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

Nils Peters
International Computer Science Institute
1947 Center Street
Berkeley, CA, USA
nils@icsi.berkeley.edu

Howard Lei
International Computer Science Institute
1947 Center Street
Berkeley, CA, USA
hlei@icsi.berkeley.edu

Gerald Friedland
International Computer Science Institute
1947 Center Street
Berkeley, CA, USA
fractor@icsi.berkeley.edu

ABSTRACT

This paper presents a system for identifying the room 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.

Categories and Subject Descriptors

S.01 [**Media Content Analysis and Processing**]: Mobile and Location-Based Media

General Terms

Experimentation, Measurement

Keywords

Room identification, Audio analysis, Room acoustics, Location estimation

1. INTRODUCTION

Most of our time is spent indoors and, as such, in reverberant environments. For extracting information from a reverberant audio stream, the human auditory system is well adapted. Based on accumulated perceptual experiences in different rooms, we can often recognize a specific environment just by listening to the audio content of a recording; e.g., we can distinguish a recording made in a reverberant church from a recording captured in a conference room.

With the emerging trend of location-based multimedia applications, such as automatic tagging of uploaded user videos, 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. Attempts to use the strength of WiFi signals to gain a better accuracy were presented, e.g., in [13]. However, in these approaches, the location must be estimated and stored as meta data at the time of the capturing process. If either GPS and WiFi coverage is insufficient, or the capturing device does not support this technology, the location cannot be estimated. In [19] an alternate method 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.

Instead we propose to analyze the audio component in multimedia data. This can be complementary to aforementioned methods as shown in [12]. Although the specific analysis of acoustical properties to predict the room environment is new (see Section 1.1), the principles of room acoustics are well understood. Rooms can be described through room impulse responses (RIR, see [9]), the "fingerprint" of a specific room. Obtaining RIRs is a time-consuming process and specific measurement signals and equipment are needed [17]. Although many applications might benefit from knowledge about the room environment, it is often too complicated or even impossible to conduct such RIR measurements. Therefore, we propose using machine learning techniques to identify rooms from ordinary audio recordings.

Besides location estimation, many other applications can benefit from knowledge about the acoustical environment. For instance automated speech recognition systems, known to be easily affected by unknown room reverberance, could adapt the recognition engine based on the identified room acoustic environment. A music recommendation system could automatically create a playlist of recordings made in a specific concert venue. In an emergency response system, the room acoustics within an emergency phone call may give additional cues beneficial for the rescue, or even expose a fake emergency call. The latter example points to law-enforcement and forensic applications.

1.1 Related work

Using machine learning techniques for identifying room acoustic properties from reverberant audio signals is a very young field of research.

A Gaussian mixture model (GMM) approach [16] estimated the room volume in reverberant speech recordings into six room classes, ranging from 40 m^3 to 18000m^3. From the four tested feature extraction approaches, the best results were achieved by computing RIR features from an esti-

MM'12, October 29–November 2, 2012, Nara, Japan.

mated RIR derived from abrupt stops in speech signals with an equal error rate or (EER) 22%. The worst EER (30%) was achieved by using Mel-Frequency Cepstral Coefficient (MFCC) features extracted from reverberant speech. In the latter, 12 MFCCs and their deltas were extracted using a 1 sec. Hamming window. In [8] three different methods to estimate the reverberation time T_{60} from reverberated speech were compared. These methods are based on the Modulation Energy Ratio, Spectral Decay Distribution, and on a maximum likelihood of a statistical model of the sound decay. In low noise conditions the latter two methods were found to provide accurate estimation to within ± 0.2 sec for $T_{60} \leq 0.8$ sec. To the author's knowledge, there are no studies for room classification using musical material.

2. METHODOLOGY

2.1 The corpus

Because no standardized dataset exists for the task of room identification, we generated a corpus from anechoic audio recordings, each filtered with a variety of impulse responses from a number of rooms. To allow reproducibility of our results, we intentionally use publicly available anechoic audio recordings and RIRs datasets. One requirement in creating the corpus was that only RIRs from real rooms were included, i.e., they are not synthesized using room acoustic modeling software or artificial reverberators. Another challenging requirement was to find publicly available RIR datasets that measured multiple RIRs in a room. This is crucial to generalize our experimental results: an RIR depends on the location of sender and receiver, therefore no RIR within a room is completely similar to another. The final set of RIRs are collected from the databases [1, 2, 4], and [18] and comprise seven rooms. For each selected room, 24 RIRs are available. Table 1 summarizes several objective RIR measures [11] and their variation across the 24 RIRs per room. Particularly interesting and potentially challenging for our approach, the datasets of *Church 1* and *Church 2* have been captured in the same room (St. Margaret's Church in York [4]), each with a different acoustical configuration. Thus they are considered as two different rooms. For *Church 1*, drapes and panels have been used to make this room suitable for lectures and speech; for *Church 2*, panels were removed to create a more reverberant space suitable for music recitals.

