Extremal Perturbations

王天昊

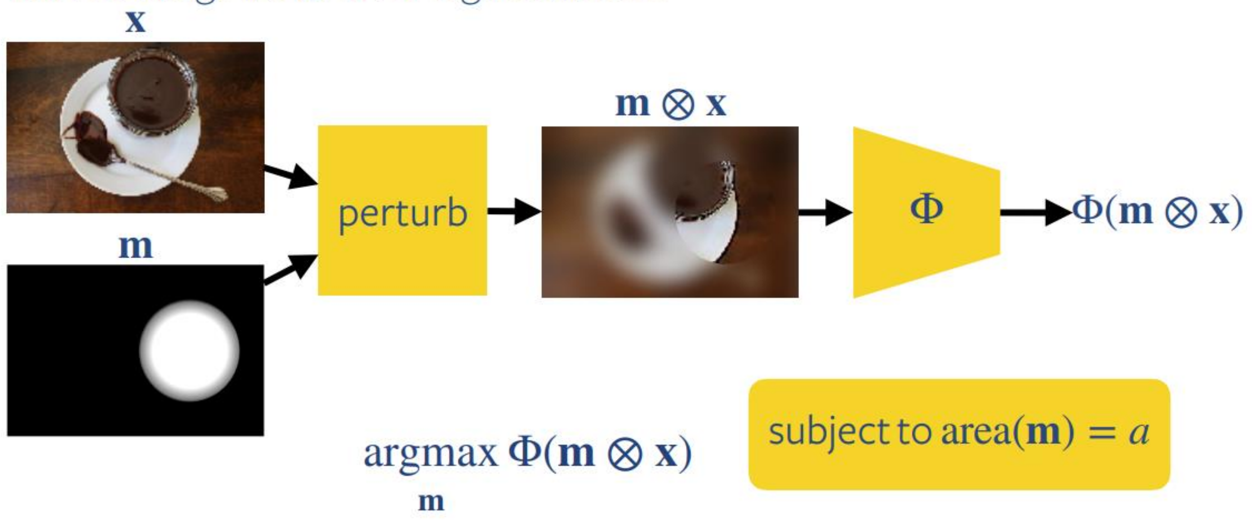
2022-08-12

Review

- What is Extremal Perturbations?

$$a^* = \min\{a : \Phi(\boldsymbol{m}_a \otimes \boldsymbol{x}) \geq \Phi_0\}.$$

We optimize the mask $\mathbf{m}$ to maximize the response of the network $\mathbf{\Phi}$ on the blurred image $\mathbf{m} \otimes \mathbf{x}$ for a given area a:


$\mathbf{x}$
$\mathbf{m}$
perturb
$\mathbf{m} \otimes \mathbf{x}$
Φ
$\Phi(\mathbf{m} \otimes \mathbf{x})$

$$\underset{\mathbf{m}}{\operatorname{argmax}}\ \Phi(\mathbf{m} \otimes \mathbf{x}) \quad \text{subject to area}(\mathbf{m}) = a$$

- R. Fong, M. Patrick and A. Vedaldi, "Understanding Deep Networks via Extremal Perturbations and Smooth Masks," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2950-2958, doi: 10.1109/ICCV.2019.00304.

Review

$$\boldsymbol{m}_a = \operatorname*{argmax}_{\boldsymbol{m}\in\mathcal{M}} \Phi(\boldsymbol{m}\otimes\boldsymbol{x}) - \lambda R_a(\boldsymbol{m}).$$

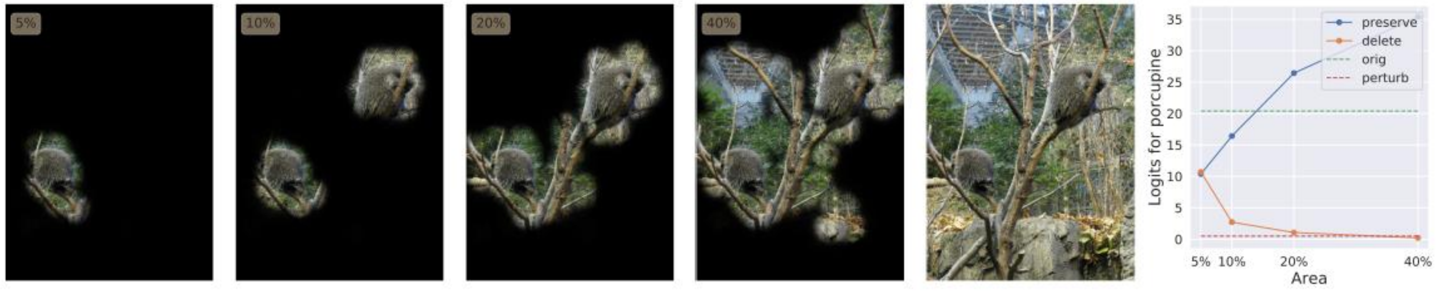


Figure 3: **Extremal perturbations and monotonic effects.** Left: "porcupine" masks computed for several areas a (a in box). Right: $\Phi(\boldsymbol{m}_a \otimes \boldsymbol{x})$ (preservation; blue) and $\Phi((1-\boldsymbol{m}_a)\otimes\boldsymbol{x})$ (deletion; orange) plotted as a function of a. At $a \approx 15\%$ the preserved region scores *higher* than preserving the entire image (green). At $a \approx 20\%$, perturbing the complementary region scores *similarly* to fully perturbing the entire image (red).

