AudioMNIST: Exploring Explainable Artificial Intelligence for audio analysis on a simple benchmark

Sören Becker , Johanna Vielhaben, Marcel Ackermann,
Klaus-Robert Müller, Sebastian Lapuschkin, Wojciech Samek

Presented by
Shamail M.,
MSc Artificial Intelligence,
Queen Mary University of London

Agenda

- ➢ Explainable Artificial Intelligence (XAI) for audio classification & LRP
- ➢ AudioMNIST: audio dataset for benchmarking
- ➢ Neural Network feature selection
- ➢ Visualisation & Audible heatmaps
- ➢ Audible explanations surpass visual for interpretability?

Audio Representations

❖ Raw waveform

- ➢ audio signal in the time domain
- ➢ represented by a waveform $x \in R^L$ - amplitude values x_t of the signal over time
- ➢ time steps between the signal values are determined by the sampling frequency f_s
- ➢ duration of the signal is L/f_s

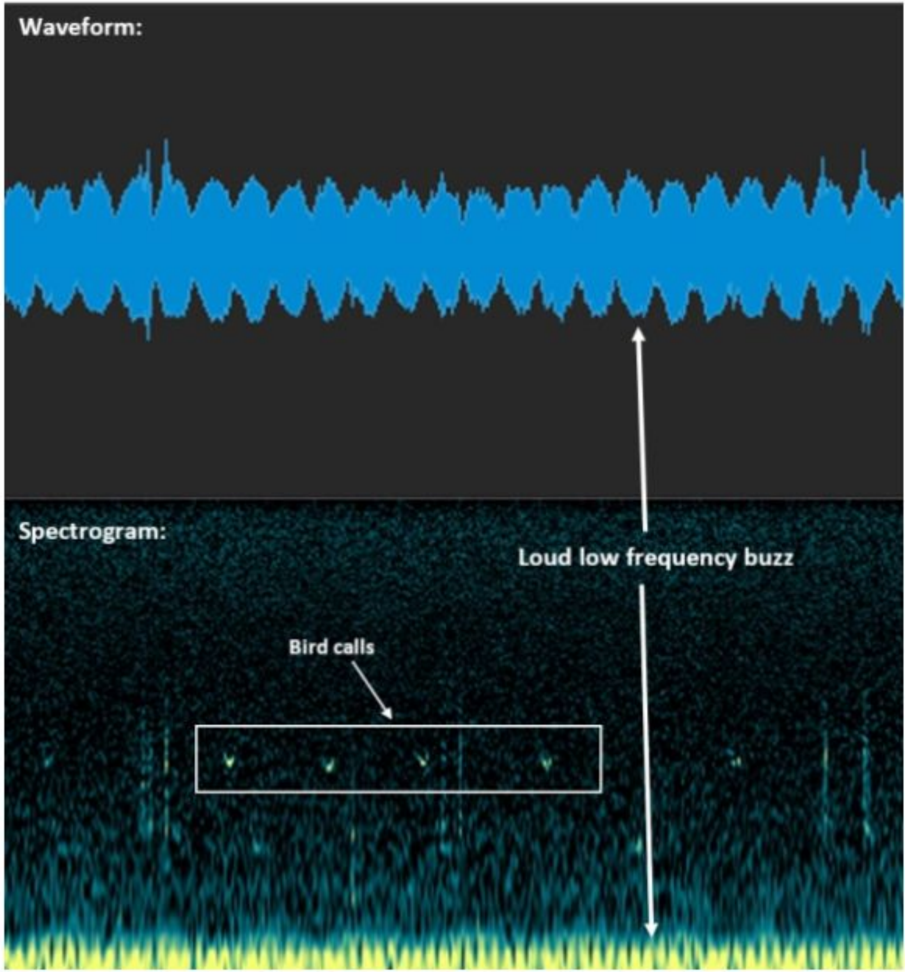
❖ Time-frequency spectrogram

- ➢ STDFT transforms the raw waveform x to its representation Y in time–frequency domain

