# Unsupervised Environment Recognition and Modeling using Sound Sensing

Arnold Kalmbach, Yogesh Girdhar, and Gregory Dudek

***Abstract*— We discuss the problem of automatically discovering different acoustic regions in the world, and then labeling the trajectory of a robot using these region labels. We use quantized Mel Frequency Cepstral Coefficients (MFCC) as low level features, and a temporally smoothened variant of Latent Dirichlet Allocation (LDA) to compute both the region models, and most likely region labels associated with each time step in the robot's trajectory. We validate our technique by showing results from two datasets containing sound recorded from 51 and 43 minute long trajectories through downtown Montreal and the McGill University campus. Our preliminary experiments indicate that the regions discovered by the proposed technique correlate well with ground truth, labeled by a human expert.**

## I. Introduction

In this paper we introduce an acoustic environment modeling framework, which can be used to automatically identify different acoustic regions in the world from a continuous audio recording of the environment, in a completely unsupervised manner. Transitions in acoustic space often correlate with transitions in other characteristic properties of the environment. Hence, for an autonomous robot, the ability to detect and recognize different acoustic regions can find use in improving many common robotic tasks. For example, sounds of people chatting could be used to trigger higher safety gains for a mobile robot, even though no person might yet be in sight; and the opposite where the ambient sounds of a breezy quite outdoor space could be used to make the robot lower its safety in the interest of higher speeds. If the robot has a model of the acoustic region it is currently immersed in, then this information could also be used to detect surprising and interesting events. For a surveillance robot, surprises in the environment could be used to request human attention, or just simply collect data at a higher resolution.

An overview of our proposed technique is shown in Fig. 1. As a robot traverses an environment, the goal is to interpret the audio as observations, which are the result of the robot being in a particular acoustic region. We would simultaneously like to come up with a description of different regions which exist in the world, and at the same time predict what region was the robot present in at each time step. This is akin to the well known Simultaneous Localization and Mapping (SLAM) problem in robotics.

Previous work [1] in environmental sound classification and recognition indicates that Mel Frequency Cepstral Coefficients (MFCC) perform well as low level features. We first compute the MFCCs for short windows of the audio,

The authors are at: Center for Intelligent Machines, McGill University, Montreal, QC H3A0E9, Canada {akalmbach,yogesh,dudek}@cim.mcgill.ca

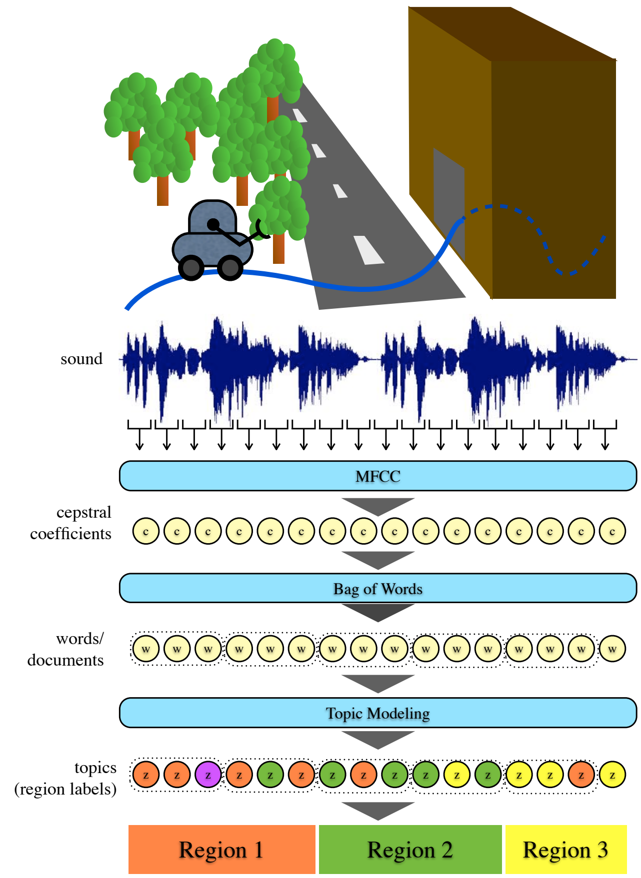



Fig. 1. *Overview*: As the robot traverses a trajectory through an unknown environment, we would like to model and recognize the different acoustic regions faced by the robot. We use quantized MFCC features and a temporally smoothened variant of LDA to discover these regions.

and then quantize MFCCs into acoustic "words" using of bag-of-words technique, by applying a fixed sized, precomputed vocabulary generated from another sound source. The acoustic words are then grouped together into documents spanning a few seconds of sound.