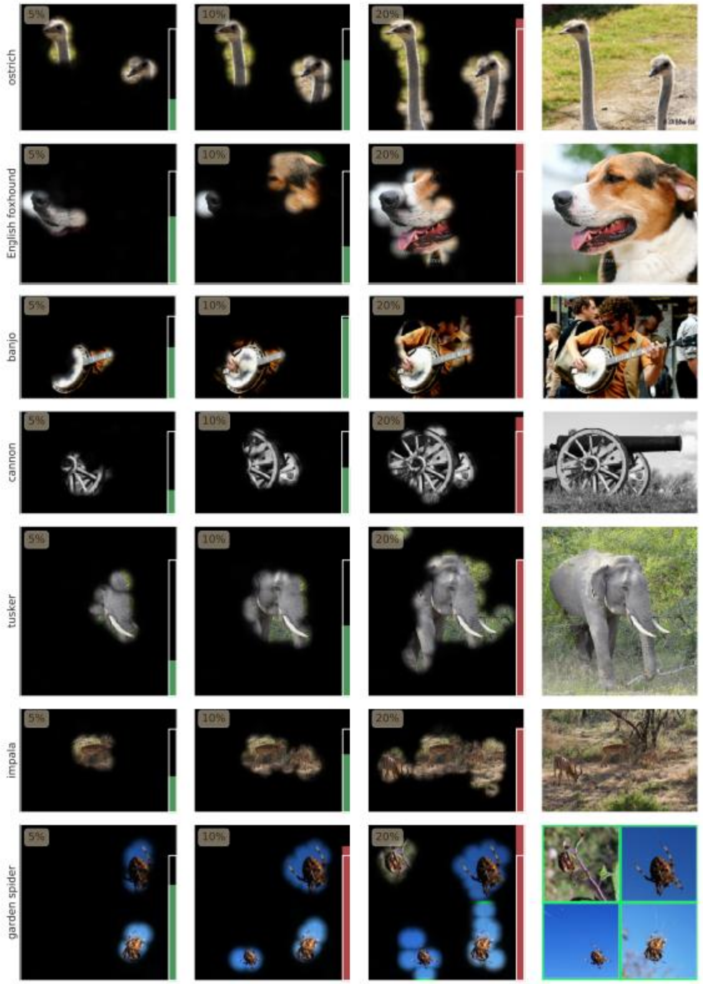


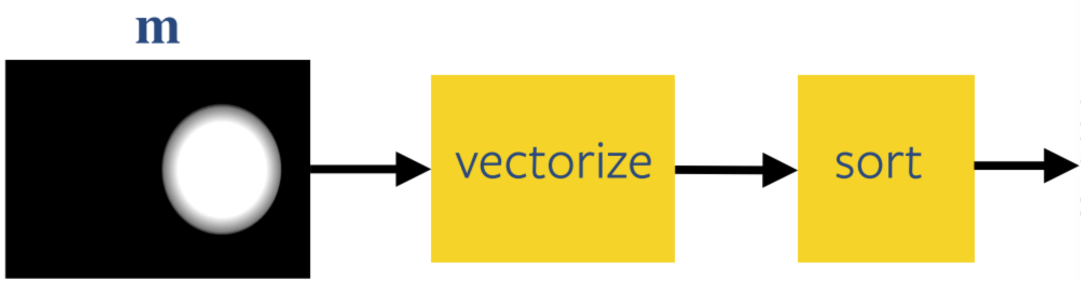
Figure 5: **Area growth.** Although each mask is learned independently, these plots highlight what the network considers to be most discriminative and complete. The bar graph visualizes $\Phi(\boldsymbol{m}_a \odot \boldsymbol{x})$ as a normalized fraction of $\Phi_0 = \Phi(\boldsymbol{x})$ (and saturates after exceeding Φ_0 by 25%).

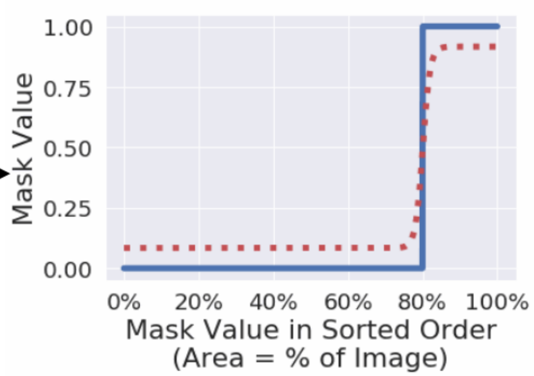
- R. Fong, M. Patrick and A. Vedaldi, "Understanding Deep Networks via Extremal Perturbations and Smooth Masks," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2950-2958, doi: 10.1109/ICCV.2019.00304.

