INFERRING ROOM SEMANTICS USING SPEECH ANALYSIS

Muhammad Ahmed Shah
Advisors: Khaled Harras and Bhiksha Raj

1 ABSTRACT

The existence of accurate semantically rich indoor maps can lead to a significant growth in indoor, location based applications. In recent year the problem of indoor mapping has received a lot of attention but, while the existing systems can generate very accurate floorpans, the do not provide semantic tags for the spaces they map. Without knowledge of the environmental context of the user location based application would remain highly limited in their efficacy.

We propose using the acoustic response generated by the user's speech to infer the characteristics of his environment. We are working toward building a system that can detect when the user is talking, records a snippet of his speech and, using learning algorithms, infers the semantics of his environment.

2 RESEARCH QUESTION AND SIGNIFICANCE

With the advancement of mobile technologies, average users are in possession of incredibly powerful devices, not only in terms of their processing power but in terms of the information that they are capable of gathering, allowing applications to become aware of the users' environment. The popularity of location-based services like location-enabled social networking, navigation and advertisement, has grown rapidly in the past decade. The more information is extracted from the available data, the more accurate, relevant and intelligent these applications would become.

Consider a day in the life of José. José has just started at his new job and he doesn't know the company building very well. He gets lost all the time and needs to ask for directions every once in a while. It gets especially embarrassing when he has to ask for directions to the restroom. To make things worse the building is undergoing renovation so areas are being closed off and some rooms are being repurposed. It would be very convenient (and much less embarrassing) if it were possible for José to view an up-to-date indoor map of the building on his phone in which the rooms had at least the high-level semantic labels and would update if certain areas were closed off or reopened. In this project we shall only tackle the problem of inferring the semantics of spaces using data available from a commodity smartphone.

The current navigation and localization technologies, such as Google Maps, use satellite services to map areas and localize users however, these technologies become extremely unreliable in indoor environments. The importance of mapping indoor environment is widely recognized and is an active area of research. The current approaches use data from cellphone sensors, such as accelerometer [2] to map the indoor environment. However, the generated map is devoid of any labels and hence has limited usefulness. Although some work has been done in leveraging multiple data sources such as wireless signals, sound, light and sensor data to label of areas on a floor plan [1][4][3], a practical and scalable solution has not been developed yet. Using wireless signal requires a dense mesh of these signals around the user which may not be readily available. Moreover, the proposed approaches require a very tedious training phase and are based on matching exact rooms rather than identifying the high-level semantics of the space

RoomSense[5] and EchoTag[6] are most closely related to our work. They leverage the fact that different environments cause the sound to be reflected differently, to identify rooms. Their results have led us to believe that using acoustics to determine the characteristics of the environment could be viable approach for inferring the semantics of the room.

We propose a light-weight system that would be able to assign a high-level semantic tag to the user's environment. Our system uses the microphones on smartphones in a scalable and non-intrusive manner to infer the semantics of the users environment. Like RoomSense and EchoTag, our approach leverages the differences in the reverberation patterns observed in different rooms. Unlike the aforementioned systems our system would not rely on a predetermined sound impulse, it would, instead, utilize speech as the input signal. We envision that our system will be capable of detecting when the user is speaking and use it as the queue to start collecting samples. By eliminating the need for direct input from the user we have made our system very convenient. Furthermore, since our approach focusses only on the assignment of high-level semantic labels, the training phase is simplified compared to that of the existing systems, such as RoomSense.

We envision the system in its entirety as the ultimate goal of the project that will extend beyond the duration of QSIURP, potentially as a senior thesis project. For the limited duration of the QSIURP grant we focused extracting as much information as possible from speech recordings conducted in several rooms in CMUQ.

