Acoustic Scene Classification

By Yuliya Sergiyenko

Seminar: Topics in Computer Music
RWTH Aachen
24/06/2015

Outline

1. What is Acoustic scene classification (ASC)
2. History
   1. “Cocktail party problem” & CASA
   2. Sets of categories
   3. DCASE challenge
3. Room Identification (RI)
   1. Room Impulse Responses
   2. Room Volume Classification
   3. State of the art
      1. MFCC
      2. GMM
   4. Competing solution
4. Conclusions

Acoustic scene classification

- **Acoustic scene classification (ASC)** - identifying the location at which the audio or video recording was made.

- Possible application fields:
  - hearing aids with automatic program adaptation:
  - speech recognition;
  - forensic analysis;
  - etc.

History

- “One of our most important faculties is our ability to listen to, and follow, one speaker in the presence of others. We may call it ‘the cocktail party problem’…” (Cherry, 1957)

- Computational auditory scene analysis (CASA) is the study of auditory scene analysis (ASA) by computational means (Bregman, 1990). The goal of the CASA system is to be able to separate sound mixtures in the same way that humans are able to do.

History

- Researchers generally define a set of categories, record samples from these environments, and treat ASC as a supervised classification problem within a closed universe of possible classes. [1]
  - Sawhney and Maes in 1997 described a simple classification of five pre-defined classes of environmental sounds: ‘people’, ‘voices’, ‘subway’, ‘traffic’, and ‘other’, via extraction of several discriminating features.

DCASE challenge

- To evaluate and compare existing algorithms in ACS the IEEE Audio and Acoustic Signal Processing (AASP) Technical Committee organized a challenge in 2013.

- The DCASE challenge dataset was especially created to provide researchers with a standardised set of recordings produced in 10 different urban environments. The soundscapes included: 'bus', 'busy-street', 'office', 'openairmarket', 'park', 'quiet-street', 'restaurant', 'supermarket', 'tube' (underground railway) and 'tubestation'.

“New problem” - Room identification

- **Room identification (RI)** - given the audio, identifying the particular room in which the recording was made.

- Beneficial in:
  - Location estimation;
  - Music recommendation;
  - Better speech recognition indoors;
  - Law-enforcement and forensics.

- Location estimation:
  - GPS data
    - only rough estimate
    - often fails indoors
  - Strength of WiFi signals (2010)
    - location must be estimated during capturing
    - needs sufficient WiFi coverage
  - Visual similarities (video)
    - changes in spatial configuration
- Music recommendation:
  - Could help create a list of recordings made at specific venue

- Better speech recognition indoors:
  - Automated speech recognition systems are affected by unknown room reverberance. Knowing the room will help adapt recognition system.

- Law-enforcement and forensics:
  - Emergency phone calls
    - more clues
    - filter out fake calls

Room Impulse Responses (RIR)

- Rooms can be described through room impulse responses – “fingerprints” of the room.[2]
- Obtaining RIRs is a very time-consuming process and specific measurement signals and equipment are needed. [3]
- It is often too complicated or even impossible to conduct such RIR measurements.

Room Volume Classification

- Room volume classification from reverberant speech signals can be useful in acoustic scene analysis applications to help in characterizing the types of rooms. [4]

- Previous work required the room impulse response (RIR) to explicitly either estimate or classify the room volume so the researches started looking for simpler approaches.

State of the art in RI

- N. Peters, H. Lei, and G. Friedand, “Name That Room: Room Identification Using Acoustic Features in a Recording”, 2012 [5]

- Ideas:
  - analyzing the audio component in multimedia data;
  - using machine learning techniques to identify rooms from ordinary audio recordings.
  - This room identification system is derived from a GMM based system using Mel-Frequency Cepstral Coefficient (MFCC) acoustic features.

GMMs

- **Mixture Models** are a type of density model which comprise a number of component functions.
- **A Gaussian Mixture Model (GMM)** is a parametric probability density function represented as a weighted sum of Gaussian component densities.
- GMMs are commonly used as a parametric model of the probability distribution of continuous measurements or features in a biometric system.