Method

Area constraint

Optimizing for a given area size is non-trivial. We do it by sorting the mask values and comparing the result to the desired 0-1 distribution $\mathbf{r}_a$:





$$L_{area} = \| \, \mathrm{vecsort}(\mathbf{m}) - \mathbf{r}_a \|^2$$

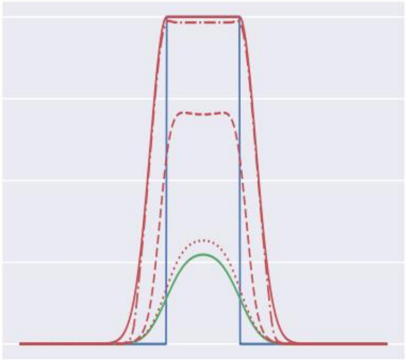
Algorithm

Pick area a and perform SGD to optimize:

$$\underset{\mathbf{m}}{\mathrm{argmax}} \; \Phi(\mathrm{smoothconv}(\mathbf{m}) \otimes \mathbf{x}) - \lambda \| \, \mathrm{vecsort}(\mathrm{smoothconv}(\mathbf{m})) - \mathbf{r}_a \|^2$$

Method

Smooth mask



— $m(v)$: mask

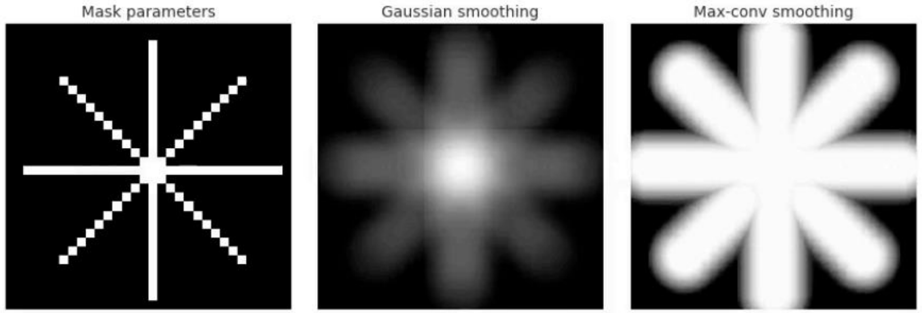
— $\mathrm{conv}(u; m; k) = \frac{1}{Z} \sum_{v \in \Omega} k(u - v)m(v)$

— $\mathrm{maxconv}(u; m; k) = \max_{v \in \Omega} k(u - v)m(v)$

··· $\mathrm{smoothconv}(u; m; k; T) = \mathrm{smax}_{v \in \Omega; T}\, k(u - v)m(v)$

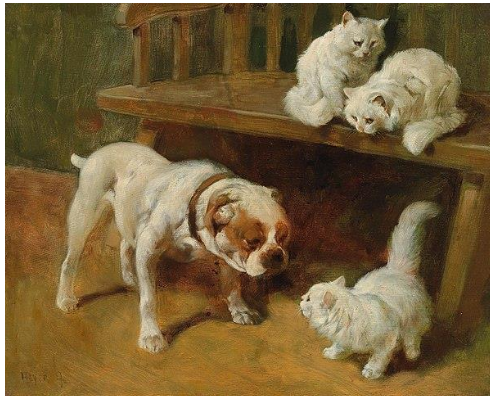
$$\mathrm{smax}_{u \in \Omega; T}\, f(u) = \frac{\sum_u f(u)\exp(f(u)/T)}{\sum_u \exp(f(u)/T)}$$

Right: comparison between original mask (L), mask after conv (M), and maxconv (R).

