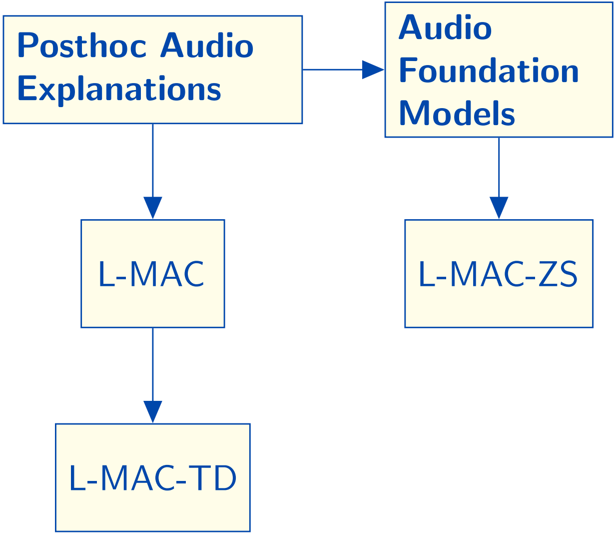
Posthoc Explanations for Audio Models

Cem Subakan

December 5, 2024

Plan


Posthoc Audio Explanations
Audio Foundation Models
L-MAC
L-MAC-ZS
L-MAC-TD

Collaborators



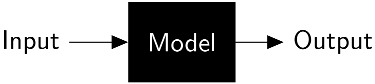


Francesco Paissan, Luca Della Libera, Eleonora Mancini, Mirco Ravanelli, Cem Subakan

Table of Contents

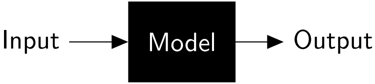
Explainable Machine Learning

- Black-box models

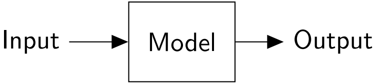


Explainable Machine Learning

- Black-box models

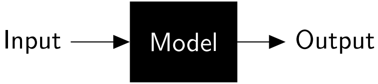


- Explainable Models

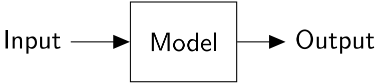


Explainable Machine Learning

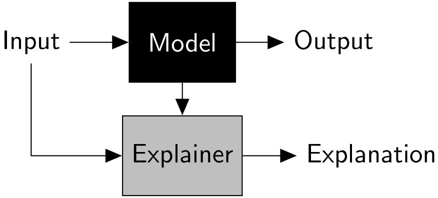
- Black-box models



- Explainable Models

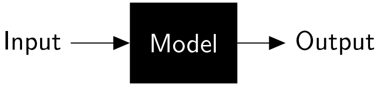


- Posthoc Explanations

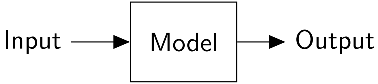


Explainable Machine Learning

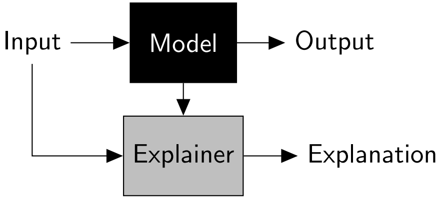
- Black-box models



- Explainable Models



- Posthoc Explanations



Desiderata: Faithful, Listenable, Understandable Explanations
Important Tool for Decision Critical Applications (e.g. Healthcare, DeepFake detection)

Neural Network Explanation

- Why does this particular input lead to that particular output?

Neural Network Explanation

- <u>Why does this particular input lead to that particular output?</u>

Neural Network Explanation

- Why does this particular input lead to that particular output?

Recording, Classified as DOG

Neural Network Explanation

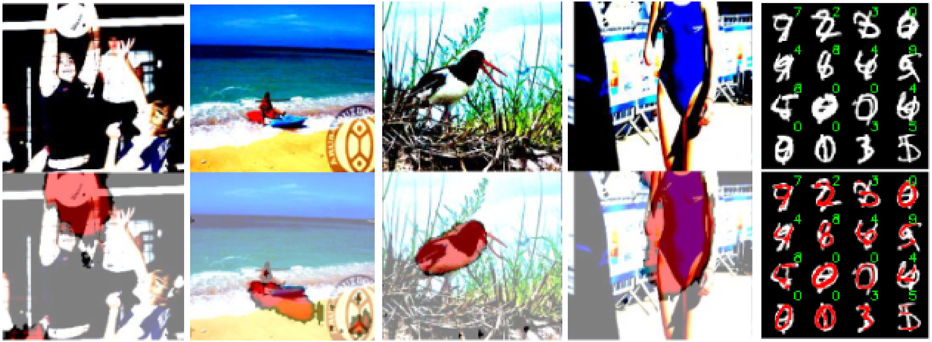
- Why does this particular input lead to that particular output?

Recording, Classified as DOG

Interpretation

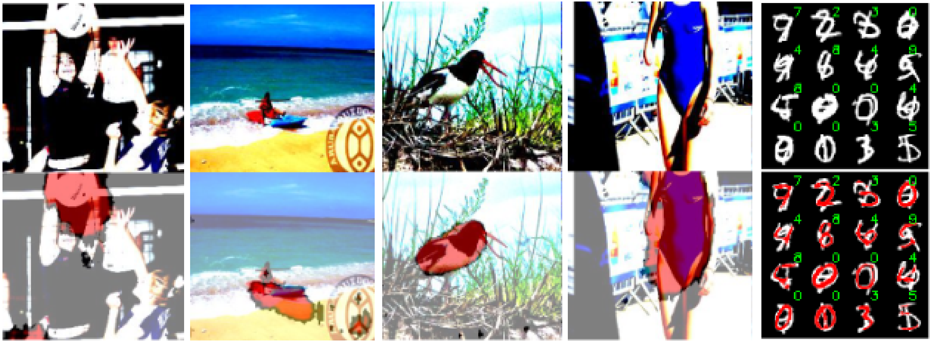
Explanations

- Saliency maps are commonly used in computer vision for producing explanations.



Explanations

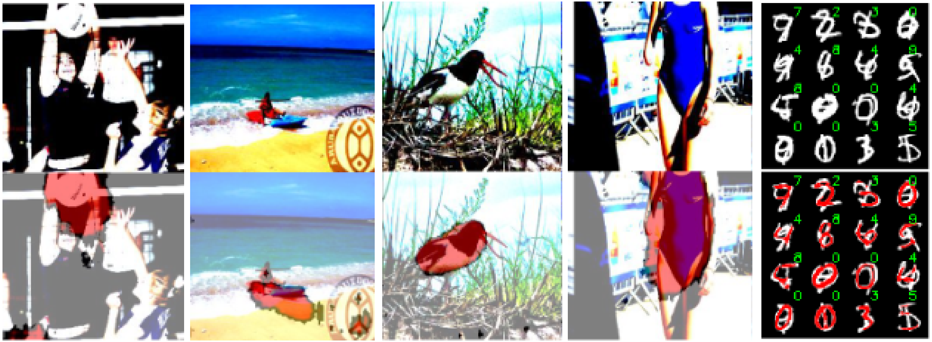
- Saliency maps are commonly used in computer vision for producing explanations.



- The explanations should **faithfully** follow the original model.

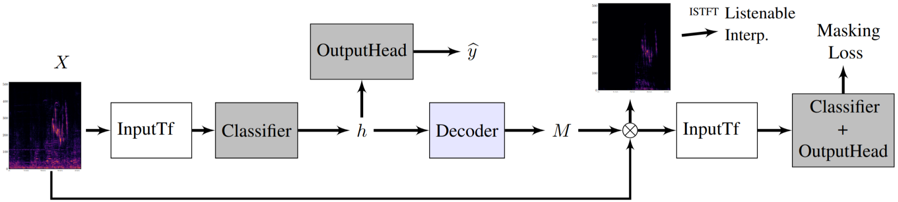
Explanations

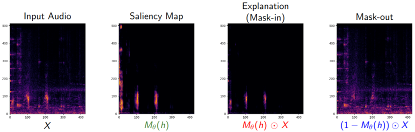
- Saliency maps are commonly used in computer vision for producing explanations.