In text analysis, Latent Dirichlet Allocation (LDA) [14] has been used to model a set of documents as a mixture of abstract topics. These topics have been shown to be consistent with human understanding of the documents. Similarly, we model different temporal windows of the recorded sound as a mixture of sounds from different regions. Textual words used in LDA are replaced by MFCC words, and we hypothesize that the textual topics can be interpreted as different acoustic regions in the world. To take advantage of temporal continuity of these "documents", we propose an extension to the LDA which produces topic label distributions which are smooth in temporal space. We show that by taking into account information from a document's temporal neighborhood, the performance of the algorithm is vastly improved.

We present results using two datasets containing sound recorded from 45 minute long trajectories through downtown Montreal and the McGill University campus. Our preliminary results indicate that the regions discovered by the proposed technique correlate well with ground truth labeled by a human expert.

## II. Related Work

In the context of robotics, Chu et al.[2] have worked on recognizing sound environments, comparing K-Nearest Neighbors, Gaussian Mixture Models, and Support Vector Machines to model a wide set of audio features, including MFCCs and others. In addition, Roy et al.[3] have used sound captured by a boom-mounted microphone, tapping on the floor for sensing the environment.

### A. Classification of Environmental Sounds

In general, the classification and recognition of environmental sounds has received substantial attention in the domain of context-aware computing systems [4] and in forensic audio surveillance [5]. These approaches use Hidden Markov Models and other supervised learning techniques to classify sound clips as belonging to one of several predefined classes [4][5][6][7]. In general, they have achieved high degrees of accuracy, however the supervised approach depends on a pre-knowledge of the kinds of sounds which will be presented. This can be problematic in robotics, where we often do not know what kinds of environment will be encountered, either because the robot is exploring them for the first time or because we do not want to have to predict where the robot will go. Previous works, however, have mostly been focused on picking the best features, rather than how to model them, and we rely on their insights in this question.

Two kinds of features have been most commonly used; the Mel Frequency Cepstral Coefficients and the MPEG-7 [8] Standard. Both feature sets were designed for other applications, and criticisms of their application to ambient sounds do exist. Specifically, both feature sets rely on spectral characteristics of the audio, which work well for speech and music, but are problematic for more non-stationary signals. Nevertheless both have been used commonly as benchmarks for features in environmental sound models[9]. In addition, it has been shown that both feature sets do give acceptable results [10]. Because of the relative complexity involved in calculating the MPEG-7 features, with the potential application of running our algorithm online in mind we have chosen to use MFCCs.

## III. Approach

Given the audio captured by the robot as it traverses an environment, we first compute the Mel Frequency Cepstral Coefficients (Section III-A) for small overlapping sound windows. We then use the bag-of-words representation for each 20 second segment of the sound (Section III-B), and describe it as a histogram over these audio words. Finally, we use a temporally smoothened topic modeling framework (Section III-C) to model the hidden random variable representing the source of the audio words.

### A. MFCC

The Mel Frequency Cepstral Coefficients are a compact representation of the spectrum of a signal, which like human perception, give stronger weight to lower frequencies than higher ones. Decorrelating the weighted spectral components using a discrete cosine transform, we get a concise description of the timbre of a window. To achieve such a weighting, we construct a filterbank $\hat{h}_l[k]$ with $L$ overlapping bands with triangular magnitude responses weighted such that each has equal area. Let $\hat{x}[k,p]$ be the DFT of input signal $x[n]$ taken on the window from $n=[k,p]$ of length $P$. Then the MFCCs are defined:

$$cc_x[m,p] = \beta(m) \sum_{l=1}^{L} \left( \sum_{k=0}^{P-1} |\hat{x}[k,p]\hat{h}_l[k]| \right) cos\left[ \frac{m\pi}{L} \left( l - \frac{1}{2} \right) \right] \tag{1}$$

where the normalization factor $\beta$ is defined:

$$\beta(m) := \begin{cases} \sqrt{\frac{1}{L}}, & m = 0 \\ \sqrt{\frac{2}{L}}, & m > 0 \end{cases} \tag{2}$$

More detail on the construction of the filterbank and on choosing appropriate parameters for the window size and number of coefficients to calculate can be found in [11].