$$Y_{k,m} = \sum_{n=0}^{N-1} x_{n+mH} \cdot w_n \cdot e^{-\frac{i\pi kn}{N}}$$

w: window function
M: length of window
H: hop size

- ➢ Allows for use of VGG & AlexNet architectures

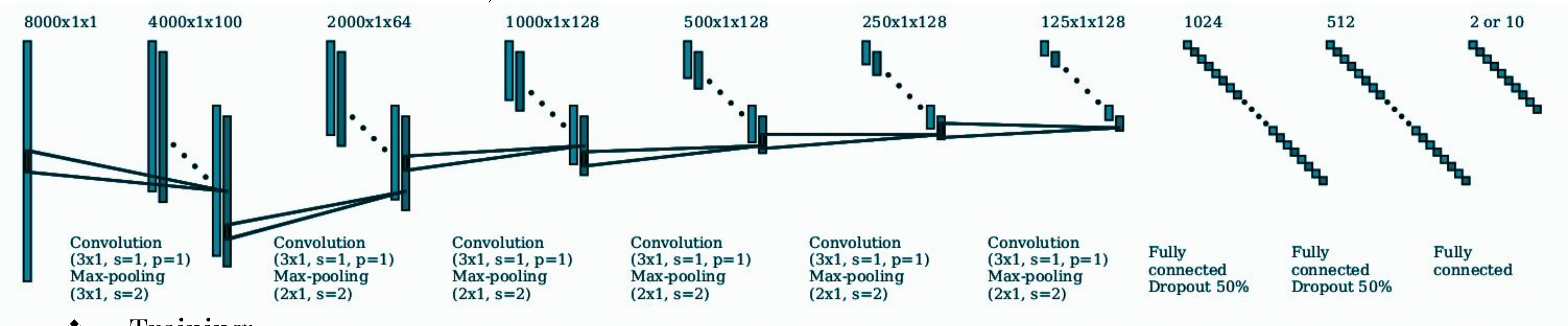


AudioMNIST & Classification

- Dataset (Becker et. al., 2018):
  - 30,000 audio recordings of English spoken digits (0-9) - 60 different speakers
  - Sampling frequency: 48 kHz
  - Meta info: age (22-61 years), gender (12 female & 48 male), origin & accent
  - Audio recordings resampled at 8kHz & zero padded
- Tasks:
  - Spoken digit recognition
  - Speaker's sex recognition
- Architecture:
  - Input: single feature map as an $(8000 \times 1 \times 1)$ tensor
  - 2 networks:
    - AudioNet - Waveform input
    - AlexNet - Spectrogram input
  - For convolution and max and pooling layers, stride is abbreviated with s and padding with `$p$`

AudioMNIST & Classification: AudioNet

- ❖ Input:
  - ➢ Dimension: (8000 × 1 × 1) tensor of raw audio data
  - ➢ Signal is normalized by the waveform’s 95th amplitude percentile (removes outliers - environmental noise)

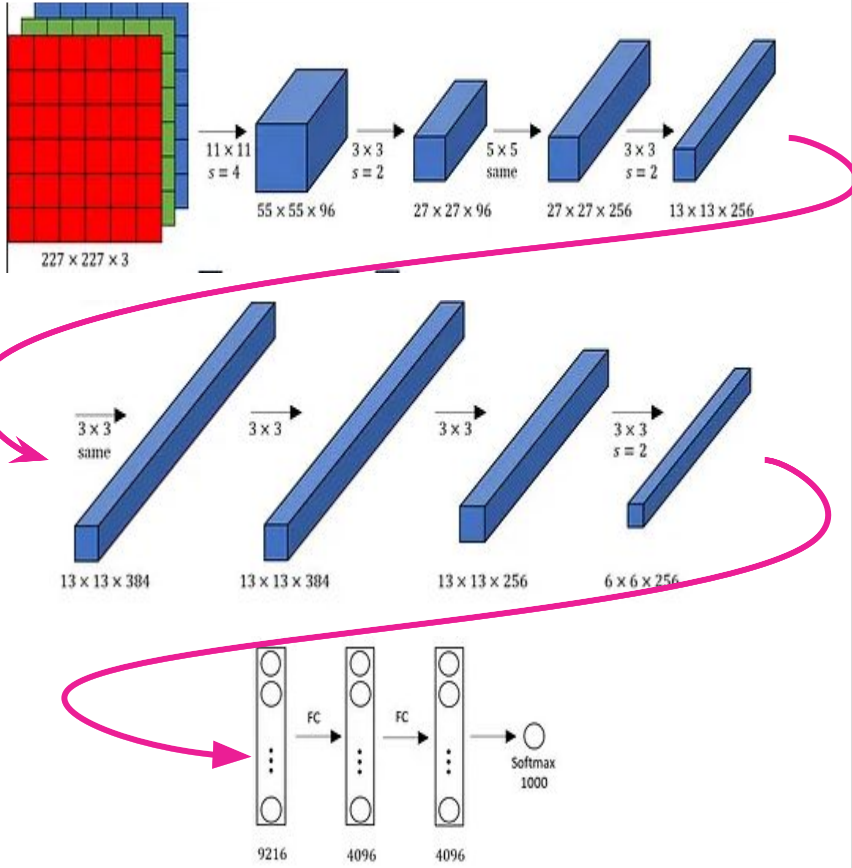


- ❖ Training:
  - ➢ trained with stochastic gradient descent
  - ➢ batch size: 100; epochs: 50000
  - ➢ Initial learning rate: 0.0001; lowered by a factor of 0.5
  - ➢ Digit classification:
    - ▪ Momentum: 0.9; learning rate was lowered every 10000 steps.
  - ➢ Gender classification:
    - ▪ 10000 epochs with the learning rate being reduced after 5000

AudioMNIST & Classification: AlexNet (Modified)

