INFERRING ROOM SEMANTICS USING CONTINUOUS ACOUSTIC MONITORING

Muhammad Ahmed Shah⋆ *Bhiksha Raj*⋆ *Khaled Harras*⋆ *Anurag Kumar*⋆

⋆ Carnegie Mellon University

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 accurate floorpans, they 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. In this paper we propose a dataset and methods for inferring room semantics using acoustic monitoring. We propose using the acoustic response generated by the user's speech, as a modality, to infer the characteristics of his environment. One important assertion in this work idea is that evidence for any semantic tag can accumulate over time, and hence as more data keeps coming in, the system should change its output.

Index Terms— Indoor Mapping, Semantics, Acoustic Monitoring

1. INTRODUCTION

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. Moreover, 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.

Most of the current navigation and localization technologies (*e.g Google Maps*) are extremely unreliable in indoor environments. The importance of mapping indoor environment is widely recognized and is an active area of research [1][2][3]. Some 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. There have also been a few works on leveraging multiple data sources such as wireless signals, sound, light and sensor data to label of areas on a floor plan [1][2][3]. However, 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.

Moreover, the area of audio analysis has mainly focused on identifying particular classes of sound events or acoustic scenes in audio recordings gained. The recent DCASE Challenge [14] includes focused on detection and classification of acoustic scenes and events. Several approaches have been proposed for audio event and scene detection tasks [16][17][18]. However, all of these works deal with relatively well defined sound classes compared to the rather murkey, even overlapping categories in semantic mapping. For example a pantry could also be a printer room, and hence identifying purpose of the indoor environment is harder. work.

There have been a few works in area of indoor environment recognition. RoomSense [5] and EchoTag [6] leverage the fact that different environments cause the sound to be reflected differently, and hence rooms can be identified by using an active sensing technique. However, RoomSense does not deal with room identification rather it focuses on room and within-room level localization. On the other hand EchoTag only focuses on identifying environments upto a square meter around the the smartphone. Furthermore these active sensing techniques with predefined sound impulses are impractical in a real world setting because it may become annoying for users if their phone is periodically emitting a high-frequency sound. Perhaps the area of work most closely related to ours is that of audio forensics. Zhao and Malik[13] attempt to identify the environments of speech recordings by separating out the noise and reverberation components from the recordings and using a SVM for the learning and classification. Malkin and Waibel [12] used a passive sensing approach with linear autoencoders to classify recordings of ambient sounds in different environments. While these works indeed perform room level identification, they do not apply them in the context of indoor mapping.

In this work we work on the problem of inferring the semantics of indoor spaces using data available from a commodity smartphone. While the problem of inferring the indoor floorplan has received much attention and the research community has produced viable solutions [2][19][20], the problem of inferring the usage of an indoor space has not been well explored. Specifically, we employ continuous acoustic monitoring to infer the current usage (semantic tag) of indoor spaces. Our approach is based on the observation that different physical features of spaces, such as dimensions, furniture and construction materials used, influences the acoustics of the space. Hence, the acoustic signature of a room can be exploited to infer information about semantic tag of that space.

When considering a practical and deployable indoor mapping application certain constraints may be relaxed while some other may be imposed. In a crowdsourced mapping system the abundance of data can be assumed and the goal of the system becomes to confidently assign a label to a space after seeing some reasonable number of samples. On the other hand, there are severe restrictions on the quality of recording devices available. Overwhelmingly the sound samples available to the system would be recorded on a smartphone and hence would be prone to excessive noise and lower fidelity. Moreover, sound quality may be compromised due to the motion and the position (for example, the phone may be in the pocket) of the recording device. Moreover, we are interested in semantic mapping where we require accumulation of evidence for labels over time since the spaces may be re-purposed. These challenges are yet to be tackled effectively by the research community.

Currently, to the best of our knowledge there are not many resources in form of public datasets in this problem domain. In this paper we attempt to address this issue by creating an audio based dataset tailor made for indoor semantic mapping problem[1]. We have compiled a large dataset of speech recordings consisting of two male speakers saying five sentences in multiple locations in different types of rooms. We are making this data public to encourage interest in using acoustic monitoring for indoor mapping and facilitate research in this area. We then propose two methods for inferring semantic tags using audio data. Our first proposed methods is based on Gaussian Mixture model as classification technique. In the second approach we use GMM based supervectors [CITE] along with Support Vector Machine as classification method. An important aspect of our work is that analyze how labels change as more evidence is accumulated in form of additional data. More concretely we want to see, for example, how would having 60s of audio data as opposed to of 30s of audio data effect our confidence in labeling a restroom, as restroom. This scenario is intended to demonstrate the real-life performance of a crowdsourced system in which there is a constant influx of data.