### B. MFCC Bag-of-Words

The bag-of-words descriptor for text document is essentially a histogram of words, where similar words contribute towards the same bin. This idea has been borrowed from image modelling, where a textual word could be replaced by visual words, as described by Sivic et al. [12]. For images, first a vocabulary is generated by extracting visual features from an unrelated dataset, and clustering these features using the $k$-means algorithm with a large $k$ value. The cluster centers can then be used as a vocabulary. Now, for a new image, we simply extract the visual features, and map each feature to the closest vocabulary word.

Similar to visual bag-of-words technique, we first extract MFCC features from a unrelated audio sample, and cluster them to generate a vocabulary. Then for the given audio sample, we simply map each of its extracted MFCC feature to the closest audio word in the vocabulary. It is important to use an unrelated audio sample to generate the vocabulary to ensure an unbiased bag-of-words descriptor.

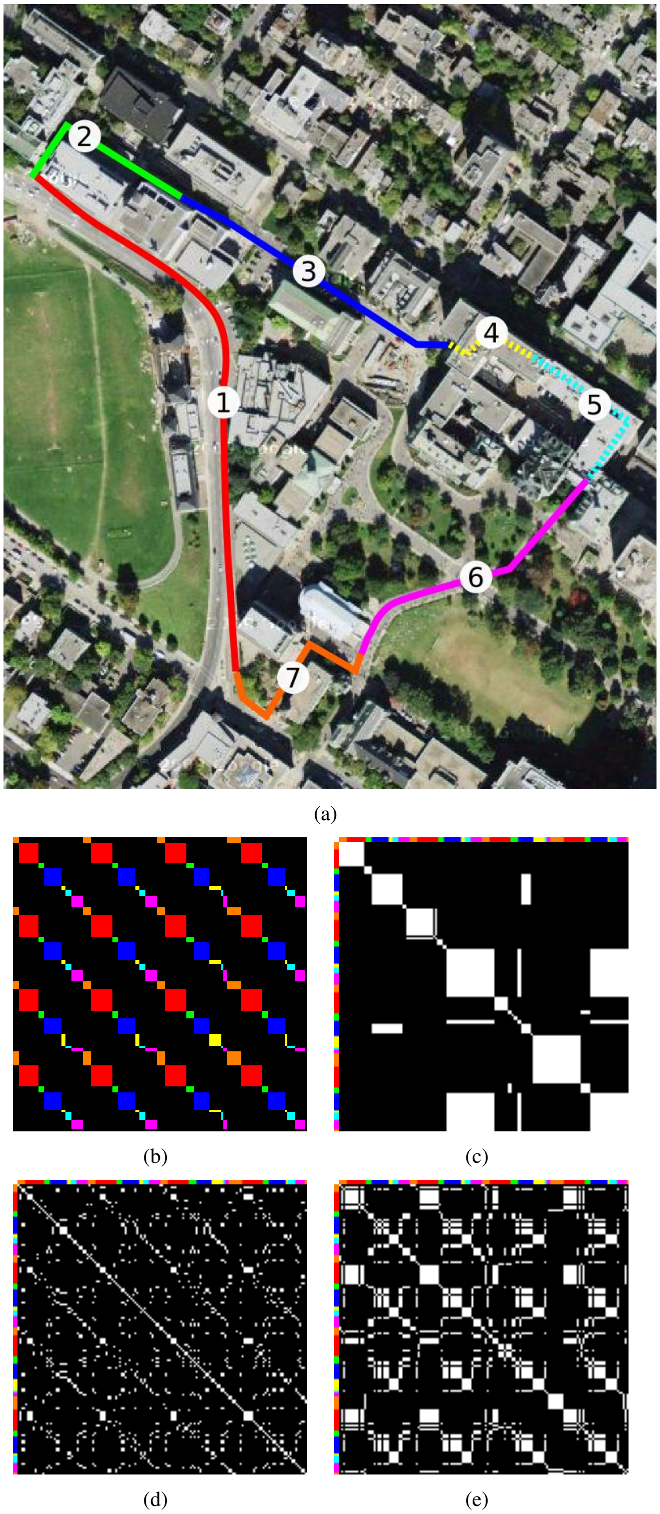



Fig. 2. *Four-loops dataset:* (a) Map showing the path traversed while recording the dataset. Different regions from the ground truth are shown with different colors. The loop was traversed four times in the dataset. (b) Ground truth similarity matrix, in which element $(i, j)$ is colored if the robot was in the same region at times $i$ and $j$. Regions are colored to match the colors in the map. (c) Best performing similarity matrix produced by matching only the most frequent MFCC words at each time step. The matrix correspond to a neighborhood size $g = 12$. (d). Similarity matrix produced by matching the most frequent topic label for each time window. Topics were generated using simple LDA, with no use of temporal neighborhood. (e). Best performing similarity matrix produced by matching the most frequent topic label for each time window, using topics generated by temporally smoothened LDA with neighborhood size $g = 4$.

