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

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

Abstract

Predictive models make mistakes and have biases. To combat both, we need to understand their predictions. Explainable AI (XAI) provides insights into models for vision, language, and tabular data. However, only a few approaches exist for speech classification models. Previous works focus on a selection of spoken language understanding (SLU) tasks, and most users find their explanations challenging to interpret. We propose a novel approach to explain speech classification models. It provides two types of insights. (i) Word-level. We measure the impact of each audio segment aligned with a word on the outcome. (ii) Paralinguistic. We evaluate how non-linguistic features (e.g., prosody and background noise) affect the outcome if perturbed. We validate our approach by explaining two state-of-the-art SLU models on two tasks in English and Italian. We test their plausibility with human subject ratings. Our results show that the explanations correctly represent the model's inner workings and are plausible to humans.

1 Introduction

As models increase in complexity, understanding how they work and the reasons behind their outputs becomes more challenging. However, this understanding is crucial for improving performance and addressing biases. Various explainable AI (XAI) techniques, such as gradient-based (Simonyan et al., 2013; Sundararajan et al., 2017; Selvaraju et al., 2022, *inter alia*) and input perturbation (Zeiler and Fergus, 2013) approaches, have been proposed to gain insights into computer vision model behavior. These techniques have been successfully applied in language (Ribeiro et al., 2016; Sanyal and Ren, 2021; Jacovi et al., 2021, *inter alia*) and tabular (Lundberg and Lee, 2017; Pastor and Baralis, 2019; Strumbelj and Kononenko, 2010) models.

Despite significant progress in XAI for vision, text, and structured data models, explanations for Spoken Language Understanding (SLU) models remain largely unexplored. Some existing approaches provide spectrogram- (Becker et al., 2018; Frommholz et al., 2023) or phoneme-based (Wu et al., 2023a) explanations. However, they are too fine-grained for broader speech tasks (e.g., intent detection or emotion recognition) where other factors interplay to convey meaning, e.g., acoustic features, linguistic aspects, and prosody. Capturing such aspects requires tailor-made explainability solutions that are also easy for human actors to understand.

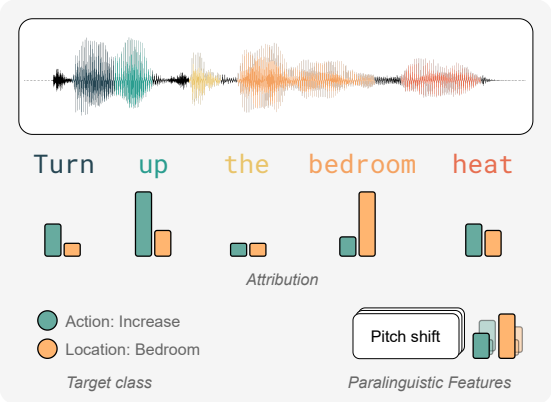

Figure 1: Explanation with word-level and paralinguistic attributes for a Fluent Speech Commands sample (Lugosch et al., 2019). Audio aligned to words represented through color. Bars show word-level attributions for target classes *Increase* (green) and *Bedroom* (orange).

Our goal is to explain predictions by describing the interaction between input utterance components and model predictions. Utterances incorporate semantic and paralinguistic information from the speaker's voice and external conditions, such as prosody and acoustics. Following Ribeiro et al.'s (2016) definition of explanation, we want to provide interpretable representations of utterances to help humans understand model behavior, addressing the following research questions (RQs).

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2221–2238

RQ1. How do we define interpretable representations by understandably describing utterances?

RQ2. How do we explain predictions at the semantic and paralinguistic levels?

We address both questions by presenting utterances as word-level audio segments with paralinguistic features (such as pitch and speaking rate). We propose a new explanation approach that provides insights on two different but complementary levels. (i) We tackle the linguistic aspect and find which parts of the spoken utterance influenced the model prediction the most (e.g., the words "Turn" and "up" for predicting that a user's request entails an "Increase" action for a voice assistant). These explanations shed light on whether and to which extent models leverage linguistic aspects, e.g., word semantics, pragmatics, or syntactic parsing when predicting the output. (ii) We measure the impact of paralinguistic features by perturbing the signal and quantifying the effect of such transformations in the predictions. Intuitively, if an alteration of a feature (say, the average pitch) changes the prediction, it indicates the model relies on that feature for the prediction. Conversely, if the model is not sensitive to that variation, that feature is irrelevant to trigger that prediction —thus also offering insights into model robustness. By observing the model's response to variations, we can identify potential vulnerabilities (e.g., sensitivity to noise injection) or biases (e.g., over- or under-reliance on prosody-related features). Building on existing XAI literature, we construct our explanations by assigning a numerical score to each input feature, whether a word or a paralinguistic feature. We argue that this representation improves overall readability and usability. Figure 1 shows a sample explanation.

We test our approach by explaining wav2vec-2.0 (Baevski et al., 2020) and XLS-R (Babu et al., 2022), two state-of-the-art speech models, on two datasets for intent classification and one for emotion recognition in English and Italian. We assess the quality of our explanations under the faithfulness and plausibility paradigms (Jacovi and Goldberg, 2020), using human subject ratings. Our results show that the explanations are faithful to the model's inner workings *and* plausible to humans. We hope that our model can provide speech researchers with a valuable tool for understanding sources of bias and errors.

Contributions. We introduce a new method for explaining speech classification models. Advancing from established XAI perturbation-based techniques, our approach is the first to study the effect of word-level audio segments and paralinguistic features on predictions. It generates easy-to-interpret visualizations that are faithful and plausible to human experts across models, languages, and tasks. We release the code at `https://github.com/elianap/SpeechXAI` to encourage future research at the intersection of SLU and interpretability.

2 Methodology

To quantify the contribution of each utterance part to a prediction, we compute word-level attribution scores as follows. First, we align the audio signal to its transcript and get word-level timestamps. Then, we use a perturbation-based technique to compute the contribution of each spoken word to the prediction by modifying the input and observing changes in the prediction. Specifically, we propose a method based on the Leave-One-Out and Local Interpretable Model-Agnostic Explanations (LIME; Ribeiro et al., 2016) techniques. We follow a similar perturbation-based approach to measure the contribution of paralinguistic aspects. Given an input utterance, we perturb the raw audio signal and measure the effect on the model prediction. We consider pitch to account for prosody, and audio stretching, background noise, and reverb levels for channel-related aspects.

We generate explanations by assigning a single numerical attribution score to each uttered word (§2.1) and paralinguistic feature (§2.2). Each score is generated via input perturbation and quantifies the entity's contribution (either a word or a paralinguistic feature) in predicting a given target class.

2.1 Word-level Audio Segment Attribution

We compute word-level contribution in two steps.

Word-level audio-transcript alignment. We extract beginning and ending timestamps for each uttered word. If no transcript or timestamp is available, we use state-of-the-art word-level time alignment models to extract them. The resulting timestamps define a set of (non-overlapping) audio segments corresponding to words in the time domain.[1] See Figure 1 (top) for an example.

Segment contribution. We compute each segment contribution by perturbing the input signal. Follow-

[1]This step filters out the parts where no word is uttered, e.g., pauses or signal tails. Since these parts do not carry semantic information, we suppose that they do not affect classification.

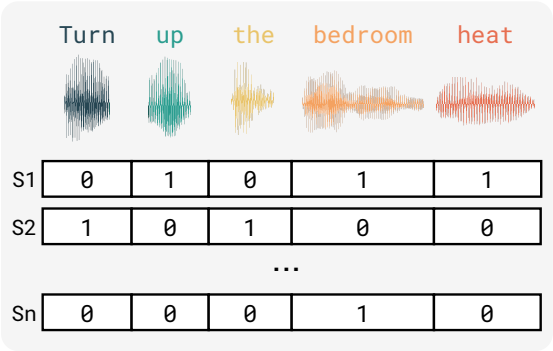

Figure 2: Word-level Time Alignment identifies audio segments to mask (top). LIME sampling process selects the segment to mask (bottom; 1 means "masked segment", $\{S_1, ..., S_n\}$ are the sampled neighbors).

ing prior work, our perturbation consists of masking one or more segments (Covert et al., 2021) by zeroing the corresponding samples in the time domain (Wu et al., 2023a).[2] Choosing which segment to mask and how the choice impacts the model prediction[3] is algorithmic-dependent. We consider Leave-One-Out and LIME, two established XAI solutions.

