Audio Event Detection for Automatic Scene Recognition

Xu Xun

Department of Computer Science and Engineering
Shanghai Jiao Tong University

June 25, 2015

Outline

Outline

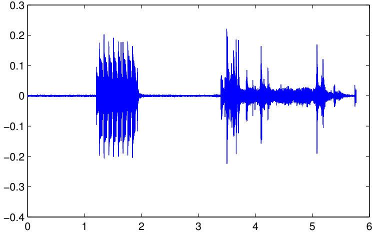
Problem Description

In this project, our problem is to recognize a scene where an audio is recorded.

play pause resume stop

In this project, our problem is to recognize a scene where an audio is recorded.

play pause resume stop



$\Rightarrow$ *office*

Problem Description

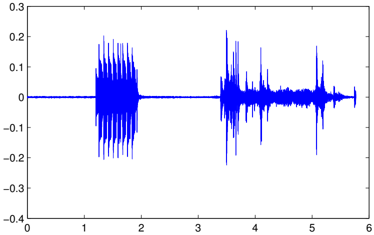
- Scene

  An acoustic environment, like *office*, *bathroom*, etc.
- Event

  A more short, primitive sound, like *phone*, *printer*, etc.

Our approach

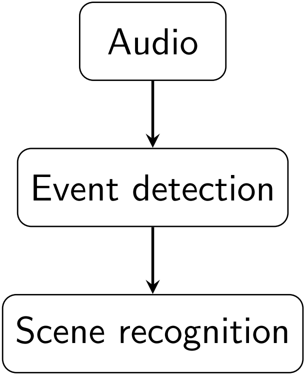
Our approach is to detect the audible events in a clip.
Then infer the scene from the detected events.



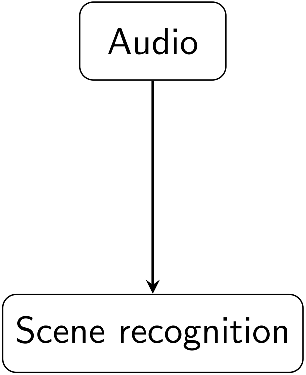
$\Rightarrow$ *phone, printer* $\Rightarrow$ *office*

Our Approach vs. Other Approaches

Our approach:



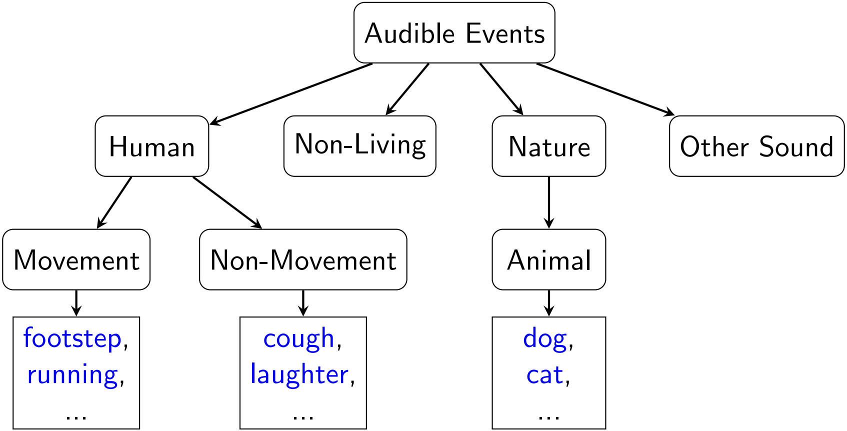
Other approaches:



Outline

Audible Event Taxonomy

We labelled common audible events into 4 classes.
There are 120 events in total.



We download the audio data for events from Sound Search Engines (SSEs).
For example, when we query "cough" in SSE:

We download clips from 1 second to 60 seconds.

We first use Minimum Statistics to calculate the noise spectrum and subtract it from the input signal.
Then we extract Mel-Frequency Cepstral Coefficients (MFCCs) from denoised signal.

Preprocess and Feature Extraction

We first use Minimum Statistics to calculate the noise spectrum and subtract it from the input signal.
Then we extract Mel-Frequency Cepstral Coefficients (MFCCs) from denoised signal.

Denoised signal
↓
Framing
↓
Fast Fourier Transform (FFT)
↓
Mel filtering
↓
Extract spectral envelope
↓
MFCCs

Preprocess and Feature Extraction

We first use Minimum Statistics to calculate the noise spectrum and subtract it from the input signal.
Then we extract Mel-Frequency Cepstral Coefficients (MFCCs) from denoised signal.

