A Simple Fusion of Deep and Shallow Learning for Acoustic Scene Classification

Eduardo Fonseca, Rong Gong and Xavier Serra

name.surname@upf.edu

Music Technology Group, UPF, Barcelona, Spain

audio commons

upf. Universitat Pompeu Fabra Barcelona

MTG Music Technology Group

Motivation

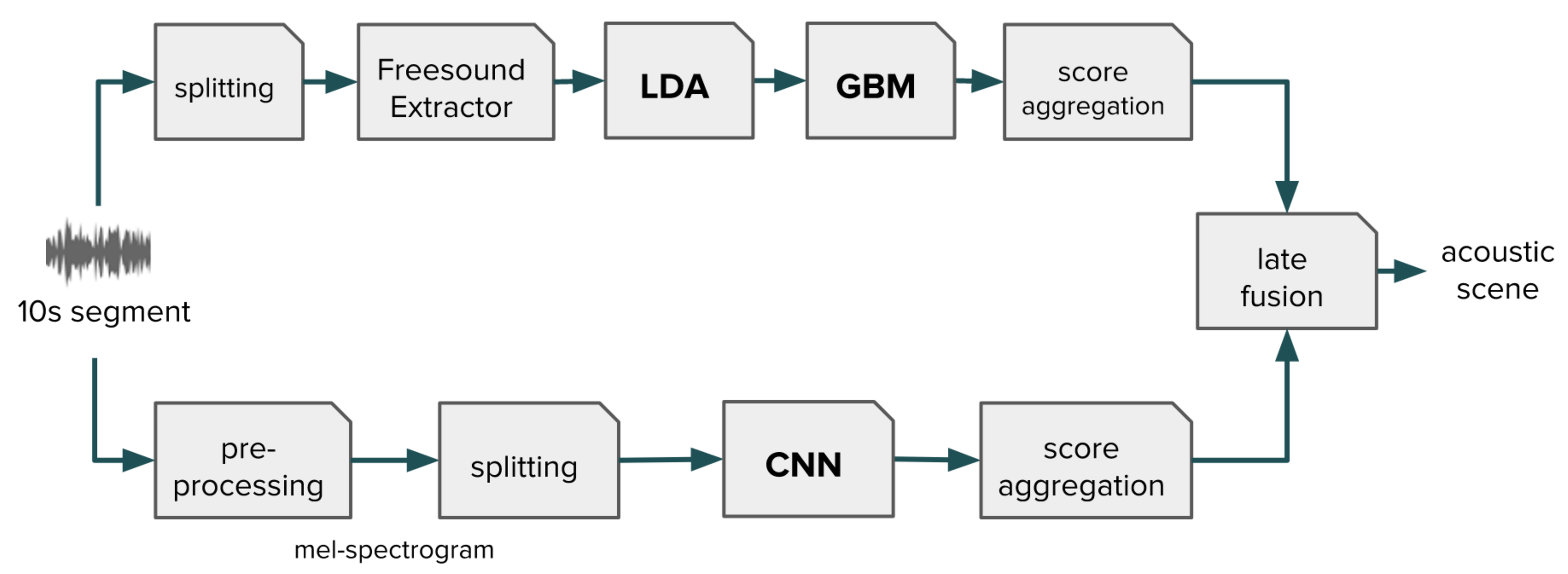
- **Task:** Recognize the environment in which an audio recording has been made
- **Applications:**
  - automatic description
  - context-aware applications
  - intelligent wearable devices

