# Transfer Learning and Bias Correction with Pre-trained Audio Embeddings


Changhong Wang[1], Gaël Richard[1], Brian McFee[2]

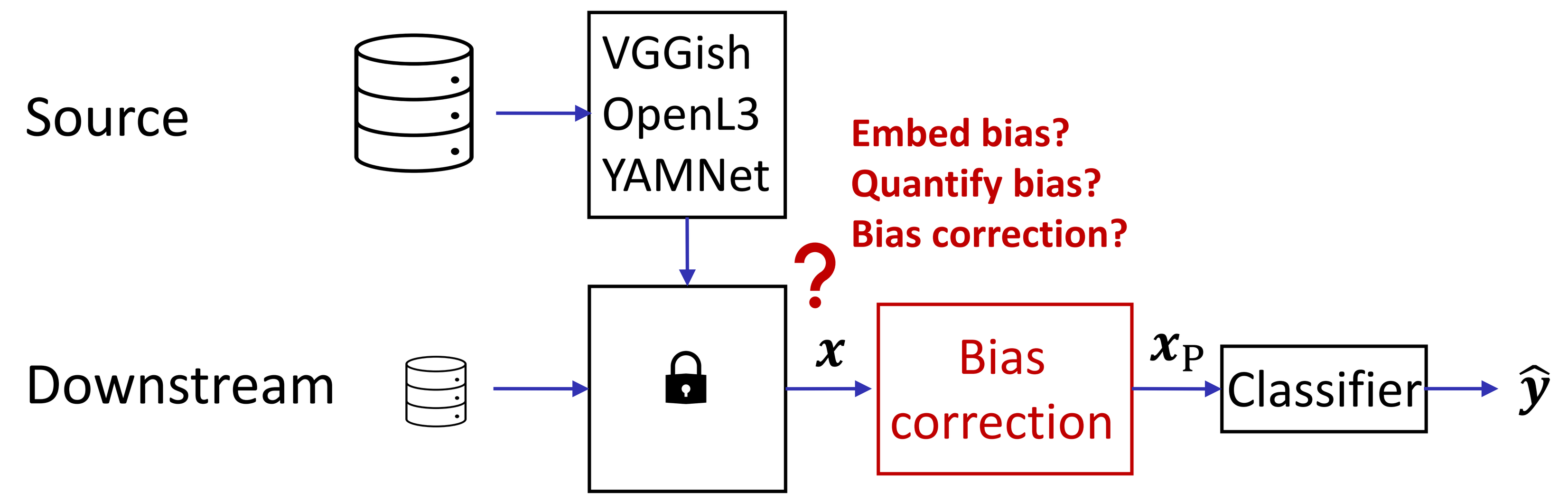
[1]LTCI, Télécom Paris, Institut Polytechnique de Paris, France