- Denoised signal
- ↓ Framing
- ↓ Fast Fourier Transform (FFT)
- ↓ Mel filtering
- ↓ Extract spectral envelope
- ↓ MFCCs

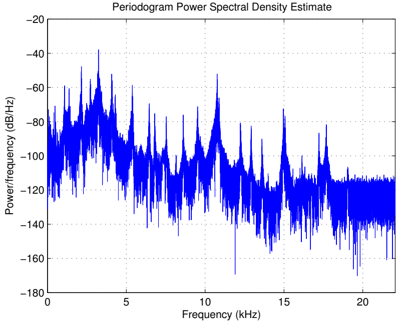


Figure: Audio in frequency domain

We use features to train Gaussian Mixture Models (GMMs).
The training is done by Expectation-Maximization (EM) algorithm.

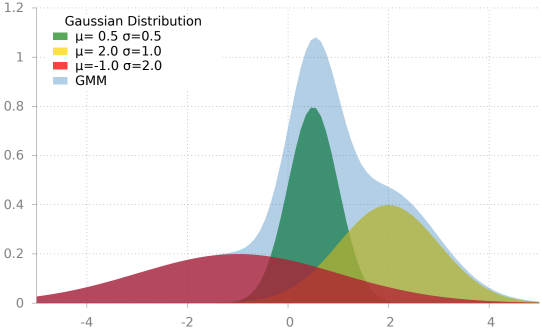


Figure: A GMM with three components

Outline

Scene-Event Relation Mining

To get the relation between scenes and audible events,
we match the context in a script with our predefined audible events.

INT. LEONARD'S BATHROOM - Night Leonard turns on the light, revealing a shower, toilet and sink. He removes toiletries from the grocery bag and places them inside.

Scene-Event Relation Mining

Table: An example of scene-event map

Scene	Top 10 events ranked by TF-IDf
bathroom	running+water, toilet, faucet, toothbrush, shower, drawer, drain, talk, paper, bowl
beach	seagull, sand, boat, talk, wave, sea, car, laughter, drink, wood, running
forest	tree, wood, dirt, talk, running, bird, river, car, leaf, grass, wind
kitchen	drawer, cutlery, microwave, dish, kettle, talk, bowl, phone, toaster, running+water
street	car, truck, subway, talk, traffic, engine, siren, phone, running, laughter

Audio Segmentation

Scene-Event map is used when we have detected the events.
We need to cut testing clips into segments for event detection.

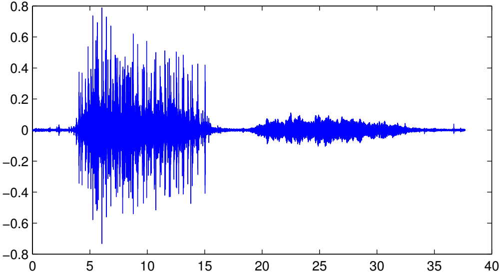


Figure: A example audio clip

We use frame energy and frequency to filter out silence and noise.

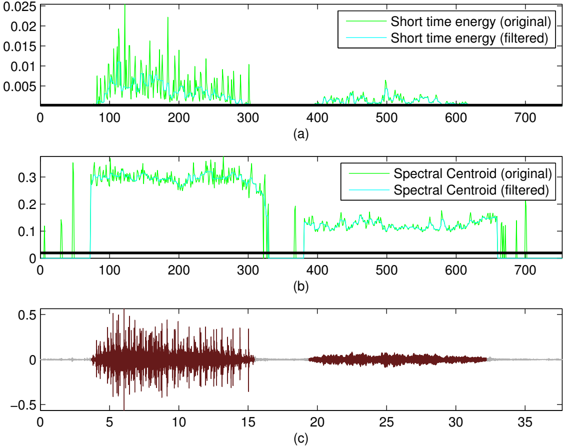
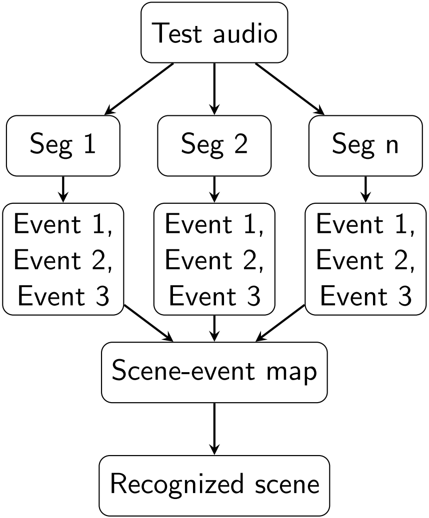


Figure: A segmentation example

For each segment, we evaluate it with our trained GMMs.
We choose the top three detected events for scene voting.