- The explanations should **faithfully** follow the original model.
- **Faithful** and **understandable** explanations are important for domains where decisions are critical!

Listenable Maps for Audio Classifiers (L-MAC)





$$\min_\theta \overbrace{\lambda_{in}\mathcal{L}_{in}(\log f(M_\theta(h) \odot X), \hat{y})}^{\text{Mask-in}}$$

$$- \overbrace{\lambda_{out}\mathcal{L}_{out}(\log f((1 - M_\theta(h)) \odot X), \hat{y})}^{\text{Mask-out}} + \overbrace{|M_\theta(h)|}^{\text{Mask Reg}}$$

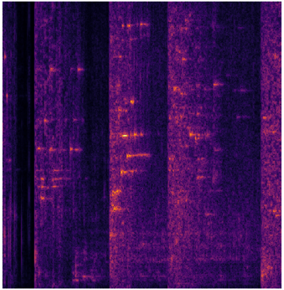
	Metric	AI (↑)	AD (↓)	AG (↑)	FF (↑)	Fid-In (↑)	SPS (↑)	COMP (↓)
Listenable (STFT→Mel)	Saliency	0.00	15.79	0.00	0.05	0.07	0.39	5.48
	Smoothgrad	0.00	15.71	0.00	0.03	0.05	0.42	5.32
	IG	0.25	15.45	0.01	0.07	0.13	0.43	5.11
	GradCAM	8.50	10.11	1.47	0.17	0.33	0.34	5.64
	Guided GradCAM	0.00	15.61	0.00	0.05	0.06	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 (ours)	**36.25**	**1.15**	**23.50**	0.20	**0.42**	**0.47**	**4.71**
	L-MAC, FT, $\lambda_g = 4$ (ours)	32.37	1.98	18.74	**0.21**	0.41	0.43	5.20
Not Listenable (Mel)	Saliency	0.00	15.81	0.00	0.10	0.07	0.39	4.53
	Smoothgrad	0.00	15.61	0.00	0.07	0.04	0.39	4.54
	IG	0.00	15.55	0.00	0.12	0.08	0.42	4.36
	GradCAM	7.00	10.93	1.04	0.17	0.29	0.34	**4.72**
	Guided GradCAM	0.125	15.40	6.67	0.08	0.07	**0.45**	4.17
	Guided Backprop	0.125	15.54	0.00	0.10	0.08	0.39	4.53
	SHAP	0.00	15.57	0.00	0.11	0.08	0.41	4.42
	L-MAC (ours)	35.63	1.59	**24.28**	0.22	**0.42**	**0.45**	4.11
	L-MAC (ours) FT, $\lambda_g = 4$	**36.13**	**1.28**	21.15	**0.23**	**0.42**	0.32	4.71

[F. Paissan, M. Ravanelli, C.Subakan; ICML 2024 (Oral)]

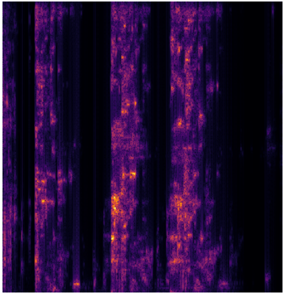
Contributions

- We develop an **understandable** and **faithful** (SOTA) posthoc explanation method for audio classifiers.

Input Audio



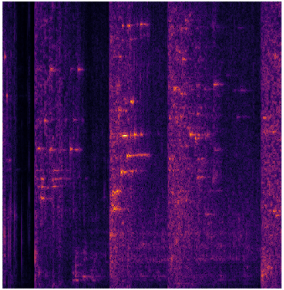
Explanation



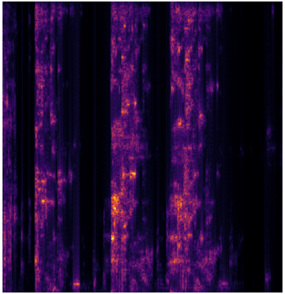
Contributions

- We develop an **understandable** and **faithful** (SOTA) posthoc explanation method for audio classifiers.

Input Audio



Explanation

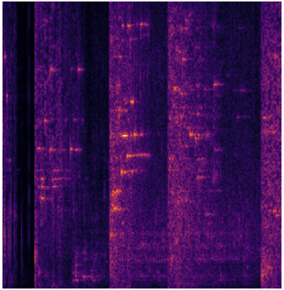


- Our method is agnostic to classifier input domain, and generates **listenable** explanations.

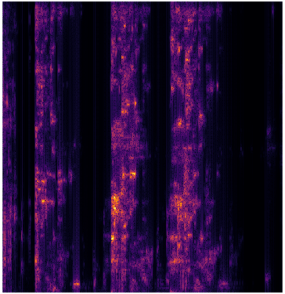
Contributions

- We develop an **understandable** and **faithful** (SOTA) posthoc explanation method for audio classifiers.

Input Audio



Explanation



- Our method is agnostic to classifier input domain, and generates **listenable** explanations.
- We propose a fine-tuning strategy that improves understandability/faithfulness trade-off.

Considerations

We would like to obtain

Considerations

We would like to obtain

- Faithful,

Considerations

We would like to obtain

- Faithful,
- Listenable,

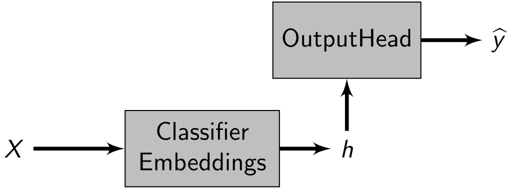
Considerations

We would like to obtain

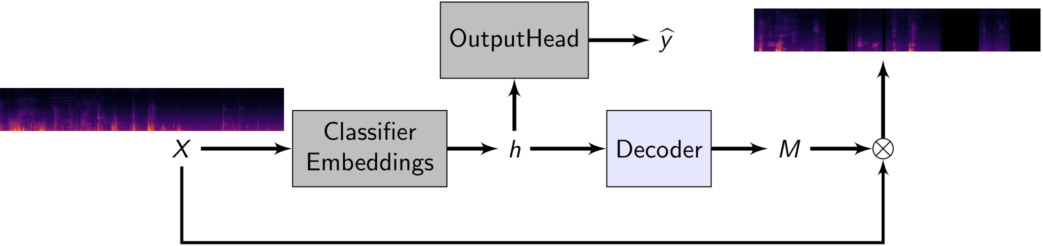
- Faithful,
- Listenable,
- Understandable

