Experiments on the DCASE Challenge 2016: Acoustic Scene Classification and Sound Event Detection in Real Life Recordings

Benjamin Elizalde[1], Anurag Kumar[1], Ankit Shah[2], Rohan Badlani[3], Emmanuel Vincent[4], Bhiksha Raj[1], Ian Lane[1]

[1]CMU, [4]Inria, [2]NIT Surathkal, [3]BITS

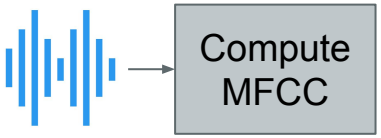
Overview

1. Task 1: Acoustic Scene Classification
    a. Approach
    b. Results on challenge
    c. Conclusions
2. Task 3: Sound Event Detection in Real Life Recordings
    a. Approach
    b. Results on challenge
    c. Conclusions

Task 1: Acoustic Scene Classification Approach

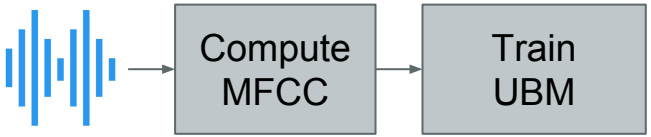
Multiclass classification of scenes using high-level features and multiple kernels for SVM

Approach



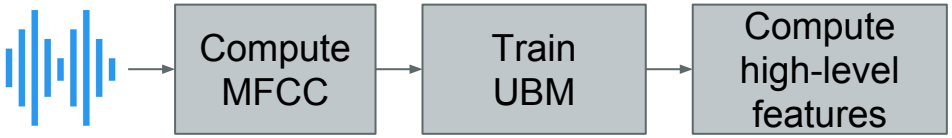
- Merge both audio channels into one.
- Compute MFCCs with 20 dim. +D+DD, 30 ms window and 50% overlap.

Approach



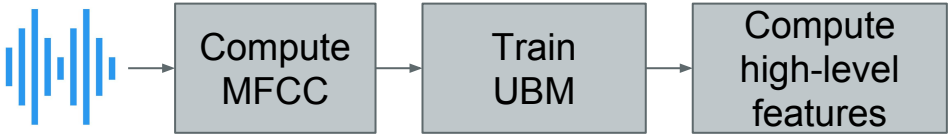
- Train multiple GMM-UBMs of different component sizes (64,128,256,512) over the MFCCs of all the training data.

Approach



- The high-level features try to capture the distribution of the scene's MFCCs.
- Compute features, labeled as Beta, by adapting the UBM to the MFCCs using MAP* for each component size.
  - Four supervectors with stacked means.
  - Four supervectors with stacked means and variance.

Approach



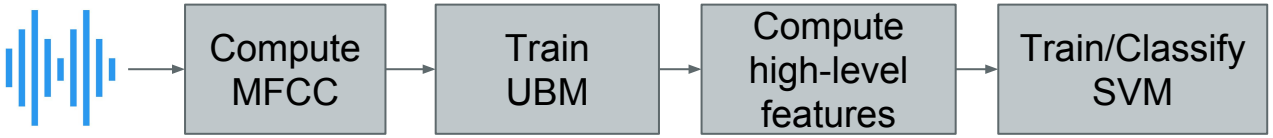
MAP*

$$\hat{\boldsymbol{\mu}}_k = \frac{n_k}{n_k + r} E_k(\boldsymbol{x}) + \frac{r}{n_k + r} \boldsymbol{\mu_k}$$

$$\hat{\boldsymbol{\sigma}}_k = \frac{n_k}{n_k + r} E_k(\boldsymbol{x}^2) + \frac{r}{n_k + r} (\boldsymbol{\sigma}_k^2 + \boldsymbol{\mu}_k^2) - \hat{\boldsymbol{\mu}}_k^2$$