- Input:
  - STFT (Hann window width 455, 420 time points overlap): dimensions 228 × 230 => cropped to 227 × 227
  - Converted to dB
  - Audio augmentation done during zero padding
- AlexNet modification
  - Input channels: 1
  - Fully connected dimensions:
    - Digit recognition: 1024, 1024, 10
    - Gender classification: 1024, 1024, 2
  - Without normalization layers
- Training:
  - Digit Recognition:
    - 5 disjoint subsets of 6000 spectrograms each
    - 5-fold cross-validation used (3-1-1 split)
  - Gender classification:
    - Dataset: 12 female speakers & 12 randomly selected male speakers
    - four disjoint subsets of 3000 spectrograms each
    - 4-fold cross-validation used (2-1-1 split)
  - trained with stochastic gradient descent
  - batch size: 100 spectrograms; epochs: 10000
  - initial learning rate: 0.001; reduced by factor of 0.5 every 2500 epochs
  - Momentum: 0.9 throughout training
  - gradients were clipped at a magnitude of 5

Original AlexNet Architecture



AudioMNIST & Classification: Results

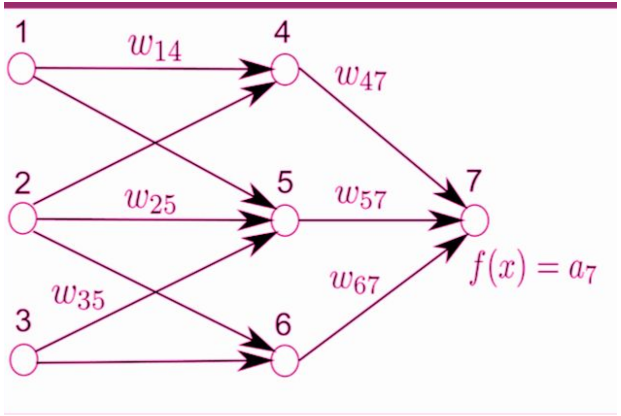
Mean accuracy ± standard deviation over data splits for AlexNet and AudioNet on the digit and sex classification tasks of AudioMNIST.

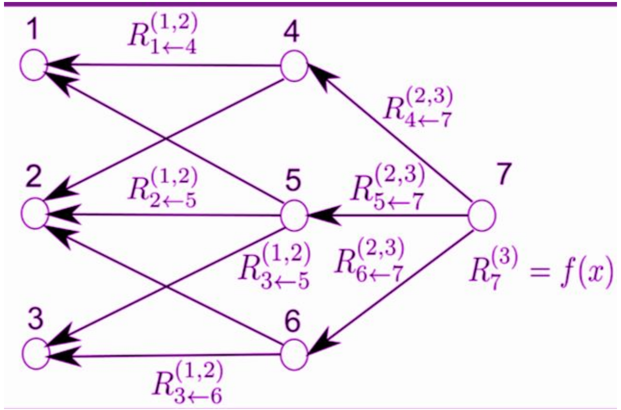
Model	Input	Task	
		Digit Classification	Sex Classification
AlexNet	spectrogram	95.82% ± 1.49%	95.87% ± 2.85%
AudioNet	waveform	92.53% ± 2.04%	91.74% ± 8.60%

Layer-wise Relevance Propagation: Post-hoc explainability

- ❖ Inspect features that impact prediction
- ❖ Starting with the output, LRP performs per-neuron decompositions and generates relevance scores R_i
- ❖ Use heatmap composed of relevance values
  - ➢ Neutral contribution: R=0
  - ➢ Positive contribution: Red colours
  - ➢ Negative Contribution: Blue colours
- ❖ Method:
  - ➢ redistribute the relevance value R_j of an upper layer neuron towards the layer inputs x_i
  - ➢ Uses pre-activation sent from input i to output j (z_{ij})
  - ➢ Score R_i at neuron i is pooled for all incoming relevance quantities $R_{i\leftarrow j}$

$$R_{i\leftarrow j} = \frac{z_{ij}}{\sum_i z_{ij}} R_j \qquad R_i = \sum_j R_{i\leftarrow j}$$