3 METHODOLOGIES

3.1 Data Collection

We use two smartphones, in different locations, to conduct our recordings. One smartphone is held by the speaker in front of his/her chest while the other is kept in their

pocket. We do this because it people keep their phones in their pocket most of the times while in plaes such as lecture halls and bathrooms therefore we wanted to avoid relying on overly optimistic results by using the handheld recordings. With that said, as the scope of our project in QSIURP is restricted to developing the best technique to extract information from the recordings, all the results and experiments mentioned in this report are done on the handheld recordings. The phones are linked over wifi such that they both start recording at the same time.

We collected audio samples from 6 types of spaces in the CMUQ building:

- Bathrooms
- Small Lecture Halls (walls < 15ft)
- Large Lecture Halls (walls > 15ft)
- Offices
- Pantries

For each type of space, depending upon the availablity, we chose 4 or 5 rooms in which we conducted our recordings. In each room we conducted a total of 100 recordings at 5 locations, the 4 corners of the room and the center. At each location we took 20 recordings on each phone. The 20 recordings consist of four sentences, each uttered five times by the speaker. Our final dataset from the CMUQ building consists of a total of 4400 recordings, of which 2200 are recorded in the handheld position and are used for the experiments are results mentioned in this report. We also took 100 recordings in a bathroom in the HBKU Male Housing complex to determine how well would our approach fare on a different building.

3.2 Techniques Used

3.2.1 Data Representation

We represented the audio files as sets of mD MFCCs with 25ms windows and 10ms overlap. For Experiments 4.1 and 4.2 we used $m = 30$ and for all the experiments henceforth we used $m = 20$.

$$v_n = x_1^n, x_2^n, ..., x_m^n$$

Where v_n is the n^{th} feature vector. We also further processed the data to get our final feature vectors. The following are the different types of feature vectors we used.

- *Raw: Unprocessed MFCCs*.
- *Difference Vectors*:
  We produce the new feature vector v'_n by appending the difference $v_n - v_{n-1}$ and $v_n - v_{n+1}$ to v_n.
  $$v'_n = v_n, v_n - v_{n-1}, v_n - v_{n+1}$$
  By doing so we take into account the temporal context of the feature vectors moreover we also normalize the vectors.
- *Vectors Normalized By Feature Means*:
  The normalized feature vectors are produced by calculating the mean of each feature within the file and then subtracting it from the feature values.
  $$F_r = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} x_1^1 & \cdots & x_m^1 \end{bmatrix} \\ \vdots \\ \begin{bmatrix} x_1^n & \cdots & x_m^n \end{bmatrix} \end{bmatrix}$$
  Where F represents the feature vectors derived from recording r.
  $$\mu = \begin{bmatrix} \frac{\Sigma_i^n x_1^i}{n} & \cdots & \frac{\Sigma_i^n x_m^i}{n} \end{bmatrix}$$
  $$F'_r = \begin{bmatrix} v_1 - \mu \\ \vdots \\ v_n - \mu \end{bmatrix}$$
  Where F'_r represents the normalized feature vectors.
- *Vectors Normalized By Mean of the First Feature*:
  The normalized feature vectors are produced by calculating the mean of the first feature within the file and then subtracting it from the feature values.
  $$\mu = \frac{\Sigma_i^n x_1^i}{n}$$
  $$v'_n = x_1^n - \mu, ..., x_m^n - \mu$$
  Where v'_n represents the n^{th} feature vector after normalization.

3.2.2 Experimental Techniques

1) *Gaussian Mixture Models (GMM):*
The room characteristics become apparent, quantitatively, as feature values. A lecture hall with its carpets and furniture may absorb sound more than a bathroom with its bare ceramic walls. This would reflect in the feature vectors for the bathroom as higher amplitude values as compared to those for the lecture halls. We can model this behaviour, like so many other phenomenon, as a gaussian distribution. The features for a class of rooms would take on certain values with high probability. Since the physical environment is very complex it is possible that the data might not be accurately modelled by a single gaussian but rather by multiple gausian distributions. Therefor we employ GMMs to model our data.