The anechoic musical recordings were taken from [4, 5, 7]. The recordings of [7] captured multiple instruments within a recording, whereas the rest of the anechoic audio files contain single instruments, e.g., trumpet, guitar, or a clarinet. We limited the sample length to 30 seconds. Forty anechoic speech recordings were taken from the EMIME speech corpus [3] and from [5] and comprise 20 different male and 20 female speaker samples of 20 seconds. All anechoic samples are musically or lexically unique within the dataset.

In total, 80 anechoic audio files and 168 RIRs are used to generate 13440 reverberant audio samples in 16 bit and 44.1 kHz. The total size of the corpus is 30 GB.

2.2 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 related audio-based tasks such as acoustic event detection [14], location identification [12], and speaker recognition [15]. 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 with an upper frequency limit of 15 kHz using HTK [20]. 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. A total of 128 mixtures and simplified factor analysis [10] are used for each GMM. The open-source ALIZE toolkit is employed for the GMM and factor analysis implementations [6].

The likelihood values for which the room of the test audio matches the room of the GMM model are known as the true scores; values for which the rooms do not match are known as the impostor scores. The system performance is based on the equal error rate (EER), which occurs at a scoring threshold where the percentage of impostor scores above the threshold equals the percentage of true scores below it.

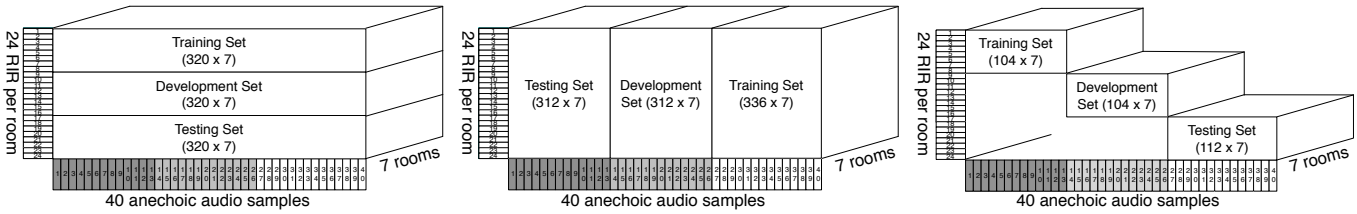
3. EXPERIMENTS AND RESULTS

Four different sets of experiments were carried out to understand the performance of our room identification system and to explore potential challenges. The first three groups of experiments explore the system's performance by using fundamentally different sets of training, testing, and development sets. All experiments are carried out using 3-fold cross validation and the averaged equal error rate (EER) is reported. All experiments are first carried out by separately testing the *Music* samples and *Speech* samples of the corpus. For the *Combined* setting, the entire corpus is used.

Table 1: Standard Acoustical Measures of the different rooms used for creating the corpus. The data shows average μ and standard deviation σ across the 24 RIRs per room.

Room + Reference	Vol $[m^3]$	EDT(A) [sec]		T_{30} [sec]		ITDG [ms]		CT [ms]		BR	
		μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
Bedroom [2]	25	0.255	0.040	0.278	0.010	1.500	0.751	13.269	4.811	1.391	0.546
Studio [1]	150	0.530	0.163	0.670	0.021	1.652	0.758	7.937	4.217	3.288	0.548
Classroom [18]	236	3.766	0.039	6.649	1.865	4.888	4.180	89.137	26.221	1.292	1.031
Church 1 [4]	3600	2.512	0.108	3.152	0.071	6.999	6.898	58.612	16.713	0.898	0.136
Church 2 [4]	3600	3.264	0.116	3.645	0.046	9.754	8.616	72.905	20.403	0.895	0.148
Great Hall [18]	unreported	4.059	0.187	5.395	2.503	3.075	1.944	59.029	17.662	1.337	0.787
Library [18]	9500	5.533	0.177	6.258	1.544	6.738	13.292	87.513	29.470	1.318	1.419