Approaches

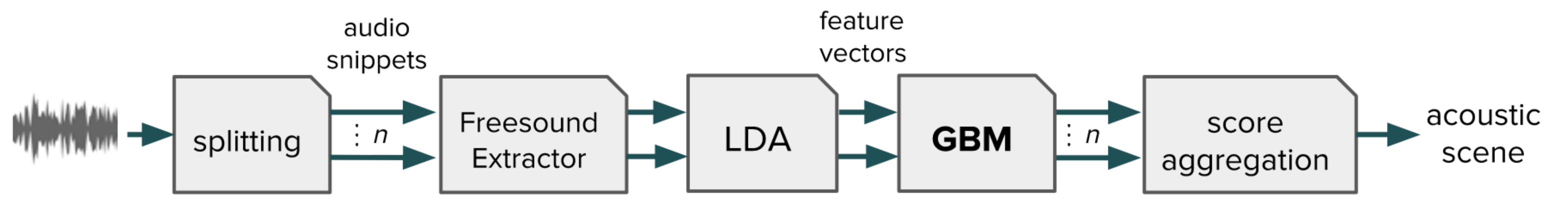
- **Feature engineering**:
  - feature extraction
  - classifier
- **Data driven:**
  - learning representations

How about combining both approaches for ASC?

Proposed System



Gradient Boosting Machine (GBM)

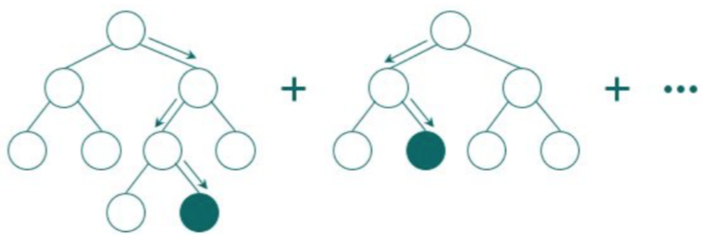


- FreesoundExtractor [1] by ESSENTIA

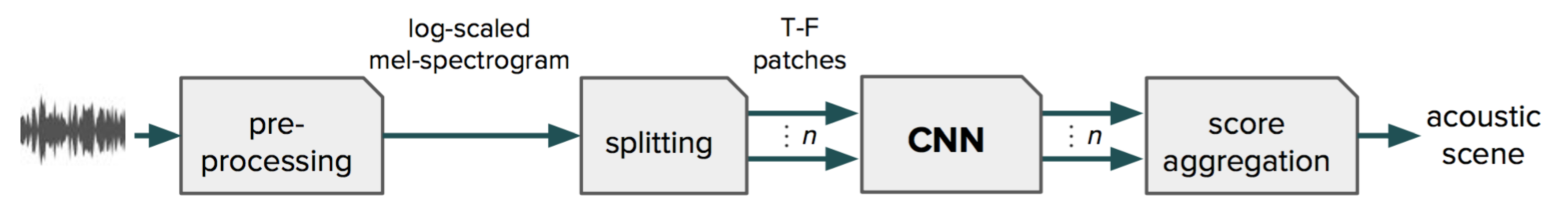
Table 1: Selected features extracted by *FreesoundExtractor*.