[2]Music and Audio Research Laboratory, New York University, USA


## Contributions

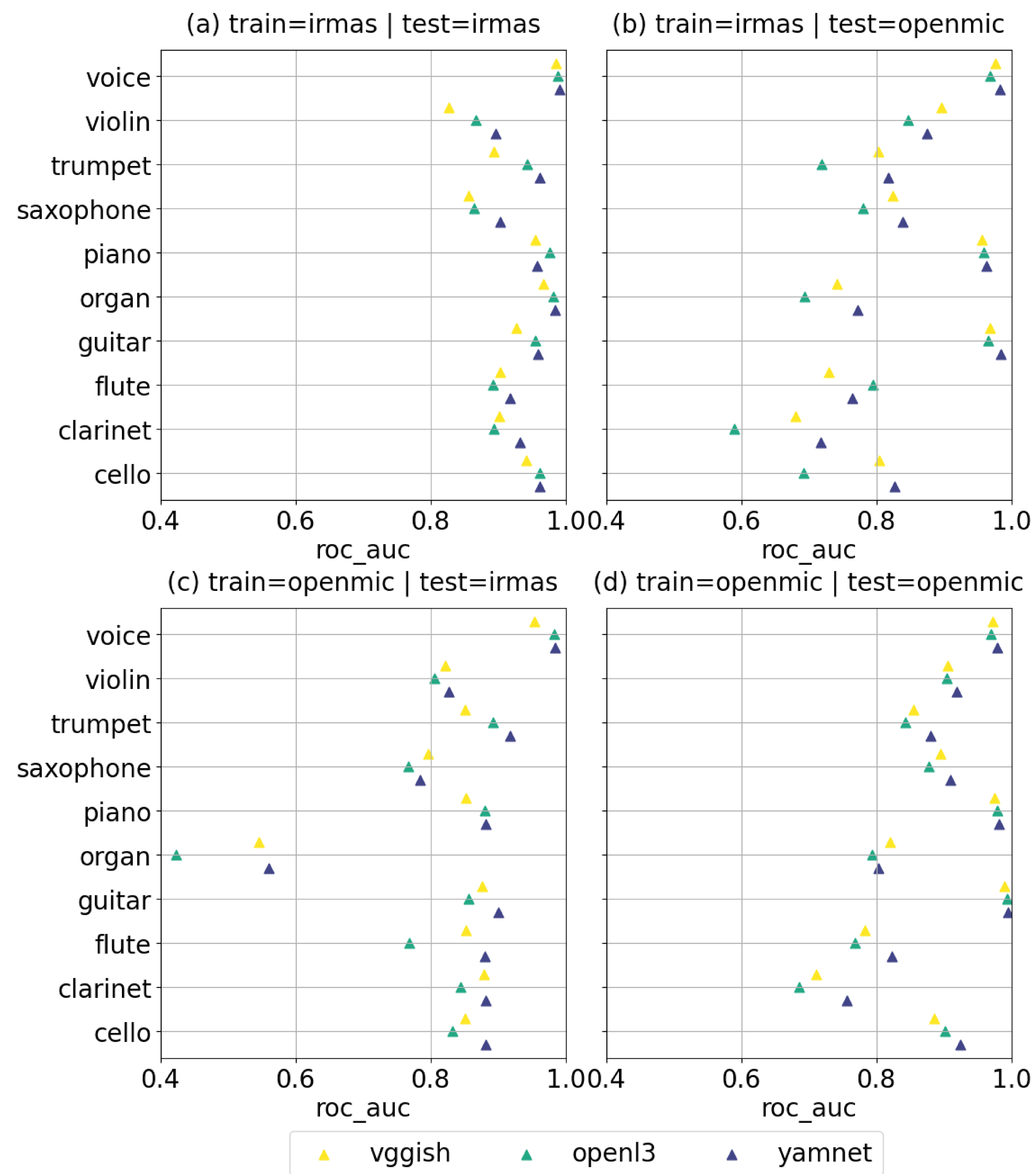
- Investigate bias propagation in transfer learning with pre-trained audio embeddings
- Identify potential sources of bias and quantify bias effects
- Propose 4 post-processing countermeasures to mitigate bias