EDT(A): A-weighted Early Decay Time; *ITDG:* Initial Time Delay Gap; *CT:* Center Time; *BR:* Bass Ratio



(a) Experiment A (b) Experiment B (c) Experiment C

Figure 1: Arrangements of training, development, and testing data for the Experiments A – C

3.1 Design

3.1.1 *Experiment A*

In this experiment, the reverberant audio files of the training sets, development sets, and testing sets are based on common anechoic audio samples. As depicted in Figure 1(a), the difference between the datasets are the different RIRs within a room the anechoic audio samples were filtered with. In total each of the three datasets comprises 320 audio samples for each of the seven rooms, resulting in 2240 audio samples per set.

3.1.2 *Experiment B*

Here, the datasets are created in opposition to Experiment A. Now the training set, development set, and testing set are created based on the same RIRs. The difference across the sets is in the source files (Figure 1(b)). Compared to Experiment A, this experiment is potentially more challenging, because the training is based on completely different anechoic source files than the model was trained on.

3.1.3 *Experiment C*

The three datasets are based on different anechoic audio samples as well as different RIRs per room. As can be seen in Figure 1(c), they have no common audio data. This scenario is closest to reality where the system estimates the room based on a completely unknown audio recording.

3.2 Results

Table 2 summarizes the averaged equal error rates (EER) for all three experiments with the different content condition *music*, *speech*, and *combined*. All results are the averaged EER of a 3-fold cross validation. Three observations can be made. Compared to the musical material, the EER of the speech content in all experiments is about twice as good. The EER of the combined condition, where testing and training datasets contained both music and speech content, is about the average of the EER for music and specch in separation. Second, the EER of Experiment C is about twice as high compared to those of Experiment A and Experiment B. Experiment A and Experiment B resulted on average in a similar EER. However, for Experiment B, where the training, development, and testing datasets differ with respect to the audio content, the variance of the EER across the three different rounds in the cross validation is considerably higher than those for Experiment A.

All experiments were also carried out using the limited feature set of pure MFCC, and MFCC+Δ. These results are not shown since they achieved a higher EER.

Figure 2 shows the confusion matrix of the normalized estimation scores of the testing data in Experiment C (music) - the experiment with the highest EER and an accuracy of 61%. For speech signals, the accuracy was 85% (not shown here). The confusion matrix clearly shows that the room identification system is able to relate audio data to the correct room. One can also see that the estimation error is not randomly distributed. Rather it depends on the (acoustical) similarities of the tested rooms. For instance, as speculated in Section 2.1, there is high confusion between the audio data associated with *Church 1* and *Church 2*. Contrarily, *Bedroom* and *Studio* are least prone to confusion.

Table 2: Resulting equal error rates (EER)

Experiment	Music	Speech	Combined
Experiment A	15.07	8.57	13.23
Experiment B	14.71	7.67	11.28
Experiment C	32.36	15.14	23.85

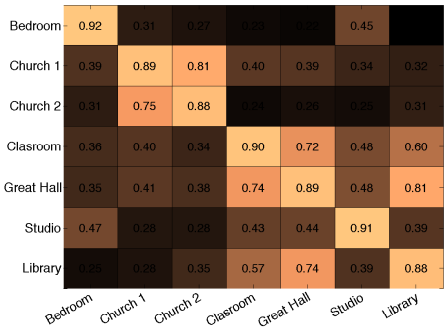


Figure 2: Confusion matrix of the estimation scores for Experiment C (music)

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 first two dimensions of the MDS are depicted in Figure 3 and clearly shows the ability of the system to separete the different rooms. Using rank correlation, we found that the first MDS dimension is well correlated with the Bass Ratio (BR) feature ($\rho(6) = -.79$), which is 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 (EDT(A)) of the RIRs ($\rho(6) = -1.0$). The EDT is based on the time in which the first 10 dB decay of the reverb occurs and is closely related to the perceived reverberance [11]. The MDS organizes the seven tested rooms in four clusters. Interestingly, these four clusters coincides with the four RIR datasets used for creating the corpus.