Leave-One-Out. Leave-One-Out consists of masking one part of the input at a time, feeding each perturbed sample to the model, and attributing each part importance by measuring how the prediction changes if the part is missing. Intuitively, sharper variations in the output signal are of higher importance.

More formally, let $\mathbf{x} \in \mathbb{R}^n$ be an audio signal and $\{x_1, .., x_n\}$ the set of n word-level audio segments within. Consider a speech classification model f applied for tasks such as intent classification or emotion recognition. Let $f(y = k|x)$ be the output probability of model f for class k given the input utterance x. We define the relevance $r(x_i) \in \mathbb{R}$ of each segment x_i to the model's prediction for a target class k as:

$$r(x_i) = f(y = k|\mathbf{x}) - f(y = k|\mathbf{x} \setminus x_i) \quad (1)$$

where $\mathbf{x} \setminus x_i$ refers to signal x with segment x_i masked.

Higher values for $r(x_i)$ indicate a greater relevance of segment x_i to the prediction. A positive score indicates that x_i contributes positively to the probability of belonging to class k. In contrast, a negative score suggests that x_i may drive the prediction toward a different class.[4] See Figure 1 (middle) for an example.

LIME. LIME approximates a classifier with a simpler, interpretable model in the "locality" of a specific instance. Roughly, the process entails sampling from the instance neighborhood, labeling every sampled neighbor with the model, and training a simpler, white-box model on the resulting set. Intuitively, the white-box model is a surrogate approximating the model being explained within the instance neighborhood.

To enable neighborhood sampling, LIME requires the input to be represented with "interpretable features", i.e., a binary representation of parts that can either be masked or not. Here, we choose word-level audio segments as such representations.[5] Figure 2 shows an example neighborhood. Notably, unlike the Leave-One-Out technique, LIME can mask multiple segments at once, allowing it to capture intersectional effects that might arise from multiple missing words. Analogously to Leave-One-Out, the relevance score indicates the magnitude and direction of the segment contribution to the class prediction.

2.2 Paralinguistic Attribution

Beyond the semantic information conveyed by words, speech includes additional paralinguistic information provided by the speaker voice or external conditions, such as pitch, speaking rate, and background noise level. We investigate the effect of paralinguistic features by leveraging ad-hoc signal perturbations.

Let $p := f : \mathbb{R}^n \to \mathbb{R}$ be a function to extract a paralinguistic measure of interest, e.g., the average pitch or signal-to-noise ratio (SNR). We transform $\mathbf{x}$ into $\widetilde{\mathbf{x}}$ such that $p(\widetilde{\mathbf{x}})$ is either higher or lower than $p(\mathbf{x})$, e.g., we shift the pitch up or increase the SNR.

To compute the impact of p on predicting the class k, we perturb $\mathbf{x}$ multiple times and average the result as follows:

[2] We prefer zeroing out segments, rather than removing them, to rule out any effect introduced by shorter recordings.

[3] In all our classification tasks, with "prediction," we refer to the normalized probability of the observed class.

[4] In the binary case, negative scores refer to the opposite class, whereas in multi-class setups, they mean *any* of the other classes.

[5] Some other choices are equal-width segments or n-grams, e.g., to account for word compounds. We leave these additional solutions to future analysis.

$$r_p(\widetilde{\mathbf{x}}) = f(y = k|\mathbf{x}) - f(y = k|\widetilde{\mathbf{x}}) \quad (2)$$

$$r(\mathbf{x}, p) = \frac{1}{|\widetilde{X}_p|} \sum_{\widetilde{\mathbf{x}} \in \widetilde{X}_p} r_p(\widetilde{\mathbf{x}}) \quad (3)$$

where $r_p(\widetilde{\mathbf{x}})$ is effect of an individual perturbation, $\widetilde{X}_p = \{\widetilde{\mathbf{x}}_1, .., \widetilde{\mathbf{x}}_t\}$ is the set of t transformed signals along p, and $r(\mathbf{x}, p)$ is the final relevance. The number t depends on the considered feature (see §3).

Each $r(\mathbf{x}, p)$ is bound between -1 and 1. High absolute values indicate that the model is sensitive to perturbations of the considered feature. In contrast, near-zero values indicate that it is otherwise robust. Moreover, positive values indicate that, on average, the perturbations reduce the prediction probability for the given class; negative ones indicate that perturbations increase it.

3 Experimental Setup

Datasets and Tasks. We evaluate our method on three datasets and two tasks: FLUENT SPEECH COMMANDS (FSC; Lugosch et al., 2019) and the Italian Intent Classification Dataset (ITALIC; Koudounas et al., 2023a) for the Intent Classification (IC) task, and IEMOCAP (Busso et al., 2008) for Emotion Recognition (ER).

FSC is a widely used benchmark dataset for the IC task. The dataset contains recordings of interactions with home voice assistants. The goal is to predict an intent as the combination of three independent predictions, i.e., an action (e.g., "increase"), an object (e.g., "heat"), and a location (e.g., "bedroom"). We focus on the test set, which comprises 3793 audio samples.

ITALIC is an IC dataset for the Italian language, including 60 unique intents. The test set consists of 1441 samples. We use the "Speaker" setup, wherein each speaker utterances belong to a single set among the train, validation, and test.

IEMOCAP is a dataset for the ER task annotated with emotion labels (i.e., happiness, anger, sadness, frustration, and a neutral state). It consists of recorded interactions between pairs of actors engaged in scripted scenarios involving ten unique actors. Among its five sessions, we consider Session '1', consisting of 942 utterances.

Models. We consider the monolingual English wav2vec 2.0 base (Baevski et al., 2020) for FSC and IEMOCAP. We use pre-existing fine-tuned checkpoints on the two datasets (Yang et al., 2021a). We use the multilingual XLS-R (Babu et al., 2022) and its fine-tuned checkpoints (Koudounas et al., 2023a) for ITALIC.

Word-level Audio Alignment and Transcription. We use WhisperX (Bain et al., 2023), a state-of-the-art multi-lingual word-level alignment and transcription model, to transcribe dataset audios and obtain word-level timestamps.

We use the gold transcriptions of the datasets to compute the word error rate (WER). WhisperX, based on Whisper (Radford et al., 2023) for generating transcriptions, achieves a WER of 1.72 on FSC, 15.77 on IEMOCAP, and 7.49 on ITALIC.

Paralinguistic Features. We consider pitch shifting, time stretching, background white noise injection, and reverberation. We provide further details on the transformations in Appendix B.2 and in our repository.

Explanation Setup. For each sample of the set, we explain the probability that the model assigns to the predicted label. Explaining the predicted class provides insights into how the model produced its prediction.

4 Results

4.1 Qualitative Evaluation

We conducted a qualitative manual evaluation. We observe local (instance-level) and global explanations (Doshi-Velez and Kim, 2017). Instance-level explanations provide insights into which features influence the model to classify a specific instance. Local explanations address questions such as: *Is it correct for the right reasons?* Or: *Was the prediction robust to a specific input perturbation?* Global explanations provide an aggregate view to grasp high-level model characteristics.

Individual level. We show the capabilities of our method by explaining wav2vec 2.0 on a FSC sample. We refer the reader to Appendix C for more examples. For a specific utterance with transcription "Turn up the bedroom heat," the model correctly predicts *increase* as the action, *heat* as the object, and *bedroom* as the location, fully identifying the intent.

Table 1 shows the word-level audio segment explanation for this utterance computed for the predicted class for each intent slot.[6] The explanation

[6]For convenience, the Table's header reports the transcribed words. However, we would like to remark that our

	Turn	up	the	bedroom	heat.
act=increase	0.250	0.545	0.260	0.139	0.021
obj=heat	0	0	0	0.014	0.550
loc=bedroom	0.002	0.006	0.087	0.997	0.323

Table 1: Example of word-level audio segment explanation; FSC dataset. The higher the value (and darker the color), the more the audio segment is relevant for the prediction.

	speed up	speed down	pitch down	pitch up	reverb	noise
act=increase	0.19	0.04	0.04	0.13	0.56	0.44
obj=heat	0	0	0	0.04	0	0.29
loc=bedroom	0.03	0.01	0.13	0.33	0.36	0.60

Table 2: Example of paralinguistic attribution $r(\mathbf{x}, p)$ for $p :=$ time stretching (speed variation), pitch shifting, and noise injection; FSC dataset, instance in Table 1. The higher the value (and darker the color), the more the model is sensitive to perturbing the feature.