Posthoc Explanations for Audio Classifiers

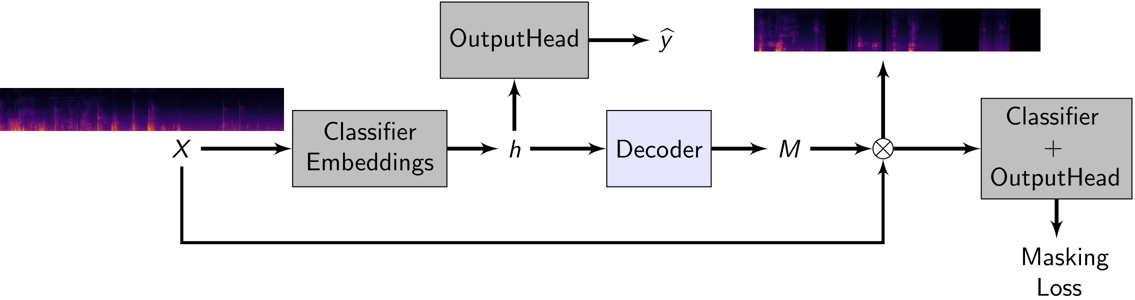
Listenable Maps for Audio Classifiers



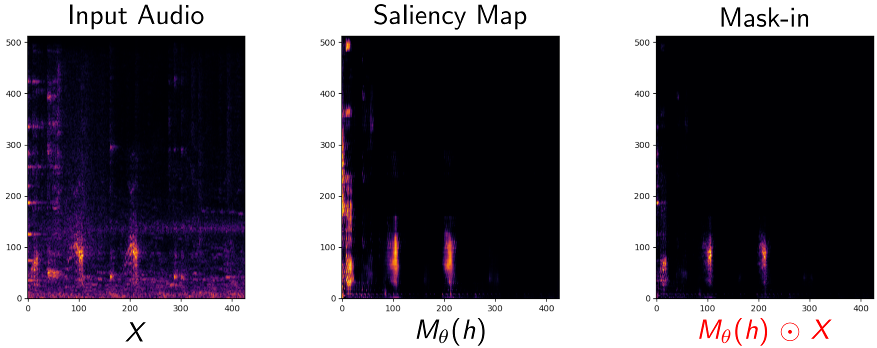
Listenable Maps for Audio Classifiers



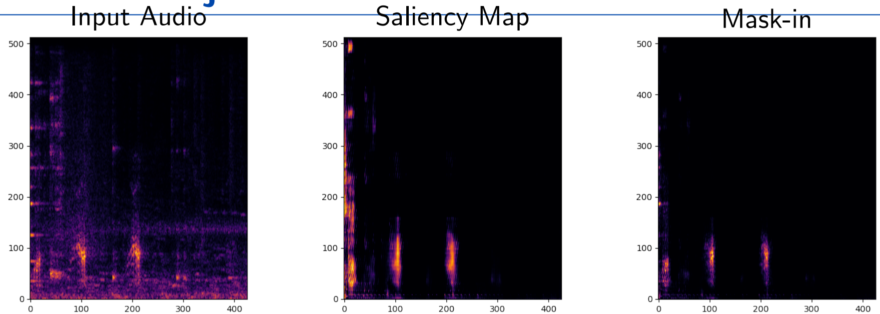
Listenable Maps for Audio Classifiers



Optimization objective



Optimization objective

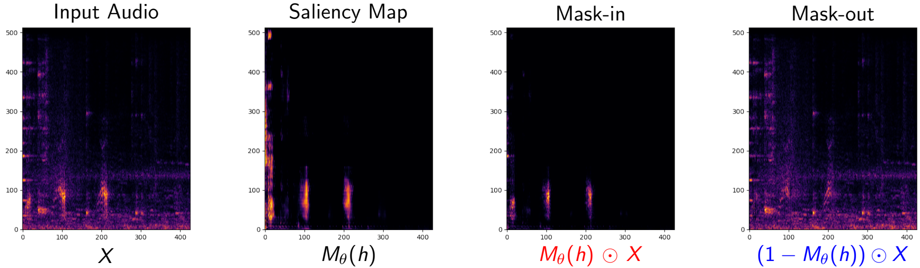


X $\qquad$ $M_\theta(h)$ $\qquad$ $M_\theta(h) \odot X$

$$\min_{\theta} \overbrace{\lambda_{in}\mathcal{L}_{in}(\log f(M_\theta(h) \odot X), \widehat{y})}^{\text{Mask-in}}$$

Maximizes the classifier agreement between the input and the explanation.

Optimization objective

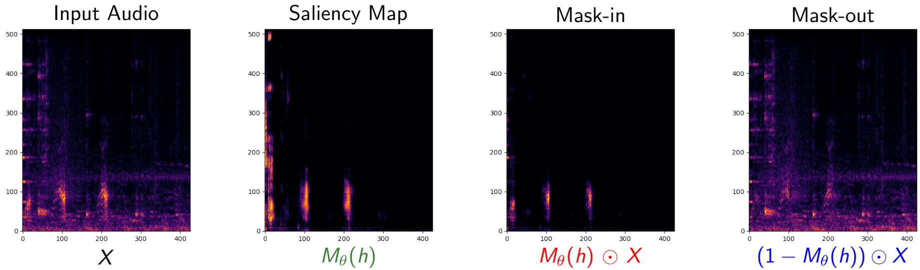


$$\min_\theta \overbrace{\lambda_{in}\mathcal{L}_{in}(\log f(M_\theta(h) \odot X), \widehat{y})}^{\text{Mask-in}}$$

$$- \overbrace{\lambda_{out}\mathcal{L}_{out}(\log f((1 - M_\theta(h)) \odot X), \widehat{y})}^{\text{Mask-out}}$$

Minimizes the classifier agreement of what is not in the explanation and the input.

Optimization objective

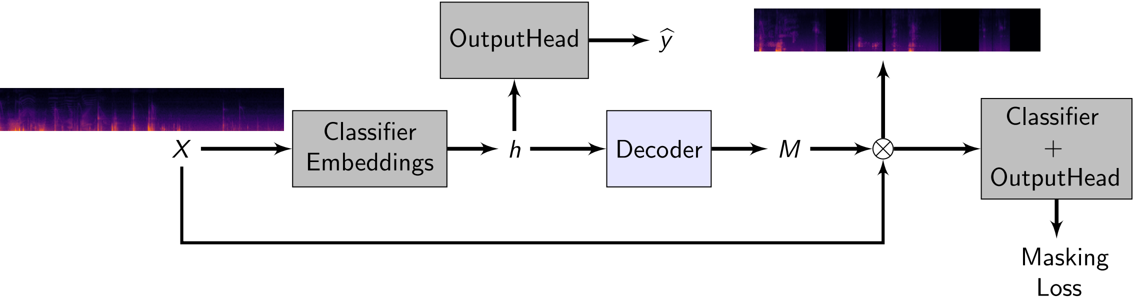


$$\min_{\theta} \overbrace{\lambda_{in}\mathcal{L}_{in}(\log f(M_\theta(h) \odot X), \widehat{y})}^{\text{Mask-in}}$$

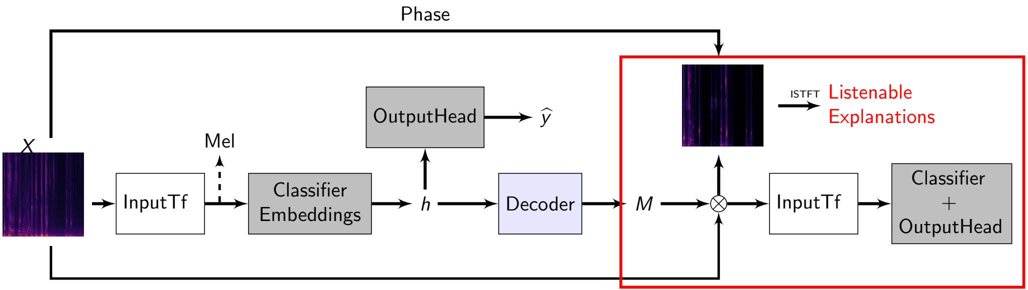
$$- \overbrace{\lambda_{out}\mathcal{L}_{out}(\log f((1 - M_\theta(h)) \odot X), \widehat{y})}^{\text{Mask-out}} + \overbrace{|M_\theta(h)|}^{\text{Mask Reg}}$$

Avoids trivial solutions.