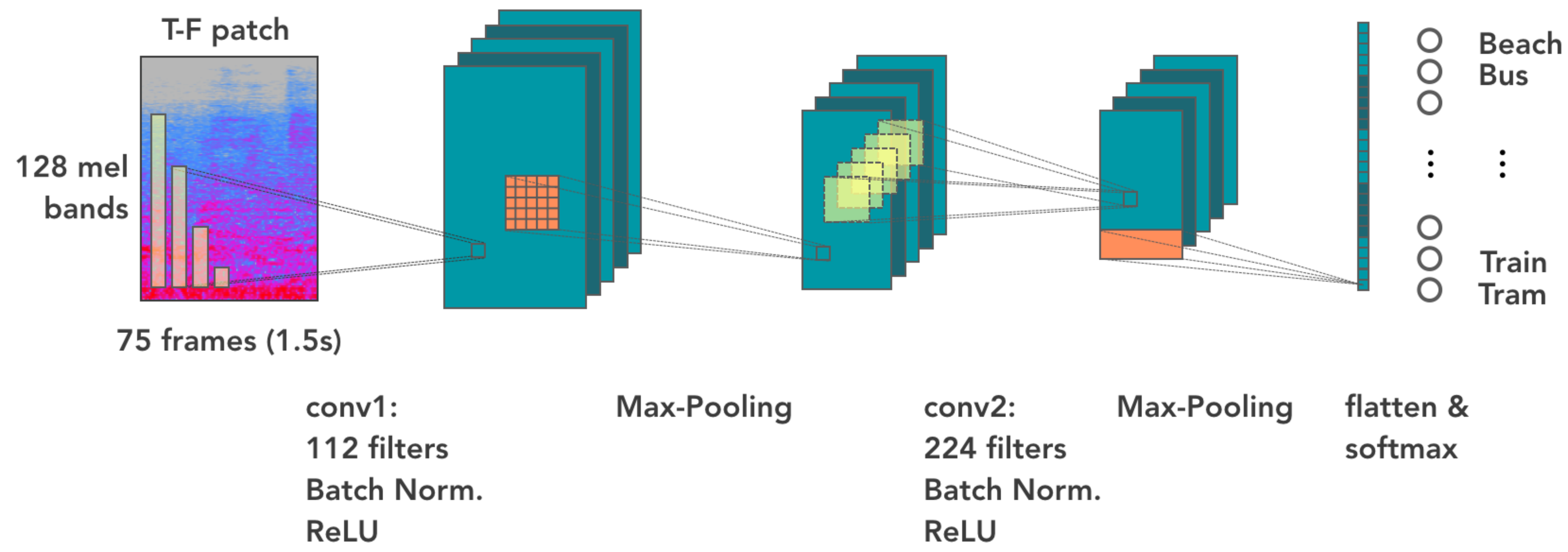
Feature name	Dim	Feature name	Dim
Bark bands energy	32	Tonal features	3
ERB bands energy	23	Pitch features	3
Mel bands energy	45	Silence rate	3
MFCC	13	Spectral features	32
HPCP	38	GFCC	13

- Gradient Boosting Machine
  - multiple decision trees
  - added iteratively
  - LightGBM [2]

