for non-expert users. More broadly, we hope that our reflections will be beneficial and thought-provoking for researchers currently working in, or entering, the field of XAI for speech models, thereby contributing to a deeper understanding of the rationale behind these models.

7. Limitations

While exploring the relationship between the informativeness of speech features and explanations, we have deliberately not delved into the needs of specific stakeholders for XAI applications. Indeed, different stakeholders present varying needs [55, 56], and to consider them is a research avenue of paramount importance for the growth of XAI. As a nascent area of investigations, however, XAI for speech is still relatively in its infancy, we thus prioritized more fundamental methodological and design decisions which prioritize a comprehensive and detailed understanding at a low level of model's rationale. Accordingly, our reflections might be more appealing for a range of users who engage with speech models and possess expertise in machine learning and/or speech analysis, ranging from developers to speech therapists assisted by speech models [56].

The balance of richness and granularity—which also relates to the interplay between accuracy and meaningfulness—is also relevant to common users who interact with speech technologies. However, investigating how explanations can be effectively communicated to and understood by these users in the context of daily speech technology use exceeds the scope of this paper and warrants further exploration.

8. Acknowledgments

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU. This paper has also received funding from the European Union's Horizon research and innovation programme under grant agreement No 101135798, project Meetween (My Personal AI Mediator for Virtual MEETtings BetWEEN People).

References

[1] A. Zytek, I. Arnaldo, D. Liu, L. Berti-Equille, K. Veeramachaneni, The Need for Interpretable Features: Motivation and Taxonomy, SIGKDD Explor. Newsl. 24 (2022) 1–13. URL: https://doi.org/10.1145/3544903.3544905. doi:10.1145/3544903.3544905.

[2] C. Munteanu, M. Jones, S. Oviatt, S. Brewster, G. Penn, S. Whittaker, N. Rajput, A. Nanavati, We need to talk: HCI and the delicate topic of spoken language interaction, in: CHI '13 Extended Abstracts on Human Factors in Computing Systems, CHI EA '13, Association for Computing Machinery, New York, NY, USA, 2013, pp. 2459–2464. URL: https://doi.org/10.1145/2468356.2468803. doi:10.1145/2468356.2468803.

[3] H. Feng, K. Fawaz, K. G. Shin, Continuous Authentication for Voice Assistants, in: Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking, MobiCom '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 343–355. URL: https://doi.org/10.1145/3117811.3117823. doi:10.1145/3117811.3117823.

[4] P. Cheng, U. Roedig, Personal Voice Assistant Security and Privacy—A Survey, Proceedings of the IEEE 110 (2022) 476–507. doi:10.1109/JPROC.2022.3153167.

[5] S. Malodia, N. Islam, P. Kaur, A. Dhir, Why Do People Use Artificial Intelligence (AI)-Enabled Voice Assistants?, IEEE Transactions on Engineering

Management 71 (2024) 491–505. doi:10.1109/TEM.2021.3117884.

[6] A. Baevski, H. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20, Curran Associates Inc., Red Hook, NY, USA, 2020, pp. 12449–12460.

[7] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed, HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units, IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021) 3451–3460. URL: https://doi.org/10.1109/TASLP.2021.3122291. doi:10.1109/TASLP.2021.3122291.

[8] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, F. Wei, WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing, IEEE Journal of Selected Topics in Signal Processing 16 (2022) 1505–1518. doi:10.1109/JSTSP.2022.3188113.

[9] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust Speech Recognition via Large-Scale Weak Supervision, in: Proceedings of the 40th International Conference on Machine Learning, ICML'23, JMLR.org, 2023, pp. 28492–28518.

[10] F. Doshi-Velez, B. Kim, Towards A Rigorous Science of Interpretable Machine Learning, 2017. arXiv:1702.08608.

[11] D. V. Carvalho, E. M. Pereira, J. S. Cardoso, Machine Learning Interpretability: A Survey on Methods and Metrics, Electronics 8 (2019). URL: https://www.mdpi.com/2079-9292/8/8/832. doi:10.3390/electronics8080832.

[12] G. Vilone, L. Longo, Notions of explainability and evaluation approaches for explainable artificial intelligence, Information Fusion 76 (2021) 89–106. URL: https://www.sciencedirect.com/science/article/pii/S1566253521001093. doi:https://doi.org/10.1016/j.inffus.2021.05.009.

[13] R. Pradhan, J. Zhu, B. Glavic, B. Salimi, Interpretable Data-Based Explanations for Fairness Debugging, in: Proceedings of the 2022 International Conference on Management of Data, SIGMOD '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 247–261. URL: https://doi.org/10.1145/3514221.3517886. doi:10.1145/3514221.3517886.

