Listenable Maps for Audio Classifiers

Francesco Paissan[1,4], Mirco Ravanelli[2,4], Cem Subakan[2,3,4]

[1]Fondazione Bruno Kessler, [2]Concordia University, [3]Laval University, [4]Mila-Québec AI Institute

Contributions

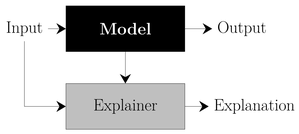
- We develop an **understandable** and **faithful** posthoc explanation method for audio classifiers.
- Our method is agnostic to the classifier input domain and generates **listenable** explanations.
- We propose a fine-tuning strategy that improves the understandability/faithfulness trade-off.

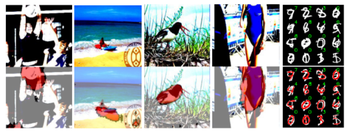
Explainable Machine Learning

- Black-box Models

- Explainable Models

- Posthoc Explanations



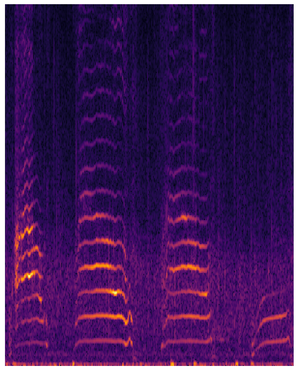


Posthoc Explanations for Audio Classifiers

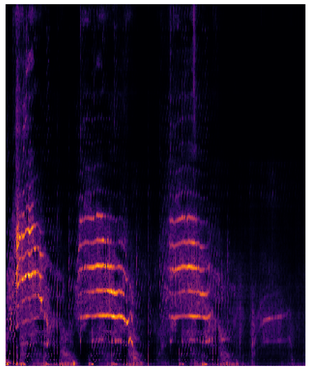
We would like to obtain

- Faithful,
- Listenable,
- Understandable

Posthoc Explanations for Audio Classifiers

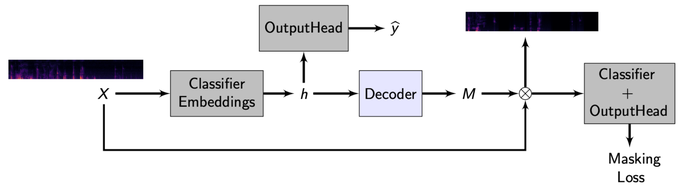


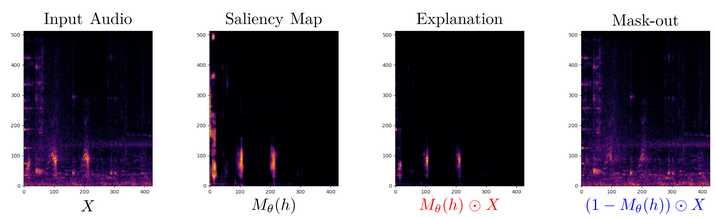
Input



Explanation

Listenable Maps for Audio Classifiers



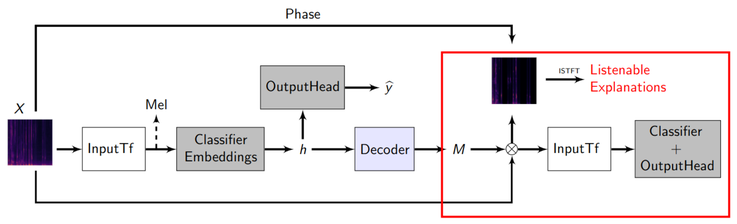


The objective function:

$$\min_\theta \overbrace{\lambda_{in}\mathcal{L}_{in}(\log f(M_\theta(h) \odot X), \hat{y})}^{\text{Mask-In}} - \overbrace{\lambda_{out}\mathcal{L}_{out}(\log f((1 - M_\theta(h)) \odot X), \hat{y})}^{\text{Mask-Out}} + \overbrace{R(M_\theta(h), X)}^{\text{Regularizer}}$$

- **Mask-In:** Maximizes the classifier agreement between the input and the explanation.
- **Mask-Out:** Minimizes the classifier agreement between the input and mask-out.
- **Regularizer:** $R(M_\theta(h), X) = \lambda_g \underbrace{\|M_\theta(h) \odot X - X_{\text{clean}}\|}_{R_1} + \lambda_s \underbrace{\|M_\theta(h)\|_1}_{R_2}$,
  - R_1: Avoids trivial solutions (e.g. all 1s).
  - R_2: Improves Understandability. Used in finetuning stage.