### C. Audio Topics

We would like to model different regions responsible for producing different MFCC words described above. Let there be $K$ different acoustic regions in the world. Let $d_t$ denote a time window around $t$, and $\{w_i^t\}$ be the set of words observed in the time window. For each word we would like to compute the most likely region responsible for producing that word. Let $z_i^t$ be this region label, responsible for producing word $w_i^t$.

Topic modeling methods were originally developed for text analysis. Hofmann [13] introduced the idea of probabilistic Latent Semantic Analysis (PLSA) for text documents, where the probability of observing word $w_i$ in a given document (or in our case, temporal window) $d_t$ was defined as:

$$\mathbb{P}(w_i^t|d_t) = \sum_{k=1}^{K} \mathbb{P}(w_i^t|z_i^t = k)\mathbb{P}(z_i^t = k|d_t), \tag{3}$$

where $w_i^t$ takes a value between $1 \ldots V$, and $z_i^t$ is the hidden region or topic label for $w_i^t$, which takes a value between $1 \ldots K$. The central idea being the introduction of a hidden variable $z$, which models the underlying topic, or the context responsible for generating the word. We can model each temporal window using a distribution $\theta_t(k) = \mathbb{P}(z_i^t = k|d_t)$ over these topics, and model each topic using a distribution $\phi_k(v) = \mathbb{P}(w = v|z = k)$ over the set of vocabulary words.

Latent Dirichlet Allocations, proposed by Blei et al. [14] places Dirichlet priors on $\theta$ and $\phi$, which have been shown to result in semantically more relevant topics. Griffiths et al.[15] subsequently proposed a collapsed Gibbs sampler for LDA, where the state is topic assignments for all the words in all the documents. The Gibbs sampler proposes to sample the topic labels from the distribution:

$$\begin{aligned}\mathbb{P}(z_i = k|\mathbf{z}_{-i}, \mathbf{w}) \propto\ & \frac{n_{k,-i}^v + \beta}{\sum_{v=1}^{V}(n_{k,-i}^v + \beta)} \cdot \\ & \frac{n_{t,-i}^k + \alpha}{\sum_{k=1}^{K}(n_{t,-i}^k + \alpha)}\end{aligned} \tag{4}$$

where $v = w_i$ is the word label, $k = z_i$ is the new topic label, $n_{k,-i}^v$ is the number of words of type $v$ with topic label $k$, excluding the current word, and $n_{t,-i}^k$ is the number of words with topic label $k$ that are in the temporal window $d_t$, excluding the current word, $\mathbf{z}_{-i}$ is the set of all topic assignments except $z_i$.