[14] C. Panigutti, R. Hamon, I. Hupont, D. Fernandez Llorca, D. Fano Yela, H. Junklewitz, S. Scalzo, G. Mazzini, I. Sanchez, J. Soler Garrido, E. Gomez, The role of explainable AI in the context of the AI Act, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 1139–1150. URL: https://doi.org/10.1145/3593013.3594069. doi:10.1145/3593013.3594069.

[15] J. Ferrando, G. Sarti, A. Bisazza, M. R. Costa-jussà, A Primer on the Inner Workings of Transformer-based Language Models, 2024. arXiv:2405.00208.

[16] M. Ancona, E. Ceolini, C. Öztireli, M. Gross, Gradient-Based Attribution Methods, in: W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, K.-R. Müller (Eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer International Publishing, Cham, 2019, pp. 169–191. URL: https://doi.org/10.1007/978-3-030-28954-6_9. doi:10.1007/978-3-030-28954-6_9.

[17] W. Samek, K.-R. Müller, Towards Explainable Artificial Intelligence, in: W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, K.-R. Müller (Eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer International Publishing, Cham, 2019, pp. 5–22. URL: https://doi.org/10.1007/978-3-030-28954-6_1. doi:10.1007/978-3-030-28954-6_1.

[18] S. Agarwal, S. Jabbari, C. Agarwal, S. Upadhyay, S. Wu, H. Lakkaraju, Towards the Unification and Robustness of Perturbation and Gradient Based Explanations, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 110–119. URL: https://proceedings.mlr.press/v139/agarwal21c.html.

[19] M. Ivanovs, R. Kadikis, K. Ozols, Perturbation-based methods for explaining deep neural networks: A survey, Pattern Recognition Letters 150 (2021) 228–234. URL: https://www.sciencedirect.com/science/article/pii/S0167865521002440. doi:https://doi.org/10.1016/j.patrec.2021.06.030.

[20] F. K. Došilović, M. Brčić, N. Hlupić, Explainable Artificial Intelligence: A Survey, in: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2018, pp. 210–215. doi:10.23919/MIPRO.2018.8400040.

[21] W. Saeed, C. Omlin, Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities, Knowledge-Based Systems 263 (2023) 110273. URL: https://www.sciencedirect.com/science/article/pii/S0950705123000230. doi:https://doi.org/10.1016/j.knosys.2023.110273.

[22] Y. Zhou, S. Booth, M. T. Ribeiro, J. Shah, Do Feature Attribution Methods Correctly Attribute Features?, Proceedings of the AAAI Conference on Artificial Intelligence 36 (2022) 9623–9633. URL: https://ojs.aaai.org/index.php/AAAI/article/view/21196. doi:10.1609/aaai.v36i9.21196.

[23] D. Qin, G. Amariucai, D. Qiao, Y. Guan, S. Fu, A Comprehensive and Reliable Feature Attribution Method: Double-sided Remove and Reconstruct (DoRaR), 2023. arXiv:2310.17945.

[24] P. J. Phillips, C. Hahn, P. Fontana, A. Yates, K. K. Greene, D. Broniatowski, M. A. Przybocki, Four Principles of Explainable Artificial Intelligence, 2021. URL: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=933399. doi:https://doi.org/10.6028/NIST.IR.8312.

[25] A. Jacovi, Y. Goldberg, Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4198–4205. URL: https://aclanthology.org/2020.acl-main.386. doi:10.18653/v1/2020.acl-main.386.

[26] Q. Lyu, M. Apidianaki, C. Callison-Burch, Towards Faithful Model Explanation in NLP: A Survey, Computational Linguistics 50 (2024) 657–723. URL: https://aclanthology.org/2024.cl-2.6. doi:10.1162/coli_a_00511.

[27] S. Becker, J. Vielhaben, M. Ackermann, K.-R. Müller, S. Lapuschkin, W. Samek, AudioMNIST: Exploring Explainable Artificial Intelligence for audio analysis on a simple benchmark, Journal of the Franklin Institute 361 (2024) 418–428. URL: https://www.sciencedirect.com/science/article/pii/S0016003223007536. doi:https://doi.org/10.1016/j.jfranklin.2023.11.038.

[28] E. Pastor, A. Koudounas, G. Attanasio, D. Hovy, E. Baralis, Explaining Speech Classification Models via Word-Level Audio Segments and Paralinguistic Features, in: Y. Graham, M. Purver (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 2221–2238. URL: https://aclanthology.org/2024.eacl-long.136.

[29] V. A. Trinh, M. Mandel, Directly Comparing the Listening Strategies of Humans and Machines, IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021) 312–323. doi:10.1109/TASLP.2020.3040545.