GMMs have been extensively used for speaker identification [7] [8], music-speech discrimination [9]. With GMMs, each sound class (room type in our case) is modeled as a mixture of several Gaussian clusters in the feature space; each sound cluster in the feature space is represented by a mean vector and a covariance matrix.

In our experiments we trained our GMMs on feature vectors from 80% of the rooms and tested it on the remaining 20%. The training process was straightforward in that we only concatenated all the feature vectors that represented the recordings from the rooms in our training set in to one file and supplied it to the GMM library for training. We used SciKitLearn's GMM module in our experiments.

During testing, our end goal was to classify the recording however the GMM classifies each feature vector independently. To bridge the gap, we

classified each feature vector independently and assigned the recording the class that was assigned to the majority of the feature vectors.

2) *KMeans-SVM Hybrid:*
In this approach we model the recordings as a collection of concepts. A very coarse physical manifestation of a concept could be the presence of a carpet in the room. The actual concepts would be much more fine grained than the example presented above and most likely would no be percievable by us therefor we employ KMeans clustering. We construct our cluster codebook using mean normalized feature vectors from the CMUQ building and use it to quantize our recordings in to single k-dimensional features in which the i^{th} feature values would represent the number of feature vectors assigned to the i^{th} cluster (concept).

$$codebook = \{c_1, ..., c_k\}$$

$$F_r = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} \rightarrow C'_r = \begin{bmatrix} c'_1 \\ \vdots \\ c'_n \end{bmatrix}$$

Where c'_i represents the closest cluster to the vector v_i.

$$F'_r = \begin{bmatrix} count(c_1) \in C'_r \\ \vdots \\ count(c_k) \in C'_r \end{bmatrix}$$

Where F'_r is our **quantized feature vector**. In our experiments we used $k = 1024$. We used SciKitLearn's module for KMeans with KMeans++ initialization.

We use the quantized feature vectors derived from the recordings taken in 80% of the rooms from each type of space for training a SVM, while the vectors from remaining rooms are used for testing. We used SciKitLearn's bindings for LibSVM with the intersection kernel.

4 Experiments

4.1 GMM: Large Lecture Halls and Small Lecture Halls As a Single Class

Initially we merged the large lecture halls and small lecture halls into a single class of lecture halls so we were dealing with 4 classes instead of the 5 classes stated in the previous section. We took 80% of the rooms for training and 20% for testing from each type of space obtained the results presented in Figure 1 and 2.

We see from the results that the 4G and 1024G GMMs yeilded the highest overall accuracy for the raw feature vectors and the difference feature vectors respectively. To validate our results we do cross validation tests. We notice that accuracy only improves marginally as we increase the number of mixtures from 64 to 1024 while the computation time increases by more than 10 times so for the cross validation tests we use the 64G GMM with the difference feature vectors. The results follow:

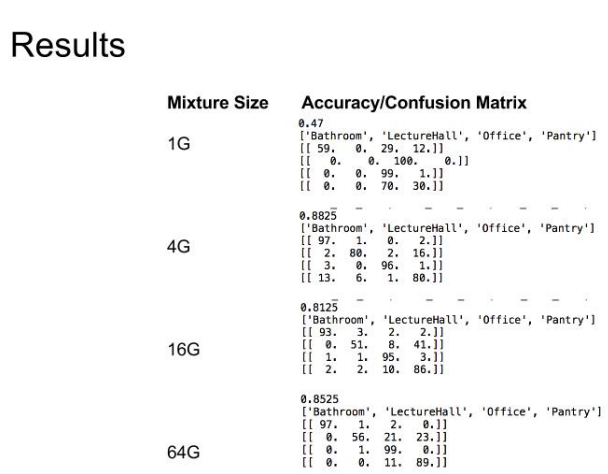

Figure 1. Results for raw feature vectors

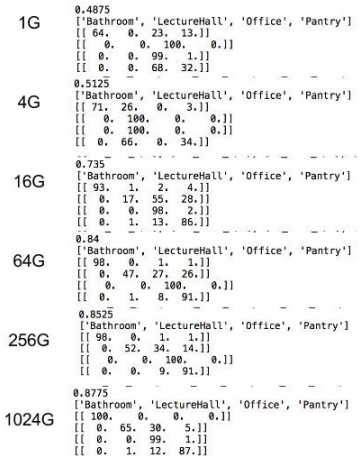

Figure 2. Results for difference feature vectors

	4G Raw	64G Difference
Maximum Accuracy (%)	92.75	86.5
Minimum Accuracy (%)	33.58	47.76
Average Accuracy (%)	65.63	73.03
Variance(Accuracies)	0.0149	0.006

From these results we see that although the maximum accuracy declined when we changed from raw feature vectors to difference feature vectors the variance in the accuracies obtained with difference combinations of the rooms in the training and test set decreased by 59.7% and the average accuracy increased by 7.4%.

From the cross validation tests we see that our randomly chosen initial spliting of the data set into testing and training set yielded overly optimistic results. We also see that further processing the feature vectors into difference feature vectors not only increased the average accuracy but also reduced the variation in the accuracies obtained for each combination of the test and training set and hence increasing the confidence we can have in our results.

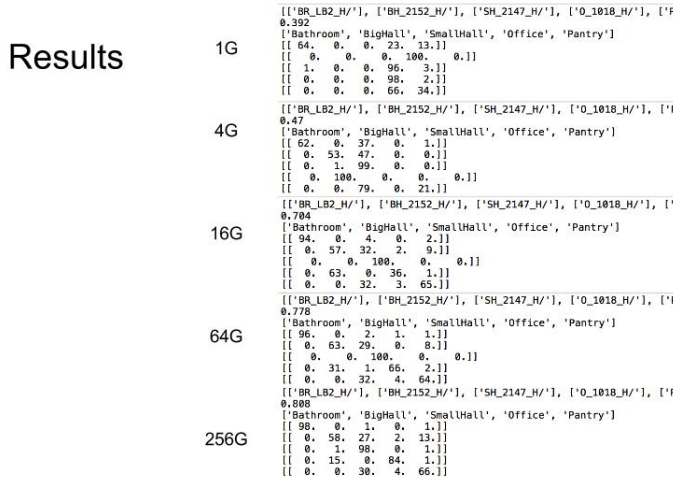

Figure 3. Results with small and large lecture halls as separate classes

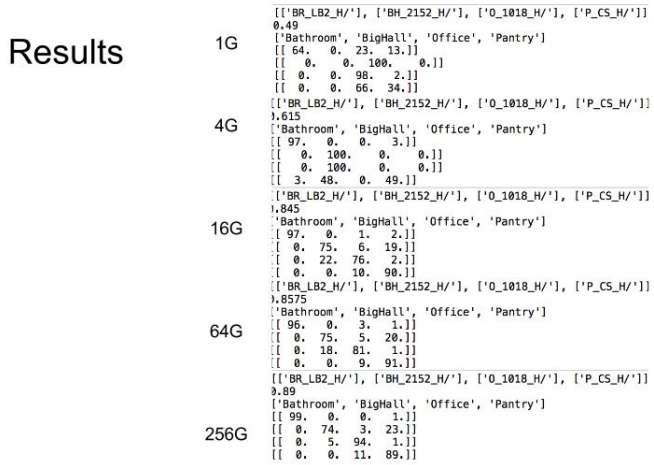

Figure 4. Results with only the large lecture halls

4.2 GMM: Large Lecture Halls and Small Lecture Halls As Different Classes

From Figures 1 and 2 it can be seen that the worst performing class for the chosen test and training set was consistently the Lecture Hall. We hypothesised that the small and large lecture halls are structurally so different that we might be able to obtain an improvement in classification accuracy if we treated them as different classes. We then re-run the same tests but with (1) small and large lecture halls as two new classes, (2) only the large lecture halls and (3) only the small lecture halls. In (2) and (3) respectively, the data for the small and large lecture halls was removed from both the training set and the test set. The results are presented in Figures 3 to 5.

