Listen to Interpret: Post-hoc Interpretability for Audio Networks with NMF

Jayneel Parekh

LTCI, Télécom Paris, Institut Polytechnique de Paris

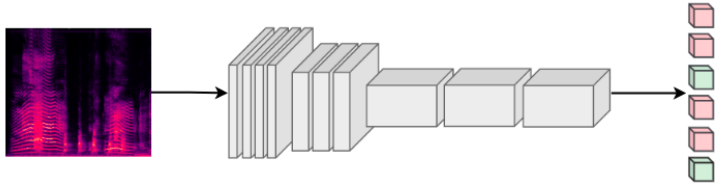
Joint work with Sanjeel Parekh, Pavlo Mozharovskyi, Florence d'Alché-Buc, Gaël Richard

NeurIPS 2022

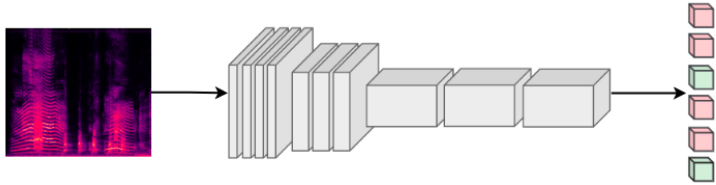
What is Interpretability?

What did the model base its decision on?

Central problem of our work



Central problem of our work



Goal: Post-hoc interpretation for the decisions of the given network

Requirements from Interpreter

Existing interpreters fail on at least one of the two points when applied on audio modality:

1. Effectively representing various audio objects composing the input when identifying relevant information for decision.

2. Providing listenable interpretations. Visual attribution maps over spectrograms are not understandable for most end-users!

Requirements from Interpreter

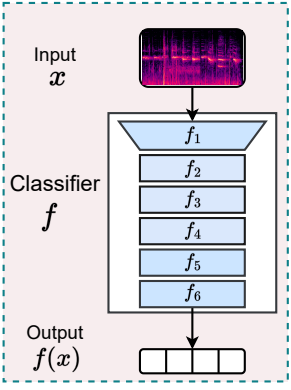
Existing interpreters fail on at least one of the two points when applied on audio modality:

1. Effectively representing various audio objects composing the input when identifying relevant information for decision.

2. Providing listenable interpretations. Visual attribution maps over spectrograms are not understandable for most end-users!

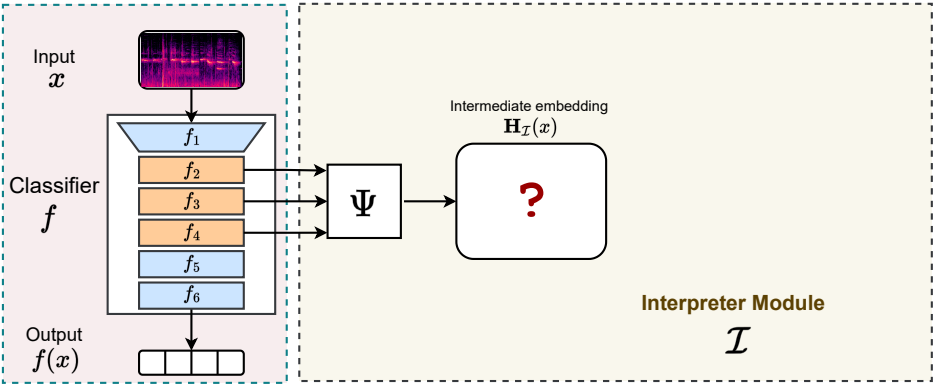
NOTE: Providing listenable interpretations is NOT the same as classical source separation or noise removal!

Design of Interpreter



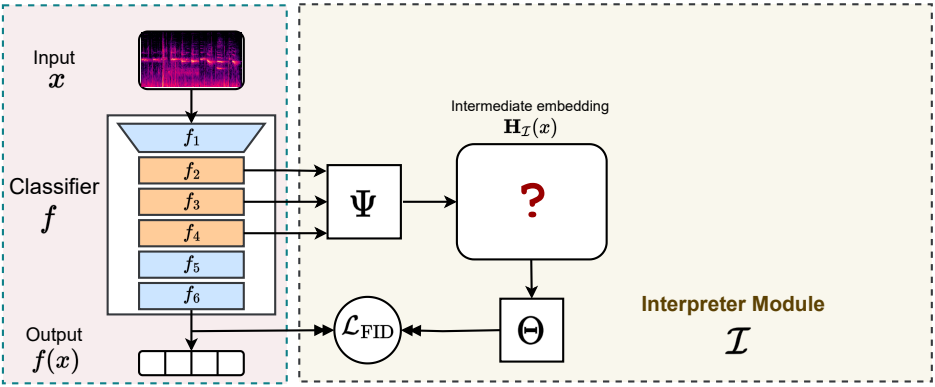
- f is the audio-processing deep network we wish to interpret.

