Audio Event Detection for Automatic Scene Recognition

Xun Xu

Department of Computer Science and Engineering
Shanghai Jiao Tong University

June 20, 2015

Outline

1. Introduction
2. Audio Event Detection
3. Scene Recognition
4. Evaluation
5. Demo

Outline

1. Introduction
   - Problem Description
   - Approach
2. Audio Event Detection
3. Scene Recognition
4. Evaluation
5. Demo

Problem Description

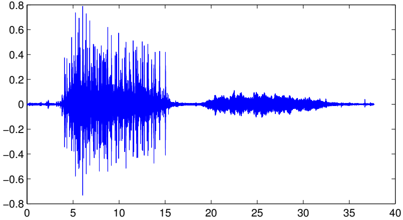
In this project, our problem is to recognize a scene where an audio is recorded. Sound example:

Play Sound

Problem Description

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

Play Sound



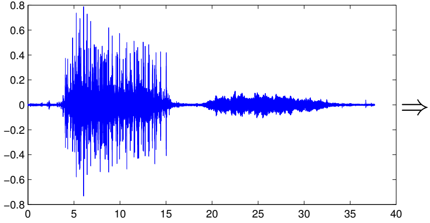
⇒ *concert*

Our Approach

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

Our Approach

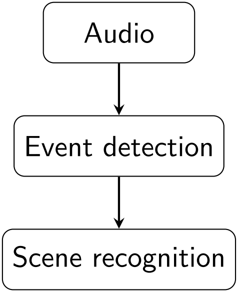
Our approach is to detect the audible events in a clip.
Then infer the scene from the detected events.



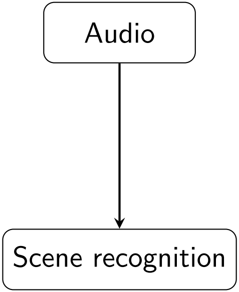
⇒ *applause, instrument* ⇒ *concert*

Our Approach vs. Other Approaches

Our approach:



Other approaches:



Outline

1. Introduction

2. Audio Event Detection
   - Audible Event Taxonomy
   - Audio Data
   - Preprocess and Feature Extraction
   - Event Model

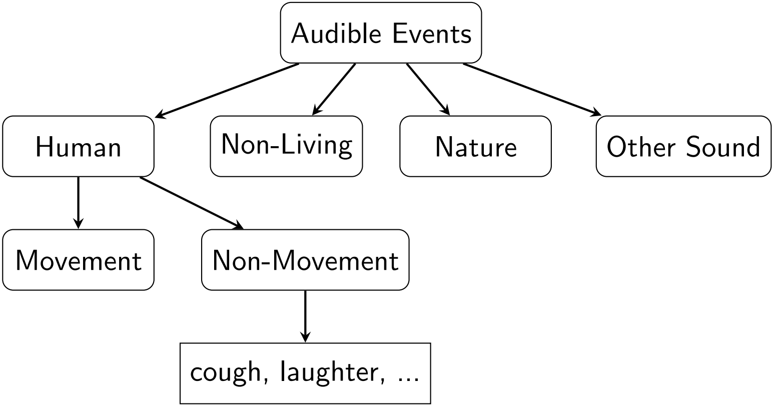
3. Scene Recognition

4. Evaluation

5. Demo

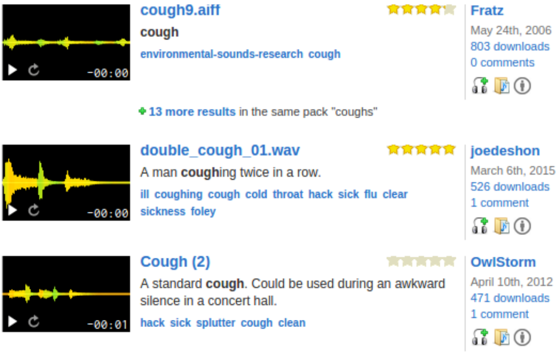
Audible Event Taxonomy

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



Audio Data

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



cough9.aiff ★★★★☆

cough

environmental-sounds-research cough

Fratz
May 24th, 2006
803 downloads
0 comments

✚ **13 more results** in the same pack "coughs"

double_cough_01.wav ★★★★★

A man **cough**ing twice in a row.

ill coughing cough cold throat hack sick flu clear sickness foley

joedeshon
March 6th, 2015
526 downloads
1 comment

Cough (2) ☆☆☆☆☆

A standard **cough**. Could be used during an awkward silence in a concert hall.

hack sick splutter cough clean

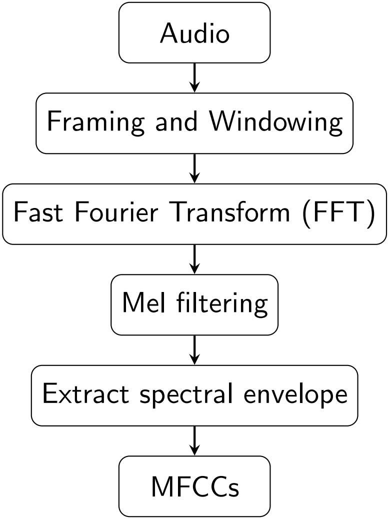
OwlStorm
April 10th, 2012
471 downloads
1 comment

Noise Reduction

denoise here

Feature Extraction

The features we extracted from audios are Mel-Frequency Cepstral Coefficients (MFCCs).