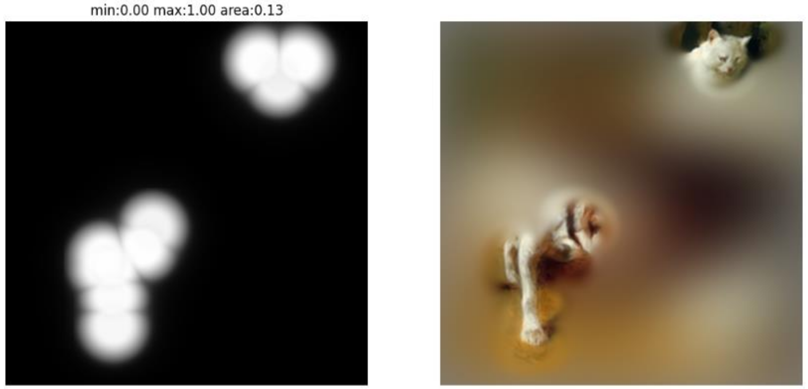


Reproduction

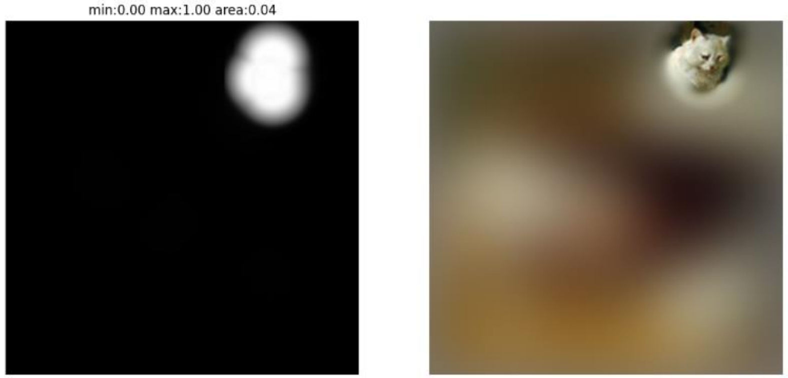
input image



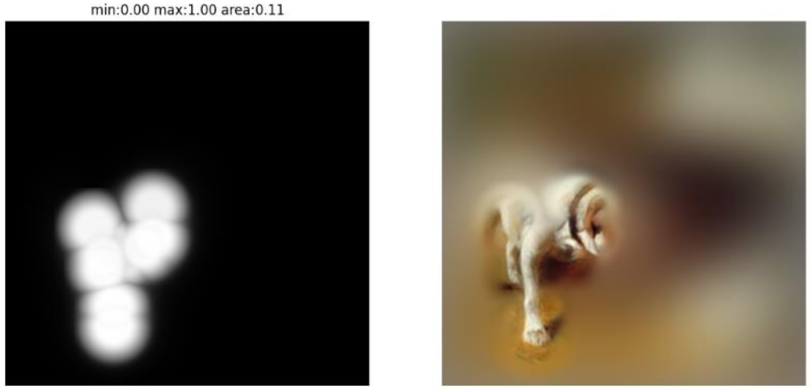
target: dog target area: 0.12



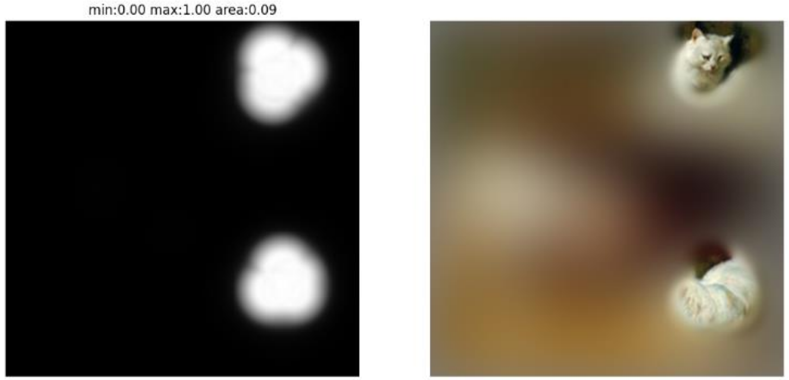
target: cat target area: 0.05



target: dog target area: 0.10



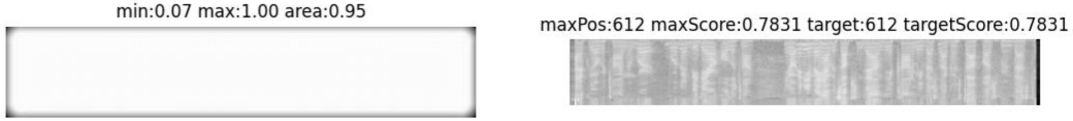
target: cat target area: 0.08



Apply to Speaker Recognition

- Single Speaker (target area: 0.20)