Design of Interpreter



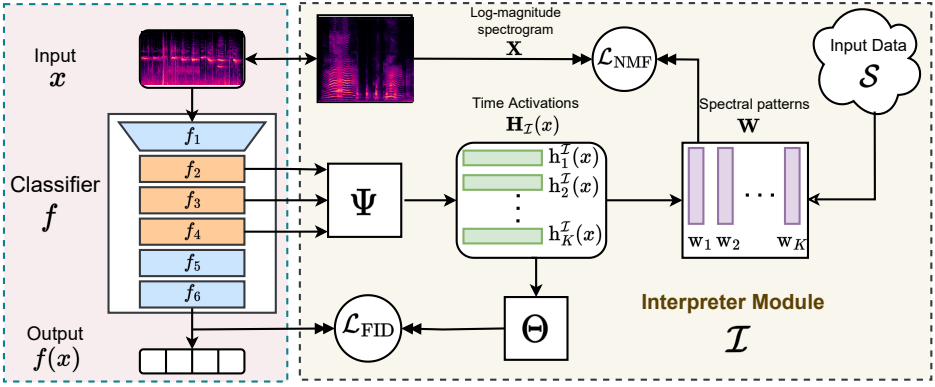
- Interpreter, implemented as neural network computes an intermediate encoding $\mathbf{H}_{\mathcal{I}}(x) = \Psi \circ f_{\mathcal{I}}(x)$, that helps interpret the decision $f(x)$ and fulfill the requirements.

Design of Interpreter: Mimicking the classifier



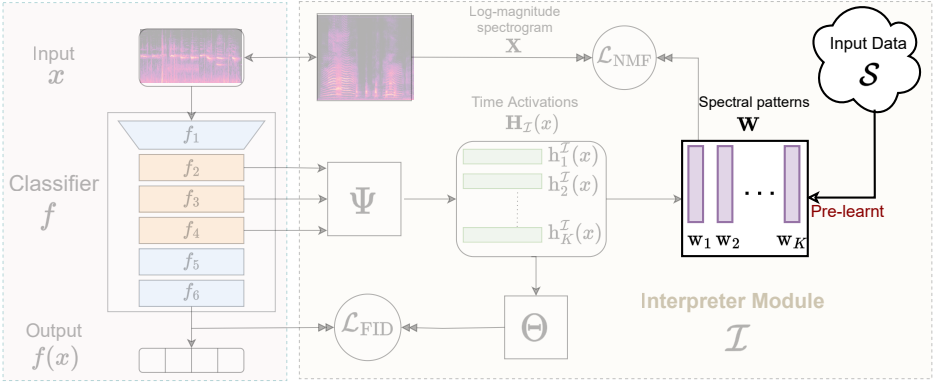
- The interpreter computes the output $\Theta \circ \mathbf{H}_{\mathcal{I}}(x)$ and mimics $f(x)$ through $\mathcal{L}_{\text{FID}}$. Shapes $\mathbf{H}_{\mathcal{I}}(x)$ to interpret classifier output.

Design of interpreter: Demystifying intermediate encoding



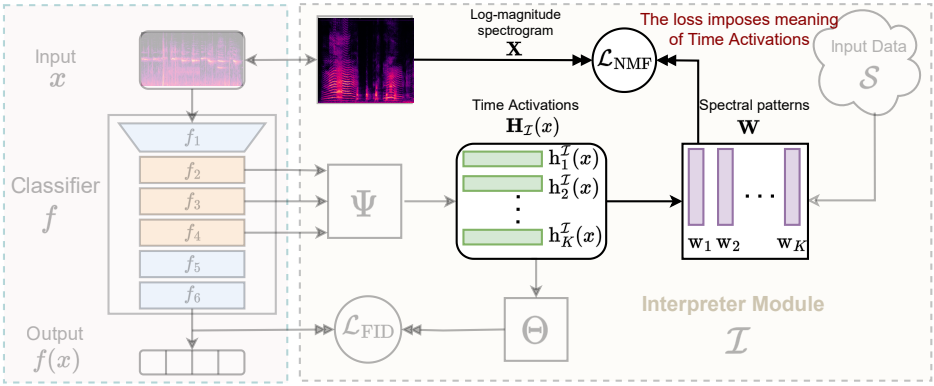
- $\mathbf{H}_{\mathcal{I}}(x) \in \mathbb{R}_+^{K \times T}$ is a non-negative matrix. We aim for it to encode presence of audio objects as activations across T time frames, for a dictionary of K spectral patterns $\mathbf{W}$.

