Auditory Scene Recognition Using Textual Knowledge

Xun Xu
Xinyu Hua

May 6, 2015

Overview

Introduction

What is Auditory Scene Recognition(ASR)?

- Recognizing context of audio clips
- Usually from a set of predefined class labels
- Example: play pause resume stop

Introduction

Possible Usage

- Crime Investigation
- Cellphone Volume Adjustment

Introduction

Related Works

- Scene Detection(Video)
- Scene Detection(Audio)
- Event Detection(Audio)

Overview

Approach

Problem Definition:

Input: An audio clip

Output: The most likely scene where the audio clip was recorded, chosen from a given set.

Approach

Intuition:

We assume scenes are composed of multiple primitive events.

By using textual knowledge to construct Scene-Event Relation, we only need to detect events, and refer back to that relation to find out the most likely scene.

Approach

Roadmap:

- Build Vocabulary
- Construct Scene Event Map
- Feature Extraction for events
- Model Building
- Scene Recognition

Approach - Build Vocabulary

Obtain Event vocabulary

1. Sound search engine Taxonomy
2. Bootstrapping to Expand
3. Filter by number of downloadable event clips

Approach - Build Vocabulary

Obtain Event vocabulary

1. Sound search engine Taxonomy
2. Bootstrapping to Expand
3. Filter by number of downloadable event clips

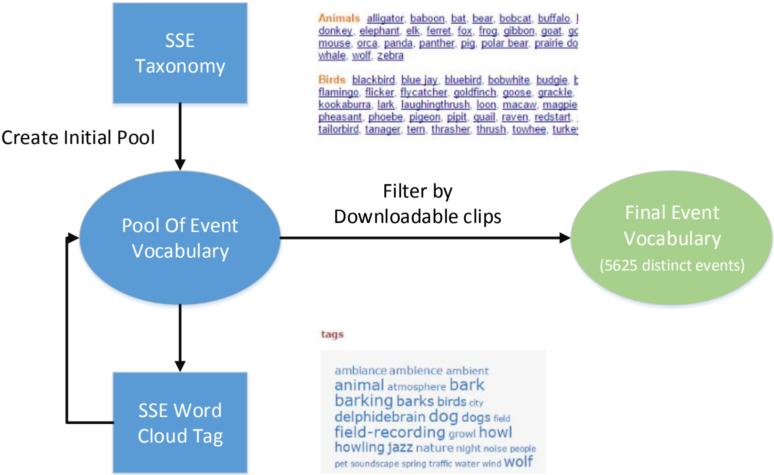
Approach - Build Vocabulary

Obtain Event vocabulary

1. Sound search engine Taxonomy
2. Bootstrapping to Expand
3. Filter by number of downloadable event clips

Approach - Build Vocabulary

Obtain Event vocabulary



Approach - Build Vocabulary

Example of Events

Vehicle, dog bark, laughter, applause, phone ring

Approach - Build Vocabulary

Obtain Scene vocabulary

1. Scene indicator in TV,Movie scripts
2. Using Stanford NLP to get the scene from a sentence
3. Sort scene to filter those appear less than 50 times

Approach - Build Vocabulary

Obtain Scene vocabulary

1. Scene indicator in TV,Movie scripts
2. Using Stanford NLP to get the scene from a sentence
3. Sort scene to filter those appear less than 50 times

Approach - Build Vocabulary

Obtain Scene vocabulary

1. Scene indicator in TV,Movie scripts
2. Using Stanford NLP to get the scene from a sentence
3. Sort scene to filter those appear less than 50 times

Approach - Build Vocabulary

Example:

```
FADE IN:
EXT. TWO-LANE HIGHWAY - SUNRISE
A dishevelled WOMAN in a business suit (27) runs down a
lonely highway in Texas hill country, moving desperately
through the thick morning fog. She's carrying a VHS
cassette. The sounds of her breathing and SHOES HITTING
the PAVEMENT ECHO into the mist.
```

↓

TWO-LANE HIGHWAY - SUNRISE

↓

HIGHWAY

Approach - Build Vocabulary

Example of scene:

Plane, Bus stop, theatre, bar, office

Approach - Construct Scene Event Map

Process the corpus

1. Sort corpus into sets of contexts
2. Parse all the contexts into Noun-Verb pairs
3. Association Network to expand

Approach - Construct Scene Event Map

Process the corpus

1. Sort corpus into sets of contexts
2. Parse all the contexts into Noun-Verb pairs
3. Association Network to expand

Approach - Construct Scene Event Map

Process the corpus

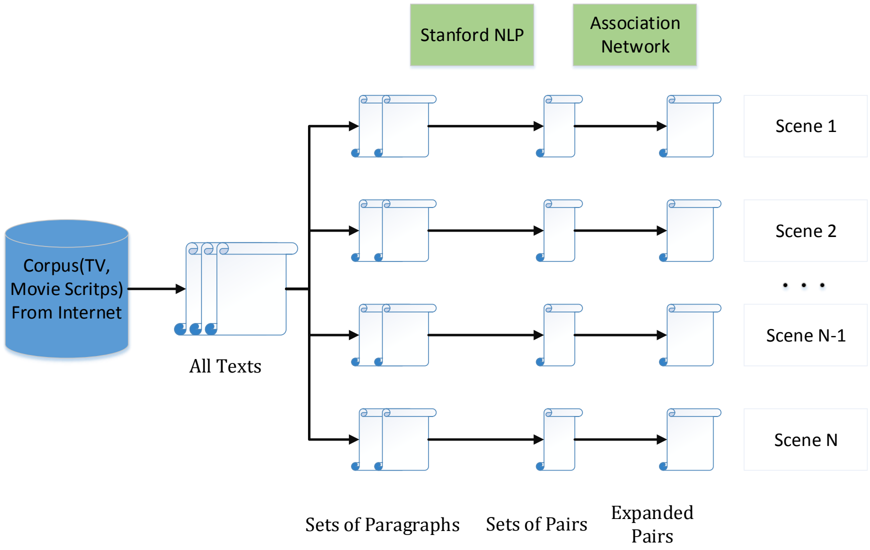
1. Sort corpus into sets of contexts
2. Parse all the contexts into Noun-Verb pairs
3. Association Network to expand

APPROACH - CONSTRUCT SCENE EVENT MAP

Process the corpus

1. Sort corpus into sets of contexts
2. Parse all the contexts into Noun-Verb pairs
3. Association Network to expand

Approach - Construct Scene Event Map



Approach - Construct Scene Event Map

Mine Scene Event Relation

1. Compute Occurrence of each event in each scene
2. Compute TFIDF
3. Construct Scene Event Map

Approach - Construct Scene Event Map

Mine Scene Event Relation

1. Compute Occurrence of each event in each scene
2. Compute TFIDF
3. Construct Scene Event Map

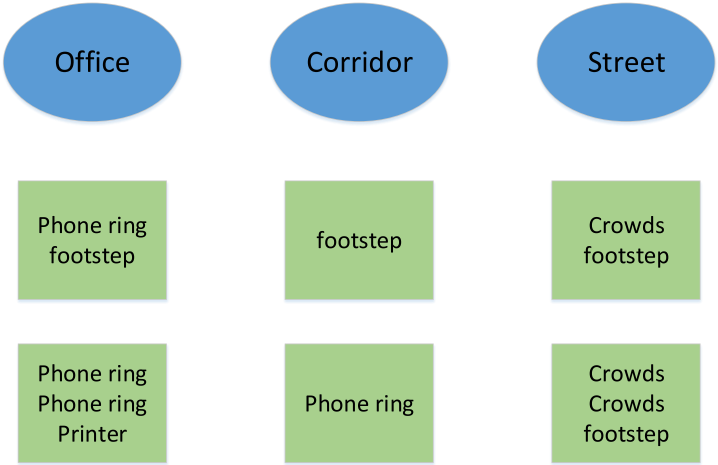
Approach - Construct Scene Event Map

Mine Scene Event Relation

1. Compute Occurrence of each event in each scene
2. Compute TFIDF
3. Construct Scene Event Map

Approach - Construct Scene Event Map

Example:



Approach - Construct Scene Event Map

Example:

	Phone ring	Footstep	Crowds	Printer
Office	0.392	0.176	0	0.778
Corridor	0.301	0.176	0	0
Street	0	0.229	0.621	0

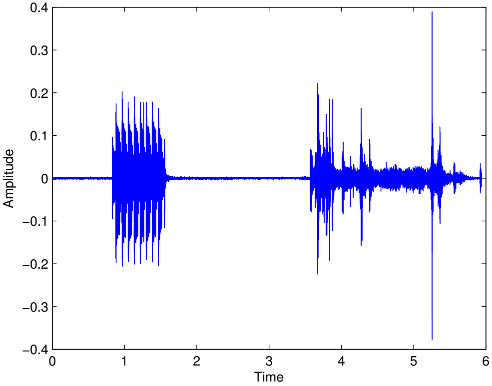
Table: TFIDF value for each Scene-Event pairs

Approach - Feature Extraction for events

- Spetrogram
- Framing and Fast Fourier Transform
- Mel-frequency Analysis
- Cepstral Analysis

Spectrogram

Here is the spectrogram of the example audio:

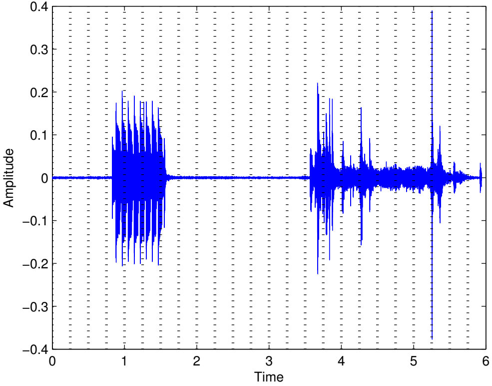


Approach - Feature Extraction for events

- Spetrogram
- Framing and Fast Fourier Transform
- Mel-frequency Analysis
- Cepstral Analysis

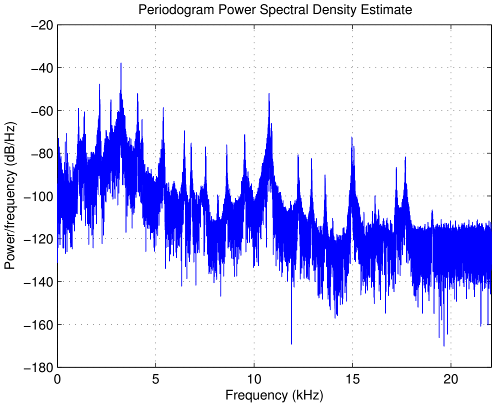
Framing and Fast Fourier Transform

Frame the audio and apply FFT on each frame.



Framing and Fast Fourier Transform

Here is the spectrum of example audio in 0-2s.

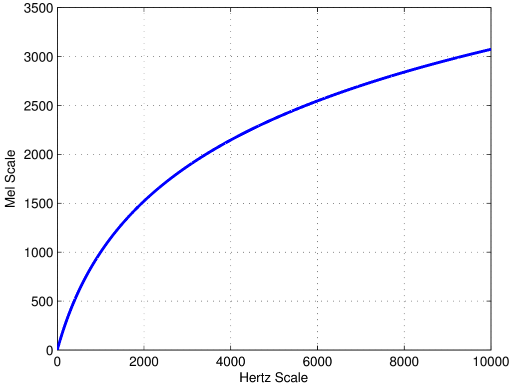


Approach - Feature Extraction for events

- Spetrogram
- Framing and Fast Fourier Transform
- Mel-frequency Analysis
- Cepstral Analysis

Mel-frequency Analysis

Perceptually, the difference between 500-1000Hz is different from 5000-5500Hz.

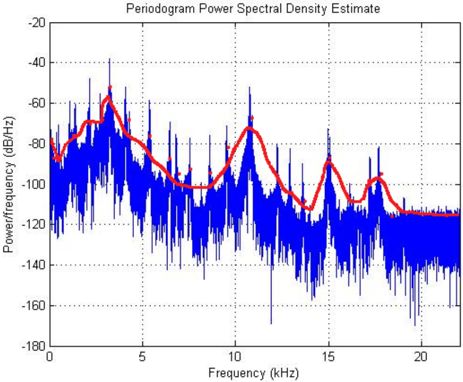


Approach - Feature Extraction for events

- Spetrogram
- Framing and Fast Fourier Transform
- Mel-frequency Analysis
- Cepstral Analysis

Cepstral Analysis

Get the envelope from spectrum.



Approach - Feature Extraction for events

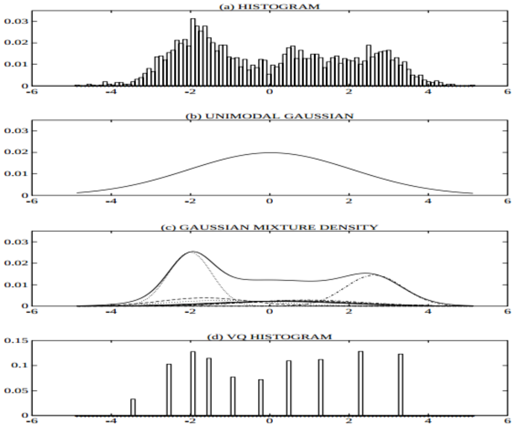
- Spetrogram
- Framing and Fast Fourier Transform
- Mel-frequency Analysis
- Cepstral Analysis

Approach - Model Building

- Model the Spectrum
- Gaussian Mixture Model
- Training GMMs

Model the Spectrum

A comparison of different models



Approach - Model Building

- Model the Spectrum
- Gaussian Mixture Model
- Training GMMs

Gaussian Mixture Model

A model with multiple gaussian distribution

$$P(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \times N(x|\mu_k, \sigma_k)$$

Approach - Model Building

- Model the Spectrum
- Gaussian Mixture Model
- Training GMMs

Training GMMs

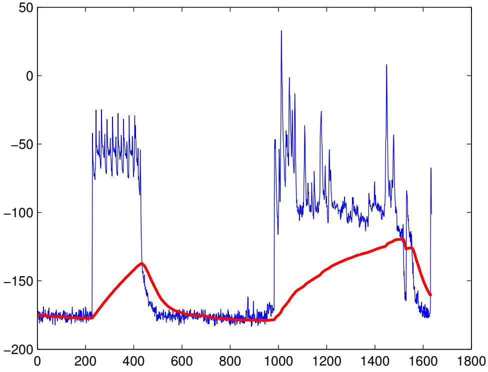
The features we extracted before are used here to train a GMM for each event.
Expectation-Maximization(EM) algorithm are used to estimate the parameters.