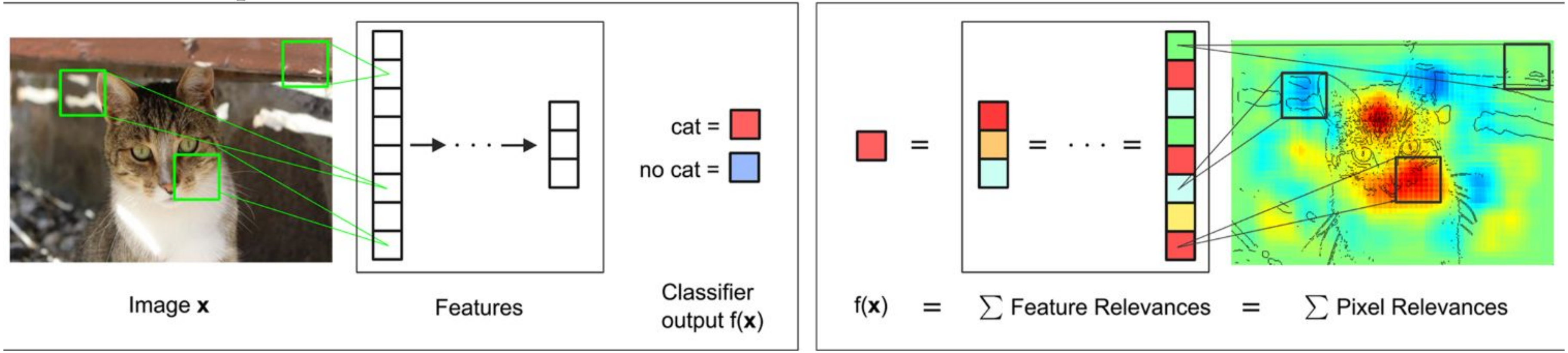




Note: initial relevance value equals the activation of the output neuron

LRP Applications

Pixel wise explanation:



Sentiment Analysis:

LRP Heatmap: this film does n't care about cleverness , wit or any other kind of intelligent humor .

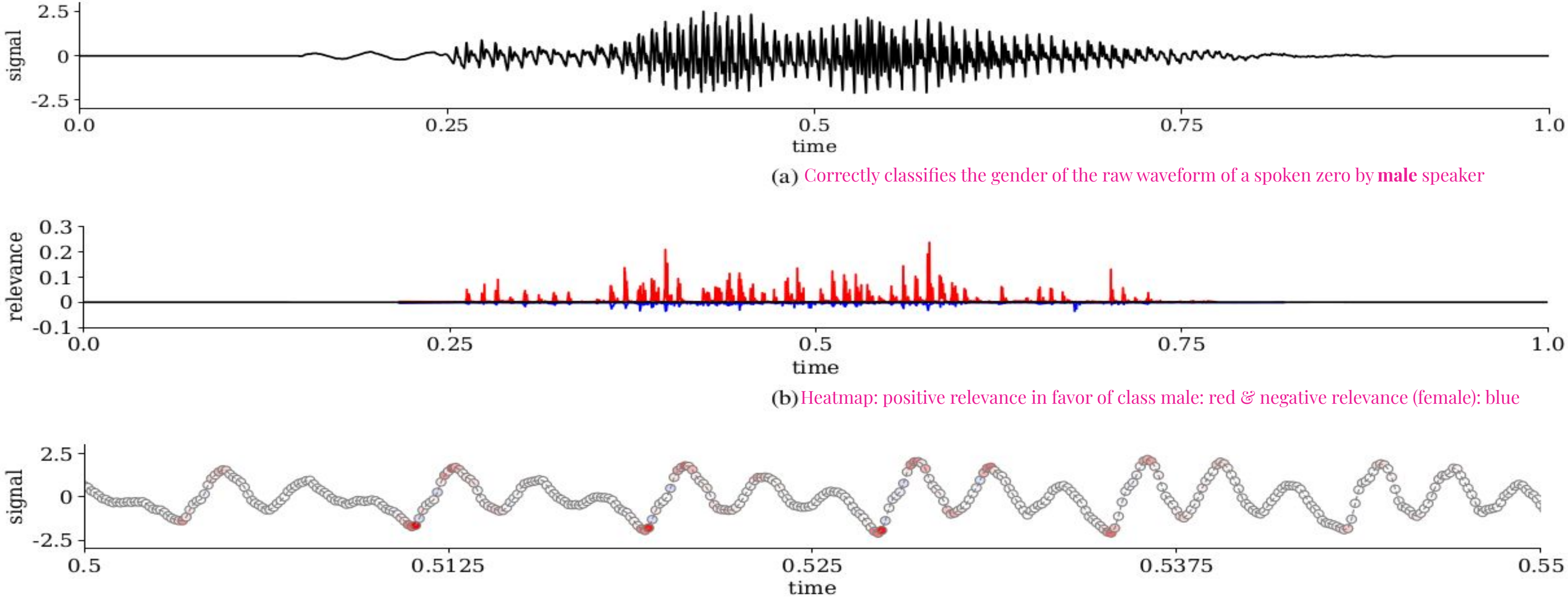
Predicted class: negative review

Scientific domains:

- Medical Imaging
- EEG Analysis
- Visualise brain activity

LRP for Audio: Visual Explanation (AudioNet)

- ❖ Relevance scores are obtained in form of an 8000 dimensional vector
- ❖ Overlay heatmap on raw waveform
- ❖ It appears that mainly samples of large magnitude are relevant for the network's classification decision



(a) Correctly classifies the gender of the raw waveform of a spoken zero by **male** speaker