The rest of the paper is organized as follows: Section 2 gives a description of our dataset; Section 3 describes our proposed approaches. We provide experiment and results in section 5 and finally we conclude in Section 6.

2. THE DATASET

Our data set consists of more than 7400 audio recordings of human speech. The speech is produced by two male undergraduate students and is recorded by two smartphones simultaneously. One of the smartphones is held by the speaker at chest, while the other is in his pocket. The speech consists of 4 short sentences. The recordings were performed in restrooms, offices, pantries, classrooms and big halls. Hence, our dataset provides 5 semantic tags in indoor environment, namely *Restrooms, Offices, Pantries, Classrooms, Big Halls*. To capture within class variation in acoustic signatures of these rooms we record in several different locations of the same class. Based upon availability we took recordings in 5 Restrooms and Offices, and 4 Classrooms, Pantries and Big Halls. In each room we took 100 recording, 3 seconds each. The speaker stood at 5 locations within the room (the corners and the center) and repeated each sentence 5 times. The recordings were obtained at times when there was low foot traffic so noise and other sounds are minimized. The recordings are taken over a single channel sampled at 44.1 KHz.

3. METHODS

3.1. Data Representation

We represented our data with 20D MFCC feature vectors. We used 64ms windows with 16ms of overlap to obtain these MFCC vectors.

3.2. Dataset Organization

We used a 4-fold cross-validation approach in all our experiments. We partioned our dataset in to four folds. Conveniently we had recordings from exactly four pantries, big halls and classrooms so we placed data from 1 room of each type in each fold. We had recordings from five bathrooms and offices so we placed data from two restrooms and pantries in one fold. In all experiments we trained on the data from 3 folds and tested on the data from the remaining fold.

Room Type	EER	AUC
Restroom	0.083	0.033
Big Hall	0.261	0.19
Classroom	0.226	0.15
Office	0.333	0.27
Pantry	0.181	0.10

Table 1. Equal Error Rate(EER) and Area under the Curve(AUC) for GMMs averaged across 4 folds

3.3. Gaussian Mixture Models

We employed a binary classification approach using GMMs, i.e. for each type of room we trained two GMMs, one trained on training data from type of room in question (G_1) and the other trained on training data from all the other rooms (G_{-1}). In our experiments we used a 64G GMM. When testing we first score each recording in the training set against both the GMMs and assign the relevant labels to them. We then partition the results into two subsets based on if the recording came from the room in question or if it came from another room, lets call these two subset S_1 and S_{-1} respectively. We will use S_1 to determine how the percentage of true positives varies as more data is seen and we will use S_{-1} to determine how the percentage of false positives varies as more data is seen. The data further subdivided into n segments, $h_{1_1}, ..., h_{n_1} \subset S_1$ and $h_{1_{-1}}, ..., h_{n_{-1}} \subset S_{-1}$, where each segment contains m recordings from our dataset. Since the recordings are all exactly 3 seconds long, each segment will have $3m$ seconds of audio data. We denote the rate of true positives and false positives just after observing the segment h_k as $T_P(k)$ and $T_N(k)$ respectively.

$$T_P(k_1) = \alpha_k(T_p(h_{(k-1)_1}))+(1-\alpha_k)\left(\frac{|\{label(r_1) = 1|r_1 \in h_{k_1}\}|}{m}\right)$$

$$T_N(k_{-1}) = \alpha_k(T_p(h_{(k-1)_{-1}}))+(1-\alpha_k)\left(\frac{|\{label(r_1) = 1|r_1 \in h_{k_{-1}}\}|}{m}\right)$$

α_k is used to weigh the accumulated evidence against the newly acquired evidence. It is calculated as follows:

$$\alpha_k = min(0.95, \frac{k-1}{k})$$

The label function is defined as follows:

$$label(r_1) = \begin{cases} 1 & \frac{P(G_1)}{P(G_{-1})} \geq t \\ 0 & o.w \end{cases}$$

t is the threshold value. In the simplest case this would 1 so that if $P(G_1)$ is equal to or even slightly greater than $P(G_{-1})$ we will label it 1. However, in our application we prioritize the purity of our labels in order to confidently assign a label to a space so we choose a t that would rescrict the amount of false positives we generate. To achieve a balance between the false positive rate and false negative rate we set t to the Equal Error Rate (EER) obtained from the Detection Error Tradeoff (DET) curves. The EER and Area Under the Curve (AUC) for each of the room trypes is presented in Table 1.