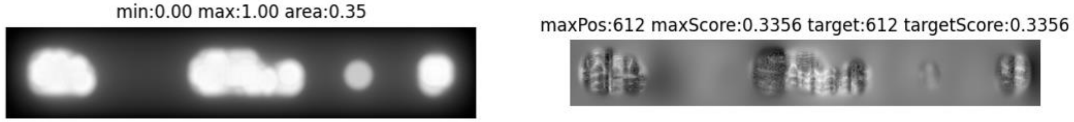
1. 1-1



min:0.07 max:1.00 area:0.95

maxPos:612 maxScore:0.7831 target:612 targetScore:0.7831

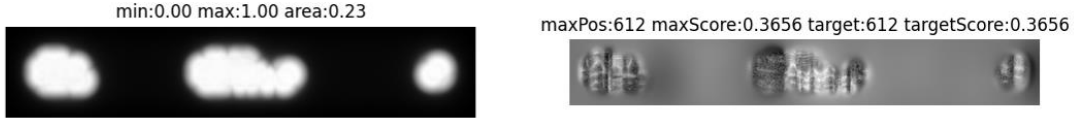
2. 1-2



min:0.00 max:1.00 area:0.35

maxPos:612 maxScore:0.3356 target:612 targetScore:0.3356

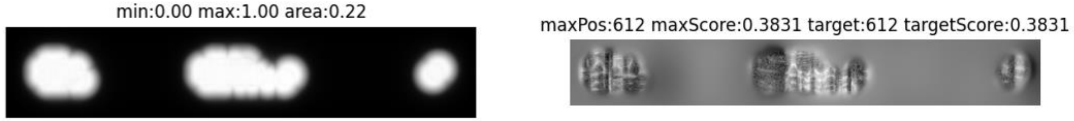
3. 1-3



min:0.00 max:1.00 area:0.23

maxPos:612 maxScore:0.3656 target:612 targetScore:0.3656

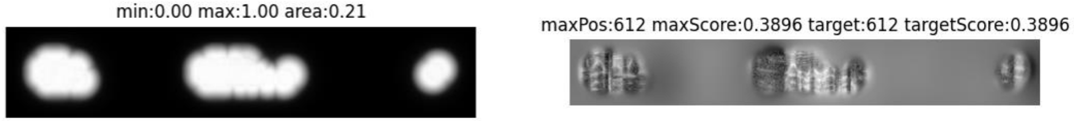
4. 1-4



min:0.00 max:1.00 area:0.22

maxPos:612 maxScore:0.3831 target:612 targetScore:0.3831

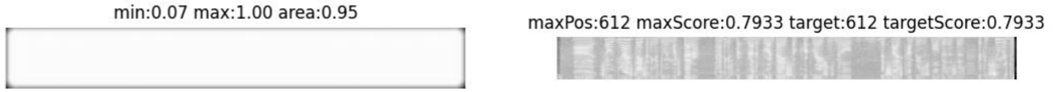
5. 1-5



min:0.00 max:1.00 area:0.21

maxPos:612 maxScore:0.3896 target:612 targetScore:0.3896

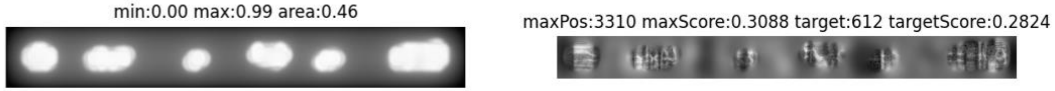
1. 2-1



min:0.07 max:1.00 area:0.95

maxPos:612 maxScore:0.7933 target:612 targetScore:0.7933

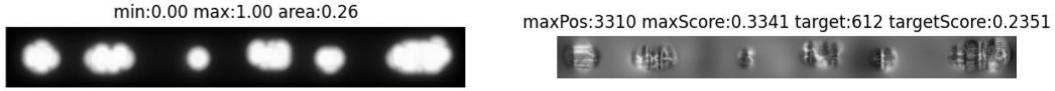
2. 2-2



min:0.00 max:0.99 area:0.46

maxPos:3310 maxScore:0.3088 target:612 targetScore:0.2824

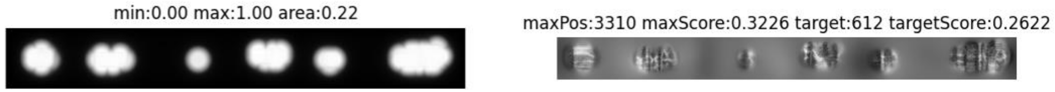
3. 2-3



min:0.00 max:1.00 area:0.26

maxPos:3310 maxScore:0.3341 target:612 targetScore:0.2351

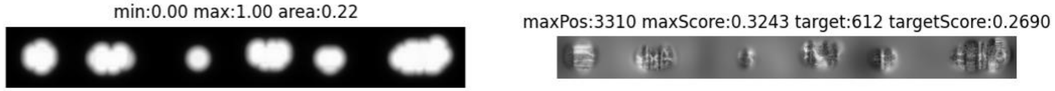
4. 2-4



min:0.00 max:1.00 area:0.22

maxPos:3310 maxScore:0.3226 target:612 targetScore:0.2622

5. 2-5



min:0.00 max:1.00 area:0.22

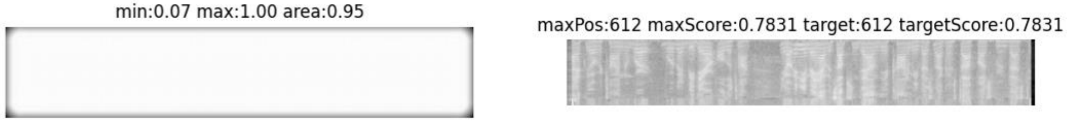
maxPos:3310 maxScore:0.3243 target:612 targetScore:0.2690

- Li, P., Li, L., Hamdulla, A., & Wang, D. (2022). Reliable Visualization for Deep Speaker Recognition. arXiv preprint arXiv:2204.03852.

Apply to Speaker Recognition

- Single Speaker (target area: 0.30)