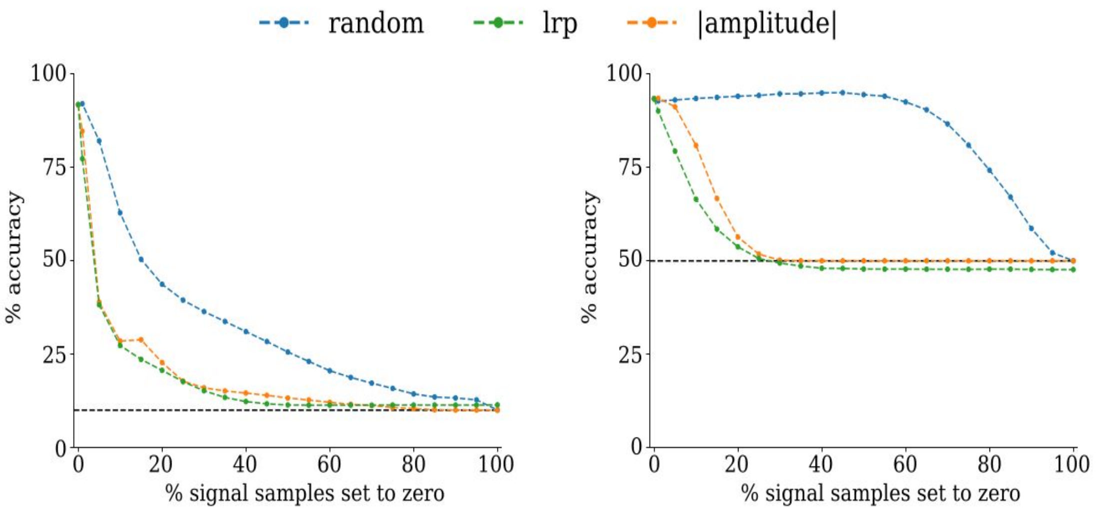
(b) Heatmap: positive relevance in favor of class male: red & negative relevance (female): blue

(c) Waveform from (a) is again visualized: samples colored according to their relevance

LRP for Audio: Feature Analysis

- Relevance-guided sample manipulation for raw waveform by pixel-flipping

- Strategies:
  - samples of the input signal are selected and flipped at random (baseline)
  - samples of the input are selected with respect to maximal absolute amplitude
    - e.g the 10% samples with the highest absolute amplitude are selected
  - samples are selected according to maximal relevance as attributed by LRP



(a) Digit Classification

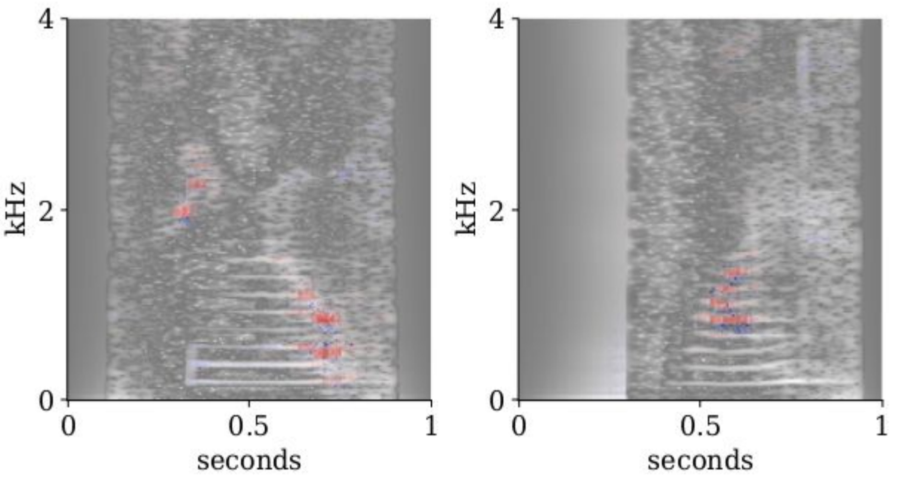
(b) Sex Classification

- Observations:
  - Decline in model performance in both the relevance-based and amplitude-based perturbation
  - the model seems to ground its inference in the high-amplitude parts of the signal

LRP for Audio: Visual Explanation (AlexNet)

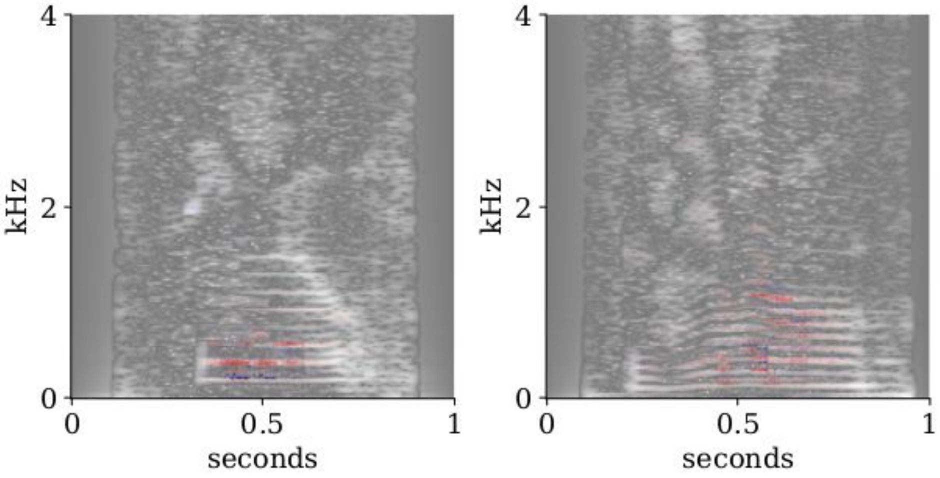
- ❖ Spectrogram: Similar to natural images
- ❖ May be hypothesized that sex classification is based on the fundamental frequency & subsequent harmonics
- ❖ Relevance maps overlayed on spectrogram