We see in Figure 3 that a large number of recordings from the pantry are being misclassified as being from the small lecture hall. This observation can also be made from figure 5. From Figure 3 we also see that a large number of recordings from the big lecture halls are also being missclassified as being from the small lecture hall. From figure 3 and 5 we see that while the small lecture hall is identified

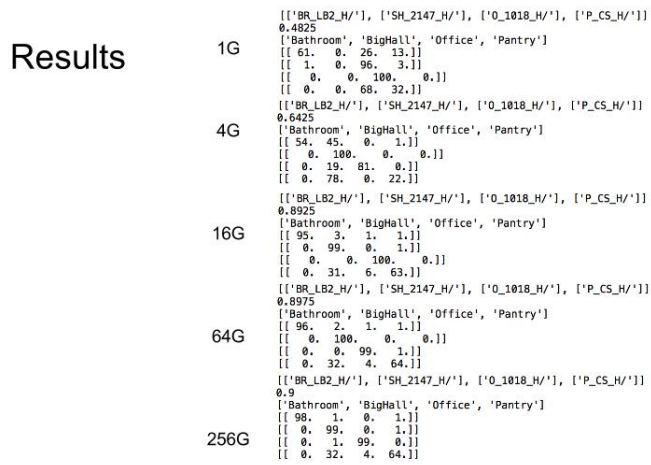

Figure 5. Results with only the small lecture halls

with high accuracy, its presence leads to a reduction in the classification accuracies of the large lecture halls and the pantry. The reason, we hypothesise, could be that the small lecture halls share characteristics of both the large lecture halls (carpets and furniture) and the pantries (dimensions).

Since only dealing with the large lecture hall produced the better results for our chosen training and testing sets we carried out cross validation tests for this set up. From Figure 4 we see that we obtain only a slight gain in overall accuracy as we increase the mixture size from 16 to 64. Although the 256G GMM performed the best, due to the required computational time we decided to use the 16G GMM for the cross validation experiments.The results are as follows:

	16G Difference
Maximum Accuracy (%)	92.23
Minimum Accuracy (%)	48.87
Average Accuracy (%)	71.46
Variance(Accuracies)	0.007

Upon closer examination of the results we find that in all our cross validation tests we obtained the lowest overall accuracy when Lecture Hall 1202, Office 1005 and Pantry 2170 were included in our test set. We hypothesise that this is so because the aforementioned office and the lecture halls have certain features that are different from other rooms in their respective classes. Lecture Hall 1202 is significantly smaller than all other large lecture halls while office 1005, unlike other offices in our dataset.

4.3 GMM: Going to 20D MFCCs

As mentioned earlier, until now we had been dealing with 30D feature vectors. Since we weren't getting the accuracies we were hoping for, we hypothesised that 30 MFCCs might be too many and some of the features would just be noise. We re-ran the the tests from experiment 4.1 and 4.2 with the new 20D dataset and got similar results. Since we did not find any appreciable difference between the results we decided to move forward with the 20D feature vectors to economise on the computation time.

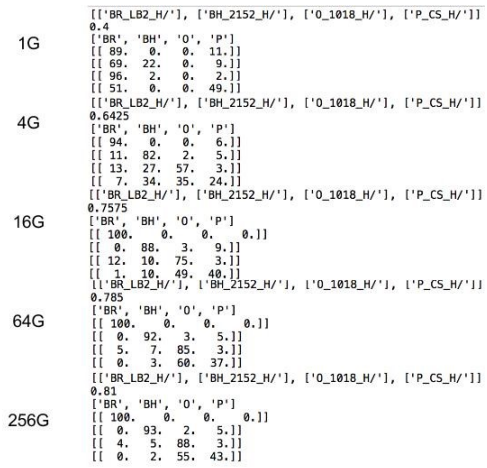

Figure 6. Results for Feature Vector Normalized by Feature Means with Only Large Lecture Halls