MFCCs

- **Mel-frequency cepstrum (MFC)** is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel-scale of frequency.

- **Mel-frequency cepstral coefficients** (**MFCCs**) are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum").

The system

- For each audio recording, one room-dependent GMM is trained for each room using MFCC features from all audio recordings associated with that room. This is done via MAP adaptation from a room-independent GMM, trained using MFCC features from all audio tracks of all rooms in the development set. During testing, the likelihood of MFCC features from the test audio tracks are computed using the room-dependent GMMs of each room in the training set. [5]

Results

- The room identification system is **able** to relate audio data to the correct room.
- The estimation error is not randomly distributed. Rather it **depends** on the **(acoustical) similarities** of the tested rooms.
- For room identification **short-term MFCC** features are found **more suitable** than-long term MFCC features.

Results

- Non-parametric multidimensional scaling (MDS) was performed on the confusion data. MDS is a technique where dissimilarities of data points are modeled as distances in a low-dimensional space. A large dissimilarity is represented by a large distance and vice versa.

- The system achieved overall accuracy of **61% for music** and **85% for speech** signals.

Competing solution

- A. H. Moore, M.Brookes and P. A. Naylor, “Roomprinting for forensic audio applications”. [6]

- Ideas:
  - Roomprint - “a set of features of a room are inferred from a recording made in the room and are compared to a set of reference roomprints in order to perform identification or verification of the recording location”
  - Roomprint can include any aspect of the room which can be explicitly measured

Main questions

- Verification: If it is claimed that a recording was made in a particular room, is there sufficient evidence to reject the claim?
- Identification: If it is known that a recording was made in one of a number of rooms, can we determine which one is most likely? []
  - covered in 2010 by Peters et al. in “Name that room”
  - differences in measurement systems or technique may have caused some of the between-room variability

Roomprints' requirements

- A roomprint must exploit features of a room which allow it to be **distinguished** from other potentially similar rooms.
- A roomprint should ideally be **invariant to the location** of the talker and microphone in the room.
- A roomprint should ideally **be invariant with time**.

- **Parameters to include:**
  - Geometric features (the size and shape of the room)
  - Room acoustics parameters
  - Environmental sounds
- **Frequency-dependent reverberation time** is investigated in this paper as a promising characteristic and used in a room identification experiment.
- **Results:**
  - an **error rate of 3.9%** has been obtained in a room identification experiment over 22 rooms (correct identification in **96%** of trials).

Conclusions

- Different state of the art approaches achieve nearly the same results.
- There is still need for improvement until any algorithm can reach the same efficiency as human in acoustic scene classification.
- Using machine learning techniques for identifying room acoustic properties is a very young field of research that is becoming more important.
- The concept of “Roomprints” is the state of the art in room identification area to this day.
- Frequency-dependent reverberation time is a good characteristic to use in a room identification process.

References

1. Daniele Barchiesi, Dimitrios Giannoulis, Dan Stowell and Mark D. Plumbley, Acoustic Scene Classification, IEEE. School of Electronic Engineering and Computer Science, November 17, 2014
2. ISO 3382-1. Acoustics – Measurement of room acoustic parameters – Part 1: Performance spaces. International Organization for Standardization (ISO), Geneva, Switzerland, 2009.
3. G. Stan, J. Embrechts, and D. Archambeau. Comparison of different impulse response measurement techniques. J. Audio Eng. Soc., 50(4):249–262, 2002. [18] R. Stewart and M. Sandler. Database of omnidirectional and B-format impulse responses. In Proc. of ICASSP, Dallas, USA, 2010.
4. N. Shabtai, B. Rafaely, and Y. Zigel. Room volume classification from reverberant speech. In Proc. of int'l Workshop on Acoustics Signal Enhancement, Tel Aviv, Israel, 2010.
5. N. Peters, H. Lei, and G. Friedand, "Name That Room: Room Identification Using Acoustic Features in a Recording," in Proc. of ACM Multimedia, Nara, Japan, 2012
6. A.H. Moore, M. Brookes, P.A. Naylor, Roomprints for forensic audio applications, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), October 20-23, 2013, New Paltz, NY