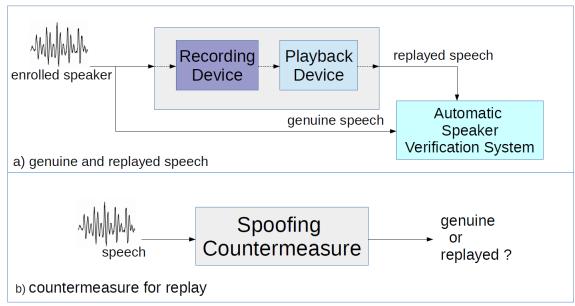
Analysing the predictions of a CNN-based replay spoofing detection system Bhusan Chettri



 Understanding the best performing model [2] of the ASVspoof 2017 challenge [1] by generating temporal and spectral explanations for its predictions using the SLIME [3] algorithm



[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

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

C centre for

■I intelligent sensing



SLIME algorithm

- Top figure: the sequence of steps used by SLIME to produce an explanation (in terms of weights) for a given input instance
- $\mathbf{x}'_i \in \{0, 1\}^{|\chi_i|}$ χi Generate Generate interpretable χ_i N_s samples sequence representation Input instance $\mathbf{z}'_{ik} \in \{0, 1\}^{|\chi_i|}$ $\mathbf{z}_{ik} \in \mathbb{R}^n$ $C(\mathbf{z}_{ik})$ Project each Generate Learn linear sample in $W_i \leq$ labels and model $\rho(\mathbf{X}_i, \mathbf{Z}_k)$ feature space distances Explanation 14 T5 T6 T7 T8 T9 T10 Perturb T 4000 $\mathbf{Z}'_i = (1, 0, 1, 1, 0, 1, 1, 1, 0, 1)$ 2 Time(sec) Time(sec) X; ₽ 4000 Input instance Interpretable representation $Z'_i = (1, 1, 0, 1, 1, 1, 0, 1, 1, 1)$ Randomly perturbed samples

Generate

Bottom figure: example of segmenting an input instance into 10 uniform temporal components (T1-T10) and generating two samples through random perturbations on these components

[2] Lavrentyeva et. al. Audio Replay Attack Detection with Deep Learning Frameworks. In Proc. Interspeech 2017. [3] Mishra et. al. Local Interpretable Model-Agnostic Explanations for Music Content Analysis In ISMIR 2017.

centre for intelligent sensing



Results and conclusion

 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. We show the significance of our analysis using two interventions

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

- The model gives more importance to the first few milliseconds for class prediction
- Demonstrated the significance of our analysis by preprocessing the test signals that lead to a predictable change in the EER

Analysing the predictions of a CNN-based replay spoofing detection system Bhusan Chettri, Saumitra Mishra, Bob L. Sturm, Emmanouil Benetos IEEE Spoken Language Technology, Athens, Greece, 18-21 December 2018