### D. Temporally Smoothened LDA

A straightforward application of Latent Dirichlet Allocations, as proposed above, requires the assumption that each temporal window has topic labels that are independent of the topic labels of its neighboring windows. This can lead to very noisy results. We propose to instead model the joint word distribution as:

$$\mathbb{P}(w_i^t|d_t) = \sum_{k=1}^{K} \mathbb{P}(w_i^t|z_i^t = k)\mathbb{P}(z_i^t = k|G(t, g)), \tag{5}$$

where $G(t,g) = \{d_{t-g}, \cdots, d_t, \cdots, d_{t+g}\}$ is the set of temporal windows in the neighborhood of $d_t$, including $d_t$. The bigger the size of this neighborhood, the more consistent the topic labels over adjacent temporal windows will be.

### *E. Region Labels for Temporal Windows*

Given the region labels $z_i^t$ for each word$w_i^t$, we would like to compute the region label $r_t$ for each temporal window $d_t$. We use the maximum likelihood estimation:

$$r_t = \operatorname*{argmax}_k \mathbb{P}(z_i^t = k | d_t) \tag{6}$$

$$= \operatorname*{argmax}_k n_t^k, \tag{7}$$

where $n_t^k$ is the number of times a word in documen $d_t$ is assigned topic label $k$.

Using these region labels for each temporal window, we can now define two different locations in a robot's trajectory to be the same regions if their region labels are the same.

## IV. Experiments

### *A. Datasets*

We recorded two datasets, 51 and 43 minutes long, corresponding to sound recorded from trajectories with loops through the McGill Campus and the surrounding downtown area of Montreal. The audio was recorded in stereo from a standard hand-held video camera at a 44.1 kHZ samplerate, while walking at approximately constant speed, and later combined into a single channel. The loops were chosen to contain varied sound environments, and contain both indoor and outdoor sounds, as well as sounds from busy and quiet environments. The map of these trajectories is shown in Fig. 2(a) and 3(a). The dotted paths segments correspond to indoor environments. The *Four-loops* dataset shown in Fig. 2 consists of four loops through the trajectory shown in the map, and the *Figure-8* dataset shown in Fig. 3 corresponds to a trajectory which is topologically equivalent to figure '8', and is looped twice.

### *B. Ground Truth*

Fig. 2(b), 3(b) show the ground truth similarity matrix, where element $(i,j)$ is colored (non black) if the robot was labeled to be in the same region at time $i$ and $j$, i.e., region labels $r_i = r_j$. The color of the element $(i,j)$ corresponds to the color of the path segments shown in Fig. 2(a) and 3(a). Thus, colored blocks in the ground truth similarity matrix correspond to sets of locations that belong to a single spatial region with a consistent acoustic fingerprint. The red squares in Fig. 2(b) thus correspond to sets of pairs of points along the trajectory that all have acoustic fingerprints that sound like the roadway marked as region 1 in Fig. 2(a).

Ground truth was produced by a human expert, by identifying points on the map where environment transitions occur. Doorways for entering and exiting buildings as well as the edges of campus were the main landmarks. Some gradual environment transitions occur in the datasets, for instance going from quiet outdoors parts of campus to busy ones. We do not try to model these gradual transitions, and instead just