- ❖ Observations:
  - ➢ Most of the relevance distributed in the lower frequency range
  - ➢ This is known discriminant features for sex in speech
  - ➢ It is difficult to link the features to higher concepts such as for instance phonemes



(a) female speaker, *zero* (b) female speaker, *one*

Gender classification



(c) female speaker, *zero* (d) male speaker, *zero*

Digit classification

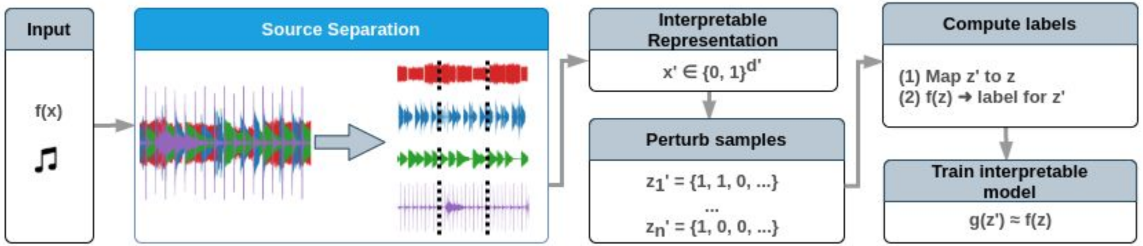
LRP for Audio: Feature Analysis

- Relevance-guided sample manipulation strategy:
  - Test set was manipulated by scaling the frequency-axis of the spectrograms:
    - Male: factor of 1.5
    - Female: factor of 0.66.
  - Manipulations match the original spectrograms of the opposite sex.
- An exact time domain signal for a modified spectrogram is not guaranteed to exist
  - approximation of the waveform corresponding to the manipulated spectrogram may be obtained via the inverse short-term Fourier transform
  - Manipulations within the thereby acquired audio signals are easily detectable for humans, as voices in the manipulated signal sound rather robotic

- Observations:
  - Accuracy of only 20.3%±12.6% on manipulated test splits
  - Identifying sex features via LRP allowed us to successfully perform transformations on the inputs that target the identified features with approximately 80% accuracy in predicting the opposite sex

LRP for Audio: Audible Explanation using AudioLIME

- LIME: Local Interpretable Model-agnostic Explanations
  - quantify relevance of components
  - explanation model = linear regression with L2 regularization
  - for MIR tasks: used rectangular regions of a spectrogram for explanations (from the task of image segmentation)
  - But, how good is this explanation for a human?
- AudioLIME: interpretability = listenability
  - A source separation algorithm (Spleeter) decomposes input audio into $d' = C \times \tau$ interpretable components (C sources, τ time segments).

