# Name That Room: Room Identification Using Acoustic Features in a Recording

Nils Peters[1,2], Howard Lei[1], 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

The 20th Anniversary ACM Multimedia 2012
29 Oct. – 2 Nov. 2012

## 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

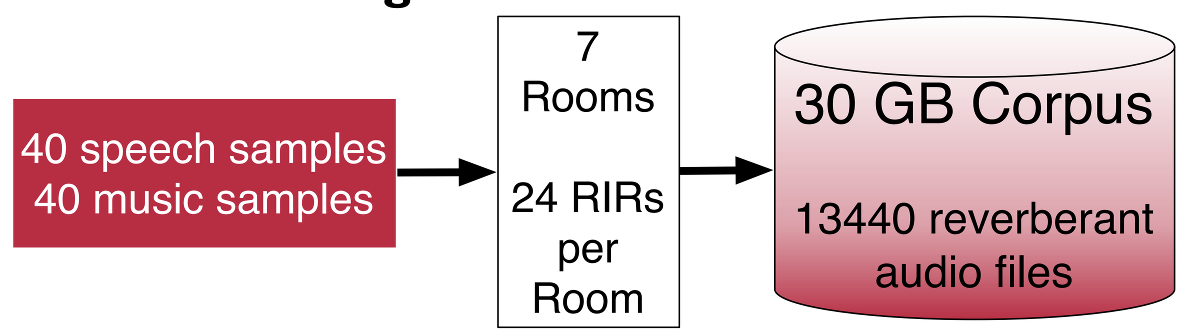
Using machine learning techniques for identifying room acoustic properties from reverberant audio signals is a very young field of research; see [3, 4].