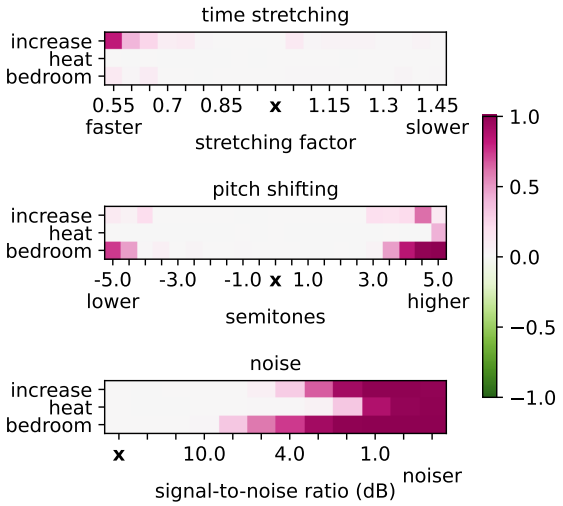

Figure 3: $r_p(\widetilde{\mathbf{x}})$ breakdown for $p :=$ time stretching, pitch shifting, and noise injection. $\mathbf{x}$ indicates the original signal. A darker red (green) color indicates a stronger drop (increase) in probability.

reveals that the segment relative to the word '*up*' is the most relevant term for the action *increase*. The words "*heat*" and "*bedroom*" increase the probability of the predictions *heat* and *bedroom*, respectively. Such explanations are reasonable and aligned with our expectations. We expand on explanation plausibility by conducting a user study with human subjects in §5.

Table 2 shows the paralinguistic explanation. Increasing the speed of the signal has a moderate effect on the prediction "action=increase" but none on the others.[7] Increasing the pitch impacts the prediction of the location and also of the action. Lowering the pitch affects the location slot. The reverberation impacts the prediction for the action slot and slightly for the location; on the other hand, the object prediction is not affected. The prediction for this instance is affected by the introduction of noise, specifically for the location and the action.

To get a finer-grained view on the effect of paralinguistic perturbation, we inspect $r(\widetilde{\mathbf{x}}_i), \forall \widetilde{\mathbf{x}}_i \in \widetilde{X}_p$, for different p, using heatmaps. Figure 3 shows $r(\widetilde{\mathbf{x}})$ when stretching the audio (i.e., increasing and decreasing speed), shifting the pitch, and injecting noise. The model's prediction are always robust to time stretching but one case. Halving the duration of the signal (i.e., making it twice as fast) makes the prediction "action=increase" drop severely. We hypothesize this effect is due to the fact that the leading phrase "Turn on" becomes hardly intelligible.

Similarly, the model is generally robust to pitch shifting. We see the prediction change only in extreme cases when the pitch is tuned up or down by three or more semitones. These cases are "location=bedroom" (pitch up and down) and "action=increase" (pitch up). Finally, the model is affected starting when the signal-to-noise ratio (SNR) reaches 10 dB. Interestingly, the effect varies across slots. The "location" prediction drops first, followed by "action" and "object" whose prediction changes after SNR is as low as 1 dB.

Tables and heatmaps provide two complementary tools for the intepretability of SLU models. The former is helpful for a first, high-level glance to understand whether specific tokens or non-linguistic features have driven the model output. Heatmaps uncover where and to which extent a model is sensitive to input perturbation. Moreover, note that our framework can easily be extended to other forms of p. We will provide easy-to-use Python implementations to facilitate and enrich such multi-faceted analysis.

Global level. We aggregate the importance scores of word audio segments or paralinguistic levels across the entire dataset to investigate the *global* influence of each component.

approach works end-to-end at the audio level, and importance scores relate to audio segments.

[7] Editing paralinguistic features has no symmetrical effects. E.g., increasing noise for a highly-noisy signal will not likely change the prediction. We expect a similar effect on prosody. Tuning the pitch up or down will contribute differently based on the original pitch of the signal.

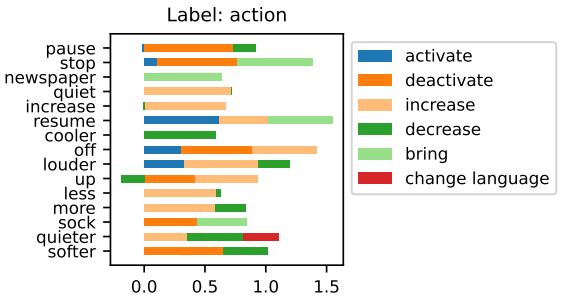

Figure 4: Top 15 most influential words, separately for each predicted class. FSC dataset, Slot: Action.

	speed		pitch		reverb	noise
	up	down	down	up		
action	0.13	0.09	0.12	0.07	0.27	0.37
object	0.07	0.05	0.07	0.04	0.17	0.43
location	0.06	0.04	0.06	0.04	0.11	0.21

Table 3: Average $r(\mathbf{x}, p)$ for $p :=$ time stretching (speed variation), pitch shifting, and noise injection; FSC dataset. The higher the score (and darker the color), the more the model is sensitive to perturbing the feature.

Figure 4 shows a summary plot for the word-level audio segment explanations of wav2vec 2.0 predictions on the FSC test set for the slot "action." We first compute the explanations for the predicted classes. Then, we aggregate audio segments corresponding to the same transcripted word after basic processing (i.e., lowercase, punctuation removal, noun singularization). We report the top 15 segments, occurring at least 5 times in the dataset, with the highest average importance. The average importance scores are represented separately for each class. Hence, the summary plot shows which spoken words are associated with which predicted class(es). From Figure 4, the importance score for spoken words such as '*newspaper*' and '*cooler*' across the entire test set is associated with a single class value. Each class ('*bring*' and '*decrease*') corresponds to a plausible value. When a term is associated with multiple labels, the summary plot can become a debugging tool. For instance, the spoken word '*pause*' is correctly linked to the predicted action '*deactivate*' but erroneously connected to '*decrease*.'

Table 3 shows the average importance score of paralinguistic explanations aggregated for each label. We observe higher average importance scores for the action label for the time stretching component, specifically when compressing the utterance duration ('*stretch down*'). The pitch transformations induce a higher change in the prediction probability for the action slot, especially when lowering the pitch. Finally, adding background noise globally impacts the model prediction.

We further investigate the complementary points of view of word-level and paralinguistic attributions in Appendix C.

4.2 Quantitative Faithfulness Evaluation

A significant research effort has been devoted to the evaluation of post-hoc explainability (Atanasova et al., 2020; Agarwal et al., 2022; DeYoung et al., 2020). Faithfulness and plausibility have been conceptualized as two crucial desiderata to make explanations meaningful and trustworthy (Jacovi and Goldberg, 2020). Faithfulness measures evaluate how accurately the explanation reflects the model's inner workings, whereas plausibility measures whether it matches human expectations.

We quantitatively evaluate the faithfulness of our word-level audio segment explanations in this section and discuss a user study for plausibility in §5. Our focus on word-level explanations is driven by token-level explainability. Building upon prior works on text-based explanations (DeYoung et al., 2020; Jacovi and Goldberg, 2020), we extend and adapt existing evaluation metrics to the specific context of audio segment explanations.

Metrics. We generalize *comprehensiveness* and *sufficiency* (DeYoung et al., 2020), two widely adopted faithfulness measures. These metrics were originally designed for token-level explanations in text classification, where explainers assign a relevance score to each token. Being in a similar setup, we use audio segments rather than tokens, leaving the metric unchanged.

Comprehensiveness evaluates whether the explanation identifies the audio segments the model "truly relied upon" to make the prediction. We measure it by progressively masking the audio segments highlighted by the explanation, observing the change in probability, and finally averaging the results. A high value of comprehensiveness indicates that the audio segments highlighted by the explanations are relevant to the prediction (see DeYoung et al. (2020) for more details).