Design of interpreter: Demystifying intermediate encoding



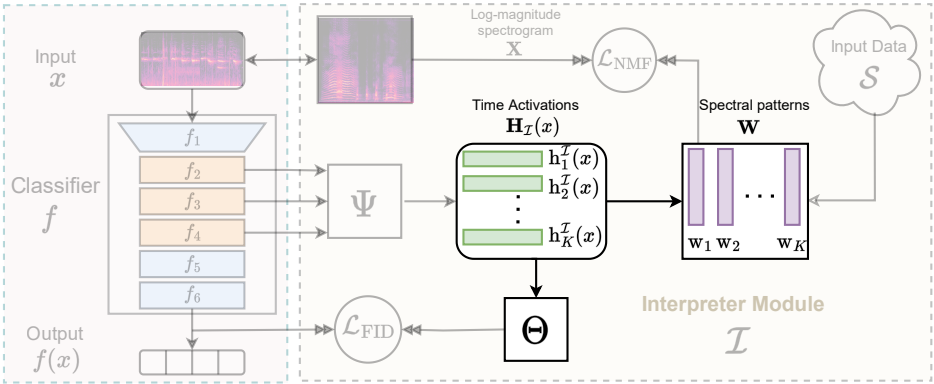
- The dictionary of spectral patterns $\mathbf{W}$ is a matrix, pre-learnt on the given dataset. Represents various audio objects/classes.

Design of interpreter: Demystifying intermediate encoding



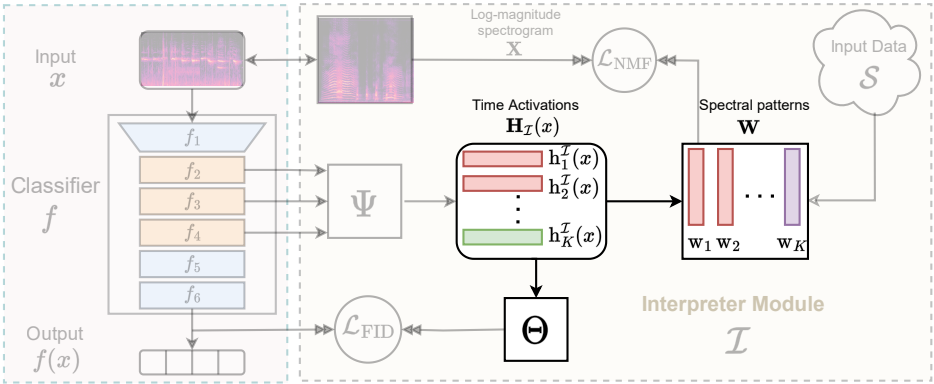
- Through $\mathcal{L}_{\text{NMF}}$ we require $\mathbf{H}_{\mathcal{I}}(x)$ to approximate input log-magnitude spectrogram as $\mathbf{X} \approx \mathbf{W}\mathbf{H}_{\mathcal{I}}(x)$, $\mathcal{L}_{\text{NMF}}(x, V_\Psi) = \|\mathbf{X} - \mathbf{W}\mathbf{H}_{\mathcal{I}}(x)\|_2^2$

Design of Interpreter: Generating listenable interpretations



- $\mathbf{W}, \mathbf{H}_{\mathcal{I}}(x)$ represent the audio objects in the input. $\Theta, \mathbf{H}_{\mathcal{I}}(x)$ identify relevant spectral patterns for decision $f(x)$. Input filtering for listenable interpretation.

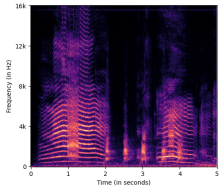
Design of Interpreter: Generating listenable interpretations



- $\mathbf{W}, \mathbf{H}_{\mathcal{I}}(x)$ represent the audio objects in the input. $\Theta, \mathbf{H}_{\mathcal{I}}(x)$ identify relevant spectral patterns for decision $f(x)$. Input filtering for listenable interpretation.

Example Interpretation

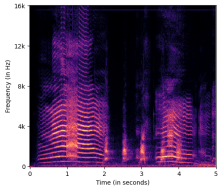
Input audio – Mix of 'CRYING-BABY' and 'DOG-BARKING'.



Classifier detects the presence of 'CRYING-BABY'

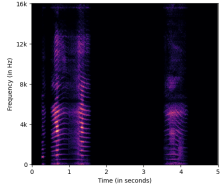
Example Interpretation

Input audio – Mix of 'CRYING-BABY' and 'DOG-BARKING'.



Classifier detects the presence of 'CRYING-BABY'

Interpretation audio for 'CRYING-BABY'.



Thank You!

Project webpage: https://jayneelparekh.github.io/listen2interpret/