Name That Room: Room Identification Using Acoustic Features in a Recording
Nils Peters1, 2, Howard Lei1, Gerald Friedland 1
1 ICSI - International Computer Science Institute, Berkeley, CA, USA
2 CNMAT - Center for New Music and Audio Technologies, UC Berkeley, CA, USA
[nils|hlei|fractor]@icsi.berkeley.edu

Abstract

We present a room identification system in an audio or video recording through the analysis of acoustical properties. The room identification system was tested using a corpus of 13440 reverberant audio samples. With no common content between the training and testing data, an accuracy of 61% for musical signals and 85% for speech signals was achieved. This approach could be applied in a variety of scenarios where knowledge about the acoustical environment is desired, such as location estimation, music recommendation, or emergency response systems.

Subject Descriptors
S.01 [Media Content Analysis & Processing]: Mobile and Location-Based Media

Keywords
Room identification, Audio analysis, Room acoustics, Location estimation

1. Introduction

With the trend of location-based multimedia applications, knowledge about the room environment is an important source of information. GPS data may only provide a rough location estimate and tends to fail inside buildings. The strength of WiFi signals can be used to gain a better accuracy [1], but this relies on WiFi signals and receiver. [2] predicts common locations by relying on identifying visual similarities (landmarks or similar interior objects). This approach does not account for changes in spatial configurations that may occur, like when new tenants or home owners move furniture or redesign their rooms.

We propose to analyze the audio component in multimedia data. This can be complementary to other methods.

1.1 Prior Art

Using machine learning techniques for identifying room acoustic properties from reverberant audio signals is a promising approach [3, 4].

1.2 Use Cases

- Room-tuning for assistive hearing aids
- Room-tuning for automated speech recognizers
- Find music performed in the same venue
- Room prediction for emergency response systems
- Forensic and law enforcement

During testing, the likelihood of MFCC features from the unmatched audio recordings are computed using all room-dependent GMMs in the training set. A total of 128 mixtures and simplified factor analysis are used for each GMM. The ALIZE toolkit is employed for the GMM and factor analysis implementations [6].

2. The Corpus

Because there is no dataset exists for the task of room identification, we created a corpus of anechoic audio recordings, each filtered with a variety of IIRs from a number of rooms. To allow reproducibility of our results, we intentionally use publicly available anechoic audio recordings and RIR datasets.

3. The Room Identification System

Our room identification system is derived from a GMM-based system using Mel-Frequency Cepstral Coefficient (MFCC) acoustic features, which have proven to be effective in tasks such as speaker recognition [5]. MFCC features CD-C19 (with 25 ms window lengths and 10 ms frame intervals), along with deltas and double-deltas (60 dimensions total), are extracted.

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.

4. Experiments and Results

Performance is measured with the Equal Error Rate (EER), a scoring threshold where the percentage of impostor scores above the threshold equals the percentage of true scores below it.

4.1 Effect of MFCC Window Size

The most prominent parameter that can influence the feature extraction process and eventually the resulting EER is the MFCC window size. Speech recognition applications historically use a window size of 25 ms. In contrast, [3] applied a 1 sec. MFCC window.

4.2 From Confusion Matrix to MDS

The confusion matrix shows that the Room ID system successfully relate audio data to the correct room. The accuracy in Experiment 3 is 85% for speech, for music 61%.

The estimation error is not random, but depends on the acoustical similarity of the tested rooms.

References