Outline

Gaussian Mixture Model distribution:

$$P(\mathbf{x}|\pi, \mu, \Sigma) = \sum_{k=1}^{M} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k), \tag{1}$$

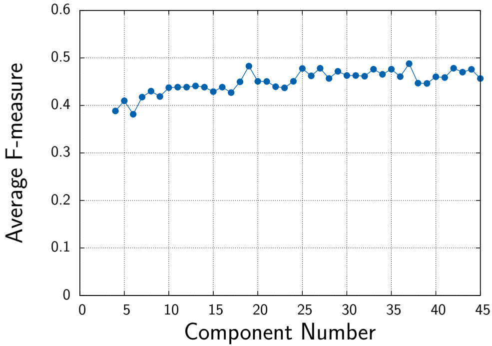


Figure: F-measure for different component number

After comparing F-measure and running time, we choose 18 as our component number.

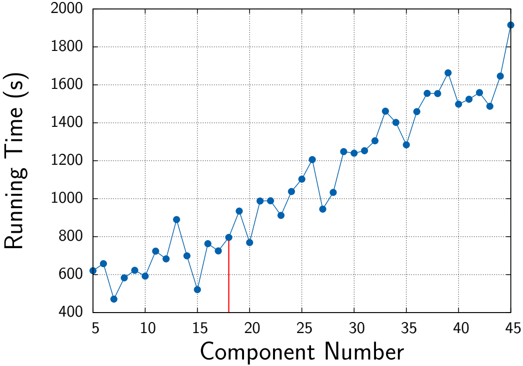


Figure: Running time for different component number

Scene Recognition Evaluation

In scene recognition, we choose 10 scenes, each scene has 10 clips. Accuracy for other 4 systems are calculated using 5-fold cross validation. Our system achieve an accuracy of 57%.

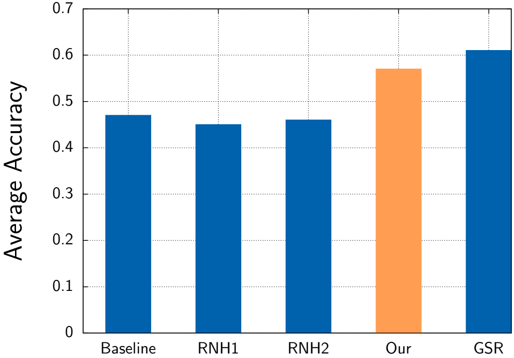


Figure: Recognition accuracy for 10 audio scenes

Scene Recognition Evaluation

Detailed result of our system with the best system *GSR*.

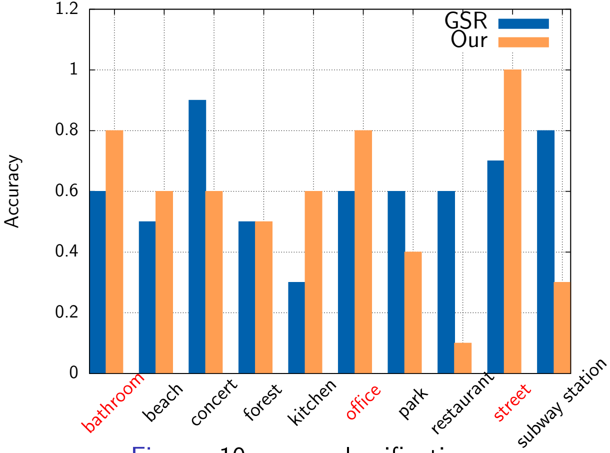


Figure: 10-scenes classification

Outline

Conclusion

- We build a scene recognition system from event detection.
- Our system has the advantange of expanding to many scenes without new scene data.
- We could outperform existing approaches in scenes where audible events are easy to capture.

Outline

Demo

Live demo for our system.

Thank you!

Any Question?