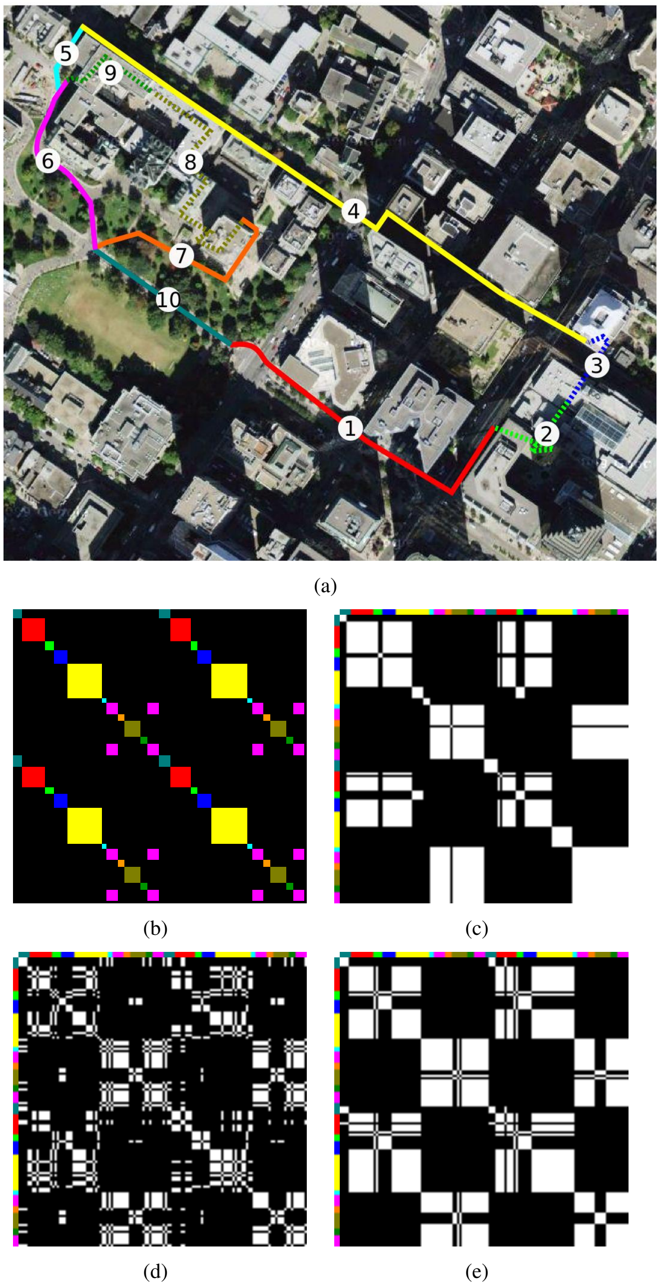
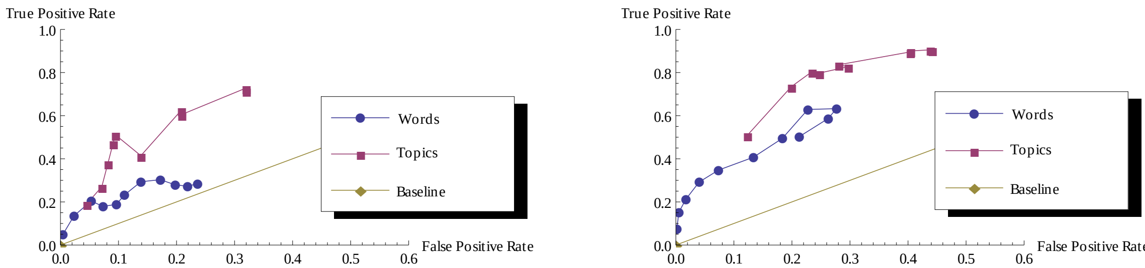



Fig. 3. *Figure-8 dataset:* (a) Map showing the path traversed while recording the dataset. Different regions from the ground truth are shown with different colors. The path is topologically equivalent to figure-8, and was looped twice. (b) Ground truth similarity matrix, in which element $(i,j)$ is colored if the robot was in the same region at times $i$ and $j$. Regions are colored to match the colors in the map. (c) Best performing similarity matrix produced by matching only the most frequent MFCC words at each time step. The matrix correspond to a neighborhood size $g = 14$. (d). Similarity matrix produced by matching the most frequent topic label for each time window. Topics were generated using simple LDA, with no use of temporal neighborhood. (e). Best similarity matrix produced by matching the most frequent topic label for each time window, using topics generated by temporally smoothened LDA with neighborhood size $g = 2$.




(a) Four-loops dataset

(b) Figure-8 dataset

Fig. 4. ROC curves for the two dataset, showing performance of the proposed technique. Each curve shows true positive rate vs false positive rates in predicting whether time-windows $i$ and $j$ correspond to the same region, for all $i$ and $j$, as we change the size of the neighborhood. The prediction is done by matching the most frequent word or region label in the temporal neighborhood of the given times. The first point on the topics curve correspond to simple LDA, with neighborhood size of zero. We see that the proposed technique (temporally smoothened LDA) produces significantly better results than MFCC words or simple LDA based predictions. First and the third points on the topics curve , and eighth point on the words curve in (b) correspond to the similarity matrices shown in Fig. 3. First and the fifth points on the topics curve , and seventh point on the words curve in (a) correspond to the similarity matrices shown in Fig. 2.

pick a single point where this transition occurs, as is in the transition from region 2 to 3 in Fig. 2(a).