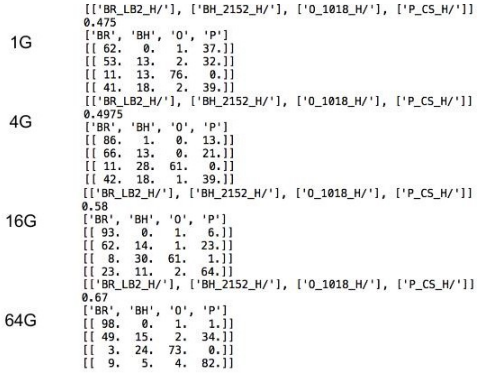

Figure 7. Results for Feature Vector Normalized by Mean of the First Feature with Only Large Lecture Halls

4.4 GMM: Normalizing the Feature Vector

In this experiment we try out two techniques for normalization i.e. normalization by feature means and normalization by the mean of the first feature. These techniques are explained in detail in Section 3.2.1. The classification accuracies are presented in Figures 6 to 8.

Figure 6 shows that in our chosen test and training set we achieved a slightly lower accuracy compared to the results in Figure 4. In order to ascertain that this normalization did indeed negatively impact the overall accuracy we performed a cross validation test using the 64G GMM.

The results are as follows:

	64G Mean Normalized
Maximum Accuracy (%)	93.5
Minimum Accuracy (%)	43.5
Average Accuracy (%)	72.12
Variance(Accuracies)	0.013

Both the maximum and Average accuracies increased from those in Experiment 4.2. We also noticed that the rooms that consistently yielded the minimum accuracies in Experiments 4.1 and 4.2 reappeared here.

We also concluded that normalizing by the mean of the first feature is not a direction we would want to explore further since it yielded abyssmal results on the same training and test set used in all the other experiments.

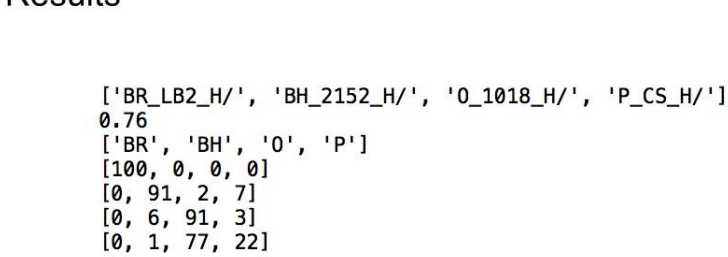

Figure 8. Results obtained with K-Means-SVM Hybrid with 1v1 multi-class classification

4.5 K-Means-SVM Hybrid: Standard One v One Multi-class Classification

The K-Means-SVM Hybrid technique is explained in detail in section 3.2.2. We construct one SVM for each space type. We use the same rooms in the training and testing sets as the previous experiments. We used $k = 1024$ for our experiments. The results are presented in Figure 8.

From Figure 8 we see that we acheive a very high accuracy in classifying all types of spaces except pantries.

4.6 K-Means-SVM Hybrid: Binary Classification

This setup is similar to the previous except instead of running the SVMs pairwise to perform multiclass classification we constructed one vs all SVMs for each class (i.e. the feature vectors of the class for which the SVM is constructed are assigned the label $+1$ while the feature vectors of all the other classes are assigned the label -1). For the experiment we used the same set of rooms in the training and test sets as the previous experiments. Since the SVMs would give a binary response, we check which SVM returned $+1$ and we assign the label of that SVM to the feature vector. If multiple classes returned $+1$ we chose the first one according to the order of the SVMs in our datastructure. If none of the SVMs returned $+1$ we would not a assign that feature vector any labels. The results are presented in Figure 9.