Sufficiency evaluates if the audio segments in the explanation are sufficient for the model to make the prediction. Differently from comprehensiveness, we preserve only the relevant audio segments and compute the prediction difference. A low score in-

Method	FSC action	FSC object	FSC location	ITALIC intent	IEMOCAP emotion
WA-L1O	0.623	0.627	0.467	0.693	**0.507**
WA-LIME	**0.638**	**0.663**	**0.481**	**0.723**	0.484
random	$0.299_{\pm 0.002}$	$0.251_{\pm 0.005}$	$0.192_{\pm 0.005}$	$0.325_{\pm 0.005}$	$0.274_{\pm 0.005}$
WA-L1O	**0.161**	0.086	0.063	0.158	0.310
WA-LIME	0.165	**0.077**	**0.054**	**0.139**	**0.264**
random	$0.483_{\pm 0.004}$	$0.447_{\pm 0.007}$	$0.338_{\pm 0.002}$	$0.558_{\pm 0.002}$	$0.454_{\pm 0.004}$

Table 4: Comprehensiveness (↑, top) and Sufficiency (↓, bottom) scores for our word attribution explanation via leave-one-out (WA-L1O), our word-level LIME (WA-LIME), and random attribution for the FSC, ITALIC, and IEMOCAP datasets, separately for each label. Best result in bold.

dicates that the audio segments in the explanations indeed drive the prediction.

Baseline. We assess the quality of explanations compared to a random explainer. The random explainer assigns a random score in the range [-1, 1] to each word audio level segment.

Results. Table 4 shows the comprehensiveness and sufficiency results for the FSC, ITALIC, and IEMOCAP datasets, separately for each label. We generated our word-level audio segment explanations with respect to the predicted class. We report our results for both the leave one out (WA-L1O) and the LIME-based (WA-LIME) methods. For the random baseline, we consider five rounds of generations, and we report average and standard deviation. The results in Table 4 show that our word-level explanations outperform the random baseline for both metrics. Moreover, WA-LIME explanations are generally more faithful.

5 Plausibility User Study

Plausibility to humans is another essential desideratum of good explanations. It measures whether explanations are reasonable, believable, and, more generally, align with human reasoning (DeYoung et al., 2020; Jacovi and Goldberg, 2020).

We conducted a user study to assess whether our approach produced plausible explanations. We focused on word-level audio segment attribution in the Intent Detection task in English and Italian. The target of our study was practitioners knowledgeable in machine learning. See Appendix D for full details.

5.1 Study Design

Quality Control. To help participants familiarize themselves with the task and to check if our definition of plausibility applies, we provided some initial questions as sanity checks and quality control. In practice, we asked participants to compare our explanations with a baseline method that assigns random scores to every word and to express a preference.[8] This task tests if the study questions are well-framed and our explanations are informative (at least over random attribution).

Plausibility Assessment. To understand the perceived plausibility of explanations, we asked participants to rate the plausibility of the explanations in absolute terms on a 4-point Likert-like scale.

Visualization Strategy. Recent evidence shows that different visualization strategies impact cognitive load, efficiency, and efficacy (Schuff et al., 2022). Our color-coded score approach (see Table 1) combines a word-level saliency map with the precise indication of the score overlaid onto it. However, we are interested in finding out if better options exist. We asked participants to compare our solution against plain word saliency maps (Arras et al., 2017; Arora et al., 2022), and bar charts for ease-of use and scalability to many examples.

5.2 Findings

Our study involved 35 participants recruited from university courses and research laboratories close to our institutions. We report here the main findings. Please refer to Appendix D for full results.

The quality control checks confirmed that our approach can provide plausible insights. In the head-to-head comparison, all participants preferred our explanation over the random one for both FSC (IC, English) and ITALIC (IC, Italian) across all

[8] We verified that all study recordings were intelligible. There is at least one Italian native speaker and a B2-level English speaker among the authors.

provided examples. Participants scored the plausibility of our explanations with 3.13/4 (std: 0.787) for FSC, and 3.37/4 (std: 0.75) for ITALIC. These scores suggest that our method generates explanations that are *highly plausible* to humans, consistent across two languages and datasets.

Regarding visualization strategies, we found statistically significant differences between the three representations ($p < 0.05$, Friedman) on FSC. Subjects preferred saliency words and color-coded tables over bar charts for identifying relevant words and found them more user-friendly when inspecting multiple explanations. They preferred our color-coded table and bar charts for comparing relative word importance. These results suggest that reporting the scores (via saliency maps with number overlays or via a bar plot) requires a lower cognitive load to compare words than saliency maps alone. Considering overall preference, *users strongly preferred our representation*, followed by saliency maps.

On ITALIC, our questions present a single target prediction (i.e., the intent) rather than the specific slots ("action," "object," or "location"). We observe statistically significant differences in the scores for comparing salient words. As for FSC, the bar plot and our color-coded table emerged as preferred methods for comparing relevant words. Overall, *participants preferred our table and the bar chart*.

For both datasets, participants preferred the bar plot and our color-coded table for comparing scores. However, other questions revealed some variation, with our color-coded table and saliency maps preferred for FSC, and our method and bar charts favored for ITALIC. One key distinction between the two datasets is the unique number of targets (or slots) explained (i.e., 3 for FSC and 1 for ITALIC). In the single-label scenario of ITALIC, users find bar plots an effective visualization. This result echoes Schuff et al. (2022)'s, suggesting that bar charts mitigate biases such as word length. In multi-label scenarios like FSC, color-coded explanations are preferred as they facilitate interpretation and comparison across labels. These findings emphasize that *the visualization strategy needs to be adapted* depending on the context and use case.

Overall, our study suggests that users find our explanations plausible and straightforward, a prerequisite to making them useful for model explanations.

6 Related Work

Few works address interpretability by design for speech and audio, like the prototypical networks (Zinemanas et al., 2021b,a) and attention-based explanations (Won et al., 2019). Most approaches focus on post-hoc interpretability to explain (already trained) models. We categorize them based on the form of the provided explanations.

Multiple studies (Montavon et al., 2019) use Layer-wise Relevance Propagation (LRP) (Bach et al., 2015), initially proposed for image classification, to explain audio analysis tasks. Most works represent explanations as time-frequency heatmaps over spectrograms (Becker et al. (2018); Frommholz et al. (2023); Colussi and Ntalampiras (2021); Arras et al. (2019), inter alia), or heatmaps over ad-hoc terms (Becker et al., 2018). While experts are familiar with spectrograms, they are challenging for laypersons to interpret.

Becker et al. (2018) also use LRP to derive relevance scores for individual samples of the input waveform. Interpreting explanations as sets of individual signal samples lacks abstraction and disregards sample context. We propose a more user-friendly and intuitive approach to explanation. Similarly, Wu et al. (2023b) assign relevance scores to audio frames for ASR, i.e., raw data bins of predefined size in the time dimension. SoundLIME (Mishra et al., 2017) applies LIME (Ribeiro et al., 2016) to equal-width segments within the time, frequency, or time-frequency domains. However, the chosen segment size affects these temporal explanations. Moreover, their explanations are not grounded in spoken words or paralinguistic information, limiting interpretability for semantic contexts like speech classification.

Wu et al. (2023a) is similar to our approach, as it tests fixed-width audio segments and audio segments aligned with phonemes. However, they require phoneme-level annotations, which limits evaluation to when such labeling is available. Moreover, their method specializes in phoneme recognition. In contrast, our approach offers a more generalized solution to any SLU classification task. We automatically derive audio segments at the word level via state-of-the-art speech transcription systems. Furthermore, to the best of our knowledge, we are the first to assess the impact of paralinguistic features on predictions in an interpretable form.

Occluding parts of the input data to measure

their impact is a well-established method in XAI (Covert et al., 2021). Different domains use diverse techniques for removing or masking parts of the data like noise addition, blurring, or masking in grey (vision), a special mask token or removing words (text), or using average values (structured data; Covert et al., 2021). For speech data, Wu et al. (2023a) generate perturbations for a LIME explainer using signal zeroing for masking phonemes. Notable efforts have analyzed speech models at the subgroup level (Dheram et al., 2022; Koudounas et al., 2023b; Zhang et al., 2022; Veliche and Fung, 2023; Koudounas et al., 2024). These works primarily concentrate on addressing fairness issues and mitigating biases. In contrast, our focus lies on enhancing the interpretability of such systems at both the individual and global levels.