Producing Listenable Explanations



Producing Listenable Explanations



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

Measuring faithfulness and understandability

- **Faithfulness**: Measures importance of explanations for classifier decisions
  - L2I-Faithfulness

$$\mathrm{FF}_n = p_{\widehat{c}}(X_n) - p_{\widehat{c}}(X_n - (X_n \odot M)),$$

Measuring faithfulness and understandability

- **Faithfulness**: Measures importance of explanations for classifier decisions
  - L2I-Faithfulness

$$\mathrm{FF}_n = p_{\widehat{c}}(X_n) - p_{\widehat{c}}(X_n - (X_n \odot M)),$$

  - Average-Increase

$$\mathrm{AI} = \frac{1}{N}\sum_{n=1}^{N}[p_{\widehat{c}}(X_n \odot M) > p_{\widehat{c}}(X_n)] \cdot 100,$$

Measuring faithfulness and understandability

- **Faithfulness**: Measures importance of explanations for classifier decisions
  - L2I-Faithfulness

$$\mathrm{FF}_n = p_{\widehat{c}}(X_n) - p_{\widehat{c}}(X_n - (X_n \odot M)),$$

  - Average-Increase

$$\mathrm{AI} = \frac{1}{N}\sum_{n=1}^{N}[p_{\widehat{c}}(X_n \odot M) > p_{\widehat{c}}(X_n)] \cdot 100,$$

  - Average-Gain

$$\mathrm{AG} = \frac{1}{N}\sum_{n=1}^{N}\frac{\max(0, p_{\widehat{c}}(X_n \odot M) - p_{\widehat{c}}(X_n))}{1 - p_{\widehat{c}}(X_n)} \cdot 100.$$

Measuring faithfulness and understandability

- **Faithfulness**: Measures importance of explanations for classifier decisions
  - L2I-Faithfulness

$$\mathrm{FF}_n = p_{\widehat{c}}(X_n) - p_{\widehat{c}}(X_n - (X_n \odot M)),$$

  - Average-Increase

$$\mathrm{AI} = \frac{1}{N}\sum_{n=1}^{N}[p_{\widehat{c}}(X_n \odot M) > p_{\widehat{c}}(X_n)] \cdot 100,$$

  - Average-Gain

$$\mathrm{AG} = \frac{1}{N}\sum_{n=1}^{N}\frac{\max(0, p_{\widehat{c}}(X_n \odot M) - p_{\widehat{c}}(X_n))}{1 - p_{\widehat{c}}(X_n)} \cdot 100.$$

  - Average-Drop

$$\mathrm{AD} = \frac{1}{N}\sum_{n=1}^{N}\frac{\max(0, p_{\widehat{c}}(X_n) - p_{\widehat{c}}(X_n \odot M))}{p_{\widehat{c}}(X_n)} \cdot 100.$$

Measuring faithfulness and understandability

- **Faithfulness**: Measures importance of explanations for classifier decisions
  - L2I-Faithfulness

$$\mathrm{FF}_n = p_{\widehat{c}}(X_n) - p_{\widehat{c}}(X_n - (X_n \odot M)),$$

  - Average-Increase

$$\mathrm{AI} = \frac{1}{N}\sum_{n=1}^{N}[p_{\widehat{c}}(X_n \odot M) > p_{\widehat{c}}(X_n)] \cdot 100,$$

  - Average-Gain

$$\mathrm{AG} = \frac{1}{N}\sum_{n=1}^{N}\frac{\max(0, p_{\widehat{c}}(X_n \odot M) - p_{\widehat{c}}(X_n))}{1 - p_{\widehat{c}}(X_n)} \cdot 100.$$

  - Average-Drop

$$\mathrm{AD} = \frac{1}{N}\sum_{n=1}^{N}\frac{\max(0, p_{\widehat{c}}(X_n) - p_{\widehat{c}}(X_n \odot M))}{p_{\widehat{c}}(X_n)} \cdot 100.$$

  - Input Fidelity

$$\text{Fid-In} = \frac{1}{N}\sum_{n=1}^{N}[\arg\max_{c} p_c(X_n) = \arg\max_{c'} p_{c'}(X_n \odot M)].$$

Measuring faithfulness and understandability

- **Faithfulness**: Measures importance of explanations for classifier decisions
  - L2I-Faithfulness

$$\mathrm{FF}_n = p_{\widehat{c}}(X_n) - p_{\widehat{c}}(X_n - (X_n \odot M)),$$

  - Average-Increase

$$\mathrm{AI} = \frac{1}{N}\sum_{n=1}^{N}[p_{\widehat{c}}(X_n \odot M) > p_{\widehat{c}}(X_n)] \cdot 100,$$

  - Average-Gain

$$\mathrm{AG} = \frac{1}{N}\sum_{n=1}^{N}\frac{\max(0, p_{\widehat{c}}(X_n \odot M) - p_{\widehat{c}}(X_n))}{1 - p_{\widehat{c}}(X_n)} \cdot 100.$$

  - Average-Drop

$$\mathrm{AD} = \frac{1}{N}\sum_{n=1}^{N}\frac{\max(0, p_{\widehat{c}}(X_n) - p_{\widehat{c}}(X_n \odot M))}{p_{\widehat{c}}(X_n)} \cdot 100.$$

  - Input Fidelity

$$\text{Fid-In} = \frac{1}{N}\sum_{n=1}^{N}[\arg\max_{c} p_c(X_n) = \arg\max_{c'} p_{c'}(X_n \odot M)].$$

Understandability

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

Understandability

$$\min_{\theta} \lambda_{in}\mathcal{L}_{in}(\log f(M_\theta(h) \odot X), \widehat{y}) - \lambda_{out}\mathcal{L}_{out}(\log f((1 - M_\theta(h)) \odot X), \widehat{y})$$
$$+ \lambda_s \underbrace{\|M_\theta(h)\|_1}_{L_1} + \lambda_g \underbrace{\|M_\theta(h) \odot X - X_{clean}\|}_{Finetuning}$$

- L_1: Avoids trivial solutions (e.g. all 1s).
- *Finetuning*: Improves Understandability.
  - Used in a second stage, selectively.

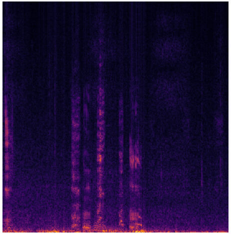
Understandability

$$\min_{\theta} \lambda_{in}\mathcal{L}_{in}(\log f(M_\theta(h) \odot X), \widehat{y}) - \lambda_{out}\mathcal{L}_{out}(\log f((1 - M_\theta(h)) \odot X), \widehat{y})$$

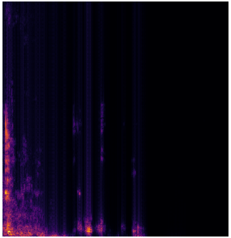
$$+ \lambda_s \underbrace{\|M_\theta(h)\|_1}_{L_1} + \lambda_g \underbrace{\|M_\theta(h) \odot X - X_{clean}\|}_{Finetuning}$$

- L_1: Avoids trivial solutions (e.g. all 1s).
- *Finetuning*: Improves Understandability.
  - Used in a second stage, selectively.

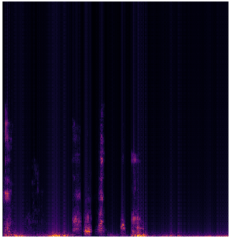
Input Audio



No finetuning



Finetuning



Experiments

- We produce explanations for classifiers trained on Sound Event Classification Datasets (**ESC50**, **US8k**).

Experiments

- We produce explanations for classifiers trained on Sound Event Classification Datasets (**ESC50**, **US8k**).
- We examine explanations on In-Domain (**ID**) and Out-of-Domain (**OOD**) cases.
  - ID: Plain datasets with data augmentation
  - OOD: Mixtures with different contaminating sources

Quantitative Results - ID

	Metric	AI (↑)	AD (↓)	AG (↑)	FF (↑)	Fid-In (↑)	SPS (↑)	COMP (↓)
Listenable (STFT→Mel)	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 (ours)	**36.25**	**1.15**	**23.50**	0.20	**0.42**	**0.47**	**4.71**
	L-MAC, FT, $\lambda_g = 4$ (ours)	32.37	1.98	18.74	**0.21**	0.41	0.43	5.20
Not Listenable (Mel)	Saliency	0.00	15.81	0.00	0.10	0.07	0.39	4.53
	Smoothgrad	0.00	15.61	0.00	0.07	0.04	0.39	4.54
	IG	0.00	15.55	0.00	0.12	0.08	0.42	4.36
	GradCAM	7.00	10.93	1.04	0.17	0.29	0.34	**4.72**
	Guided GradCAM	0.125	15.40	6.67	0.08	0.07	**0.45**	4.17
	Guided Backprop	0.125	15.54	0.00	0.10	0.08	0.39	4.53
	SHAP	0.00	15.57	0.00	0.11	0.08	0.41	4.42
	L-MAC (ours)	35.63	1.59	**24.28**	0.22	**0.42**	**0.45**	4.11
	L-MAC (ours) FT, $\lambda_g = 4$	**36.13**	**1.28**	21.15	**0.23**	**0.42**	0.32	4.71