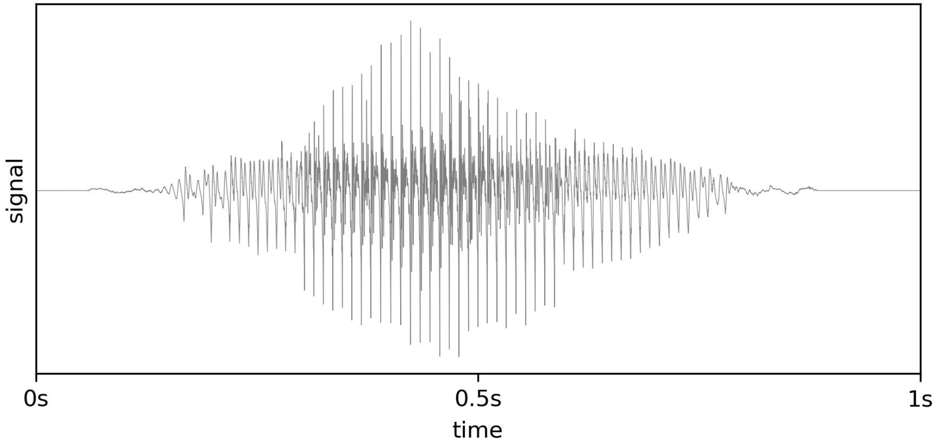


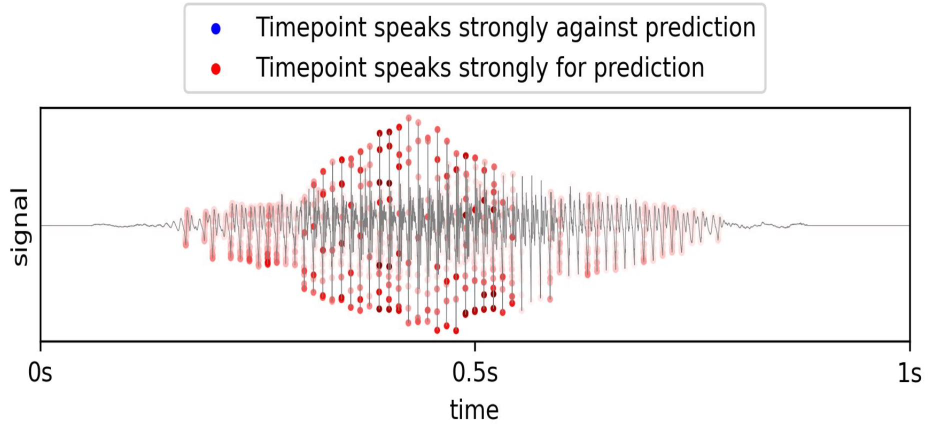
- AudioLIME preserves fundamental aspects of audio: explanations are listenable
- Play top-most relevance source segments

- To evaluate method on music tagging systems: feed the explanation back into the tagger and see if the prediction changes
- Tagger should make the same prediction when only passing the top k components, and a different prediction otherwise
- Randomly picked 100 examples
- For each example we create several explanations for the top predicted tag.
- Eg1: AI predicted tag "female vocalist" [in the top 3 selected components were the separated vocals with a female singer]
- Eg2: AI predicted the tag "rock", [in the top components we hear a drumset and a distorted guitar => associated with rock music]
- This gives audioLIME the ability to train on interpretable and listenable features.

LRP for Audio: Audible Explanation by Becker et al. (2023)

- Audio segmentation & Source separation issues:
  - not always possible!
  - only works with a limited number of source types
  - may introduce artifacts
- Visual explanation: Insufficient to communicate model reasoning for prediction
- Element-wise product between the raw waveform and the heatmap
- Cancels undesired variability of the explanations induced by the specific choice of the source separation algorithm
- Present either positive or negative relevance
- Limitation: Only for time domain