### 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

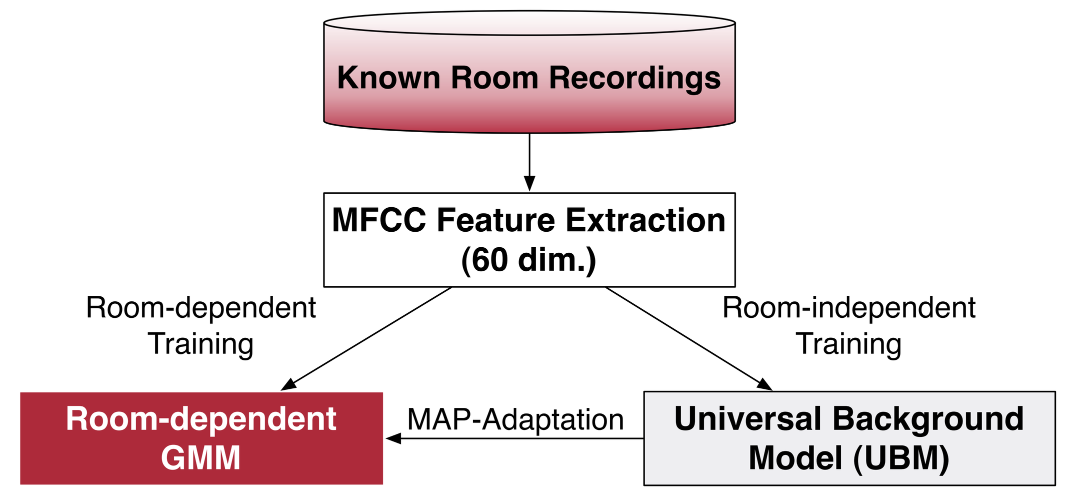
## 2. The Corpus

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

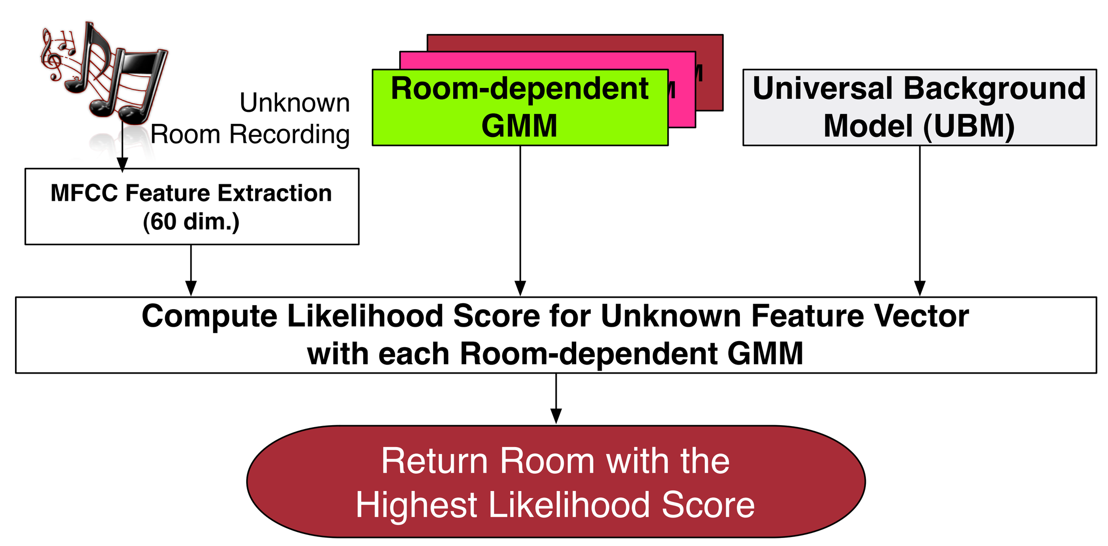



## 3. The Room Identification System

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 C0-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.

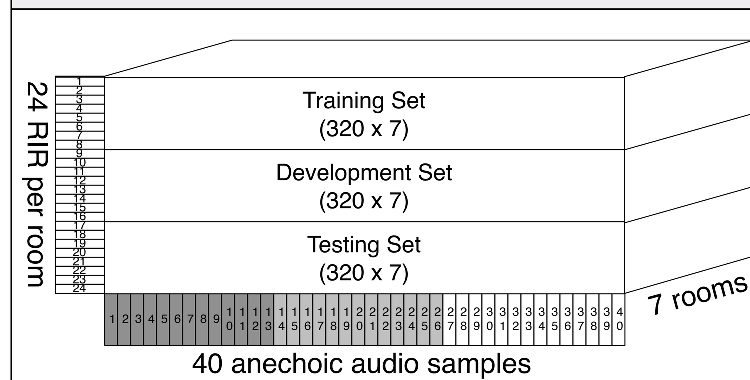
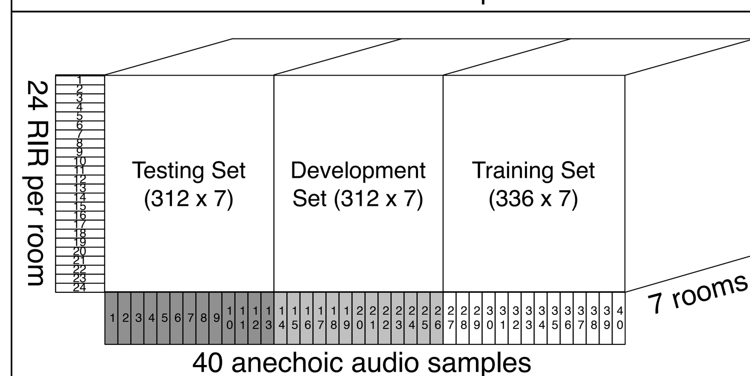
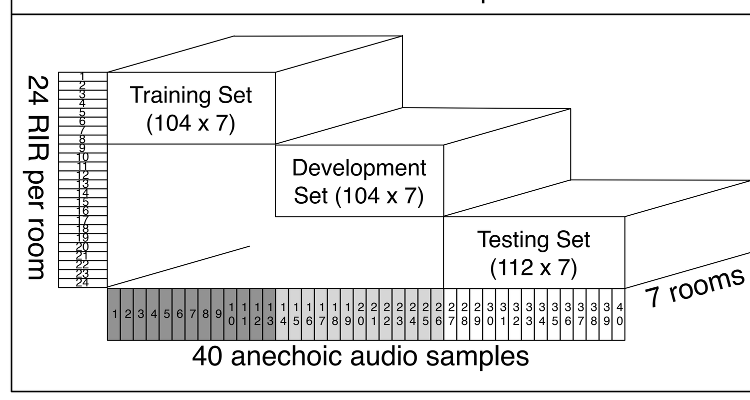