## 1. Transfer learning with pre-trained audio embeddings



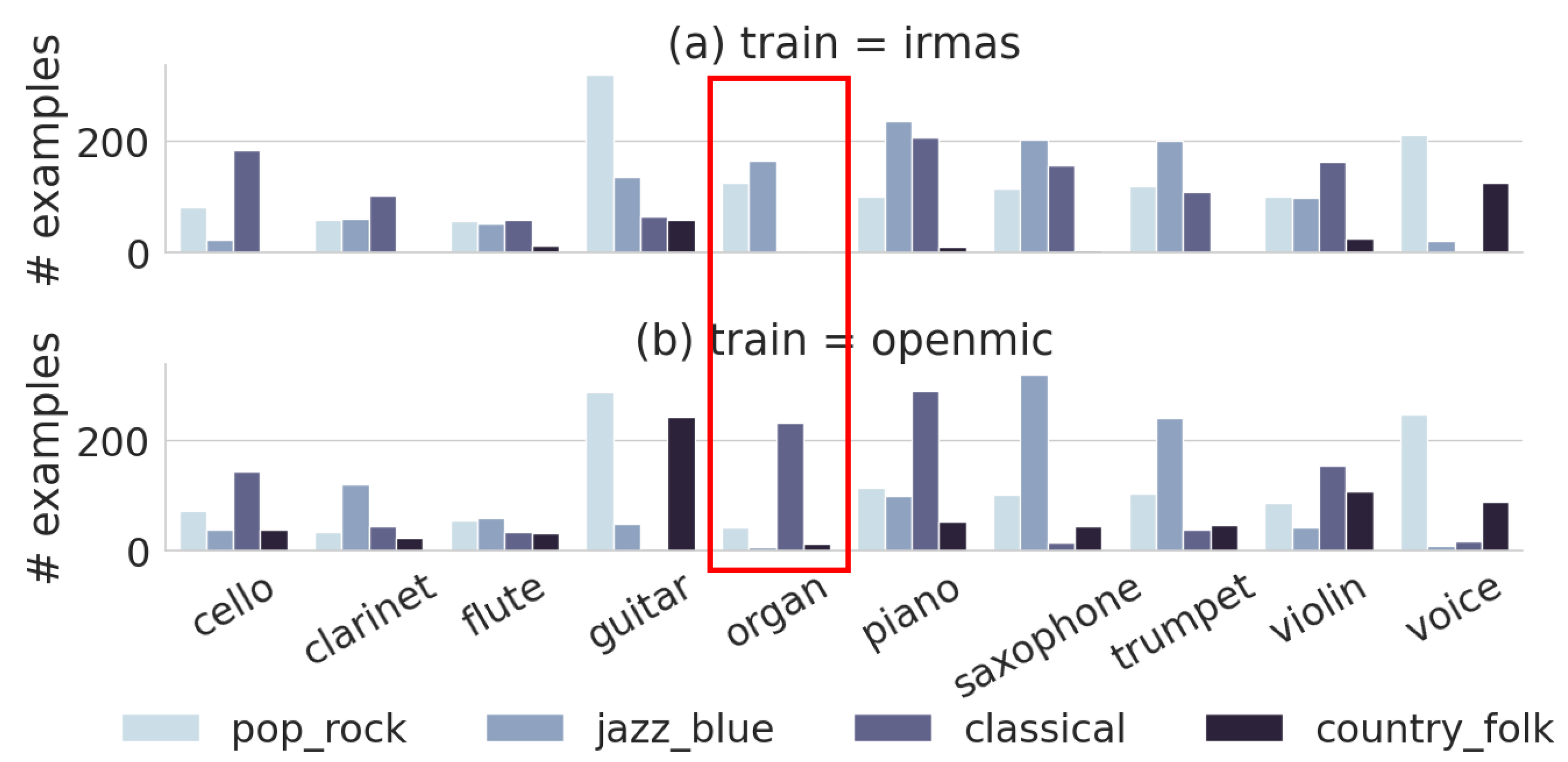

### ➢ Domain sensitivity

- Downstream task: instrument recognition (10 classes)
- Datasets: OpenMIC-2018, IRMAS
- Classifier: binary logistic regression



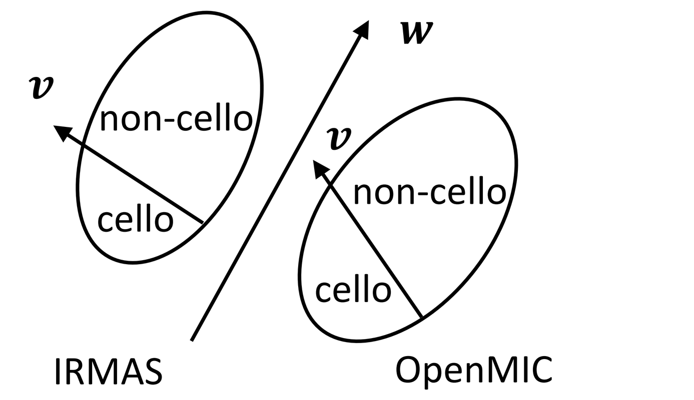

### ➢ Source of bias

- Dataset identity, genre distribution, etc.