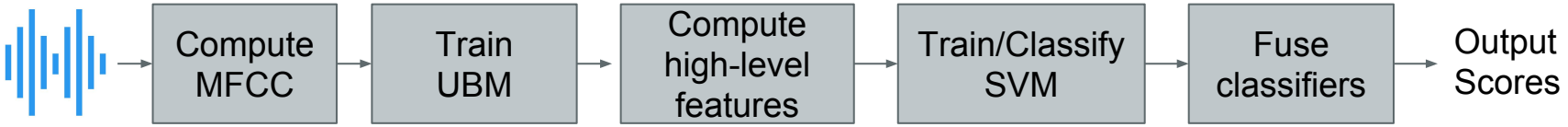
- The high-level features try to capture the distribution of the scene's MFCCs.
- Compute features, labeled as Beta, by adapting the UBM to the MFCCs using MAP* for each component size.
  - Four supervectors with stacked means.
  - Four supervectors with stacked means and variance.

Approach



- Train SVM with Linear and RBF kernels for each of the 8 supervectors, thus 16 systems in total.

Approach



- Fuse classifiers based on maximum prediction vote.

Task 1: Acoustic Scene Classification Results

Overall results on the challenge

	Accuracy %	
Run	**Development**	**Evaluation**
Baseline	72.5	77.2
Best	89.9[H]	89.7[H]

Overall results on the challenge

	Accuracy %	
Run	**Development**	**Evaluation**
Baseline	72.5	77.2
Best	89.9[H]	89.7[H]
7th → Our approach	78.9	85.9

Class-wise results

Conclusions

- GMM-based representations capture well the content of long scene recordings.
  - For example, i-vectors which got the best performance.

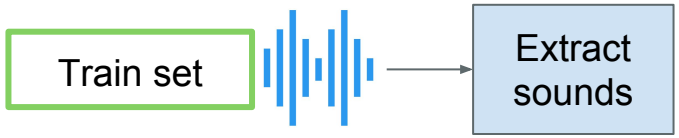
- Fusion of different beta vectors and kernels improved performance.

- We tried Alpha features which are histograms, similar to Bag-of-Audio-Words representations in combination with 7 kernels, but didn't help.

Task 3: Sound Event Detection in Real Life Recordings
Approach

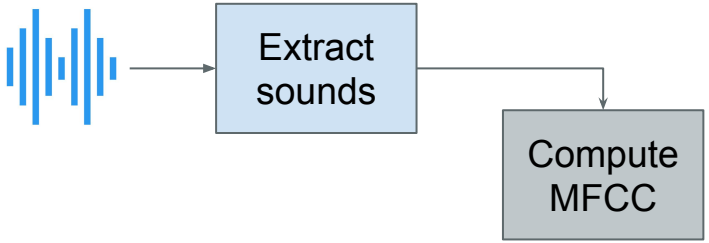
Multiclass classification of sound events in scenes, including a generic sound event and time-perturbation

Approach: Standard



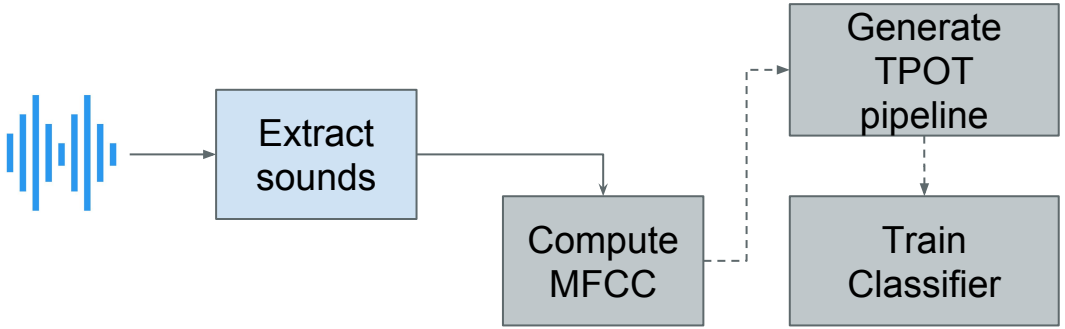
- Create a pipeline for each scene: Home and Residential Area.
- For training: extract the sound events from the scene recordings to have isolated files.