### C. Algorithm Evaluation

We first generated two vocabularies by clustering MFCC features from the two datasets, and then used the vocabulary from the first dataset to generate MFCC words for the second, and vice versa. Each MFCC word corresponds to 92 millisecond window of the sound, with a 50% overlap with the previous window. We then grouped these words into "documents", each representing 20 seconds of sound, with no overlap. The *Four-loops* dataset has 151 such documents, and the *Figure-8* dataset we has 128 documents. We ran the temporally smoothened LDA on these document-sets with varying neighborhood size $g$ mentioned in Eq. 5. For each document, we then compute the region label $r_t$ by counting the most popular topic label in that document. Now, for each pair of documents, we compare the corresponding region labels, and mark the corresponding times to belong to the same region if the region labels match.

We experimented with neighborhoods size $g = 0 \cdots 10$, and computed the true positive rates (TPR) and false positive rates (FPR) resulting from comparison with the ground truth matrix. TPR refers to the fraction of true positives similarity matches out of all positive results returned by the algorithm. Similarly, FPR refers to the fraction of false positive similarity matches out of all negative matches returned by the algorithm. An ideal algorithm has TPR of 1.0 and FPR of 0.0. The resulting plots of TPR vs FPR (known as a ROC curve) are shown in Fig. 4(a), 4(b). Fig. 2(e), 3(e) show the similarity matrices with the best performance, chosen by their distance from the baseline performance on the ROC curve. Fig. 2(d), 3(d) correspond to region similarity matrix for neighborhood size $g = 0$, referring to the use of simple LDA, with no information sharing between neighboring documents, shown here for comparison. These matrices correspond to the first point on the "topics" ROC curve.

To show the advantage of using temporally smoothened LDA over the relatively lower level bag-of-words representation, we also computed the region assignment matrix using just the word distributions. Similar to region assignment for topics, we mark the region label for each time step by the most popular MFCC word in the document and its neighborhood. Now if a pair of timesteps have the same region label, then they are marked to be from the same region. Fig. 2(c), 3(c) show the best case similarity matrices.

## V. Results

Our experiments indicate that the use of temporally smoothened LDA to discover acoustic regions in the world perform significantly better than using simple LDA, or MFCC words. Fixing the false positive rate to be less than 0.25, Table I shows the best detection accuracy for each algorithm for both the datasets.

TABLE I
OVERALL RESULTS

| Dataset | Best Detection Accuracy (False positive rate $< 0.25$) | | | | | |
|---|---|---|---|---|---|---|
| | MFCC-words | | LDA | | Temp. Smooth. LDA | |
| | TPR | FPR | TPR | FPR | TPR | FPR |
| Four-loops | 0.29 | 0.15 | 0.19 | 0.04 | 0.61 | 0.21 |
| Figure-8 | 0.63 | 0.23 | 0.50 | 0.12 | 0.80 | 0.23 |

We see that in both datasets, the temporally smoothened topics-based region labels outperformed the top-word-based region labels. It should be noted that the reported false positive rate is probably higher than the actual false positive rate because similar sounding regions at physically different locations were marked as distinct in our ground truth. For example, regions (1) and (4) from the *Figure-8* dataset are both recorded from busy streets, and sound the same, and as result are detected to be the same region by the proposed algorithm (Fig. 3(e)).

Looking at the document similarity matrices produced by the system for the two datasets (Fig. 2(e), 3(e)), we can see that the algorithm is successfully able to detect loop closure at many, but not all locations. This points to a potential application to loop-closure problem in situations where standard vision based loop closure techniques are not adequate or need to be augmented.