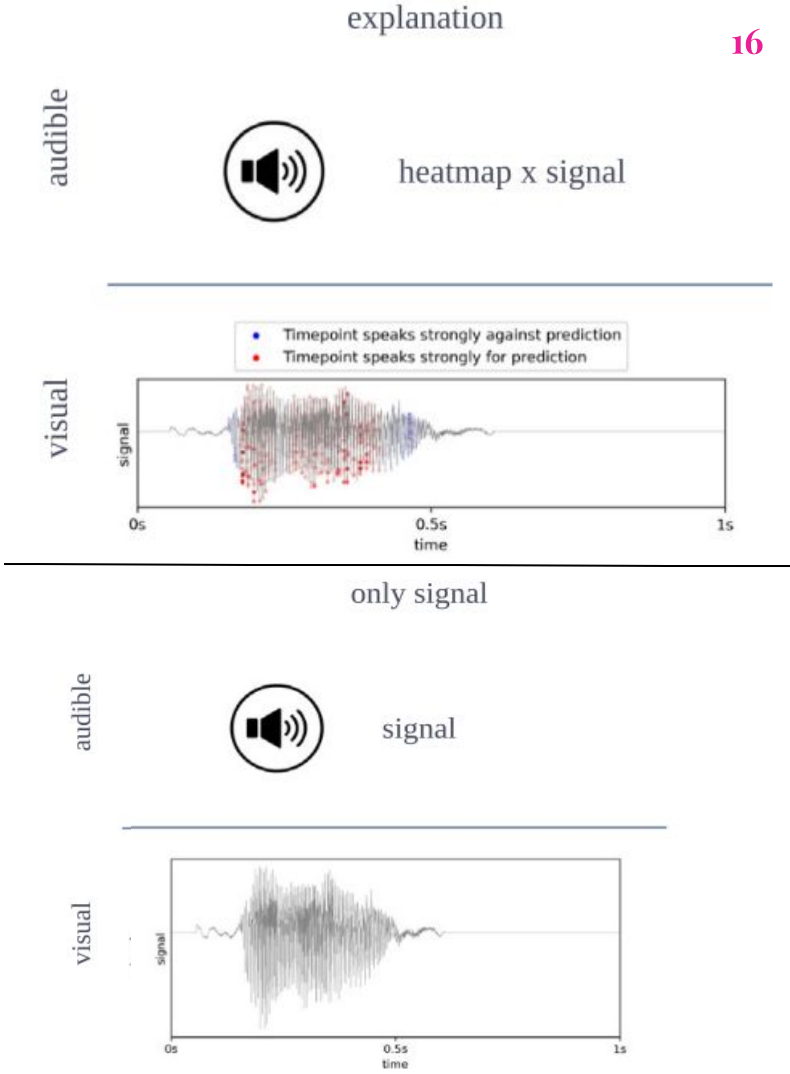




$\mathrm{ReLU}(\boldsymbol{R}) \odot \boldsymbol{x}$

Becker et al. Case Study

- Investigate: which explanation format is the most interpretable to humans: audible or visual explanations?
- Design of the user study:
  - The user was presented with either a visual or audible explanation.
  - As a baseline we present faux explanations that entail only the signal itself.
  - The user was asked to predict the model prediction based on the explanation
- Presented both the modulated or overlayed signal with relevance scores as well as solely the signal, for both the audible and visual explanation formats
- Chose 10 random samples where the model prediction is correct and 10 random samples where the model is predicting incorrectly



Becker et al. Case Study: Evaluation

- **Informedness** for each class measures how *informed* the user is about the positive and negative model predictions for this class based on the explanation
  - TP/P - FP/N
- **Markedness** measures the *trustworthiness* of the user's prediction of positive and negative model predictions for this class
  - TP/(TP+FP) - FN/(TN+FN)
- Positive values imply that the user is informed correctly by the explanation and their prediction can be trusted
- Negative values imply that the user is informed incorrectly and

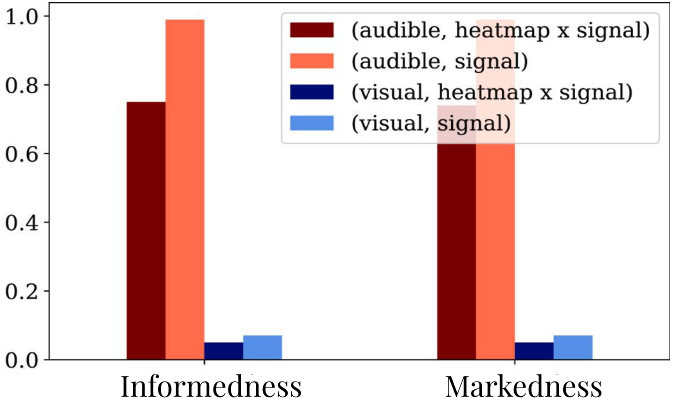


Fig: Correct model classification

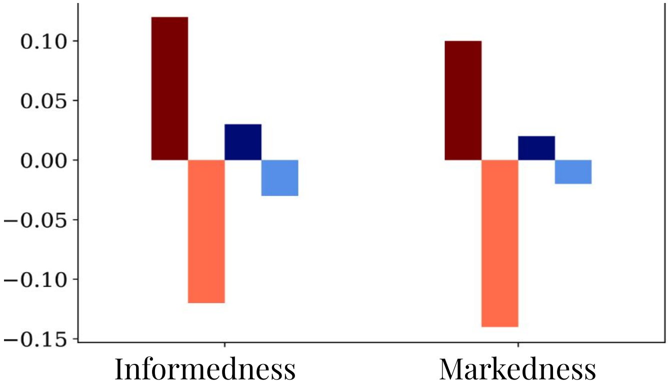


Fig: Incorrect model classification

- informedness and markedness are higher for the audible signal than for the actual audible explanation. It is possible that the model's classification strategy deviates from the user's classification strategy
- across all digits classes, both informedness and markedness have the lowest value for the samples correctly classified as a 'nine'
  33% users predicted the model classified the digit as a 'nine' and 32% predicted that the model classified it as a 'five'.
  In the explanation, only the common syllable, the 'i' is audible.

- audible explanations show a markedly greater informedness and markedness than their visual counterpart
- there is still room for improvement in terms of the interpretability of audible explanations
- only the signal show negative informedness and markedness, as the user is informed incorrectly about the model prediction

Conclusion & Future Work

- Networks are highly reliant on features marked as relevant by LRP
- For classifications based on raw waveforms LRP showed that the networks' decisions depend on a relatively small fraction of the data
  - networks focus mainly on the envelope, i.e., the "global shape", of the signal
- Through a user study, we have conclusively shown that audible explanations exhibit superior interpretability
- Next:
  - Apply LRP to more complex audio datasets to gain a deeper insight into classification decisions
  - Improve the interpretability of audible explanations, by using concept-based XAI methods in
    - J. Vielhaben, S. Bluecher, N. Strodthoff, **Multi-dimensional concept discovery (MCD)**: A unifying framework with completeness guarantees, Trans. Mach. Learn. Res. (2023).
    - R. Achtibat, M. Dreyer, I. Eisenbraun, S. Bosse, T. Wiegand, W. Samek, S. Lapuschkin, **From attribution maps to human-understandable explanations through concept relevance propagation**, Nat. Mach. Intell. 5 (9) (2023) 1006–1019