Quantitative Results - ID

	Metric	AI (↑)	AD (↓)	AG (↑)	FF (↑)	Fid-In (↑)	SPS (↑)	COMP (↓)
Listenable (STFT→Mel)	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 (ours)	**36.25**	**1.15**	**23.50**	0.20	**0.42**	**0.47**	**4.71**
	L-MAC, FT, $\lambda_g = 4$ (ours)	32.37	1.98	18.74	**0.21**	0.41	0.43	5.20
Not Listenable (Mel)	Saliency	0.00	15.81	0.00	0.10	0.07	0.39	4.53
	Smoothgrad	0.00	15.61	0.00	0.07	0.04	0.39	4.54
	IG	0.00	15.55	0.00	0.12	0.08	0.42	4.36
	GradCAM	7.00	10.93	1.04	0.17	0.29	0.34	**4.72**
	Guided GradCAM	0.125	15.40	6.67	0.08	0.07	**0.45**	4.17
	Guided Backprop	0.125	15.54	0.00	0.10	0.08	0.39	4.53
	SHAP	0.00	15.57	0.00	0.11	0.08	0.41	4.42
	L-MAC (ours)	35.63	1.59	**24.28**	0.22	**0.42**	**0.45**	4.11
	L-MAC (ours) FT, $\lambda_g = 4$	**36.13**	**1.28**	21.15	**0.23**	**0.42**	0.32	4.71

Quantitative Results - ID

	Metric	AI (↑)	AD (↓)	AG (↑)	FF (↑)	Fid-In (↑)	SPS (↑)	COMP (↓)
Listenable (STFT→Mel)	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 (ours)	**36.25**	**1.15**	**23.50**	0.20	**0.42**	**0.47**	**4.71**
	L-MAC, FT, $\lambda_g = 4$ (ours)	32.37	1.98	18.74	**0.21**	0.41	0.43	5.20
Not Listenable (Mel)	Saliency	0.00	15.81	0.00	0.10	0.07	0.39	4.53
	Smoothgrad	0.00	15.61	0.00	0.07	0.04	0.39	4.54
	IG	0.00	15.55	0.00	0.12	0.08	0.42	4.36
	GradCAM	7.00	10.93	1.04	0.17	0.29	0.34	**4.72**
	Guided GradCAM	0.125	15.40	6.67	0.08	0.07	**0.45**	4.17
	Guided Backprop	0.125	15.54	0.00	0.10	0.08	0.39	4.53
	SHAP	0.00	15.57	0.00	0.11	0.08	0.41	4.42
	L-MAC (ours)	35.63	1.59	**24.28**	0.22	**0.42**	**0.45**	4.11
	L-MAC (ours) FT, $\lambda_g = 4$	**36.13**	**1.28**	21.15	**0.23**	**0.42**	0.32	4.71

- Finetuning does not harm faithfulness significantly.
- Generating listenable explanations does not decrease the alignment with the classifier.

Quantitative Results - ID

	Metric	AI (↑)	AD (↓)	AG (↑)	FF (↑)	Fid-In (↑)	SPS (↑)	COMP (↓)
Listenable (STFT→Mel)	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 (ours)	**36.25**	**1.15**	**23.50**	0.20	**0.42**	**0.47**	**4.71**
	L-MAC, FT, $\lambda_g = 4$ (ours)	32.37	1.98	18.74	**0.21**	0.41	0.43	5.20
Not Listenable (Mel)	Saliency	0.00	15.81	0.00	0.10	0.07	0.39	4.53
	Smoothgrad	0.00	15.61	0.00	0.07	0.04	0.39	4.54
	IG	0.00	15.55	0.00	0.12	0.08	0.42	4.36
	GradCAM	7.00	10.93	1.04	0.17	0.29	0.34	**4.72**
	Guided GradCAM	0.125	15.40	6.67	0.08	0.07	**0.45**	4.17
	Guided Backprop	0.125	15.54	0.00	0.10	0.08	0.39	4.53
	SHAP	0.00	15.57	0.00	0.11	0.08	0.41	4.42
	L-MAC (ours)	35.63	1.59	**24.28**	0.22	**0.42**	**0.45**	4.11
	L-MAC (ours) FT, $\lambda_g = 4$	**36.13**	**1.28**	21.15	**0.23**	**0.42**	0.32	4.71

- Finetuning does not harm faithfulness significantly.
- Generating listenable explanations does not decrease the alignment with the classifier.
- We have comparable structural metrics.

Quantitative Results - OOD (Audio Mixtures)

	Metric	AI (↑)	AD (↓)	AG (↑)	FF (↑)	Fid-In (↑)	SPS (↑)	COMP (↓)
Listenable (STFT→Mel)	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 (ours)	**60.63**	**4.82**	**35.85**	**0.39**	**0.81**	**0.94**	**9.61**
	L-MAC FT, $\lambda_g = 4$ (ours)	50.75	6.73	26.00	**0.39**	0.78	0.84	10.51
Not Listenable (Mel)	Saliency	0.38	31.64	0.01	0.15	0.12	0.77	9.17
	Smoothgrad	0.25	31.66	0.01	0.14	0.11	0.79	9.03
	IG	0.12	31.52	0.01	0.19	0.19	0.84	8.62
	GradCAM	19.88	18.85	4.67	0.34	0.69	0.66	9.49
	Guided GradCAM	0.00	31.68	0	0.14	0.12	0.89	10.24
	Guided Backprop	0.38	31.64	0.01	0.15	0.12	0.77	9.16
	SHAP	0.25	31.60	0.00	0.17	0.15	0.82	8.81
	L-MAC (ours)	60.25	**4.84**	**34.72**	**0.44**	0.80	**0.90**	**8.29**
	L-MAC - FT, $\lambda_g = 4$ (ours)	**60.75**	**4.84**	29.34	**0.44**	**0.83**	0.64	9.38

Quantitative Results - OOD (Audio Mixtures)

	Metric	AI (↑)	AD (↓)	AG (↑)	FF (↑)	Fid-In (↑)	SPS (↑)	COMP (↓)
Listenable (STFT→Mel)	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 (ours)	**60.63**	**4.82**	**35.85**	**0.39**	**0.81**	**0.94**	**9.61**
	L-MAC FT, $\lambda_g = 4$ (ours)	50.75	6.73	26.00	**0.39**	0.78	0.84	10.51
Not Listenable (Mel)	Saliency	0.38	31.64	0.01	0.15	0.12	0.77	9.17
	Smoothgrad	0.25	31.66	0.01	0.14	0.11	0.79	9.03
	IG	0.12	31.52	0.01	0.19	0.19	0.84	8.62
	GradCAM	19.88	18.85	4.67	0.34	0.69	0.66	9.49
	Guided GradCAM	0.00	31.68	0	0.14	0.12	0.89	10.24
	Guided Backprop	0.38	31.64	0.01	0.15	0.12	0.77	9.16
	SHAP	0.25	31.60	0.00	0.17	0.15	0.82	8.81
	L-MAC (ours)	60.25	**4.84**	**34.72**	**0.44**	0.80	**0.90**	**8.29**
	L-MAC - FT, $\lambda_g = 4$ (ours)	**60.75**	**4.84**	29.34	**0.44**	**0.83**	0.64	9.38

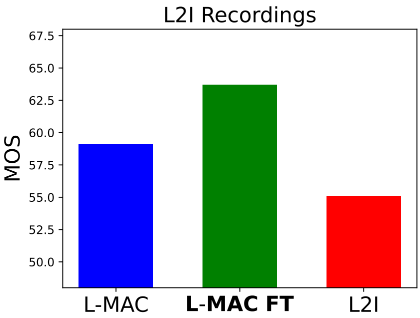
We observe the same outcome on US8k as well! (See the appendix)

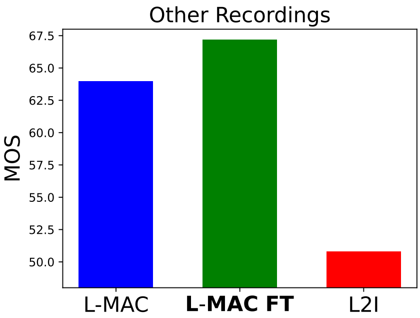
User Study

1. How well does the explanation correspond to the part of the input audio associated with the given class?
2. While evaluating, please pay attention to audio quality also.