During **testing**, the **likelihood of MFCC features** from the unknown audio recordings are computed using all room-dependent GMMs in the training set.

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].

## 4. Experiments and Results

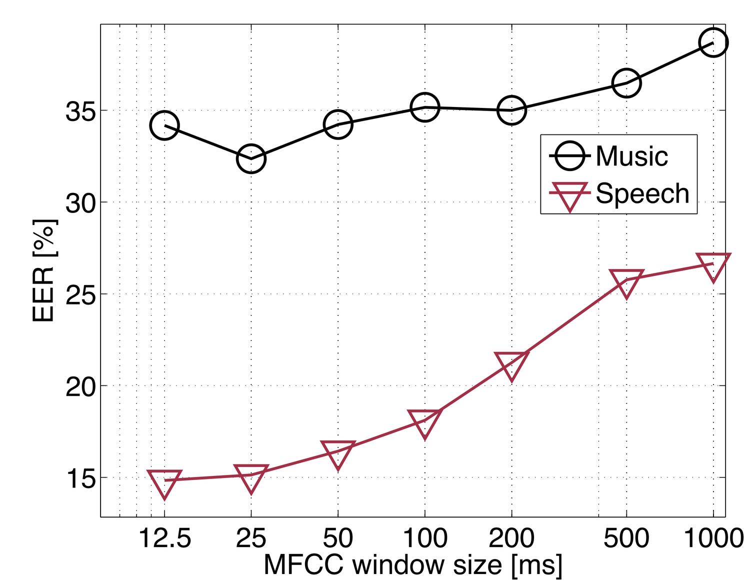
Performance is measured with the **Equal Error Rate** (EER), a scoring threshold where the percentage of impostor scores above the threshold equals the percentage of true scores below it.

| Experiment | EER Music | EER Speech | EER Combined |
|---|---|---|---|
|   | 15.07 | 8.57 | 13.23 |
|   | 14.71 | 7.67 | 11.28 |
|   | 32.36 | 15.14 | 23.85 |

Resulting equal error rates (EER) for the different experiments

### 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.




Effect of MFCC window size on the EER

## Acknowledgments

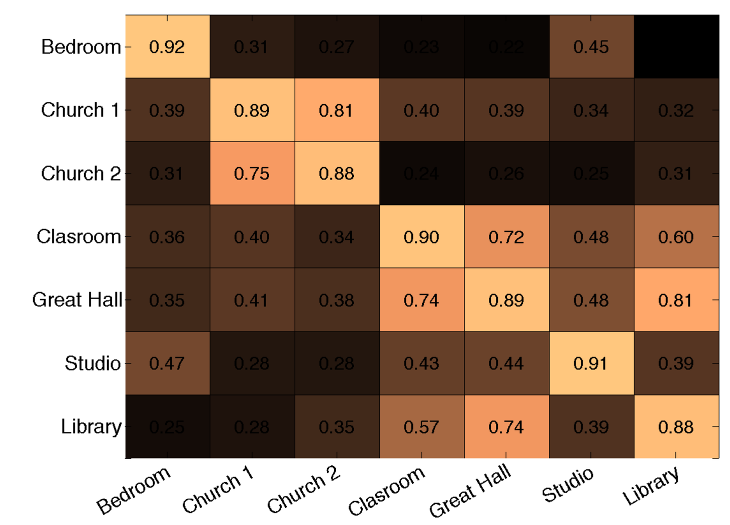
Nils Peters is supported by the German Academic Exchange Service (DAAD). Support comes also from Microsoft (Award #024263), Intel (Award #024894), and matching U.C. Discovery funding (Award #DIG07-10227).