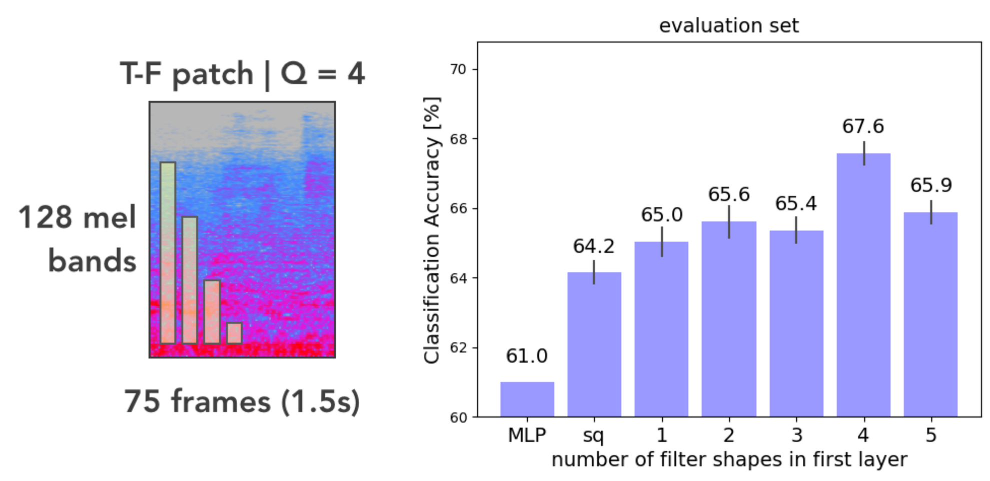


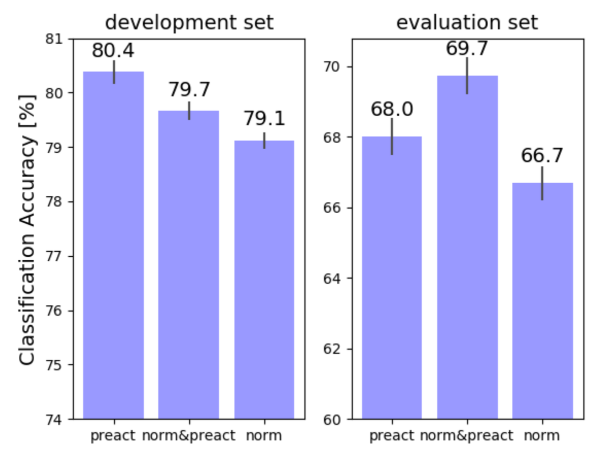
Convolutional Neural Networks (CNN)





- Design of convolutional filters:
  - **spectro**-temporal patterns for ASC?
  - multiple **vertical** filter shapes
- Pre-activation [5]:
  - adding Batch Norm & ReLu **before** first convolution