7 Conclusion

We proposed a new perturbation-based explanation method for speech classification models using word-level audio segments and paralinguistic features. The experimental evaluation highlighted the ability of our approach to cope with different tasks, models, and languages. Our analysis revealed that word-level attributions accurately identified the spoken words influencing both local and global predictions, aligning well with user expectations. Meanwhile, paralinguistic attributions shed light on how non-linguistic features such as prosody and speaking conditions impacted predictions. Our findings showed that the generated word-level explanations are faithful to the model's inner workings. Moreover, a comprehensive user study proved that the generated explanations are plausible to human experts. Users found our explanation representation intuitive for pinpointing relevant words for predictions and examining multiple explanations—a crucial aspect for ensuring their utility in model explanations.

Future work could further explore explanation assessments, such as the subjective usefulness of explanations and their helpfulness for users for specific practical applications.

Limitations

Our work has some technical and design limitations. From the technical perspective, the two explanation methods we adopt to compute word-level segment attributions have known limitations. The leave-one-out method masks one-word segment at a time, thus not considering the intersectional effect of multiple masked words. LIME, instead, may suffer from instability since it relies on random sampling to generate perturbed samples around the instance to explain, and the sample size can affect explanations. We plan to experiment with different masking strategies and include other explanation methods. Moreover, word-level explanations might not be the most helpful explanations in specific speech classification tasks, e.g., spoken language identification or speaker identification. We are accounting for this limitation by evaluating paralinguistic contributions, but we will also explore new methods. We will also investigate the impact of the perturbation techniques and third-party speech libraries on paralinguistic attributions. From the experimental design perspective, we currently focus the evaluation of explanations on word-level segment attributions due to their closeness to token-level attributions and the solid literature for their evaluation. We intend to explore novel methods for evaluating paralinguistic contributions. The faithfulness measures we adopted are based, as the proposed explanation methods, on perturbation-based criteria. While the relative comparison of these methods holds, we note the intrinsic connection between these explainers and evaluation measures. We plan to explore alternative evaluation strategies for faithfulness assessment.

Ethical Statement

Our approach builds on pre-existing language technologies, including alignment and transcription models. However, such tools achieve uneven performance across different languages and sociodemographic groups (Adda-Decker and Lamel, 2005; Radford et al., 2023; Gu et al., 2023, *inter alia*). Whisper (Radford et al., 2023), the model we use for audio transcription, reports a drop in speech recognition capabilities for languages with fewer training instances and with high linguistic distance from the high-resource Indo-European languages used in the training set. Combining these methods has the potential to increase biases. Since our approach is directly reliant on these tools, our method is likely to work better for predominant languages and social groups. Our work should therefore be taken as a starting point for further activities to test and, where necessary, broaden its applicability.

We would also point out some ethical dual-use considerations. Paralinguistic attributions uncover if models are sensitive to signal perturbations. Malicious actors could take advantage of these vul-

nerabilities to manipulate or craft audio signals, potentially resulting in adversarial attacks. Yet, practitioners can use paralinguistic attributions as a proactive tool for robustness assessment.

Acknowledgments

This work is partially supported by the FAIR - Future Artificial Intelligence Research funded by the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013), the spoke "FutureHPC & BigData" of the ICSC - Centro Nazionale di Ricerca in High-Performance Computing, Big Data and Quantum Computing funded by the European Union - NextGenerationEU, the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 949944, INTEGRATOR), and a MUR FARE 2020 initiative under grant agreement prot. R20YSMBZ8S (INDOMITA). This manuscript reflects only the authors views and opinions, neither the European Union nor the European Commission can be considered responsible for them. DH and GA are members of the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis. EP, AK, and EB are members of the Database and Data Mining Group of Politecnico di Torino.

References

Martine Adda-Decker and Lori Lamel. 2005. Do speech recognizers prefer female speakers? In *Ninth European Conference on Speech Communication and Technology*.

Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. 2022. Openxai: Towards a transparent evaluation of model explanations. *Advances in Neural Information Processing Systems*, 35:15784–15799.

Siddhant Arora, Danish Pruthi, Norman Sadeh, William W Cohen, Zachary C Lipton, and Graham Neubig. 2022. Explain, edit, and understand: Rethinking user study design for evaluating model explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5277–5285.

Leila Arras, José Arjona-Medina, Michael Widrich, Grégoire Montavon, Michael Gillhofer, Klaus-Robert Müller, Sepp Hochreiter, and Wojciech Samek. 2019. Explaining and interpreting lstms. *Explainable ai: Interpreting, explaining and visualizing deep learning*, pages 211–238.

Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. " what is relevant in a text document?": An interpretable machine learning approach. *PLOS ONE*, 12(8):1–23.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proc. Interspeech 2022*, pages 2278–2282.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Sören Becker, Marcel Ackermann, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. 2018. Interpreting and explaining deep neural networks for classification of audio signals. *arXiv preprint arXiv:1807.03418*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Ebrahim (Abe) Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Marco Colussi and Stavros Ntalampiras. 2021. Interpreting deep urban sound classification using layer-wise relevance propagation. *arXiv preprint arXiv:2111.10235*.

Ian C Covert, Scott Lundberg, and Su-In Lee. 2021. Explaining by removing: A unified framework for model explanation. *The Journal of Machine Learning Research*, 22(1):9477–9566.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to

evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Pranav Dheram, Murugesan Ramakrishnan, Anirudh Raju, I-Fan Chen, Brian King, Katherine Powell, Melissa Saboowala, Karan Shetty, and Andreas Stolcke. 2022. Toward fairness in speech recognition: Discovery and mitigation of performance disparities. In *Proc. Interspeech 2022*, pages 1268–1272.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Milton Friedman. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701.

Annika Frommholz, Fabian Seipel, Sebastian Lapuschkin, Wojciech Samek, and Johanna Vielhaben. 2023. Xai-based comparison of input representations for audio event classification. *arXiv preprint arXiv:2304.14019*.

Xiangming Gu, Wei Zeng, and Ye Wang. 2023. Elucidate gender fairness in singing voice transcription. *arXiv preprint arXiv:2308.02898*.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. 2021. Contrastive explanations for model interpretability. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1597–1611, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Iver Jordal, Araik Tamazian, Emmanouil Theofanis Chourdakis, Céline Angonin, Tushar Dhyani, askskro, Nikolay Karpov, Omer Sarioglu, Baker-Bunker, kvilouras, Enis Berk Çoban, Florian Mirus, Jeong-Yoon Lee, Kwanghee Choi, MarvinLvn, SolomidHero, and Tanel Alumäe. 2023. iver56/audiomentations: v0.33.0.

Alkis Koudounas, Moreno La Quatra, Lorenzo Vaiani, Luca Colomba, Giuseppe Attanasio, Eliana Pastor, Luca Cagliero, and Elena Baralis. 2023a. Italic: An italian intent classification dataset. In *Interspeech 2023, 24rd Annual Conference of the International Speech Communication Association, Dublin, Ireland, 20-24 August 2023*. ISCA.

Alkis Koudounas, Eliana Pastor, Giuseppe Attanasio, Luca de Alfaro, and Elena Baralis. 2024. Prioritizing data acquisition for end-to-end speech model improvement. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Alkis Koudounas, Eliana Pastor, Giuseppe Attanasio, Vittorio Mazzia, Manuel Giollo, Thomas Gueudre, Luca Cagliero, Luca de Alfaro, Elena Baralis, and Daniele Amberti. 2023b. Exploring subgroup performance in end-to-end speech models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. 2019. Speech model pre-training for end-to-end spoken language understanding. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association*, pages 814–818.

Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Emanuel Metzenthin. 2020. Lime for time. `https://github.com/emanuel-metzenthin/Lime-For-Time`.

Saumitra Mishra, Bob L Sturm, and Simon Dixon. 2017. Local interpretable model-agnostic explanations for music content analysis. In *ISMIR*, volume 53, pages 537–543.

Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. 2019. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 193–209.

Eliana Pastor and Elena Baralis. 2019. Explaining black box models by means of local rules. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, SAC '19, page 510–517, New York, NY, USA. Association for Computing Machinery.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Soumya Sanyal and Xiang Ren. 2021. Discretized integrated gradients for explaining language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10285–10299, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hendrik Schuff, Alon Jacovi, Heike Adel, Yoav Goldberg, and Ngoc Thang Vu. 2022. Human interpretation of saliency-based explanation over text. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 611–636, New York, NY, USA. Association for Computing Machinery.