### 4.2 From Confusion Matrix to MDS

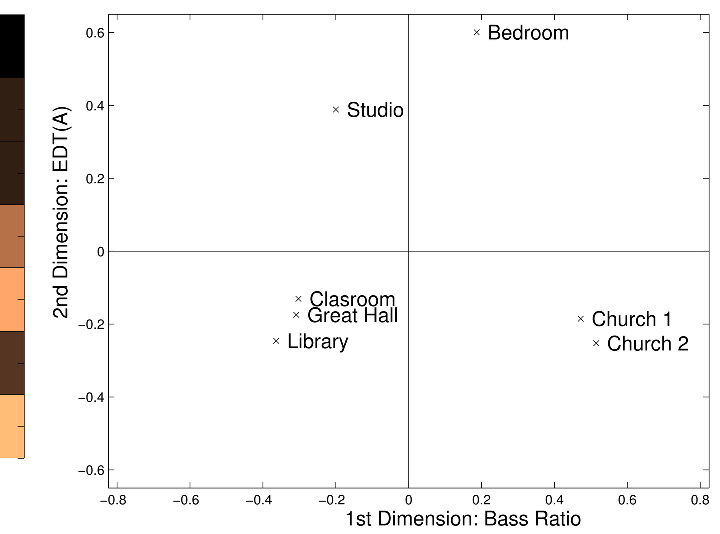
The confusion matrix shows that the Room ID system successfully relate audio data to the correct room.

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.




Confusion Matrix of the estimation scores for Experiment 3 (music)




Non-parametric MDS analysis of the same data (Exp. 3, music)

Non-parametric multidimensional scaling (MDS) was performed on the confusion data.

The first MDS dimension is well correlated with the **Bass Ratio** ($\rho(6) = -.79$), the ratio of the low-frequency reverb time compared to the mid-frequency reverb time. The second MDS dimension is perfectly correlated with the **A-weighted Early Decay Time** of the RIRs ($\rho(6) = -1.0$), the time in which the first 10 dB decay of the reverb occurs and is closely related to the perceived reverberance.

## 5. Conclusion and Future Work

We have presented a system for identifying the room in an audio or video recording based on MFCC-related acoustical features. With no common content between the training and testing data, the system achieved overall accuracy of 61% for music and 85% for speech signals.

To potentially improve the accuracy for music content, we want to explore **additional features** such as those based on the **modulation spectrogram**.

## References

[1] E. Martin, O. Vinyals, G. Friedland, and R. Bajcsy. *Precise indoor localization using smart phones.* ACM MM, 2010.
[2] A. Ulges and C. Schulze. *Scene-based image retrieval by transitive matching.* ACM ICMR, 2011.
[3] N. Shabtai, B. Rafaely, and Y. Zigel. *Room volume classification from reverberant speech.* IWAENC, 2010.
[4] N. D. Gaubitch, H. W. Löllmann, M. Jeub, T. H. Falk, P. A. Naylor, P. Vary, and M. Brookes. *Performance comparison of algorithms for blind reverberation time estimation for speech.* IWAENC, 2012.
[5] D. Reynolds, T. Quatieri, & R. Dunn. *Speaker verification using adapted gaussian mixture models.* Digital signal processing, 10(1- 3), 2000.
[6] J. Bonastre, F. Wils, & S. Meignier. *ALIZE, a free toolkit for speaker recognition.* ICASSP 2005.