Approach - Scene Recognition

- Audio Segmentation
- Event Detection for Segments
- Infer Scene from Events

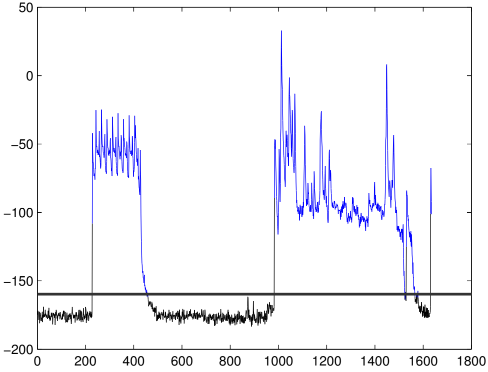
Audio Segmentation

We use a filter to smooth the frame energy



Audio Segmentation

Using the last value of smoothed line as threshold to segment:



Audio Segmentation

Exponential filter:

$$Y(n) = (1 - \alpha) \times Y(n-1) + \alpha \times X(n)$$

Choosing α:

$$\begin{cases} Y(n) \leq Y(n-1) & \alpha = 1/50 \\ Y(n) > Y(n-1) & \alpha = 1/500 \end{cases}$$

APPROACH - SCENE RECOGNITION

- Audio Segmentation
- Event Detection for Segments
- Infer Scene from Events

Event Detection for Segments

We apply GMMs to segments and find the events which have the highest score for features.

Approach - Scene Recognition

- Audio Segmentation
- Event Detection for Segments
- Infer Scene from Events

Infer Scene from events

Use TFIDF scores as weight for voting.

	Phone ring	Footstep	Crowds	Printer
Office	0.392	0.176	0	0.778
Corridor	0.301	0.176	0	0
Street	0	0.229	0.621	0

Table: TFIDF value for each Scene-Event pairs

Assume we detect "Phone ring" and "Printer" in the example audio.
Office: $0.392 + 0.778 = 1.17$
Corridor: 0.301
Street: 0

Overview

Introduction

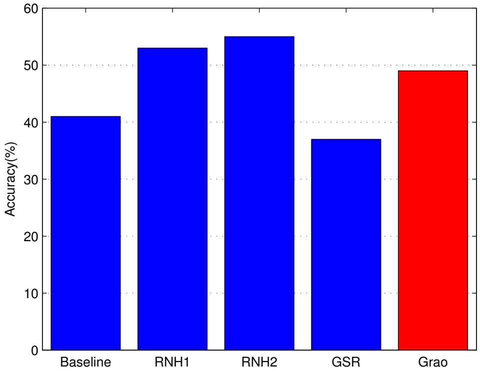
Approach

Evaluation

Problems and Improvements

Evaluation

We have performed a 5 scene classification task on our system and other 4 systems.
20 clips for each scene, and a five-fold cross validation are conducted for other four systems.



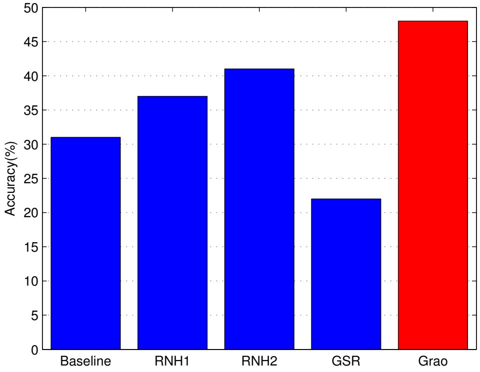
Evaluation

Table: Recognition Accuracy for 5 Audio Scenes

	bathroom	kitchen	office	restaurant	street	average
baseline	55%	25%	20%	50%	55%	41%
RNH1	55%	55%	35%	80%	40%	53%
RNH2	55%	45%	50%	70%	55%	55%
GSR	70%	15%	15%	75%	10%	37%
Grao	65%	35%	**75**%	5%	**65**%	49%

Evaluation

If ”restaurant” was removed:



Overview

Introduction

Approach

Evaluation

Problems and Improvements

Problems and Improvements

- Hard to control the quality of event vocabulary(Granularity of primitive events)
- Fail to detect multiple events occuring at the same time
- Perform bad under noisy environment.