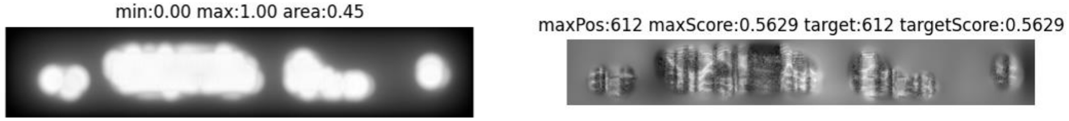
1. 1-1



min:0.07 max:1.00 area:0.95

maxPos:612 maxScore:0.7831 target:612 targetScore:0.7831

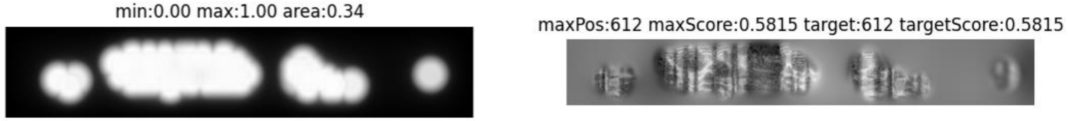
2. 1-2



min:0.00 max:1.00 area:0.45

maxPos:612 maxScore:0.5629 target:612 targetScore:0.5629

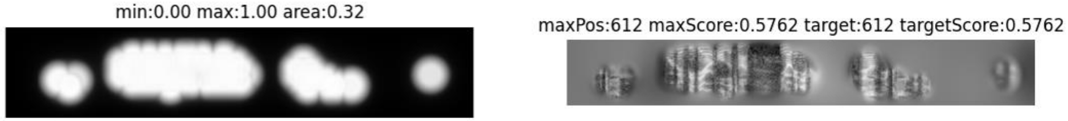
3. 1-3



min:0.00 max:1.00 area:0.34

maxPos:612 maxScore:0.5815 target:612 targetScore:0.5815

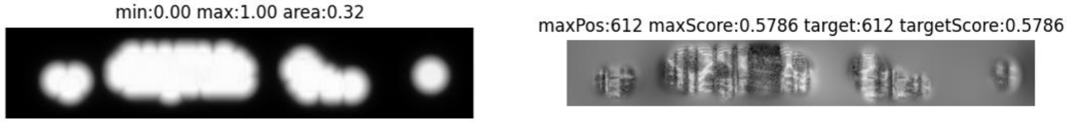
4. 1-4



min:0.00 max:1.00 area:0.32

maxPos:612 maxScore:0.5762 target:612 targetScore:0.5762

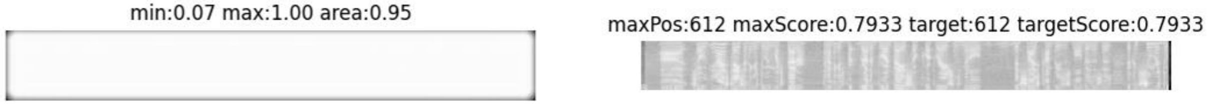
5. 1-5



min:0.00 max:1.00 area:0.32

maxPos:612 maxScore:0.5786 target:612 targetScore:0.5786

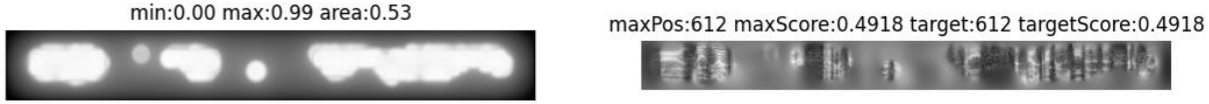
1. 2-1



min:0.07 max:1.00 area:0.95

maxPos:612 maxScore:0.7933 target:612 targetScore:0.7933

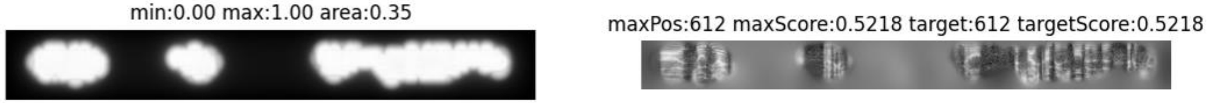
2. 2-2



min:0.00 max:0.99 area:0.53

maxPos:612 maxScore:0.4918 target:612 targetScore:0.4918

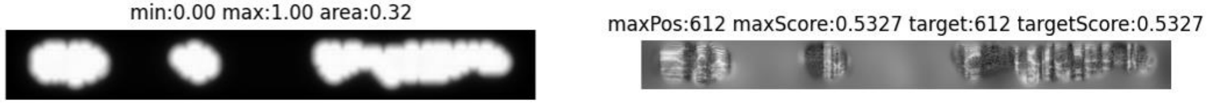
3. 2-3



min:0.00 max:1.00 area:0.35

maxPos:612 maxScore:0.5218 target:612 targetScore:0.5218

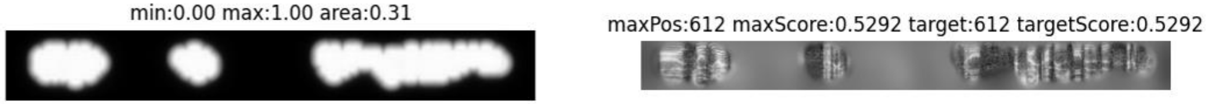
4. 2-4



min:0.00 max:1.00 area:0.32

maxPos:612 maxScore:0.5327 target:612 targetScore:0.5327

5. 2-5



min:0.00 max:1.00 area:0.31

maxPos:612 maxScore:0.5292 target:612 targetScore:0.5292

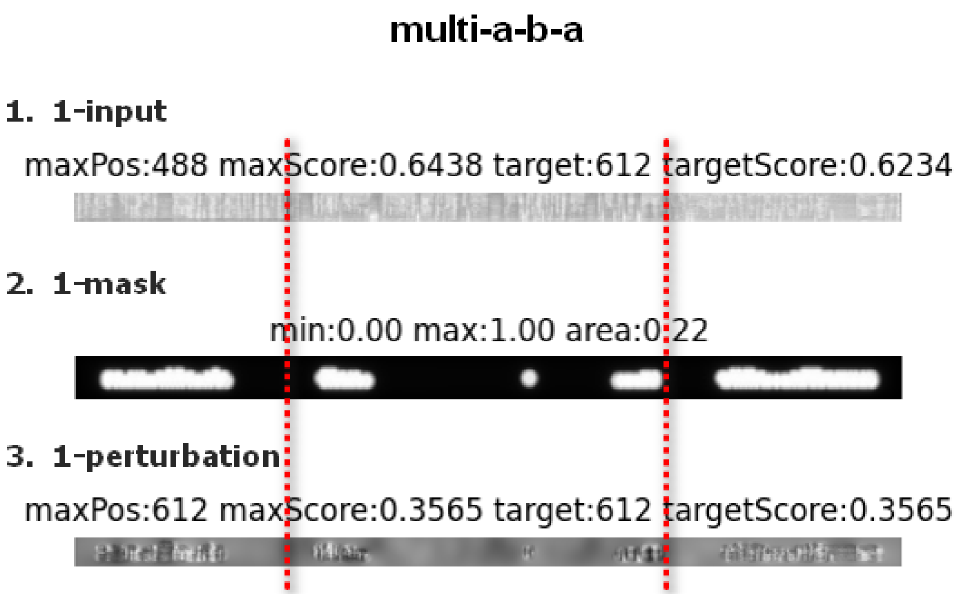
Some thoughts

- Is it right?

- Why did the score drop so much?

