Singing Voice Detection
Using VGGish Embeddings

Shayenne Moura, Marcelo Queiroz
Computer Music Research Group - University of São Paulo (Brazil)
Acknowledgments

Thank you all for making this work possible and visible!
Music Information Retrieval

interdisciplinary science of retrieving information from music

musicology, psychology, signal processing, informatics, machine learning, computational intelligence or some combination of these.
Goal!

Singing Voice Detection

Classify polyphonic audio segments as singing/non-singing
Target Sources
MedleyDB: A Dataset of Multitrack Audio for Music Research

Welcome to the companion website for MedleyDB, a dataset of annotated, royalty-free multitrack recordings. MedleyDB was curated primarily to support research on melody extraction, addressing important shortcomings of existing collections. For each song we provide melody 10 annotations as well as instrument activations for evaluating automatic instrument recognition. The dataset is also useful for research on tasks that require access to the individual tracks of a song such as source separation and automatic mixing.
Approaches for ML tasks with audio

Common approach
Approaches for ML tasks with audio

Transfer learning approach

[Diagram showing the transfer learning approach with a DCNN and EMBEDDINGS leading to a prediction output]
Mel Frequency Cepstral Coefficients (MFCC)

A handcrafted audio representation feature commonly used for voice related tasks MFCC
VGGish network

- VGG-inspired acoustic model in Hershey et. al. (2017)
- Trained on a preliminary version of YouTube-8M
- Embeddings: 128-dimensional audio features extracted at 1Hz
A feature obtained from data using deep learning which theoretically preserves relevant information.
Support Vector Machine
Random Forest
Evaluation

- Quantitative: classification accuracy
- Qualitative: musical content
Method

**Target Sources:** female singer, male singer, vocalists, and choir

**Dataset:** MedleyDB multitrack

**Features:** VGGish embeddings and MFCC

**Classifiers:** SVM and Random Forest

**Evaluation:** quantitative and qualitative
Preliminary Results
Quantitative Evaluation: VALIDATION

Classification Accuracy on Validation Set with SVM

- MFCC
- VGGish

C-value
Quantitative Evaluation: 10 SPLITS
Quantitative Evaluation: VALIDATION

Classification Accuracy on Validation Set with Random Forest

- MFCC
- VGGish
Quantitative Evaluation: TEST

Classification Accuracy on Test Set

Classifier

Baseline
Best SVM
Best RF

MFCC
VGGish

Accuracy
Qualitative Evaluation

**MFCC**

**VGGish**
Observations

- Some sound sources are frequently present when singing voice is active
- Most confused: Piano and guitar with drums
- Synthesizers are detected as singing voice
Conclusions

VGGish features increase classification accuracy by 8 points compared to MFCC

Future directions
- Evaluate effect of more training data using VGGish features
- Combine VGGish features with other features
- Evaluate using cross validation
Thanks!

Any questions?

Come to my poster!

shayenne.moura@usp.br