[30] K. Markert, R. Parracone, M. Kulakov, P. Sperl, C.-Y. Kao, K. Böttinger, Visualizing Automatic Speech Recognition – Means for a Better Understanding?, in: Proc. 2021 ISCA Symposium on Security and Privacy in Speech Communication, 2021, pp. 14–20. doi:10.21437/SPSC.2021-4.

[31] X. Wu, P. Bell, A. 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.

[32] X. Wu, P. Bell, A. 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), 2024, pp. 10296–10300. doi:10.1109/ICASSP48485.2024.10445989.

[33] N. Audibert, C. Fougeron, Intra-speaker phonetic variation in read speech: comparison with inter-speaker variability in a controlled population, in: Interspeech 2022, ISCA, Incheon, South Korea, 2022, pp. 4755–4759. URL: https://hal.science/hal-03852142. doi:10.21437/Interspeech.2022-10965.

[34] J. Clark, C. Yallop, An Introduction to Phonetics and Phonology, B. Blackwell, Oxford, UK, 1990.

[35] G. Yule, The Study of Language, 7 ed., Cambridge University Press, 2020.

[36] K. N. Stevens, Acoustic Phonetics, The MIT Press, 2000.

[37] N. H. van Schijndel, T. Houtgast, J. M. Festen, Effects of degradation of intensity, time, or frequency content on speech intelligibility for normal-hearing and hearing-impaired listeners, The Journal of the Acoustical Society of America 110 (2001) 529–542. URL: https://doi.org/10.1121/1.1378345. doi:10.1121/1.1378345.

[38] K. N. Stevens, Acoustic correlates of some phonetic categories, The Journal of the Acoustical Society of America 68 (1980) 836–842. doi:10.1121/1.384823.

[39] J. Honey, Sociophonology, John Wiley & Sons, Ltd, 2017, pp. 92–106. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781405166256.ch6. doi:https://doi.org/10.1002/9781405166256.ch6.

[40] D. Hirst, Speech Prosody: from Acoustics to Interpretation, Springer Berlin, Heidelberg, 2024.

[41] C. T. Best, The Diversity of Tone Languages and the Roles of Pitch Variation in Non-tone Languages: Considerations for Tone Perception Research, Frontiers in Psychology 10 (2019). URL: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2019.00364. doi:10.3389/fpsyg.2019.00364.

[42] F. Alías, J. C. Socoró, X. Sevillano, A Review of Physical and Perceptual Feature Extraction

Techniques for Speech, Music and Environmental Sounds, Applied Sciences 6 (2016). URL: https://www.mdpi.com/2076-3417/6/5/143. doi:10.3390/app6050143.

[43] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, S. Poria, A review of deep learning techniques for speech processing, Information Fusion 99 (2023) 101869. URL: https://www.sciencedirect.com/science/article/pii/S1566253523001859. doi:https://doi.org/10.1016/j.inffus.2023.101869.

[44] S. Communication, L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, P.-A. Duquenne, H. Elsahar, H. Gong, K. Heffernan, J. Hoffman, C. Klaiber, P. Li, D. Licht, J. Maillard, A. Rakotoarison, K. R. Sadagopan, G. Wenzek, E. Ye, B. Akula, P.-J. Chen, N. E. Hachem, B. Ellis, G. M. Gonzalez, J. Haaheim, P. Hansanti, R. Howes, B. Huang, M.-J. Hwang, H. Inaguma, S. Jain, E. Kalbassi, A. Kallet, I. Kulikov, J. Lam, D. Li, X. Ma, R. Mavlyutov, B. Peloquin, M. Ramadan, A. Ramakrishnan, A. Sun, K. Tran, T. Tran, I. Tufanov, V. Vogeti, C. Wood, Y. Yang, B. Yu, P. Andrews, C. Balioglu, M. R. Costa-jussà, O. Celebi, M. Elbayad, C. Gao, F. Guzmán, J. Kao, A. Lee, A. Mourachko, J. Pino, S. Popuri, C. Ropers, S. Saleem, H. Schwenk, P. Tomasello, C. Wang, J. Wang, S. Wang, SeamlessM4T: Massively Multilingual & Multimodal Machine Translation, 2023. URL: https://arxiv.org/abs/2308.11596. arXiv:2308.11596.

[45] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, K. Vesely, The Kaldi Speech Recognition Toolkit, in: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, IEEE Signal Processing Society, 2011, pp. 1–4. IEEE Catalog No.: CFP11SRW-USB.