Apply to Speaker Recognition

- Multi Speaker (target area: 0.20)



multi-a-b-a

1. 1-input

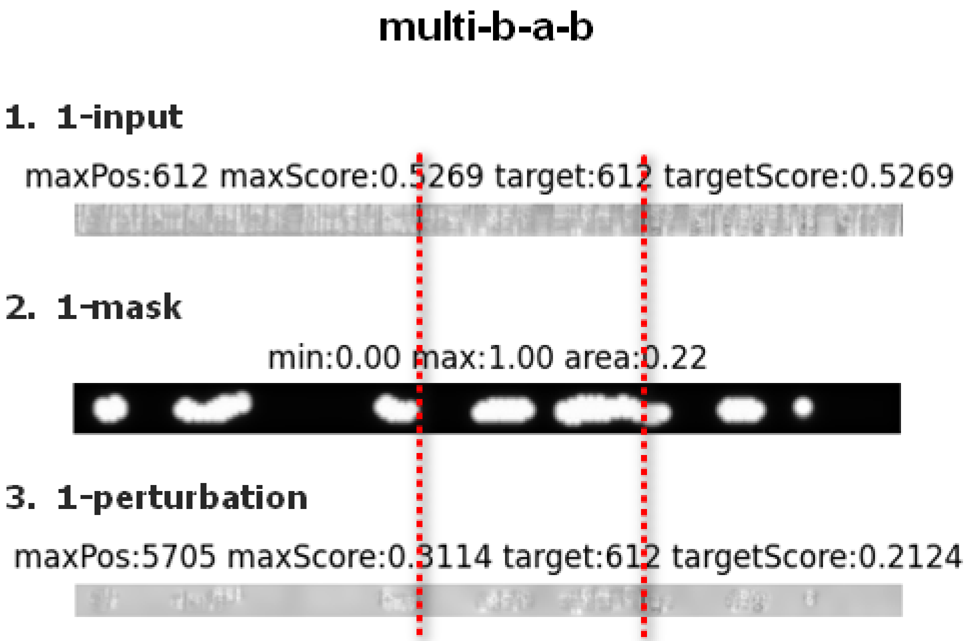
maxPos:488 maxScore:0.6438 target:612 targetScore:0.6234

2. 1-mask

min:0.00 max:1.00 area:0.22

3. 1-perturbation

maxPos:612 maxScore:0.3565 target:612 targetScore:0.3565



multi-b-a-b

1. 1-input

maxPos:612 maxScore:0.5269 target:612 targetScore:0.5269

2. 1-mask

min:0.00 max:1.00 area:0.22

3. 1-perturbation

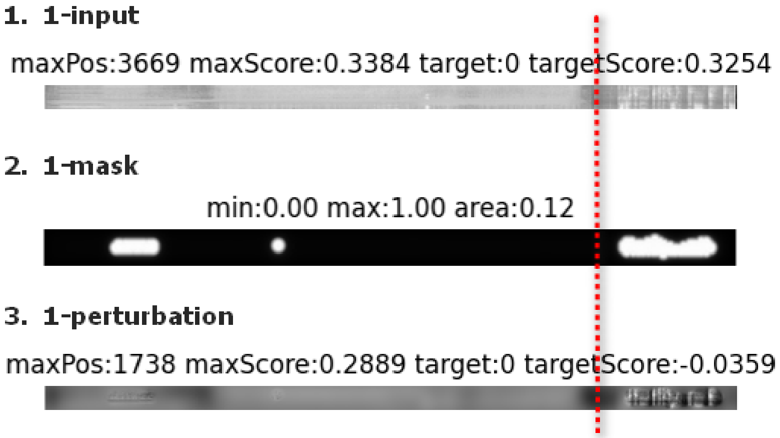
maxPos:5705 maxScore:0.3114 target:612 targetScore:0.2124

Some thoughts

- Hypothesis:
  - There are some commonalities between different speakers.
  - The mask is coarse-grained, while the speaker's common information is relatively discrete, and the personality information is relatively continuous.

Apply to Speaker Recognition

- Single speaker combined noise, silent section (target area: 0.10)



1. 1-input

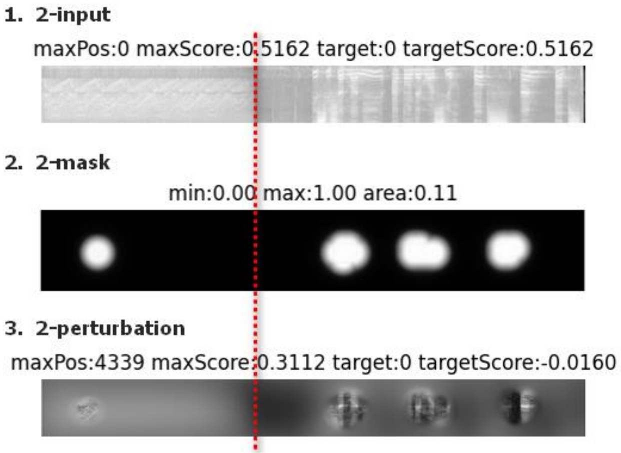
maxPos:3669 maxScore:0.3384 target:0 targetScore:0.3254

2. 1-mask

min:0.00 max:1.00 area:0.12

3. 1-perturbation

maxPos:1738 maxScore:0.2889 target:0 targetScore:-0.0359



1. 2-input

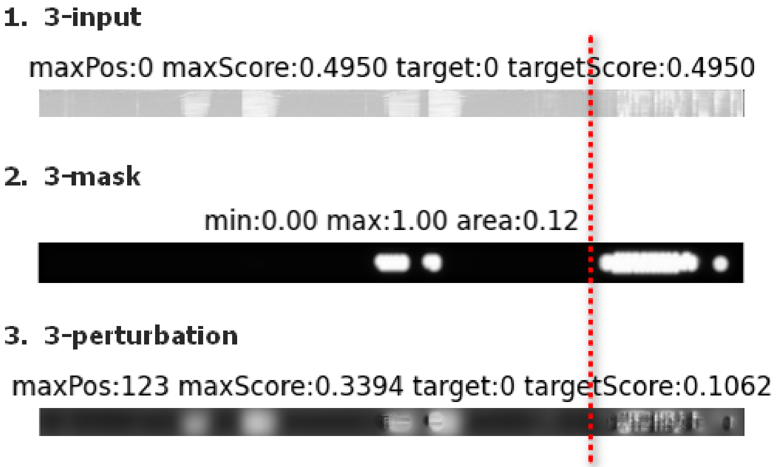
maxPos:0 maxScore:0.5162 target:0 targetScore:0.5162