RR Selvaraju, M Cogswell, A Das, R Vedantam, D Parikh, and D Batra. 2022. Grad-cam: Visual explanations from deep networks via gradient-based localization. arxiv 2016. *arXiv preprint arXiv:1610.02391*.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Erik Strumbelj and Igor Kononenko. 2010. An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research*, 11:1–18.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Adam Twardoch. 2023. Audiostretchy. `https://github.com/twardoch/audiostretchy`.

Irina-Elena Veliche and Pascale Fung. 2023. Improving fairness and robustness in end-to-end speech recognition through unsupervised clustering. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In *Breakthroughs in Statistics: Methodology and Distribution*, pages 196–202. Springer.

Minz Won, Sanghyuk Chun, and Xavier Serra. 2019. Toward interpretable music tagging with self-attention. *arXiv preprint arXiv:1906.04972*.

Xiaoliang Wu, Peter Bell, and Ajitha Rajan. 2023a. Can we trust explainable ai methods on asr? an evaluation on phoneme recognition. *arXiv preprint arXiv:2305.18011*.

Xiaoliang Wu, Peter Bell, and Ajitha Rajan. 2023b. Explanations for automatic speech recognition. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. 2021a. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*.

Yao-Yuan Yang, Moto Hira, Zhaoheng Ni, Anjali Chourdia, Artyom Astafurov, Caroline Chen, Ching-Feng Yeh, Christian Puhrsch, David Pollack, Dmitriy Genzel, Donny Greenberg, Edward Z. Yang, Jason Lian, Jay Mahadeokar, Jeff Hwang, Ji Chen, Peter Goldsborough, Prabhat Roy, Sean Narenthiran, Shinji Watanabe, Soumith Chintala, Vincent Quenneville-Bélair, and Yangyang Shi. 2021b. Torchaudio: Building blocks for audio and speech processing. *arXiv preprint arXiv:2110.15018*.

MD Zeiler and R Fergus. 2013. Visualizing and understanding convolutional networks. arxiv. *arXiv preprint arXiv:1311.2901*.

Yuanyuan Zhang, Yixuan Zhang, Bence Mark Halpern, Tanvina Patel, and Odette Scharenborg. 2022. Mitigating bias against non-native accents. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2022, pages 3168–3172.

Pablo Zinemanas, Martín Rocamora, Eduardo Fonseca, Frederic Font, and Xavier Serra. 2021a. Toward interpretable polyphonic sound event detection with attention maps based on local prototypes. In *6th Workshop on Detection and Classification of Acoustic Scenes and Events, DCASE 2021, Barcelona, Spain, 15-19 nov. 2021, pp. 50-54*. Universitat Pompeu Fabra.

Pablo Zinemanas, Martín Rocamora, Marius Miron, Frederic Font, and Xavier Serra. 2021b. An interpretable deep learning model for automatic sound classification. *Electronics*, 10(7).

A WhisperX

We adopt WhisperX (Bain et al., 2023) to extract beginning and ending timestamps for each uttered word. WhisperX builds upon Whisper (Radford et al., 2023), a state-of-the-art speech model trained on diverse audio data enabling multilingual speech recognition, translation, and language identification capabilities. While the original Whisper model performs speech transcription at a high level of accuracy, it only provides utterance-level timestamp annotations. WhisperX improves upon this by leveraging Whisper's foundation and incorporating additional techniques to achieve word-level timestamp precision. It applies voice activity detection to isolate speech segments from non-speech periods. Forced phoneme alignment is then used to map the acoustic features of spoken utterances

to their constituent words at a finer temporal granularity. WhisperX thus provides the word-level timing data needed for our analysis while retaining Whisper's high performance in speech processing domains.

B Implementation Details

In this section, we describe the libraries and parameters used to generate the explanations via input perturbation on word-level audio segments (B.1) and on paralinguistic features (B.2).

B.1 Word-Level Attribution

To compute word-level contribution, we generalize two approaches from XAI literature, L1O (Leave One Out) and LIME. For both approaches, we introduce perturbations through selective zeroing (i.e., silencing) of audio segments in the time domain as in (Wu et al., 2023a).

For LIME, we use the 'LIME FOR TIME' library (Metzenthin, 2020). We customize it to deal with word-level segments and zeroing as the form of perturbation. The original implementation relies on equal-width splits defined by the number of segments the input would be split. Instead, we consider word-level audio segments and split the audio based on the timestamps derived from WhisperX. We generate neighbor samples of the instance to explain by masking random audio segments. In our version, we mask these segments by setting their values to zero. In our experiments, we set the number of generated perturbed samples to 1000. For the other settings, we use default ones (e.g., Ridge linear model as the interpretable model and random selection of the segments to mask).

B.2 Paralinguistic Attribution

We analyze the impact of paralinguistic aspects by introducing targeted perturbations to the utterances. Specifically, we apply custom audio manipulations and then compute the variations in the class prediction probability. In our analysis, we experimented with time stretching, noise addition, pitch modification, and reverberation effects.

Time stretching. We use the AudioStretchy library (Twardoch, 2023), which enables high-quality time-scale modification via Time-domain harmonic scaling (TDHS). This allows adjustment of speech rate without impacting pitch contour or formant structure evolution over time. More in detail, downward time scaling progressively shortens utterance length from the original to 55% in decrements of 5%, spanning the range from 55-95% duration. Similarly, upward time scaling progressively lengthens utterance length from the original to 145% in increments of 5%, spanning the range from 105-145% duration.

Noise. We leverage the noise addition transformation from Torchaudio (Yang et al., 2021b) library, which scales and adds noise according to a specified signal-to-noise ratio (SNR). In our experiments, we add white background noise at various SNR levels. Specifically, noise is introduced with the SNR ranging from 40 decibels (dB) down to 0.1 dB, decreasing in steps of 2.5, 5, or 10 dB depending on the level. This results in 11 noise perturbation conditions with SNR values of 40, 20, 10, 7.5, 5, 4, 3, 2, 1, 0.5 and 0.1 dB. As a result, we worsen the clarity of the speech from a high SNR of 40 dB down to 0.1 dB to evaluate the model sensitivity to noise.

Pitch shifting. We use the pitch shift function of the Torchaudio library and vary the number of steps to shift the input waveform. This allows us to isolate the effect of pitch variation independently from other temporal factors like time-stretching or speed changes. We apply both downward and upward modulation of the utterance fundamental frequency (f0). Downward pitch scaling lowers the f0 within the range of -0.5 to -5 semitones in decrements of 0.5 semitones. This progressively shifts the semitones lower by up to 5 semitones. Conversely, upward pitch scaling raises the f0 within the [0.5, 5] semitones range, with increments of 0.5 semitones. This progressively transposes the utterance up by a maximum of 5 semitones, effectively shifting the pitch closer to one full semitone higher than the largest downward value. By systematically altering the pitch up and down within these controlled bounds, we aim to evaluate the model's invariance to changes in vocal prosody that may occur naturally due to differences among speakers.

Reverberation. We apply the room impulse response generator from Audiomentations (Jordal et al., 2023) to introduce a reverberation. It models a cuboid room with parameterized dimensions, absorption, configurable source, and microphone placements to simulate natural reverberant effects. We systematically varied the dimensions of the virtual room environment. Specifically, we model room width (x-axis), depth (y-axis), and height (z-axis) within the range of 3 to 7 meters, altering

	It's	really	romantic.
class=happy	-0.02	0.24	0.04

Table 5: Example of word-level audio segment explanation; IEMOCAP dataset.

	speed up	speed down	pitch down	pitch up	reverb	noise
class=happy	0.33	0.12	0.82	0.26	0.94	0.74

Table 6: Example of paralinguistic attribution; IEMOCAP dataset, instance in Table 5.

each dimension incrementally by 1 meter. This produces a total of 5 unique room-size configurations for evaluation. The absorption coefficient, which determines how room surfaces absorb sound, is held constant at 0.1. This setting effectively simulated a typical office or residential room space rather than a sound-treated studio, which would require a higher coefficient above 0.4. Additionally, we established a minimum distance of 0.5 meters between any sound source, microphone, or reflecting surface to control for closely spaced early reflections vs. larger room reverberation effects. By varying these parameters, the perturbation aims to assess model robustness when processing audio captured or generated in enclosed environments that naturally differ in size.