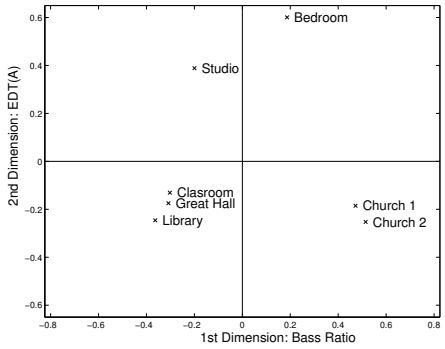


Figure 3: MDS analysis of the data shown in Fig. 2

3.3 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, [16] applied a 1 sec MFCC window size.

Using the design of Experiment C and by varying the MFCC window size from 12.5 ms to 1 sec, we measured the effect on the EER. Figure 4 shows that a larger window size leads to a higher EER. On average, the lowest EER was achieved with a size of 25 ms. This finding suggests that for room identification short-term MFCC features are more suitable than-long term MFCC features.

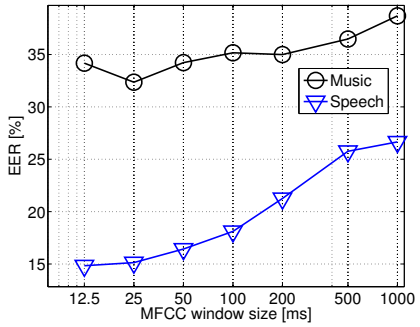


Figure 4: Effect of MFCC window size on the EER

4. 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. Using a 30 GB corpus with more than 13000 reverberant audio samples from seven different rooms, this GMM-based system was tested under various conditions. With no common content between the training and testing data, the system achieved overall accuracy of 61% for music and 85% for speech signals. Moreover, with common content between the training and testing data, the error is halved. These results are very promising and show the feasibility of using implicit audio cues for identifying the acoustical environment in a video or audio recording. To potentially improve the accuracy for music content, we want to explore additional features such as those based on the modulation spectrogram. We plan to train our system on large scale real-world audio and video datasets from Flickr and YouTube for identifying concert venues and other indoor environments.

5. ACKNOWLEDGMENTS

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

6. REFERENCES

[1] Mardy database: `http://www.commsp.ee.ic.ac.uk/~sap/uploads/data/MARDY.rar`.

[2] `http://www.1-1-1-1.net`.

[3] `http://www.emime.org/participate/emime-bilingual-database`.

[4] `http://www.openairlib.net`.

[5] Bang & Olufsen. Music for Archimedes. Audio CD.

[6] J. Bonastre, F. Wils, and S. Meignier. ALIZE, a free toolkit for speaker recognition. In *Proc. of ICASSP*, volume 1, pages 737–740. IEEE, 2005.

[7] Denon. Anechoic orchestral music recording. Audio CD, 1995.

[8] 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. In *Proc. of int'l Workshop on Acoustics Signal Enhancement*, Aachen, Germany, 2012.

[9] ISO 3382-1. *Acoustics – Measurement of room acoustic parameters – Part 1: Performance spaces.* International Organization for Standardization (ISO), Geneva, Switzerland, 2009.

[10] P. Kenny and P. Dumouchel. Experiments in speaker verification using factor analysis likelihood ratios. In *Proc. of Odyssey*, 2004.

[11] H. Kuttruff. *Room Acoustics*. Spon Press, London, UK, 2009.

[12] H. Lei, J. Choi, and G. Friedland. Multimodal city-verification on flickr videos using acoustic and textual features. In *Proc. of ICASSP*, Kyoto, JP, 2012.

[13] E. Martin, O. Vinyals, G. Friedland, and R. Bajcsy. Precise indoor localization using smart phones. In *Proceedings of the international conference on Multimedia*, pages 787–790. ACM, 2010.

[14] R. Mertens, H. Lei, L. Gottlieb, and G. Friedland. Acoustic super models for large scale video event detection. In *Proc. of ACM Multimedia Workshop on Social Media*, Arizona, USA, 2011.

[15] D. Reynolds, T. Quatieri, and R. Dunn. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1-3):19–41, 2000.

[16] 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.

[17] 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.

[19] A. Ulges and C. Schulze. Scene-based image retrieval by transitive matching. In *Proc. of the ICMR*, pages 47:1–47:8, Trento, Italy, 2011. ACM.

[20] S. Young et al. The HMM toolkit (HTK), 1995.