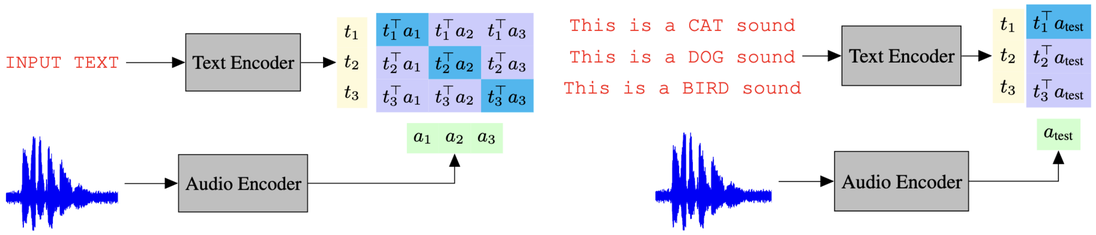
Listenable Maps for Zero-Shot Audio Classifiers

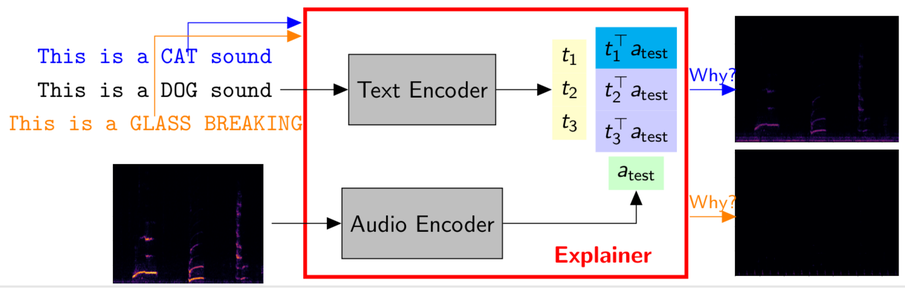
Francesco Paissan[1,2], Luca Della Libera[3,2] Mirco Ravanelli[3,2], Cem Subakan[4,3,2]

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

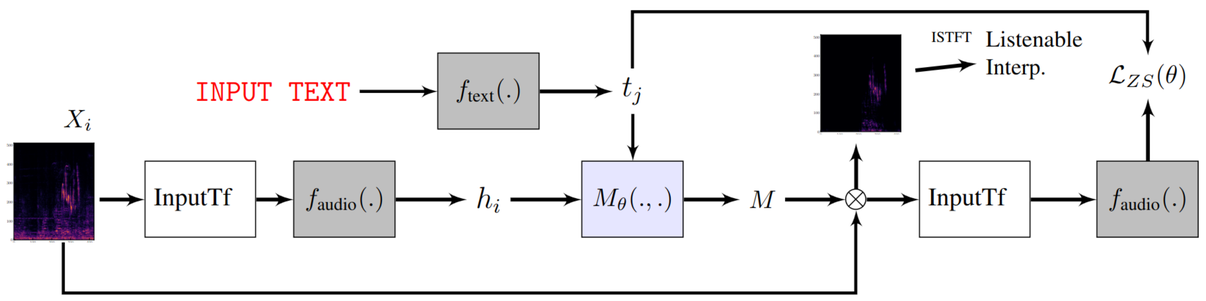
Text-Audio Representations



Explaining Zero-Shot Classifiers



Proposed Method



- **Desiderata:** *Faithful, Understandable/Listenable Explanations, Sensitivity to Prompts*
- **Challenge:** There is no faithfulness signal to train the masking network $M_\theta(\cdot)$.
- **The loss function:**

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

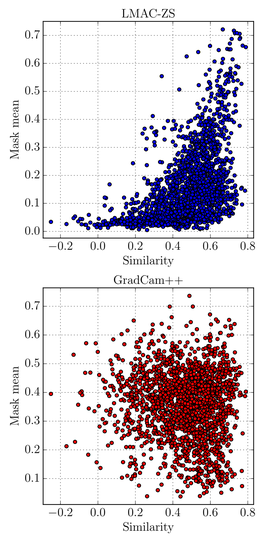
  - Similarity Matching: Text-audio similarities before and after masking.
  - Mask Regularization: Prevents trivial solutions through sparsity.
  - Prompt Diversity: Different text prompts should give different masks.

Experimental Results

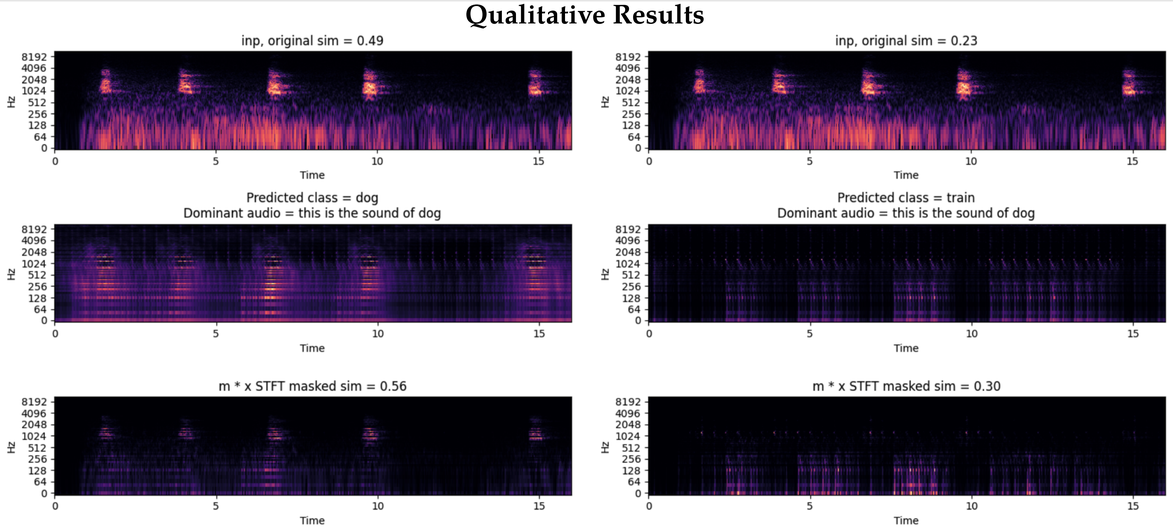
Quantitative Results

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
ScoreCAM	29.97	12.14	8.82	0.70	0.75	0.32	12.59	0.41
GScoreCAM	29.64	8.56	6.62	**0.79**	0.84	0.36	12.52	0.39
LMAC-ZS	**43.35**	**4.29**	**10.57**	0.78	**0.90**	0.65	11.86	0.1

In-Domain: ESC-50 and UrbanSound8K for In-Domain.
Out-of-Domain: Contamination with Speech, White Noise, Mixtures.



Qualitative Results



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**, and **Sensitive** to prompts.

Check out the code and audio samples!

SpeechBrain