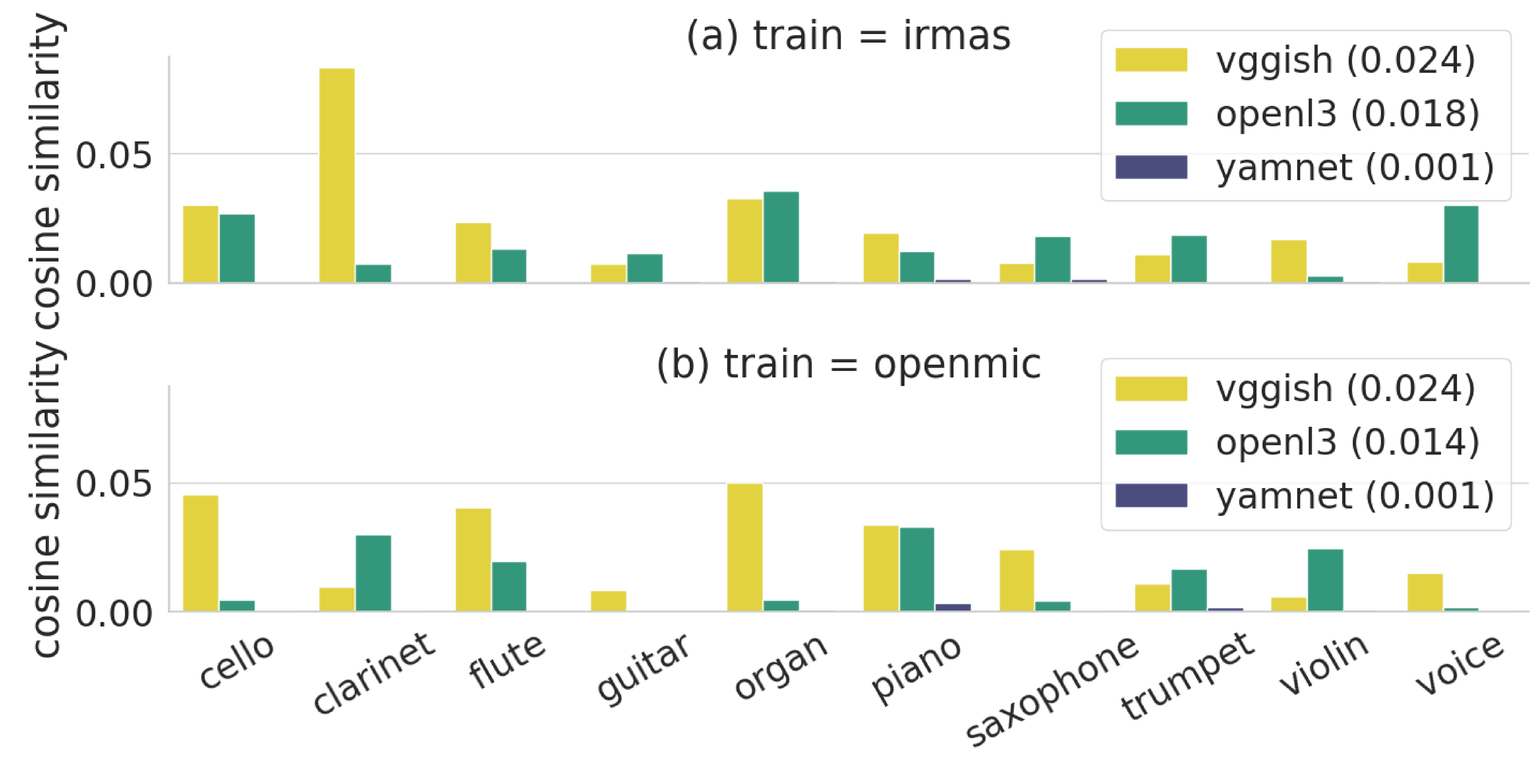



## 2. Quantifying bias effects

- Cosine similarity between domain separation and instrument recognition $c(\boldsymbol{w},\boldsymbol{v}) = \frac{\langle \boldsymbol{w},\boldsymbol{v}\rangle}{\|\boldsymbol{w}\|\times\|\boldsymbol{v}\|}$




$\boldsymbol{v}$: Instrument separation direction (binary logistic regression)
$\boldsymbol{w}$: domain separation direction (linear discriminant analysis, LDA)




## 3. Bias correction

### ➢ Single bias correction (LDA)

- Project out undesirable separation direction:

$$\boldsymbol{x}_{\mathrm{P}} := (\mathbf{I} - \boldsymbol{w}\boldsymbol{w}^{\mathrm{T}})\boldsymbol{x}$$

$\boldsymbol{x}$: original embedding, $\boldsymbol{x}_{\mathrm{P}}$: processed embedding, $\mathbf{I}$: unit matrix

### ➢ Multiple bias correction (mLDA)

- Extract domain separation direction in genre-space: $\boldsymbol{w}_{\boldsymbol{g}}$
- Collect $\boldsymbol{w}_{\boldsymbol{g}}$ into a matrix $W$
- Factorizing $W$ by reduced singular vector decomposition (SVD) for orthogonal basis $V$: $W = U\Sigma V^{\mathrm{T}}$

$$\boldsymbol{x}_{\mathrm{P}} := (\mathbf{I} - VV^{\mathrm{T}})\boldsymbol{x}$$

### ➢ Nonlinear bias correction (KLDA, mKLDA)

- Generalize by explicit kernel approximation

$$\langle f(\boldsymbol{w}), f(\boldsymbol{v})\rangle \approx k(\boldsymbol{w},\boldsymbol{v})$$