User Study

1. How well does the explanation correspond to the part of the input audio associated with the given class?
2. While evaluating, please pay attention to audio quality also.





- **Recording 1:**
  - **L-MAC**
  - **L2I [NeurIPS'22]**
- **OOD (Speech):**
  - **L-MAC**
  - **L2I [NeurIPS'22]**

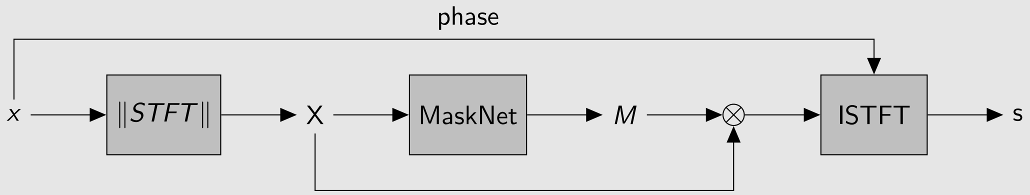
Conclusions

- We proposed a SOTA **posthoc explanation** method for audio classifers.
- Our method is agnostic to classifier input representation.
- Our method provides **understandable**, **listenable** and **faithful** explanations both in ID and OOD cases.
- Our code is available in SpeechBrain.

Table of Contents

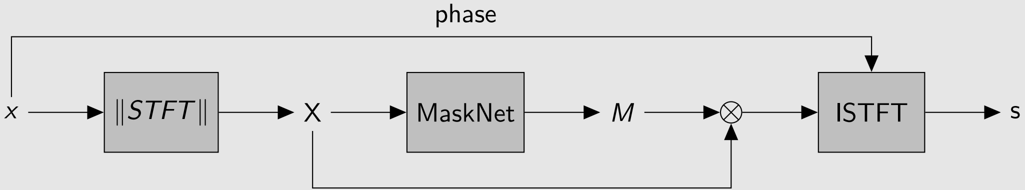
Masking in Frequency Domain vs Learnt Domain

Classical Frequency Domain Magnitude Masking

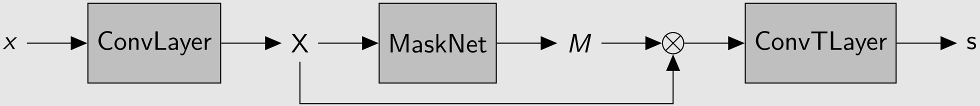


Masking in Frequency Domain vs Learnt Domain

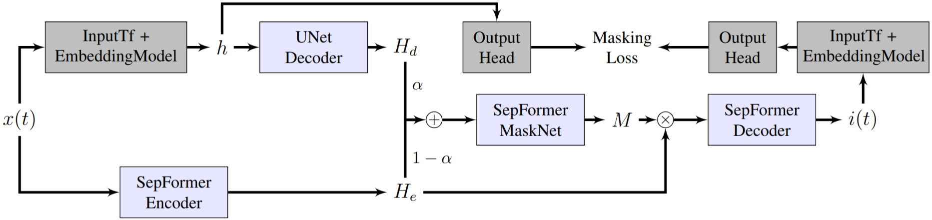
Classical Frequency Domain Magnitude Masking

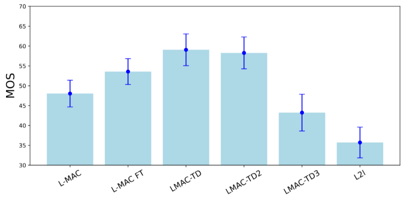


Learnable Domain Masking



L-MAC in Time Domain





Metric	AI (↑)	AD (↓)	AG (↑)	FF (↑)	Fid-In (↑)	SPS (↑)	COMP (↓)
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	36.25	**1.15**	23.50	0.20	0.42	0.47	**4.71**
L-MAC, FT, $\lambda_g = 4$	32.37	1.98	18.74	0.21	0.41	0.43	5.20
LMAC-TD, $\alpha = 1.00$ (ours)	66.00	2.62	22.39	**0.42**	0.87	**0.86**	10.50
LMAC-TD, $\alpha = 0.75$ (ours)	**69.75**	2.10	**28.07**	**0.42**	**0.91**	**0.86**	10.53
LMAC-TD, $\alpha = 0.00$ (ours)	46.50	5.55	11.86	**0.42**	0.86	0.80	10.88

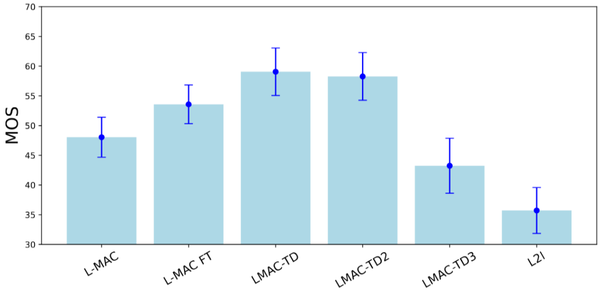
[E. Mancini, F. Paissan, M. Ravanelli, C. Subakan; Submitted to ICASSP 2025]

User Study

1. How well does the explanation correspond to the part of the input audio associated with the given class?
2. While evaluating, please pay attention to audio quality also.

User Study

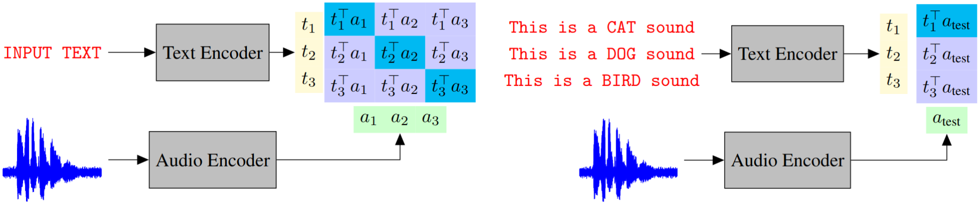
1. How well does the explanation correspond to the part of the input audio associated with the given class?
2. While evaluating, please pay attention to audio quality also.



- **Recording 1:**
  - **LMAC-TD**
  - **L-MAC**
  - **L2I [NeurIPS'22]**
- **Recording 2:**
  - **LMAC-TD**
  - **L-MAC**
  - **L2I [NeurIPS'22]**

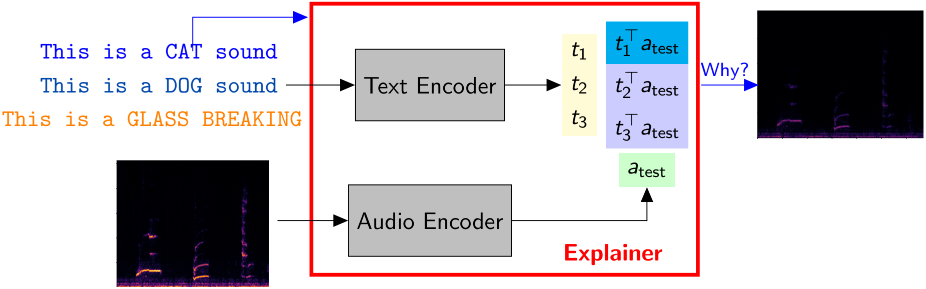
Table of Contents

Text-audio foundation models

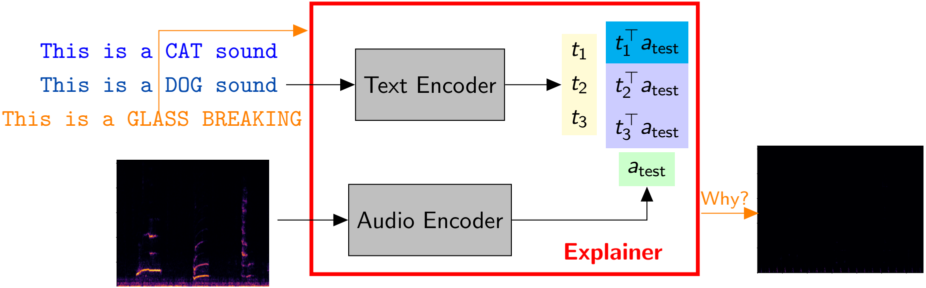


[Elizalde et al., CLAP: Learning Audio Concepts from Natural Language Supervision, ICASSP 2023]