Approach: Standard



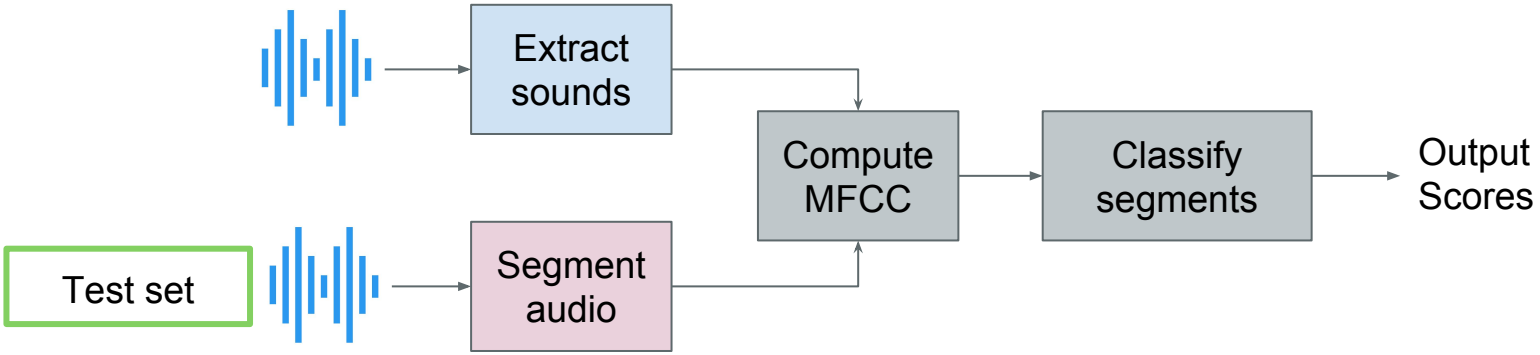
- Merge both audio channels into one..
- Compute MFCCs with 13 dim. +D+DD, 30 ms window and 50% overlap.
    - Also explored: Separable Gabor Filter Banks, Scatnet and Stacked with and without PCA.

Approach: Standard



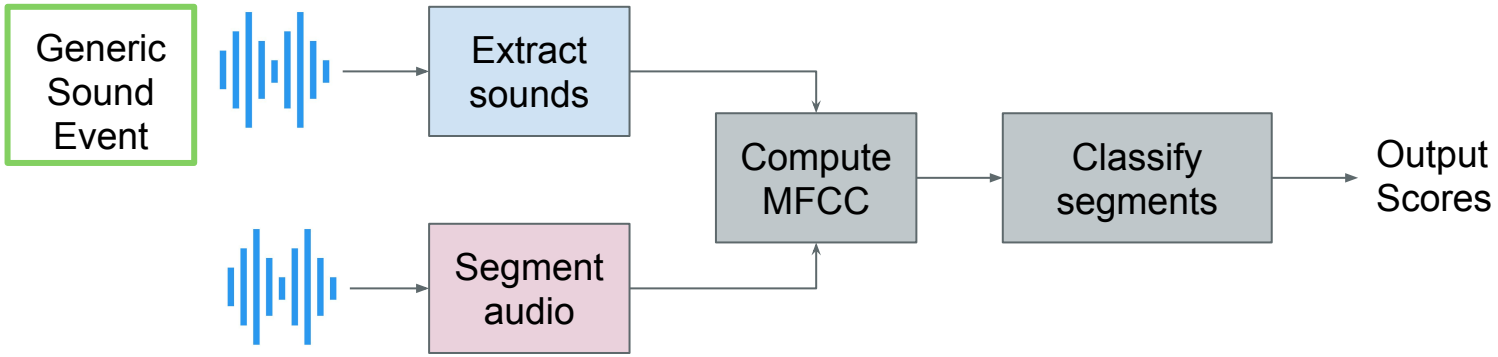
- Tpot is a toolbox that automatically creates machine learning pipelines.
- Tpot (v4) tests and optimizes 12 classifiers such as SVMs, Decision Trees, K-Neighbors, Logistic Regression, etc.
- Main selections: Random Forest, Logistic Regression, Gradient Boosting