$f$: explicit nonlinear transformation

| Debiasing method | Global bias correction: Within-domain IR-IR | Global: Within-domain OP-OP | Global: Cross-domain OP-IR | Global: Cross-domain IR-OP | Class-wise bias correction: Within-domain IR-IR | Class-wise: Within-domain OP-OP | Class-wise: Cross-domain OP-IR | Class-wise: Cross-domain IR-OP |
|---|---|---|---|---|---|---|---|---|
| *VGGish* | *91.6* | *87.95* | *82.82* | *83.81* | *91.60* | *87.95* | *82.82* | *83.81* |
| VGGish-LDA | 91.60 | 87.99 | 82.99 (**+0.18**) | 83.82 (0.0) | 91.60 | 87.94 | 82.93 (**+0.12**) | 83.85 (+0.03) |
| VGGish-mLDA | 91.45 | 87.98 | 82.70 (-0.11) | 83.30 (-0.51) | 91.56 | 87.87 | 83.13 (**+0.31**) | 83.66 (-0.16) |
| VGGish-K | 92.24 | 88.08 | 82.57 (-0.25) | 83.67 (-0.14) | 92.24 | 88.08 | 82.57 (-0.25) | 83.67 (-0.14) |
| VGGish-KLDA | 92.24 | 88.08 | 82.58 (-0.24) | 83.67 (-0.14) | 92.22 | 88.07 | 82.70 (-0.12) | 83.78 (-0.04) |
| VGGish-mKLDA | 92.22 | 88.15 | 82.42 (-0.39) | 83.70 (-0.11) | 92.26 | 88.08 | 82.70 (-0.11) | 83.76 (-0.05) |
| *OpenL3* | *93.26* | *87.16* | *80.56* | *80.13* | *93.26* | *87.16* | *80.56* | *80.13* |
| OpenL3-LDA | 93.26 | 87.16 | 80.56 (+0.01) | 80.15 (+0.02) | 93.24 | 87.18 | 80.59 (+0.04) | 80.38 (**+0.26**) |
| OpenL3-mLDA | 93.11 | 87.16 | 80.67 (**+0.12**) | 79.93 (-0.20) | 93.09 | 87.23 | 80.57 (+0.02) | 80.62 (**+0.50**) |
| OpenL3-K | 93.89 | 87.91 | 79.46 (-1.09) | 81.23 (**+1.11**) | 93.89 | 87.91 | 79.46 (-1.09) | 81.23 (**+1.11**) |
| OpenL3-KLDA | 93.89 | 87.84 | 79.03 (-1.53) | 81.23 (**+1.11**) | 93.96 | 87.91 | 79.99 (-0.57) | 81.79 (**+1.66**) |
| OpenL3-mKLDA | 93.88 | 87.88 | 79.56 (-1.00) | 81.20 (**+1.07**) | 94.04 | 87.83 | 79.97 (-0.59) | 81.32 (**+1.19**) |
| *YAMNet* | *94.65* | *89.74* | *85.01* | *85.47* | *94.65* | *89.74* | *85.01* | *85.47* |
| YAMNet-LDA | 94.65 | 89.74 | 85.01 (0.0) | 85.47 (0.0) | 94.65 | 89.74 | 85.02 (0.0) | 85.47 (0.0) |
| YAMNet-mLDA | 94.65 | 89.74 | 85.01 (0.0) | 85.47 (0.0) | 94.65 | 89.74 | 85.02 (0.0) | 85.46 (0.0) |
| YAMNet-K | 93.83 | 89.24 | 85.87 (**+0.86**) | 84.56 (-0.91) | 93.83 | 89.24 | 85.87 (**+0.86**) | 84.56 (-0.91) |
| YAMNet-KLDA | 93.83 | 89.23 | 85.87 (**+0.86**) | 84.56 (-0.91) | 93.63 | 89.24 | 86.00 (**+0.99**) | 84.76 (-0.70) |
| YAMNet-mKLDA | 93.79 | 89.19 | 85.72 (**+0.71**) | 84.43 (-1.04) | 93.79 | 89.34 | 85.53 (**+0.51**) | 84.60 (-0.87) |

code: github. com/changhongw/audio-embedding-bias

## Conclusion

- Training regime of embeddings, e.g. self-supervised training is more prone to overfitting a domain
- Class-vocabulary alignment between source and downstream task
- Require identifying populations to treat as equivalent