[46] K. Ikarous, The Encoding of Vowel Features in Mel-Frequency Cepstral Coefficients, in: A. Vietti, L. Spreafico, D. Mereu, V. Galatà (Eds.), Il parlato nel contesto naturale [Speech in the natural context], Officinaventuno, Milano, 2018, p. 9–18. URL: https://doi.org/10.17469/O2104AISV000001.

[47] B. McFee, M. McVicar, D. Faronbi, I. Roman, M. Gover, S. Balke, S. Seyfarth, A. Malek, C. Raffel, V. Lostanlen, B. van Niekirk, D. Lee, F. Cwitkowitz, F. Zalkow, O. Nieto, D. Ellis, J. Mason, K. Lee, B. Steers, E. Halvachs, C. Thomé, F. Robert-Stöter, R. Bittner, Z. Wei, A. Weiss, E. Battenberg, K. Choi, R. Yamamoto, C. Carr, A. Metsai, S. Sullivan, P. Friesch, A. Krishnakumar, S. Hidaka, S. Kowalik, F. Keller, D. Mazur, A. Chabot-Leclerc, C. Hawthorne, C. Ramaprasad, M. Keum, J. Gomez, W. Monroe, V. A. Morozov, K. Eliasi, nullmightybofo, P. Biberstein, N. D. Sergin, R. Hennequin, R. Naktinis, beantowel, T. Kim, J. P. Åsen, J. Lim, A. Malins, D. Hereñú, S. van der Struijk, L. Nickel, J. Wu, Z. Wang, T. Gates, M. Vollrath, A. Sarroff, Xiao-Ming, A. Porter, S. Kranzler, Voodoohop, M. D. Gangi, H. Jinoz, C. Guerrero, A. Mazhar, toddrme2178, Z. Baratz, A. Kostin, X. Zhuang, C. T. Lo, P. Campr, E. Semeniuc, M. Biswal, S. Moura, P. Brossier, H. Lee, W. Pimenta, librosa/librosa: 0.10.1, 2023. URL: https://doi.org/10.5281/zenodo.8252662. doi:10.5281/zenodo.8252662.

[48] M. Ravanelli, Y. Bengio, Interpretable Convolutional Filters with SincNet, 2019. arXiv:1811.09725.

[49] M. Angrick, C. Herff, G. Johnson, J. Shih, D. Krusienski, T. Schultz, Interpretation of convolutional neural networks for speech spectrogram regression from intracranial recordings, Neurocomput. 342 (2019) 145–151. URL: https://doi.org/10.1016/j.neucom.2018.10.080. doi:10.1016/j.neucom.2018.10.080.

[50] H. Fayyazi, Y. Shekofteh, IIRI-Net: An interpretable convolutional front-end inspired by IIR filters for speaker identification, Neurocomput. 558 (2023). URL: https://doi.org/10.1016/j.neucom.2023.126767. doi:10.1016/j.neucom.2023.126767.

[51] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, A. Y. Ng, Deep Speech: Scaling up end-to-end speech recognition, 2014. arXiv:1412.5567.

[52] P. Lamere, P. Kwok, W. Walker, E. Gouvea, R. Singh, B. Raj, P. Wolf, Design of the CMU Sphinx-4 Decoder, in: Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003), 2003, pp. 1181–1184. doi:10.21437/Eurospeech.2003-382.

[53] G. Sarti, N. Feldhus, L. Sickert, O. van der Wal, Inseq: An Interpretability Toolkit for Sequence Generation Models, in: D. Bollegala, R. Huang, A. Ritter (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 421–435. URL: https://aclanthology.org/2023.acl-demo.40. doi:10.18653/v1/2023.acl-demo.40.

[54] V. Miglani, A. Yang, A. Markosyan, D. Garcia-Olano, N. Kokhlikyan, Using Captum to Explain Generative Language Models, in: L. Tan, D. Milajevs, G. Chauhan, J. Gwinnup, E. Rippeth (Eds.), Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023), Association for Computational Linguistics, Singapore, 2023, pp. 165–173. URL: https://aclanthology.org/2023.nlposs-1.19. doi:10.18653/v1/2023.nlposs-1.19.

[55] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, A. Sesing, K. Baum, What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research, Artificial Intelligence 296 (2021) 103473. URL: https://www.sciencedirect.com/science/article/pii/S0004370221000242. doi:https://doi.org/10.1016/j.artint.2021.103473.

[56] M. Calvano, A. Curci, A. Pagano, A. Piccinno, Speech Therapy Supported by AI and Smart Assistants, in: R. Kadgien, A. Jedlitschka, A. Janes, V. Lenarduzzi, X. Li (Eds.), Product-Focused Software Process Improvement, Springer Nature Switzerland, Cham, 2024, pp. 97–104.