Some insight that we garner from this experiment is that while a significant number of feature vectors were *un*classified, not many we *mis*classified (except in the case of pantries). In a real-world setting it might be acceptable to leave some recordings unclassified, since we would probably be crowdsourcing the data and would have access to a huge volume of recordings.

5 Conclusions and Future Work

We have demonstrated from our experiments that it is feasible to use voice recordings taken from smartphones as

	GMM(30D)				GMM(20D)					KMeans + SVM (1024D)			
	raw		diff		raw	diff	norm		C0	1v1	multiclass	Physical	Binary SVMs
LH = SH + BH	88.25 (4G)	65.63 (4G)*	87.75 (1024G)	73.03 (64G)*									
BH			89.0 (256G)	71.46 (16G)*	84.0(16G)	90.5 (64G)	81.0 (256G)	72.12 (64G)*	67.0 (64G)	76	75.75	74.5	64.5
SH			90.0 (256G)		78.25(64G)	85.0 (64G)	76.25 (256G)						
SH & BH			80.8 (256G)										

Results

```
['BR_LB2_H/', 'BH_2152_H/', 'O_1018_H/', 'P_CS_H/']
0.645
93
['BR', 'BH', 'O', 'P']
[100, 0, 0, 0]
[0, 74, 1, 3]
[0, 2, 63, 2]
[0, 0, 41, 21]
```

Figure 9. Results obtained with K-Means-SVM Hybrid with Binary Classification

a feature in room semantic identification tasks. While our results are not strong enough to suggest that voice recording can be used independently we are optimistic that when comibined with other modalities such as location, visual and sensor data we could achieve the desired level of accuracy.

Moving on with this project we plan to improve our data set by including a more diverse set of recordings while also looking into more sophisticated techniques to analyse these recordings. By diversifying our set of recordings we mean to include recordings from more speakers and obtain the recordings in different buildings. So far our recordings consist of a single speaker and are taken exclusively in the CMUQ building. Therefore we can not be confident that our results would hold if a different speaker was introduced or the location was changed. We would also need to look into more techniques to analyze this data in order to achieve consistent classification accuracy of over 90%.

We also plan to include more modalities such as location, visual and sensor data in our system to augment the results from audio analysis. Our results from further exploration of voice recordings as means of room identification will guide our decision to either retain voice samples as the prime modality in our Semantic Identification task or to relegate it to same status as other modalities.

6 References

1) Bao, Xuan, et al. "PinPlace: associate semantic meanings with indoor locations without active fingerprinting." Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM, 2015.
2) Alzantot, Moustafa, and Moustafa Youssef. "Crowdinside: automatic construction of indoor floorplans." Proceedings of the 20th International Conference on Advances in Geographic Information Systems. ACM, 2012.
3) Elhamshary, Moustafa, and Moustafa Youssef. "SemSense: Automatic construction of semantic indoor floorplans." Indoor Positioning and Indoor Navigation (IPIN), 2015 International Conference on. IEEE, 2015.
4) Azizyan, Martin, Ionut Constandache, and Romit Roy Choudhury. "SurroundSense: mobile phone localization via ambience fingerprinting." Proceedings of the 15th annual international conference on Mobile computing and networking. ACM, 2009.
5) Rossi, Mirco, et al. "RoomSense: an indoor positioning system for smartphones using active sound probing." Proceedings of the 4th Augmented Human International Conference. ACM, 2013.
6) D. Reynolds and A. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. IEEE Trans. on Speech and Audio Processing, 3(1):72–82, 1995.
7) E. Scheirer and M. Slaney. Construction and evaluation of a robust multi-feature speech/music discriminator. In Proc. IEEE ICASSP, 1997.
8) A. P. Schmidt and et al. Reduced-rank spectra and minimum entropy priors for generalized sound recognition. In Music Classification and Identification System. www.trevorstone.org/school/ MusicRecognitionDatabase.pdf.