Approach: Standard



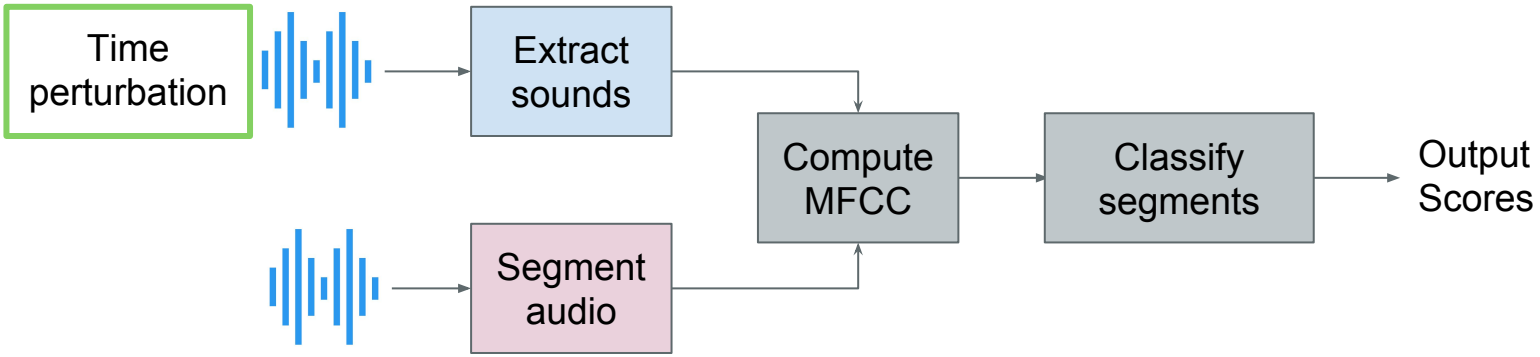
- For testing: the scene recording was trimmed into contiguous one-second segments.
- Compute MFCCs and then classify each segment.

Approach: Generic



- The generic class was created to address out-of-vocabulary sound events.
- The class is specific to each scene.
- The number of audio files corresponds to the average of files per class.

Approach: Perturbation



- The time-perturbation is intended to address robustness and compensate for small data.
- The signal was speeded up to 2x and slowed down to 0.3x to output 13 different versions of the original file.

Task 3: Sound Event Detection in Real Life Recordings
Results

Results on challenge

	Overall Segment-based Error Rate	
Run	Development	Evaluation
Baseline	0.91	0.87
Best	0.91^{S}	0.80^{S}

Results on challenge

	Overall Segment-based Error Rate	
Run	**Development**	**Evaluation**
Standard (S)	0.84	1.05
Baseline	0.91	0.87
Best	0.91^{S}	0.80^{S}

Results on challenge

	Overall Segment-based Error Rate	
Run	**Development**	**Evaluation**
Standard (S)	0.84	1.05
S+Generic	0.81	1.07

Baseline	0.91	0.87
Best	0.91^{S}	0.80^{S}

Results on challenge

	Overall Segment-based Error Rate	
Run	**Development**	**Evaluation**
Standard (S)	0.84	1.05
S+Generic	0.81	1.07
9th → S+Generic+Perturbation	0.76	0.961

Baseline	0.91	0.87
Best	0.91^{S}	0.80^{S}

Results on challenge

	Overall Segment-based Error Rate	
Run	**Development**	**Evaluation**
Standard (S)	0.84	1.05
S+Generic	0.81	1.07
9th → S+Generic+Perturbation	0.76	0.961
*S+Perturbation	0.74	0.96