Audio

Framing and Windowing

Fast Fourier Transform (FFT)

Mel filtering

Extract spectral envelope

MFCCs

Event Model

We use Gaussian Mixture Models to train the features.

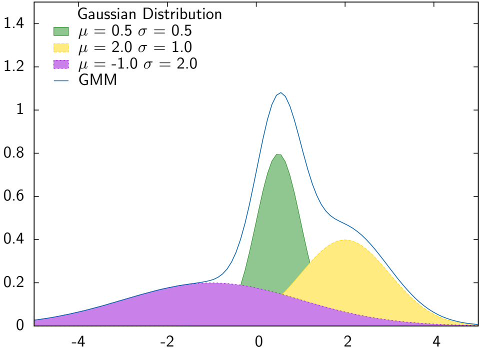


Figure: GMM with three components

Outline

1. Introduction
2. Audio Event Detection
3. Scene Recognition
   - Scene-Event Map
   - Audio Segmentation
   - Scene Inference
4. Evaluation
5. Demo

Scene Extraction

We use the scripts for movies, plays and TV series to extract the scenes.

Below is a script example. We call it a *context*, including a scene, and some descriptive sentences.

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 Extraction

We use Natural Language Process (NLP) tools to process a context, and eliminate the following type of words:

- Person names
- Time indicator
- Adjective, determiner, ...

Scene Extraction

We use Natural Language Process (NLP) tools to process a context, and eliminate the following type of words:

- Person names
- Time indicator
- Adjective, determiner, ...

Table: Top 10 Occurred Scenes

Scene	Occurrence
house	3537
office	3259
apartment	2919
room	2580
bedroom	2257
car	1699
street	1622
kitchen	1431
living room	1374
tardis	1259

Scene-Event Relation Mining

We first match audible events in a context, and count their occurrence.

> 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

Based on the idea of Term-Frequency-Inverse Document Frequency (TFIDF), we calculate two scores of an event e, to a scene s.

1. $TF = log(1 + f(e, s))$
   $f(e, s)$ is the number of contexts e appears in all contexts under scene s.
2. $IDF = 1 + log(\frac{N}{N_e})$
   N is the number of scenes. N_e is the number of scene that event e appears.

These two scores are then multiplied, and used as the importance of an event to a scene.

$$TFIDF = TF \times IDF \tag{1}$$

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
concert	piano, applause, crowd, chorus, child, cry, talk
forest	tree, wood, dirt, talk, running, bird, river, car, leaf, grass, wind
kitchen	drawer, cutlery, microwave, dish, kettle, talk, bowl, phone, toaster, running+water
office	desk, drawer, page+turn, talk, phone, printer, paper, chair, leaf, typewriter
park	talk, car, tree, laughter, dog, child, grass, crowd, running, phone
restaurant	talk, drink, laughter, phone, car, leaf, paper, dish, ring, chair, write
street	car, truck, subway, talk, traffic, engine, siren, phone, running, laughter
subway station	subway, train, car, tube, talk, pace, crowd, metal, phone, vehicle

Audio Segmentation

In testing, we segment the audio into smaller parts for event detection. We set two thresholds based on the following two features:

1. Frame Energy
   The averaged energy of a frame, calculated as:

$$E_i = \frac{\sum_{n=1}^{N} (x_i(n))}{N} \tag{2}$$

2. Spectral Centroid
   The "center" of frequency, calculated as:

$$C_i = \frac{\sum_{k=1}^{N} k \times Amp(k)}{\sum_{k=1}^{N} Amp(k)} \tag{3}$$

Audio Segmentation

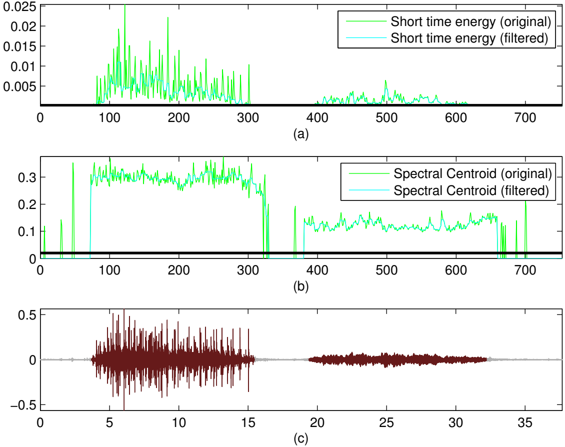


Figure: A segmentation example

Scene Inference

Outline

1. Introduction
2. Audio Event Detection
3. Scene Recognition
4. Evaluation
   - Event Detection Evaluation
   - Scene Recognition Evaluation
5. Demo

Event Detection Evaluation

Scene Recognition Evaluation

Outline

1. Introduction
2. Audio Event Detection
3. Scene Recognition
4. Evaluation
5. Demo

Demo

Live demo for our system.

Thank you!

Any Question?