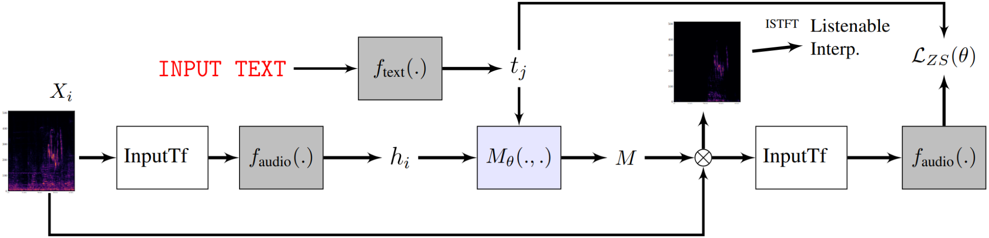
Listenable Maps for Zero-Shot Audio Classifiers



Listenable Maps for Zero-Shot Audio Classifiers



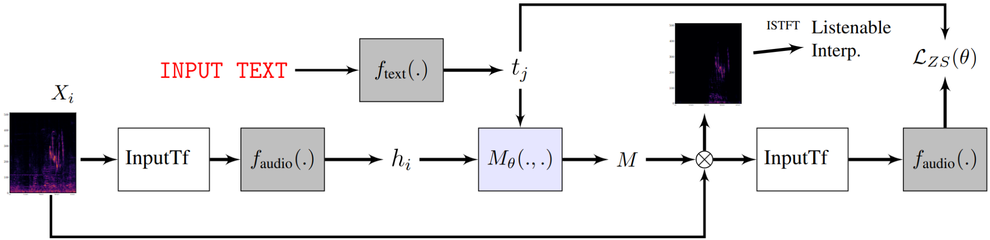
LMAC-ZS: Listenable Maps for Zero-Shot Audio Classifiers



- LMAC-ZS estimates **listenable** and **faithful** explanations for zero-shot audio classifiers.

[F. Paissan, L.D. Libera, M. Ravanelli, C. Subakan, NeurIPS 2024]

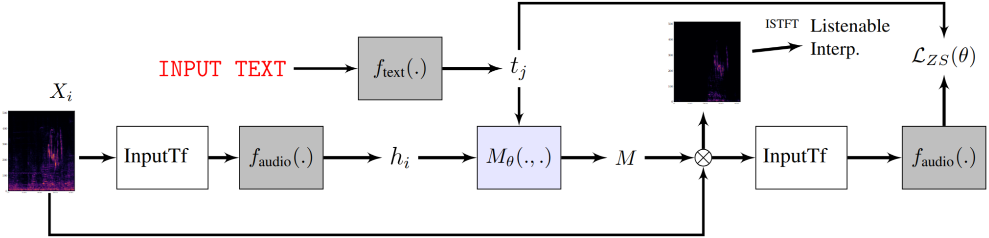
LMAC-ZS: Listenable Maps for Zero-Shot Audio Classifiers



- LMAC-ZS estimates **listenable** and **faithful** explanations for zero-shot audio classifiers.
  - **Challenge:** No classifier for faithfulness signal!

[F. Paissan, L.D. Libera, M. Ravanelli, C. Subakan, NeurIPS 2024]

LMAC-ZS: Listenable Maps for Zero-Shot Audio Classifiers



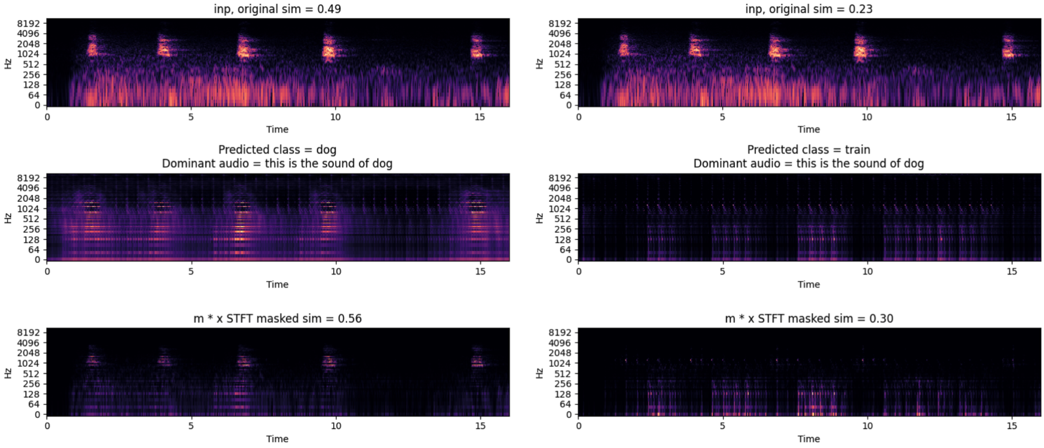
- LMAC-ZS estimates **listenable** and **faithful** explanations for zero-shot audio classifiers.
  - **Challenge:** No classifier for faithfulness signal!
- But we can measure cross-modal similarities:

$$\mathcal{L}_{\text{ZS}}(\theta) = \overbrace{\sum_{i,j} \left| C_{i,j} - t_i^\top f_{\text{audio}}\Big(M_\theta(t_i, h_j) \odot X_{\text{audio},j}\Big)\right|}^{\text{Similarity Matching}} + \overbrace{\lambda_1 \left\| M_\theta(t_i, h_j) \right\|_1}^{\text{Mask Regularization}} + \overbrace{\lambda_2 \sum_i D(X_{\text{audio},i})}^{\text{Prompt Diversity}}.$$

[F. Paissan, L.D. Libera, M. Ravanelli, C. Subakan, NeurIPS 2024]

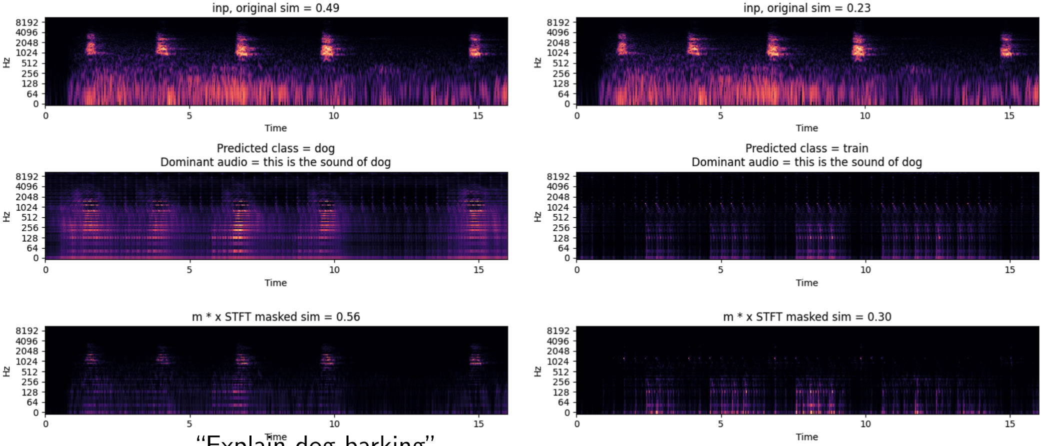
Qualitative Results

$$D(X_{\text{audio},i}) = \sum_{j;j\neq i} \left\| t_i^\top t_j - f_{\text{audio}}\left(X_{\text{audio},i} \odot M_\theta(t_i, h_i)\right)^\top f_{\text{audio}}\left(X_{\text{audio},i} \odot M_\theta(t_j, h_i)\right) \right\|.$$