## VI. CONCLUSIONS

We have presented an acoustic topic modeling framework suitable for discovering regions in the world, based on the changing sound environment around a robot. The proposed technique works in a unsupervised way, and can therefore be used in situations where there is limited knowledge about the places the robot will encounter. We use quantized Mel Frequency Cepstral Coefficients (MFCC) as low level features, and a temporally smoothened variant of Latent Dirichlet Allocation (LDA) to compute both, the region models, and most likely region labels associated with each time step. Our experiments with two, over 45 minute long datasets, show that the proposed technique does better than using simple LDA in matching the performance of a human expert labeled similarity map. We were able to achieve a true positive detection rate of 0.80, while having a false positive rate of 0.23 on one dataset, and true positive rate of 0.61 and false positive rate of 0.21 for the other.

In our ongoing work, we are looking at ways to make the proposed technique work online and in realtime, which would allow the system to be used by a robot exploring previously unknown territories. An online spatiotemporal topic modeling framework such as ROST [**?**] could be used to enable this. The probabilistic nature of the pose matches that ensue from our algorithm suggest that it should be straightforward to combine these estimates with those from other sensing media, such as vision, using traditional information filtering methods. We have not done that in this paper since we have chosen to focus on the core issues in interpreting audio data, but this should be a fruitful application of our approach.

## REFERENCES


[1] M. Cowling and R. Sitte, "Comparison of techniques for environmental sound recognition," *Pattern Recognition Letters*, vol. 24, no. 15, pp. 2895–2907, Nov. 2003. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0167865503001478

[2] S. Chu and S. Narayanan, "Where am I? Scene recognition for mobile robots using audio features," in *IEEE Conference on Multimedia and Expo*, 2006, pp. 885–888. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4036742

[3] N. Roy and G. Dudek, "Surface Sensing and Classification for Efficient Mobile Robot Navigation," in *IEEE International Conference on Robotics and Automation*, 1996, pp. 1224–1228.

[4] L. Ma, B. Milner, and D. Smith, "Acoustic environment classification," *ACM Transactions on Speech and Language Processing*, vol. 3, no. 2, pp. 1–22, July 2006. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1149290.1149292

[5] R. C. Maher and J. Studniarz, "Automatic Search and Classification of Sound Sources in Long-Term Surveillance Recordings," *AES 46th International Conference*, pp. 14–17, 2012.

[6] G. Wichern and J. Xue, "Segmentation, indexing, and retrieval for environmental and natural sounds," *Audio, Speech, and . . .*, vol. 18, no. 3, pp. 688–707, 2010. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5410056

[7] S. Kim and P. Georgiou, "Supervised acoustic topic model for unstructured audio information retrieval," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2010. [Online]. Available: http://worshipersam.net/www/pages/pdf_papers/apsipa2010-supervisedLDA.pdf

[8] D. Mitrovic, "Analysis of the data quality of audio descriptions of environmental sounds," *Journal of Digital Information Management*, 2007. [Online]. Available: http://www.freepatentsonline.com/article/Journal-Digital-Information-Management/186470812.html

[9] M. Cowling and R. Sitte, "Recognition of environmental sounds using speech recognition techniques," *Pattern Recognition Letters*, vol. 24, no. 15, pp. 2895 – 2907, 2003. [Online]. Available: http://www.springerlink.com/index/G742814W731866V7.pdf

[10] G. Muhammad, Y. a. Alotaibi, M. Alsulaiman, and M. N. Huda, "Environment Recognition Using Selected MPEG-7 Audio Features and Mel-Frequency Cepstral Coefficients," in *2010 Fifth International Conference on Digital Telecommunications*. Ieee, June 2010, pp. 11–16.

[11] B. Sturm, M. Morvidone, and L. Daudet, "Musical instrument identification using multiscale mel-frequency cepstral coefficients," no. 1, pp. 0–4, 2010. [Online]. Available: http://vbn.aau.dk/files/48186471/melfreqovercomplete.pdf

[12] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *In Proc. ICCV*, 2003, pp. 1470–1477.

[13] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," *Machine Learning*, vol. 42, no. 1, pp. 177–196, 2001.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[15] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. Suppl 1, pp. 5228–5235, 2004. [Online]. Available: http://www.pnas.org/content/101/suppl.1/5228.abstract