C Additional Qualitative Evaluations

We report additional results on the word-level and paralinguistic attributions. We focus on the emotion recognition task for which the content by itself is not sufficient to convey the meaning and the emotion.

Individual level. Tables 5 and 6 show the word-level (via LIME) and paralinguistic attributions for an instance of IEMOCAP. Figure 5 further focuses on the paralinguistic aspect. Wav2vec 2.0 correctly predicts the emotion label of the instance as '*happy*'. The speaker pronounced the sentence 'It's really romantic' with a cheerful tone while laughing. At the content level, the relevant word is 'really'. We then may wonder how the model is sensitive to variation of paralinguistic features for this instance. The model is highly sensitive when introducing reverberation and noise (see Table 6). Then, it is highly sensitive to shifting down the pitch. When we lower the pitch, the probability of the class 'happy' drops from 0.95 of the original

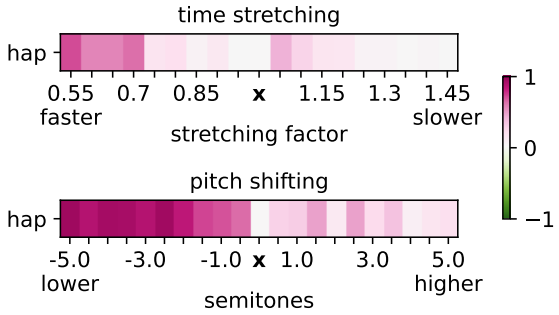

Figure 5: $r_p(\widetilde{\mathbf{x}})$ breakdown for $p :=$ time stretching, and pitch shifting; same instance **x** of Table 5.

	What	am	I	gonna	do,	huh?
class=sad	0.07	0.02	0.04	0.18	0.10	0.20

Table 7: Example of word-level audio segment explanation; IEMOCAP dataset.

	speed up	speed down	pitch down	pitch up	reverb	noise
class=sad	0.97	0.81	0.01	0.17	0.98	0.09

Table 8: Example of paralinguistic attribution; IEMOCAP dataset, instance in Table 7.

recording to 0.329 when shift of -0.5 semitones and to 0.017 when we shift of 5. We report the drop in Figure 5. In all these cases, the model labels the perturbed instance to the 'neutral' class.

Table 7 and 8 show the word-level attributions (again via LIME) and paralinguistic ones for another instance of IEMOCAP. The model correctly predicts the sentence as belonging to the '*sad*' emotion. At the word level (Table 7), '*huh?*' and '*gonna*' are the most important words. At the paralinguistic level, other than the introduction of reverberation, the time-stretching transformation highly induces a change in the prediction probabilities. As we can also observe from the heatmap in Figure 6, all time-stretching transformations induce a drop in the prediction probability of the '*sad*' class. Hence, the model is highly sensitive to the perturbation of the speaking rate for this instance. The speaking rate is an important characteristic of communicating sad emotions, making the model sensitivity of this feature on the emotion label plausible. Moreover, we observe a change in the prediction probability when increasing the pitch (for values $\geq$3 in Figure 6). The prediction probability drops by 0.38 when we increase the

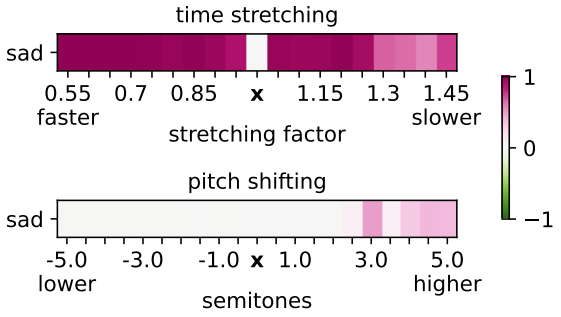

Figure 6: $r_p(\widetilde{\mathbf{x}})$ breakdown for $p :=$ time stretching, and pitch shifting; same instance **x** of Table 7.

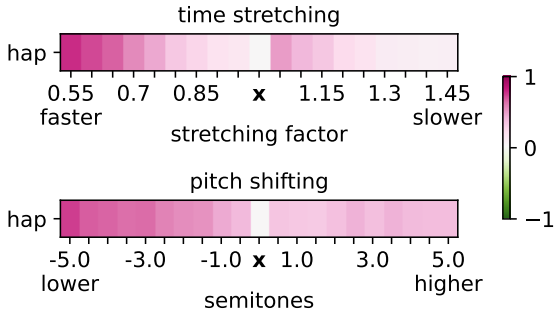

Figure 7: Average $r_p(\widetilde{\mathbf{x}})$ for 50 instances of the IEMOCAP dataset assigned to the '*happy*' class for $p :=$ time stretching and pitch shifting.

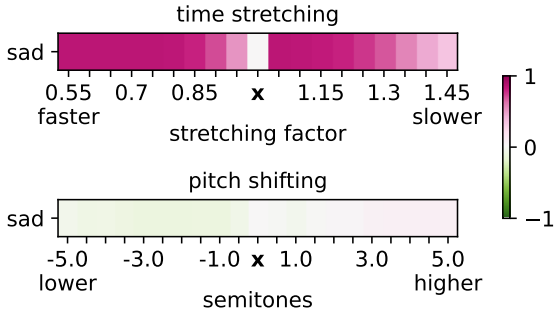

Figure 8: Average $r_p(\widetilde{\mathbf{x}})$ for 50 instances of the IEMOCAP dataset assigned to the '*sad*' class for $p :=$ time stretching and pitch shifting.

pitch by 5 octaves (specifically, it goes from 0.98 to 0.6, and the probability assigned to the '*happy*' class increases). We argue this change is plausible for this instance prediction: the higher pitch could be generally associated with other emotions such as happiness.

Global level. Finally, we may wonder whether the higher sensitivity of the pitch for the '*happy*' class and of the time stretching for the '*sad*' class is local to the two instances we analyzed, or it is observed for multiple instances. We compute the prediction difference when varying the perturbation for 50 instances predicted as '*happy*' and for 50 as '*sad*'. We then aggregate the scores separately for the two classes and compute the average. Figures 7 and 8 show the average prediction differences for the '*happy*' and '*sad*' classes, respectively. The model confirms to be sensitive to the pitch for the class '*happy*'. Moreover, it is also sensitive when time stretching the audio, especially when increasing the speed. For the '*sad*' class, the model confirms to be sensitive to variations of the speaking rate via time stretching. Shifting the pitch has a negligible impact: raising the pitch induces, on average, a slight drop in the prediction probability of the '*sad*' class, while lowering it causes a slight increase.

These analyses show that paralinguistic attributions can be a valid tool to inspect and understand the model behavior.

D Plausibility User Study

The user study is available at https://forms.gle/vuWpm7ha6r3BRt6w8. The link was released in October 2023 via email. Participation was voluntary and not compensated. We did not collect any personally identifiable information on participant subjects.

D.1 Setup.

We first ask about preferences in the visualization, separately from the rest. This design should avoid biases since all other visualizations use our color-code representation. We use a scale from 1 to 4 for all ratings to encourage participants to take a clear stance, discouraging the selection of a neutral score. The entire study typically requires around 20 minutes to complete.

We provide to the participants the following definition for *plausibility*: "Plausibility reflects how explanations are aligned with human reasoning and how they are convincing to humans. If the explanation of a system prediction sounds reasonable, clear, and like something a person would consider, then they are considered plausible." We then instruct the participant to consider the following aspects when evaluating plausibility "(i) Does the score assigned to the word align with my expectations? (ii) Would the explanations I would provide as a human match those of the model? (iii) Does the explanation seem reasonable and believable? (iv) Does it make sense within the context of the problem?'.

