Analysing the Predictions of a CNN-Based Replay Spoofing Detection System

Bhusan Chettri, Saumitra Mishra, Bob L. Sturm, Emmanouil Benetos

{b.chettri, saumitra.mishra, emmanouil.benetos}@qmul.ac.uk, bobs@kth.se

1 Introduction

- Replay spoofing attack involves playing back pre-recorded audio samples to an Automatic Speaker Verification (ASV) system.
- The vulnerability of ASV systems to such attacks has been acknowledged and studied [1], but there has been little or no research into what such systems are actually learning to discriminate.
- We analyse a CNN-based replay spoofing detection system by generating temporal and spectral explanations for its predictions using the SLIME [3] algorithm.
- We demonstrate the significance of our analysis from an attacker and an ASV administrator perspective by raising and lowering the equal error rate (EER) respectively.

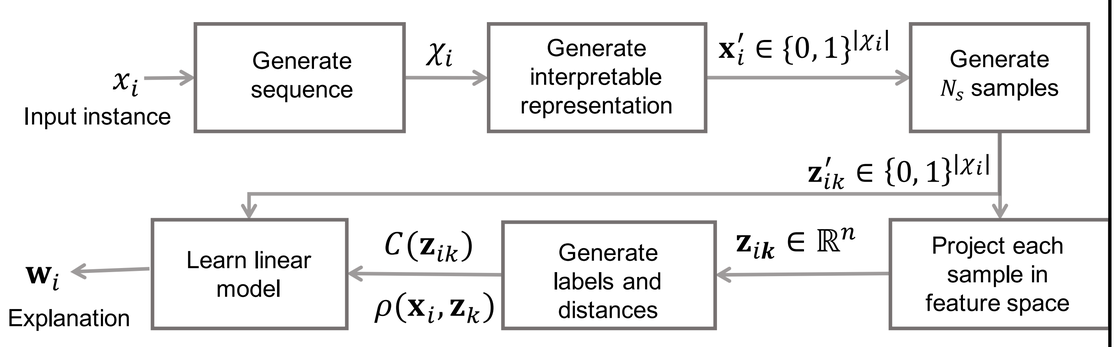
2 System description

- **Dataset**: ASVspoof 2017 dataset that was released as a result of the second ASV spoofing and countermeasures challenge [1].
- **Input**: unified 4 seconds log power spectrogram.
- **Model**: CNN adapted from light-CNN [2], the best performing model of the ASVspoof 2017 challenge.
- **Performance**: evaluated in terms of the equal error rate (EER)

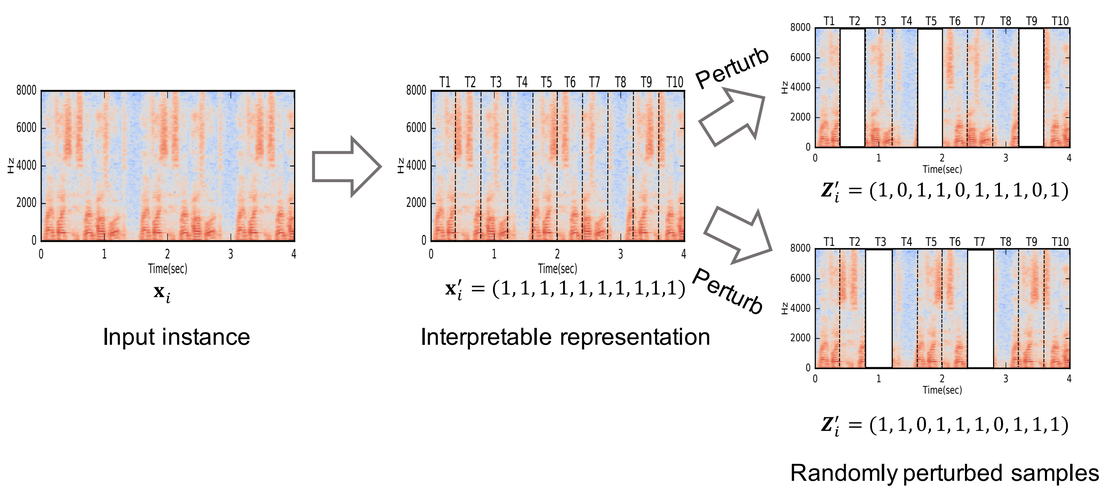
Dev EER	Eval EER
7.6 %	10.6 %

3 SLIME algorithm [3]

- SLIME uses the following sequence of steps to produce an explanation (in terms of weights $\mathbf{w}_i$) for a given input instance x_i.



- *Example*: segmenting an input x_i into 10 uniform temporal components (T1-T10) and generating two samples through random perturbations on these components.



4 Explaining the predictions

We generate explanations for predictions of the *most **confidently classified spoof** instances* in the Train, Dev and Eval subsets.

- **Temporal explanations**: we use 10 temporal components T1-T10 each of 400 ms.

Instance id	Top 4 explanations	Corresponding weights
T_1002124	T1, T10, T7, T8	0.34, 0.27, 0.01, 0.01
D_1001596	T1, T10, T5, T7	0.51, 0.12, 0.01, 0.01
E_1014008	T1, T10, T4, T5	0.35, 0.21, 0.01, 0.01

- **Spectral explanations**: we use 10 spectral components F1-F10, each of 813 Hz bandwidth except for F10 (683 Hz).

Instance id	Top 4 explanations	Corresponding weights
T_1002124	F3, F5, F1, F2	0.11, 0.11, 0.1, 0.1
D_1001596	F7, F2, F4, F8	0.11, 0.1, 0.1, 0.1
E_1014008	F3, F6, F1, F5	0.15, 0.14, 0.14, 0.13

We repeat the above process for all the confidently classified spoof instances in the dataset and make the following **observations**:

- *While the model use information across all the frequency bands, more emphasis is given on the first and the last temporal components (**T1, T10**) for spoofing detection.*

5 Interventions

We show the significance of our analysis using two interventions.

- **Intervention I**: Replace T1 and T10 of confidently classified spoof instances by T1 of the most confident genuine instance.
- **Intervention II**: Remove samples from the start of misclassified spoof audio files to ensure that speech occurs in the first 400 ms.

	Dev EER %	Eval EER %
I: Break the system	7.6 $\rightarrow$ 34.13	10.6 $\rightarrow$ 29.76
II: Protect the system	7.6 $\rightarrow$ 5.9	10.6 $\rightarrow$ 7.8

6 Conclusion

- We use SLIME algorithm to analyse an adapted state-of-the-art CNN model for replay spoofing detection on the ASVspoof 2017 2.0 dataset [4]. We find that the model gives more importance to the first few milliseconds for class prediction.
- We further demonstrate the significance of our analysis by pre-processing the test signals that lead to a predictable change in the EER. We aim to extend this analysis across different replay conditions of the ASVspoof 2017 dataset.

[1] Kinnunen et. al. The ASVspoof 2017 Challenge: Assessing the Limits of Audio Replay Attack Detection in the Wild. In *Proc. Interspeech 2017.*

[2] Lavrentyeva et. al. Audio Replay Attack Detection with Deep Learning Frameworks. In *Proc. Interspeech 2017*, Pages 82–86, August, 2017.

[3] Mishra et. al. Local Interpretable Model-Agnostic Explanations for Music Content Analysis In *ISMIR 2017*.

[4] Delgado et. al. ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements. In *Speaker Odyssey 2018*.

IEEE Spoken Language Technology (SLT) workshop, 18-21 December, Athens, Greece.