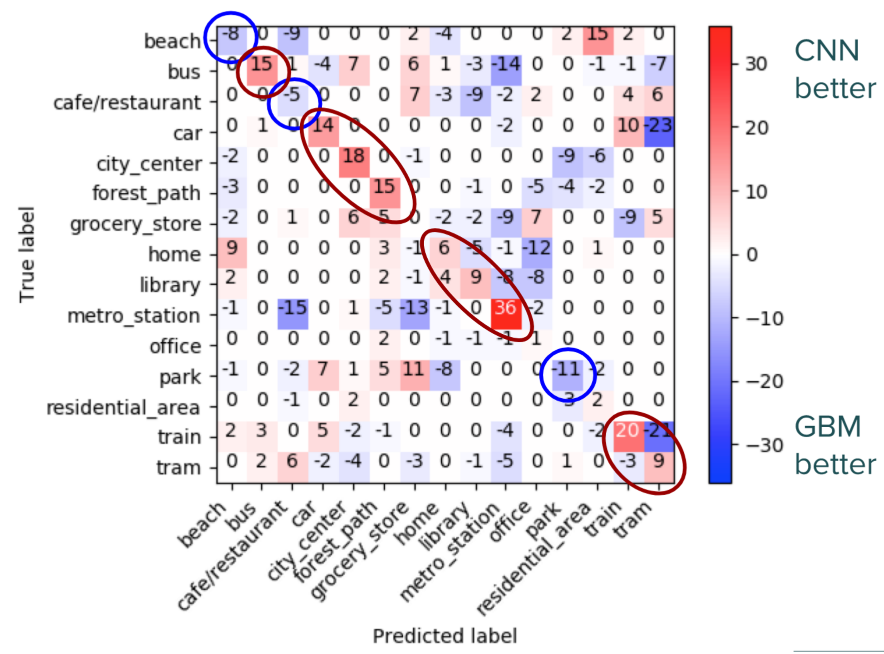


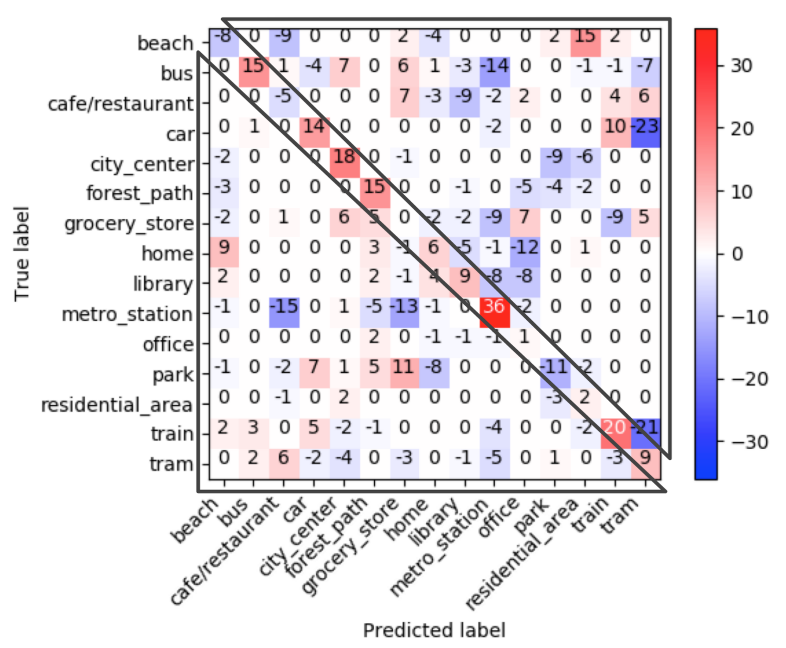


Dataset

- **TUT Acoustic Scenes 2017** [3]:
  - 15 classes with audio segments of 10s
  - **development**: 312 segments/class & 4-fold cross-validation setup
  - **evaluation**: 108 segments/class
  - mismatch between dev/eval due to different recording conditions

Confusion Matrix Analysis: CNN - GBM





Results

Table 8. Acoustic scene classification accuracy (%).

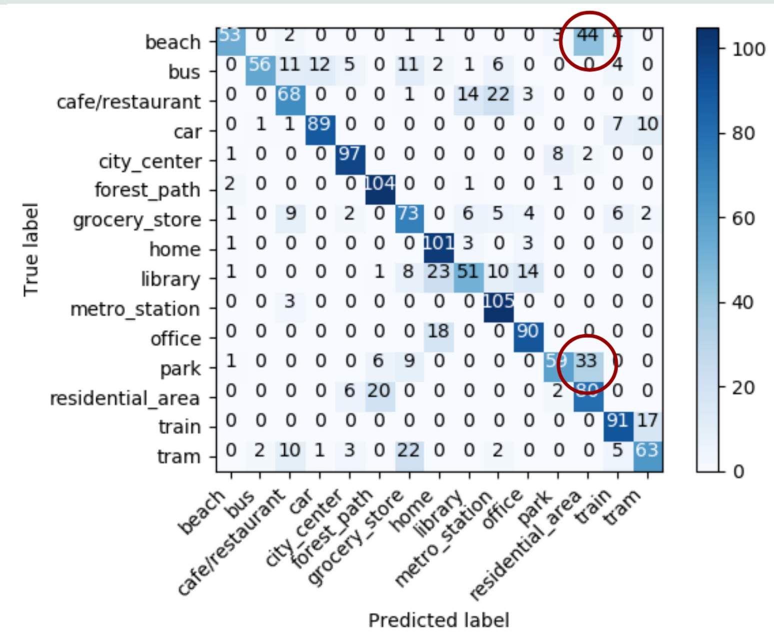
System	dev set*	eval set**
MLP baseline	74.8	61.0
CNN	79.7	69.7
GBM	81.1	63.6
Fusion	83.3	**72.8**

* 4-fold cross-validation
** training on full dev set

- Our method is still outperformed by some submissions to Detection and Classification of Acoustic Scenes and Events, 2017, Task 1 [4]
- But the proposed approach is simpler in comparison:
  - GANs, ensembles of 4 or more systems, data augmentation, etc.

Late Fusion

- CNN: softmax activation values
- GBM: prediction probabilities
- Late fusion approach:
  - means + argmax
  - *stacking* with **logistic regression**



Conclusions & Future work

- Simplicity of models:
  - GBM + out-of-box feature extractor
  - CNN + domain knowledge
- Simple late fusion approach

providing complementary info

- How to improve?
  - individual models & measures against overfitting
  - fusion approach: join (learned) representations

References & Resources

[1] http://essentia.upf.edu/documentation/freesound_extractor.html

[2] https://github.com/Microsoft/LightGBM

[3] Mesaros et al. TUT database for acoustic scene classification and sound event detection. EUSIPCO, 2016

[4] http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/

[5] Han et al. CNNs with binaural representations and background substraction for ASC. DCASE 2017