Producing Listenable Explanations



$$\textit{Listenable Explanation} = \text{ISTFT}\left((M_\theta(h) \odot X)e^{jX_{\text{phase}}}\right)$$

Evaluation Metrics

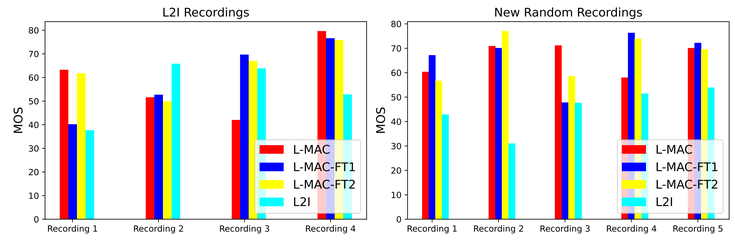
- **Faithfulness**: Measures importance of explanations for classifier decisions
  - L2I-Faithfulness, Average-Increase, Average-Gain, Average-Drop, Input Fidelity
- **Structural metrics**: Measures the understandability of the explanations
  - Sparseness, Complexity

Experimental results

Quantitative Analysis

	Metric	AI (↑)	AD (↓)	AG (↑)	FF (↑)	Fid-In (↑)	SPS (↑)	COMP (↓)
ID	Saliency	0.00	15.79	0.00	0.05	0.07	0.39	5.48
	Smoothgrad	0.00	15.71	0.00	0.03	0.05	0.42	5.32
	IG	0.25	15.45	0.01	0.07	0.13	0.43	5.11
	GradCAM	8.50	10.11	1.47	0.17	0.33	0.34	5.64
	Guided GradCAM	0.00	15.61	0.00	0.05	0.06	0.44	5.12
	Guided Backprop	0.00	15.66	0.00	0.05	0.06	0.39	5.47
	L2I, RT=0.2	1.63	12.78	0.42	0.11	0.15	0.25	5.50
	SHAP	0.00	15.79	0.00	0.05	0.06	0.43	5.24
	L-MAC (NotListenable)	35.63	1.59	**24.28**	**0.22**	**0.42**	0.45	**4.11**
	L-MAC, FT, $\lambda_g = 4$ (ours)	32.37	1.98	18.74	0.21	0.41	0.43	5.20
	L-MAC (ours)	**36.25**	**1.15**	23.50	0.20	**0.42**	**0.47**	4.71
OOD	Saliency	0.62	31.73	0.07	0.06	0.12	0.76	11.06
	Smoothgrad	0.12	31.84	0.00	0.06	0.13	0.83	10.66
	IG	0.37	31.15	0.03	0.12	0.26	0.87	10.22
	L2I	5.00	25.65	1.00	0.20	0.35	0.52	10.99
	GradCAM	14.12	17.62	7.46	0.25	0.00	0.91	9.66
	Guided GradCAM	0.00	31.74	0.00	0.07	0.11	0.89	10.24
	Guided Backprop	0.63	31.73	0.07	0.06	0.11	0.76	11.06
	SHAP	0.00	31.81	0.00	0.07	0.14	0.84	10.58
	L-MAC (NotListenable)	60.25	**4.84**	34.72	**0.44**	0.80	0.90	**8.29**
	L-MAC FT, $\lambda_g = 4$ (ours)	50.75	6.73	26.00	0.39	0.78	0.84	10.51
	L-MAC (ours)	**60.63**	4.82	**35.85**	0.39	**0.81**	**0.94**	9.61

User Study



Conclusions

- We proposed a state-of-the art posthoc explanation technique.
- Our method is agnostic to classifier input representation.
- Our method provides **understandable**, **listenable** and **faithful** explanations both in ID and OOD cases.
- Our finetuning strategy consistently improves the explanation quality, while marginally affecting faithfulness.
- This work has been carried out within the SpeechBrain project. Our results are fully reproducible, and the trained model weights are available.

SpeechBrain

Check out the project webpage!