[1] Dataset is publicly available https://github.com/shinigami1494/QSIURP2016

Room Type	EER	AUC
Restroom	0.33	0.21
Big Hall	0.45	0.44
Classroom	0.34	0.36
Office	0.46	0.48
Pantry	0.36	0.33

Table 2. Equal Error Rate(EER) and Area under the Curve(AUC) for SVMs averaged across 4 folds

3.4. GMM Supervectors with SVM Classifier

We pool all the training data, irrespective of the room type it came from to construct a universal GMM G_U. Then for each individual recording in both the training and test set we use Maximum Aposteriori (MAP) adaption [21][22] to adapt the parameters of G_U to the MFCC vectors of the recording in order to produce a single high-dimensional supervector.

As with GMMs, we applied a binary classification approach here as well. For each room type we train a binary SVM using a linear kernel. We calculate the rate of true positives and false positives over time using the equations for $T_P(k)$ and $T_N(k)$ respectively. We however slightly change the $label$ function because unlike GMMs, instead of providing probabilities for both the positive and negative classes the SVM outputs a single value that represents the score for the class +1.

$$label(r_1) = \begin{cases} 1 & P(class = 1) \geq t \\ 0 & o.w \end{cases}$$

t, like for the GMMs, is the equal error rate determined from the DET curves for the SVM classifier. The EER and Area Under the Curve (AUC) for each of the room trypes is presented in Table 2.

4. EXPERIMENTS AND RESULTS

We evaluate our approaches using 4 fold cross validation. Since our data set contains 5 rooms of type restroom and office one fold contains two rooms of these types. Though this doesn't effect the training phase, we remove these rooms from the testing phase since they extend the span of the observation window and for some time intervals the only evidence we have are the recordings from this fold resulting in anomalous rises and falls in the true positive and false positive rates. We use the difference between the true positive rate and false positive rate to evaluate how confidently can we be in assigning a particular label to a space. We also present DET curves and the area under them as baseline results for individual sample based classification.

4.1. GMMs

We see, from Figure 1, that the rate of false positives is significantly lower than that of true prositives. This is encouraging because it means that it is unlikely for rooms to be misclassified but may remain unclassified. In the scenario we envisioned this should not be a problem because we would have ample data so we could afford to discard some. From these results we see that the restroom has the best and most consistent performance. This was to be expected since the acoustice response generated by a restroom is very dissimilar one generated by the other room types. We also notice that the variance in the rate of false positive goes down as more recordings are added so could expect our confidence to converge eventually to a stable value. This is contrary to our expectation that as more recordings are added our confidence would have increased. We could have obtained even higher true positive rates had we not modified the classification theshold to use the EER, however this would come at the cost of a higher false positive rate. In our application leaving the room is to be preferred to assigning an incorrect label to it therefore we shall want to minimize false positives as much as possible. Figure 2 presents the DET curve for the restroom class. Fold 3 follows the axes and has AUC = 0 so it was perfectly classified.

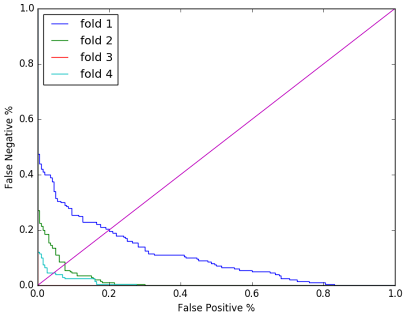

Fig. 2. DET Curve for restroom using GMMs

4.2. GMM Supervectors with SVM

The results we obtained from using an SVM to classify GMM supervectors were below our expectations. We based our choice of approach on the success that GMM Supervectors have seen in speaker identification tasks and several other acoustic classification tasks. As presented in Table 2, the average AUC across 4 folds for the SVMs is significantly higher than that for the GMMs. This is reflected in the true positive and false negative rates over time, presented in Figure 3. Exept the restroom all classes performed poorly with the false positive rates approaching or even surpassing the true positive rates.

5. CONCLUSION

We have demonstrated from our experiments that it is feasible to use voice recordings taken from smartphones as 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. We also, using our experimental results, demonstrate that the GMM supervectors may not be as suited to room semantic identification as they are to other applications such as speaker recognition.

We believe that this is a promising field that has yet to recieve much attention. Automated semantic tagging is a huge step in developing robust and practical indoor mapping solution. By making our dataset public and providing baseline results we hope to encourage and support the research community willing to engage in research in this field.

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.

