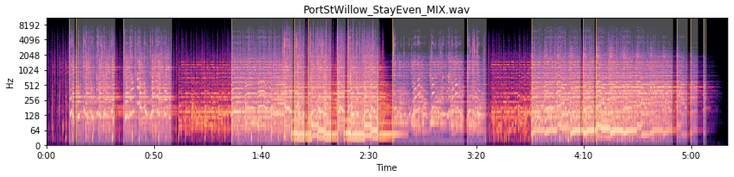
# VGGish embeddings and perceptual features
## *for Singing Voice Detection*

**Shayenne Moura** - (shayenne.moura@usp.br) - Computer Music Research Group, Universidade de São Paulo

## 1. Objective

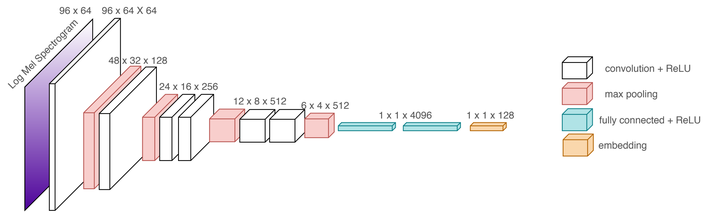
Explore results using learned representations against perceptually-motivated features




Classify polyphonic audio segments as singing/non-singing

## 2. VGGish

- 128-dimensional audio features extracted at 1Hz
- VGG-inspired acoustic model in Hershey et. al. (2017)
- Trained on a preliminary version of YouTube-8M



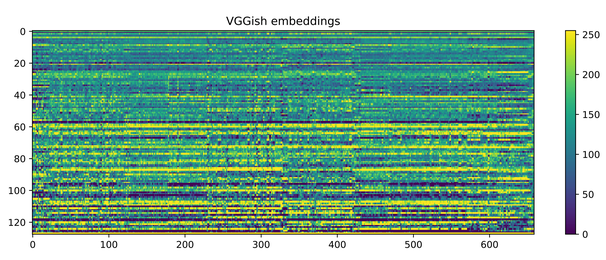

## 3. Dataset

- 61 songs containing singing voice from MedleyDB
- 10 splits for train/test: 70%/30%
- Artist conditional split

## 4. Method

Comparing singing voice classification using VGGish versus using perceptually-motivated features

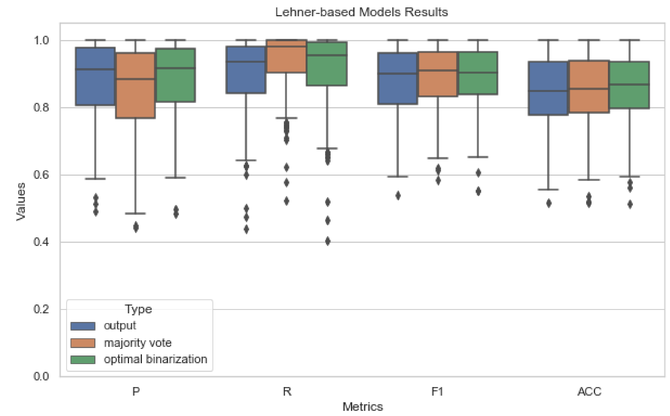
- *Target Sources*: female and male singer, vocalists, and choir
- *Features*: VGGish *versus* MFCC, VV, Fluct, SF, and SC
- *Classifiers*: Random Forest
- *Outputs*: Original, Majority Vote, and Optimal Binarization
- *Evaluation*: quantitative and qualitative

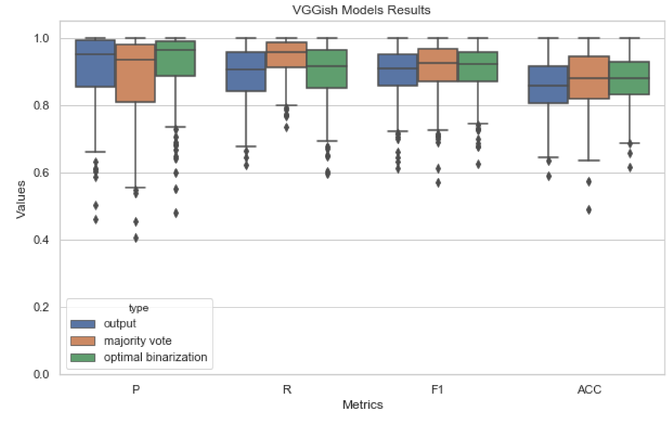



## 5. Results

Metrics related to the optimal binarization output for each genre on test set

| genre | Classical | Jazz | Musical Theatre | Pop | Rock | Singer/ Songwriter | World/ Folk |
|---|---|---|---|---|---|---|---|
| | Perceptual features models | | | | | | |
| ACC | **0.94** | 0.94 | **0.93** | 0.83 | 0.83 | 0.80 | 0.97 |
| P | 0.97 | 0.93 | 0.94 | 0.86 | 0.84 | 0.86 | 0.99 |
| *R* | 0.94 | 0.98 | 0.99 | 0.93 | 0.88 | 0.87 | 0.97 |
| *F1* | 0.95 | 0.95 | 0.96 | 0.89 | 0.86 | 0.85 | 0.98 |
| | VGGish embeddings models | | | | | | |
| *ACC* | 0.92 | **0.97** | 0.88 | **0.88** | **0.86** | **0.84** | **0.99** |
| *P* | 0.96 | 0.97 | 0.97 | 0.92 | 0.90 | 0.90 | 0.99 |
| *R* | 0.94 | 0.98 | 0.91 | 0.91 | 0.87 | 0.87 | 0.99 |
| *F1* | 0.95 | 0.98 | 0.93 | 0.92 | 0.88 | 0.88 | 0.99 |




## 6. Conclusions

VGGish features have comparable classification accuracy relative to perceptual features without specialization for voice recognition

*Future directions*:

- Add a pitch recognition phase to perform singing voice transcription

The author acknowledges the support of CNPq and the WiMIR participation grant.