Category	FSC Bar	FSC Words	FSC Table	ITALIC Bar	ITALIC Words	ITALIC Table
Identify words (↑)	$2.54_{\pm 0.74}$	$\mathbf{3.54}_{\pm 0.51}$	$3.40_{\pm 0.85}$	$3.54_{\pm 0.61}$	$3.51_{\pm 0.51}$	$\mathbf{3.66}_{\pm 0.54}$
Compare word (↑)	$3.37_{\pm 0.77}$	$2.60_{\pm 0.81}$	$\mathbf{3.63}_{\pm 0.60}$	$\mathbf{3.66}_{\pm 0.68}$	$2.74_{\pm 0.82}$	$3.63_{\pm 0.65}$
Inspect multiple (↑)	$2.57_{\pm 1.01}$	$\mathbf{3.34}_{\pm 0.73}$	$3.29_{\pm 0.83}$	$\mathbf{3.34}_{\pm 0.76}$	$3.29_{\pm 0.75}$	$3.29_{\pm 0.62}$
Overall preference rank (↓)	$2.51_{\pm 0.70}$	$1.94_{\pm 0.73}$	$\mathbf{1.54}_{\pm 0.74}$	$1.89_{\pm 0.83}$	$2.20_{\pm 0.83}$	$\mathbf{1.80}_{\pm 0.68}$

Table 9: User study - Effectiveness of the visualization. Average and standard deviation scores for the four questions.

Category	FSC Bar, Words	FSC Bar, Table	FSC Word, Table	ITALIC Bar, Words	ITALIC Bar, Table	ITALIC Word, Table
Identify words	**<0.0001**	**0.0001**	0.2291	0.4284	0.1425	0.1126
Compare words	**0.0008**	0.0752	**<0.0001**	**<0.0001**	0.4278	**0.0001**
Inspect multiple	**0.0016**	**0.0010**	0.3867	0.4364	0.3583	0.4762
Overall preference rank	**0.0083**	**0.0001**	0.0518	0.0802	0.3187	**0.0297**

Table 10: User study - Effectiveness of the visualization. p-value of the pairwise Wilcoxon test. The statistically significant pairwise differences are in bold (p-value<0.05).

Category	FSC	ITALIC
Identify words	**<0.0001**	0.3577
Compare words	**<0.0001**	**<0.0001**
Inspect multiple	**0.0092**	0.7030
Overall preference rank	**0.0002**	0.2564

Table 11: User study - Effectiveness of the visualization. p-value of the Friedman test. The statistically significant differences are in bold (p-value<0.05).

D.2 Task, datasets, and explainer

We focus our study on a single task, Intent Classification, to simplify the user experience and enhance participant understanding. We include explanations from the FSC and ITALIC datasets to cover both the English and Italian languages. This inclusion enables the assessment of the plausibility of our methods across different linguistic contexts. Moreover, the choice of the task and datasets enables the assessment in the multi-label and single-label classification scenarios. While for ITALIC we target the intent alone, for FSC we simultaneously predict and explain the action, object, and location. Multi-label explanation evaluation introduces an additional complexity for the participants; we consider this additional load in our assessment.

We focus on LIME explanations since they obtained higher faithfulness.

D.3 Assessments

Quality Control. We compare our explanation with the random baseline to assess the reliability of the explanations. For a given audio, we presented to the participants an explanation generated by our approach and one randomly generated. We then asked users to indicate which explanation they found more plausible. This initial question allows us to determine whether users perceive our explanations as indeed more plausible than random ones. Moreover, it helps prevent potential biases in interpretation. By offering a clear choice between our explanations and random ones, we guide users to focus on what they expect and find most suitable.

For each dataset, we presented two explanations (ours and the random baseline) for the same recording visualized in a color-coded table.

Plausibility Assessment. We asked the participants to rate an explanation's plausibility level on a scale from 1 to 4, with 1 indicating no plausibility at all and 4 indicating very high plausibility.

We presented ten explanations for ten distinct recordings for each dataset, visualized in a color-coded table.

Visualization Strategy. In our work, we visualize explanations as a color-coded score table that combines the heat map representation of word-level saliency explanations and the precise indication of the score. We assess how users find this visualization effective compared to word-level saliency

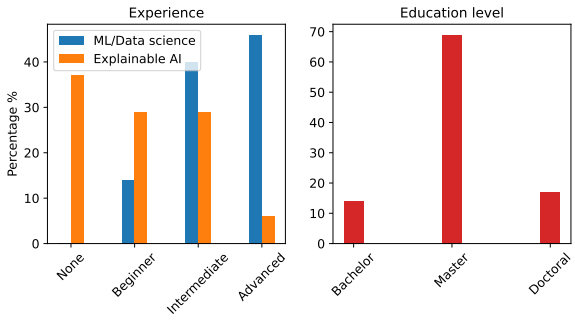

Figure 9: Statistics of participants. Expertise in machine learning and/or data science and in Explainable AI (left) and education level (right).

explanations as color-coded highlighted words and bar plots showing the importance of each word. For each visualization, we asked them to quantify with a score from 1 to 4 the following aspects: (i) how easy the visualization allows them to identify the words that push (or pull from) the prediction, (ii) how easy it is to compare the relative importance among words, (iii) how much would be easy to inspect many explanations. Finally, we asked the participants to express their overall preferences among these visualization options.

For each dataset, we presented the explanation of a single recording visualized in the three forms: bar chart, saliency map, and color-coded table.

D.4 Participants

Participation Requisites. To be eligible to participate in the study, individuals must have a minimum proficiency level of B2 or higher in both English and Italian.

Sociodemographic Statistics. We also asked participants, as optional questions, their age (a numerical integer) and gender (male, female, non-binary, undeclared) of participants. Table 12 reports the distribution of gender and age.

Level of Expertise. We collect information on participants' expertise in machine learning (ML) and/or data science and their familiarity with explainable AI. We categorize the expertise into four levels: None, Beginner (Limited knowledge or experience), Intermediate (Moderate knowledge or experience), and Advanced (Extensive knowledge or experience). We then collect data on the education level of the participants.

Figure 9 presents the statistics in percentage form. Most participants have moderate or advanced knowledge or experience in ML and data science, while most have no experience with explainable AI. The majority of the participants hold a Master's degree as their highest level of education.

D.5 Result Details

Table 9 shows the average and standard deviation scores for the four questions for the assessment of the effectiveness of the visualization.

We test the statistical significance of the score and their relative ranking. We use the Friedman (Friedman, 1937) and the Wilcoxon signed-rank (Wilcoxon, 1992) tests.

Given a question and the scores for the three representations, we use the Friedman test to test the null hypothesis that scores have the same distribution. For p-values lower than a significance level (that we set at 5%), we reject the null hypothesis and say that the three scores differ. Table 11 shows the Friedman test over the scores for the three visualizations (bar chart, saliency words, and color-coded table). The scores of the participants differ for all questions on the multi-label dataset FSC. For ITALIC, the scores differ significantly when evaluating the representation efficacy for comparing words.

With the Wilcoxon signed-rank test, we test the null hypothesis that two paired samples come from the same distribution. If the obtained p-value is lower than our confidence threshold of 5%, we reject the null hypothesis, and we say that there is a difference in scores between the two groups. We compute the Wilcoxon test for each question and for each pair of visualizations. Table 10 reports the p-values of the pairwise Wilcoxon test.

For FSC, the scores show statistically significant differences between the bar plot representation and color-coded ones (both saliency words and our color-coded table) when evaluating their ability to identify words, inspect multiple explanations, and consider overall aspects. Participants preferred saliency words and our table for these tasks and scored them similarly. Participants preferred our color-coded table and the bar plot over saliency words to compare relative importance among words.

For ITALIC, participants preferred, as for FSC, our table and the bar plot over saliency words for comparing relative word importance and scored them similarly. The other significant difference is in the overall preference ranked: users preferred our color-coded table over saliency maps.

We argue that these differences in preference

Gender	Male	Female	Non-Binary	Undeclared
	68.57%	31.43%	0%	0%
Age	≤ 25	[26-29]	≥ 30	Undeclared
	31.43%	60.0%	8.57%	0.0%

Table 12: Gender and age distribution of the participants.

between the two datasets can be reconducted to a key distinction between the two: multi-label (FSC) vs single-label (ITALIC) scenario.

E CO2 Emission Related to Experiments

Experiments were conducted using a private infrastructure, which has a carbon efficiency of 0.29 kgCO_2eq/kWh. A cumulative of 60 hours of computation was performed on hardware of type RTX A6000 (TDP of 300W). Total emissions are estimated to be 5.22 kgCO_2eq of which 0 percent were directly offset. Estimations were conducted using the MachineLearning Impact calculator presented in Lacoste et al. (2019).