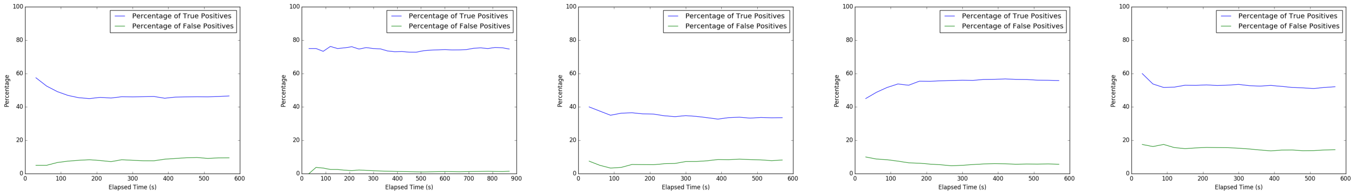

Fig. 1. True Positive and False Positive Rate, from left to right, for Big Hall, Restroom, Office, Pantry and classrooms using GMMs

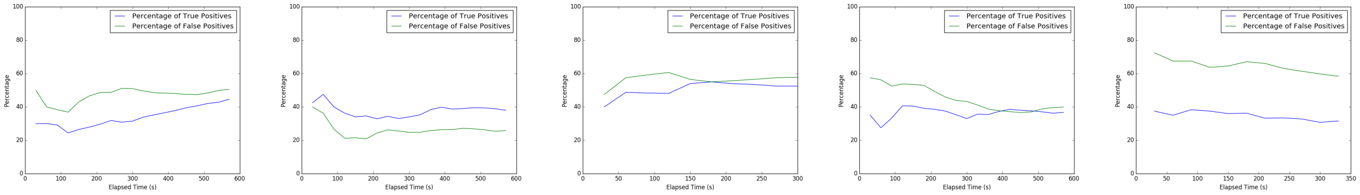

Fig. 3. True Positive and False Positive Rate, from left to right, for Big Hall, Restroom, Office, Pantry and classrooms using GMM Supervectors with SVMs

Eventhough we have more than 7400 recordings, the length of time they cover is not sufficient to draw robust inferrences yet this is most evident in the results for the classroom in Section 4 which has a constant 0.25 rate of true positive, this is because the classifier classified every recording the same.Therefore we plan on enlisting people on campus to install our android application and provide us with much needed data.

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.
9. Lee, Keansub, Daniel PW Ellis, and Alexander C. Loui. "Detecting local semantic concepts in environmental sounds using markov model based clustering." 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2010.
10. Rajapakse, Menaka, and Lonce Wyse. "Generic audio classification using a hybrid model based on GMMs and HMMs." 11th International Multimedia Modelling Conference. IEEE, 2005.
11. Lee, Keansub, and Daniel PW Ellis. "Audio-based semantic concept classification for consumer video." IEEE Transactions on Audio, Speech, and Language Processing 18.6 (2010): 1406-1416.
12. Malkin, Robert G., and Alex Waibel. "Classifying user environment for mobile applications using linear autoencoding of ambient audio." Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.. Vol. 5. IEEE, 2005.
13. Zhao, Hong, and Hafiz Malik. "Audio recording location identification using acoustic environment signature." IEEE Transactions on Information Forensics and Security 8.11 (2013): 1746-1759.
14. http://www.cs.tut.fi/sgn/arg/dcase2016/
15. Campbell, William M., Douglas E. Sturim, and Douglas A. Reynolds. "Support vector machines using GMM supervectors for speaker verification." IEEE signal processing letters 13.5 (2006): 308-311.
16. Pancoast, Stephanie, and Murat Akbacak. "Bag-of-Audio-Words Approach for Multimedia Event Classification." Interspeech. 2012.
17. Kons, Zvi, and Orith Toledo-Ronen. "Audio event classification using deep neural networks." INTERSPEECH. 2013.

18. Ballan, Lamberto, et al. “Deep networks for audio event classification in soccer videos.” ICME. 2009.

19. Shin, Hyojeong, Yohan Chon, and Hojung Cha. “Unsupervised construction of an indoor floor plan using a smartphone.” IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42.6 (2012): 889-898.

20. Gao, Ruipeng, et al. “Jigsaw: Indoor floor plan reconstruction via mobile crowdsensing.” Proceedings of the 20th annual international conference on Mobile computing and networking. ACM, 2014.

21. J. Gauvain and C. Lee, “Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains,” Speech and audio processing, IEEE Trans. on, 1994.

22. F. Bimbot and et al., “A tutorial on text-independent speaker verification,” EURASIP journal on applied signal processing, vol. 2004, pp. 430–451, 2004.