2. 2-mask

min:0.00 max:1.00 area:0.11

3. 2-perturbation

maxPos:4339 maxScore:0.3112 target:0 targetScore:-0.0160



1. 3-input

maxPos:0 maxScore:0.4950 target:0 targetScore:0.4950

2. 3-mask

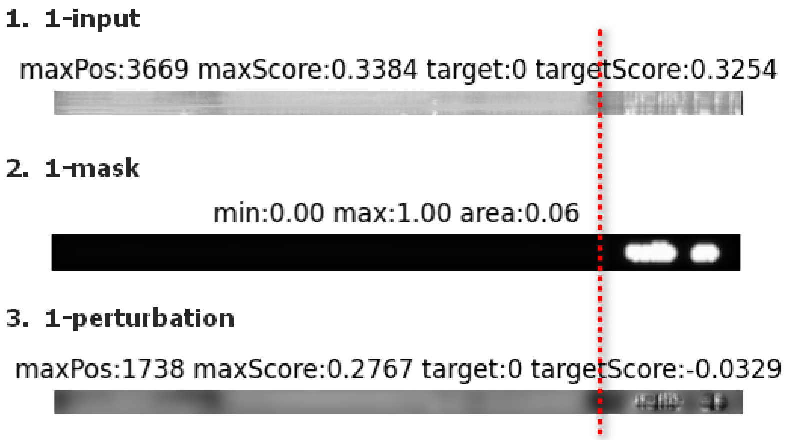
min:0.00 max:1.00 area:0.12

3. 3-perturbation

maxPos:123 maxScore:0.3394 target:0 targetScore:0.1062

Apply to Speaker Recognition

- Single speaker mixed noise, silent section (target area: 0.05)



1. 1-input

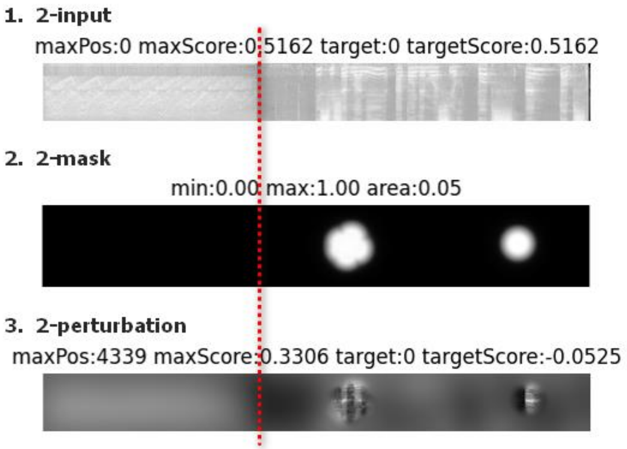
```
maxPos:3669 maxScore:0.3384 target:0 targetScore:0.3254
```

2. 1-mask

```
min:0.00 max:1.00 area:0.06
```

3. 1-perturbation

```
maxPos:1738 maxScore:0.2767 target:0 targetScore:-0.0329
```



1. 2-input

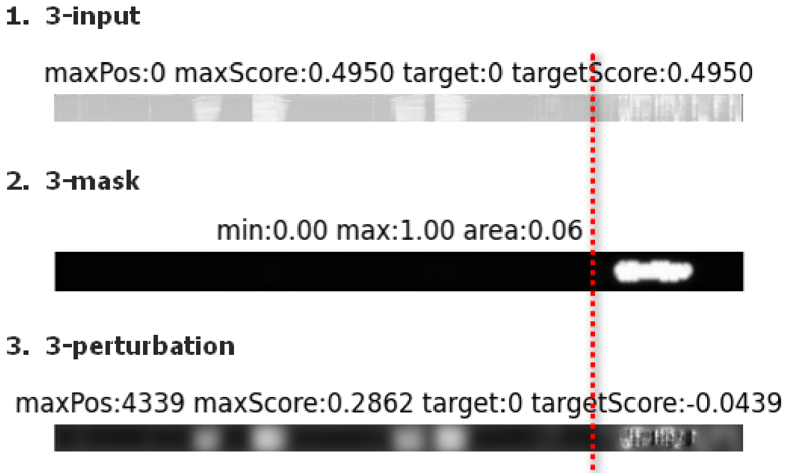
```
maxPos:0 maxScore:0.5162 target:0 targetScore:0.5162
```

2. 2-mask

```
min:0.00 max:1.00 area:0.05
```

3. 2-perturbation

```
maxPos:4339 maxScore:0.3306 target:0 targetScore:-0.0525
```



1. 3-input

```
maxPos:0 maxScore:0.4950 target:0 targetScore:0.4950
```

2. 3-mask

```
min:0.00 max:1.00 area:0.06
```

3. 3-perturbation

```
maxPos:4339 maxScore:0.2862 target:0 targetScore:-0.0439
```

Some thoughts

- The key areas are all within the speaker's time domain segment.
- The blurred part affects the score.
- Initially, the hypothesis is verified.

Next work

- Adjust granularity
- Quantitative experiments for hypothesis

Thanks!