Qualitative Results

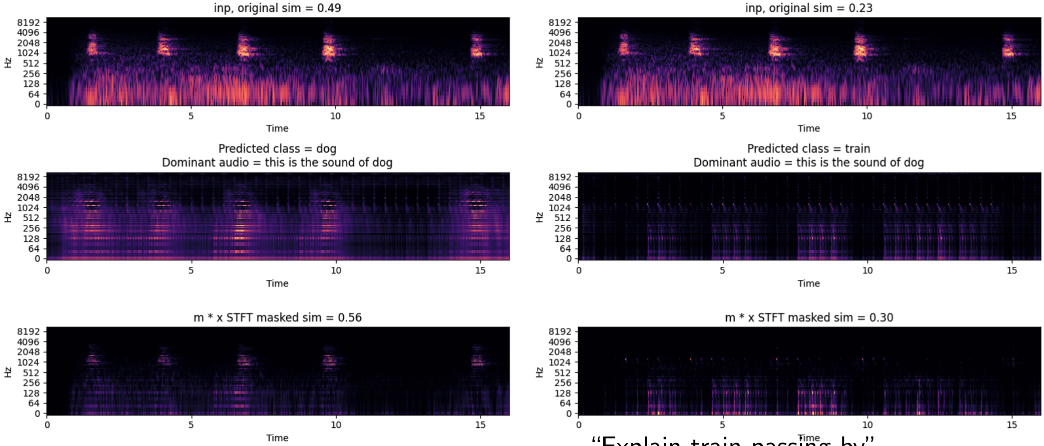
$$D(X_{\text{audio},i}) = \sum_{j;j\neq i} \left\| t_i^\top t_j - f_{\text{audio}}\left(X_{\text{audio},i} \odot M_\theta(t_i, h_i)\right)^\top f_{\text{audio}}\left(X_{\text{audio},i} \odot M_\theta(t_j, h_i)\right) \right\|.$$



"Explain dog barking"

Qualitative Results

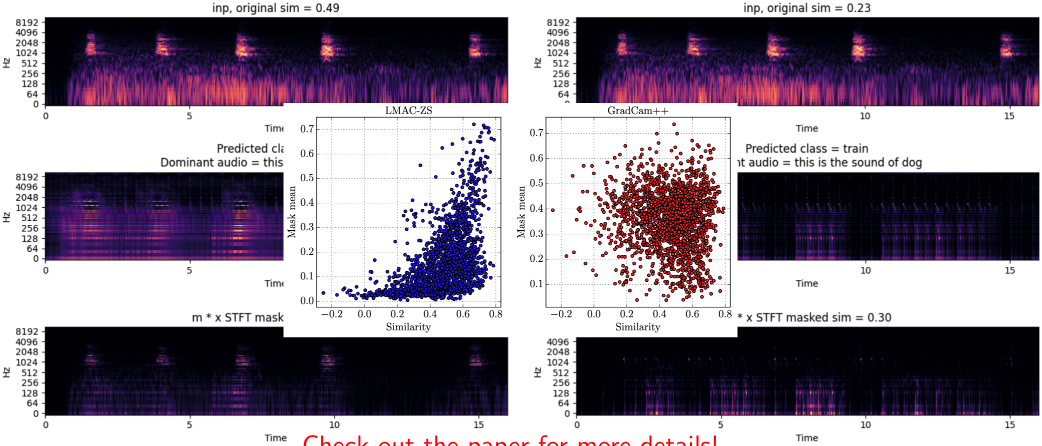
$$D(X_{\text{audio},i}) = \sum_{j;j\neq i} \left\| t_i^\top t_j - f_{\text{audio}}\left(X_{\text{audio},i} \odot M_\theta(t_i, h_i)\right)^\top f_{\text{audio}}\left(X_{\text{audio},i} \odot M_\theta(t_j, h_i)\right) \right\|.$$



“Explain train passing by”

Qualitative Results

$$D(X_{\text{audio},i}) = \sum_{j;j\neq i} \left\| t_i^\top t_j - f_{\text{audio}}\left(X_{\text{audio},i} \odot M_\theta(t_i, h_i)\right)^\top f_{\text{audio}}\left(X_{\text{audio},i} \odot M_\theta(t_j, h_i)\right) \right\|.$$



Check out the paper for more details!

Quantitative Results

Metric	AI (↑)	AD (↓)	AG (↑)	FF (↑)	Fid-In (↑)	SPS (↑)	COMP (↓)	MM
	ZS classification on ESC50, Mel-Masking, 80.7% accuracy							
Gradcam	2.90	45.85	1.01	0.28	0.19	0.71	9.52	0.15
GradCam++	8.45	35.07	3.19	0.50	0.39	0.41	10.32	0.35
SmoothGrad	0.50	52.76	0.12	0.024	0.036	0.301	10.52	0.039
IG	0.25	53.47	0.054	0.064	0.022	0.57	10.09	0.037
LMAC-ZS	**23.45**	**17.12**	**10.31**	**0.51**	**0.68**	**0.80**	**9.12**	0.17
	ZS classification on ESC50, STFT-Masking, 78.9% accuracy							
GradCam	20.30	23.75	7.77	0.78	0.58	0.72	**11.54**	0.14
GradCam++	32.50	8.97	7.95	0.79	0.84	0.41	12.41	0.35
SmoothGrad	6.95	32.75	2.85	0.78	0.47	0.53	11.98	0.0001
IG	16.10	21.51	6.05	**0.79**	0.65	**0.74**	11.58	0.0095
LMAC-ZS	**43.35**	**4.29**	**10.57**	0.78	**0.90**	0.65	11.86	0.1

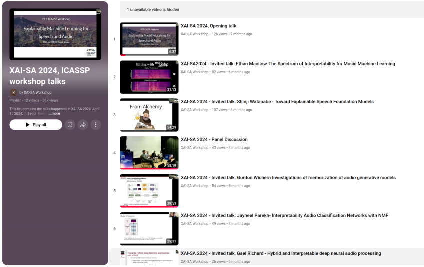
Conclusions

- First decoder-based explainability technique for zero-shot classifiers.
- Extensive faithfulness evaluation shows that LMAC-ZS aligns with CLAP predictions.
- The generated explanations are:
  - **Listenable**
  - **Faithful**
  - **Sensitive** to prompts.

Check out the code and audio samples

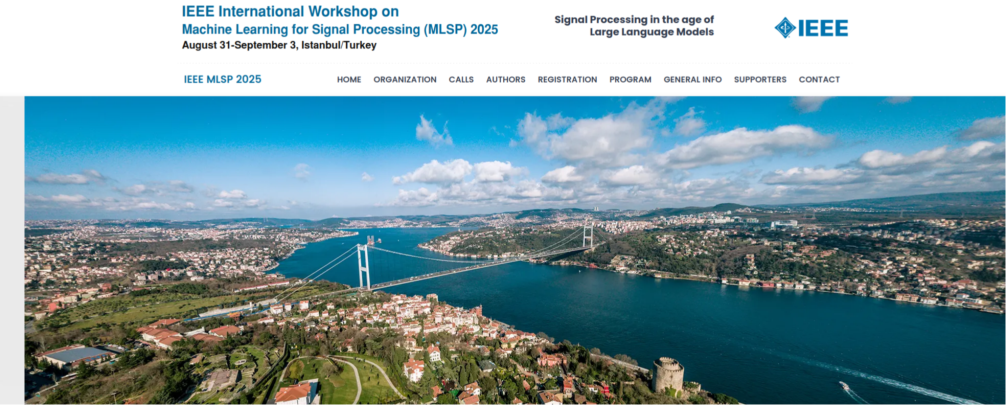
XAI-SA, ICASSP 2024 Workshop

- ICASSP 2024 Workshop, Explainable AI for Speech and Audio




IEEE MLSP 2025

- We are general chairing MLSP 2025!



- `2025.ieeemlsp.org`

Thanks for listening!