Baseline	0.91	0.87
Best	0.91^{S}	0.80^{S}

Results on challenge

	Segment-based Error Rate (Class-based average)	
Run	Home	Residential
Baseline	0.97	1.31
Best	0.98^{S}	0.98^{D}

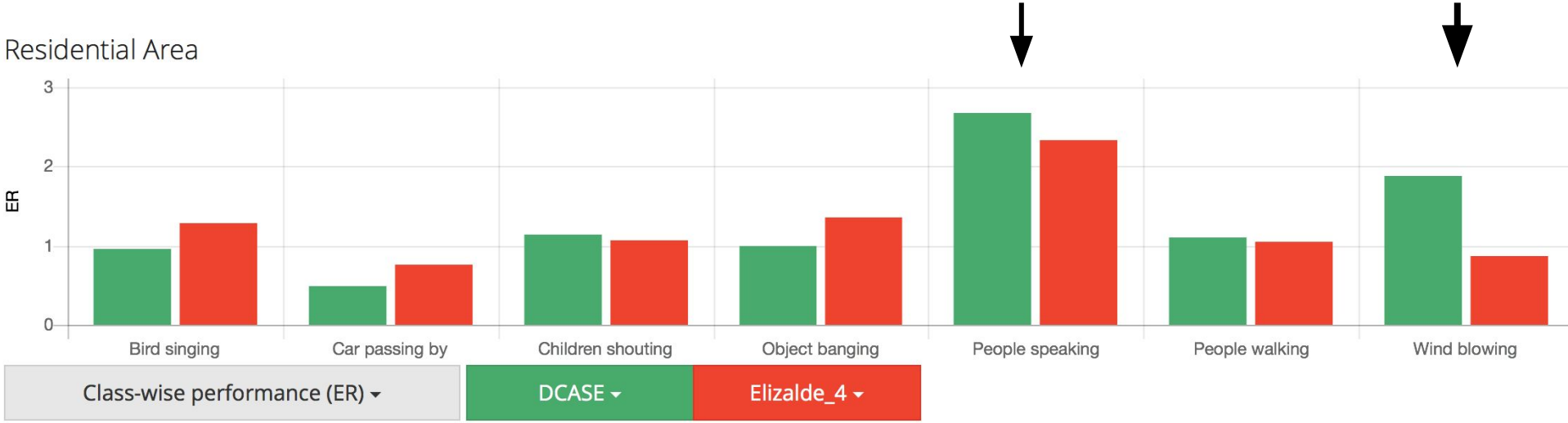
Results on challenge

	Segment-based Error Rate (Class-based average)	
Run	**Home**	**Residential**
Standard (S)	1.12	0.99
S+Generic	1.92	2.08
S+Generic+Perturbation	1.52	1.24
*S+Perturbation	1.03	**0.89**

Baseline	0.97	1.31
Best	0.98^{S}	0.98^{D}

Class-wise results on Home

Class-wise results on Residential Area



Conclusions

- Classifier exploration improved performance, unlike feature exploration.
- The generic sound event could be an option for out-of-vocabulary-words.
- Time-perturbation significantly improved our system.

- Large-scale sound event detection using weakly-labeled data.

Questions?

bmartin1@andrew.cmu.edu

Per-scene results on the challenge

Scene	Rank - ex aequo
Residential Area	1 tiers
Forest, Grocery Store, Metro, Office	2
Train, Home, City Center, Bus	3
Beach, Cafe, Car, Tram	4
Library, Park	5

Sound-events results on challenge

S+Generic+Perturbation		
Home	**Residential**	**Rank ...**
Cupboard	Wind Blowing	1
Glass Jingling	Children Shouting	2
...	...	...
Water Tap Running	People Speaking	9
Object Impact	Bird Singing	13