Recognizing Kitchen Sounds for Monitoring Behavior

David Moore, Adam Williams, and Allen Hanson
University of Massachusetts Amherst

Background

As people age, they often require some form of supervision in order to live safely in their own home. Computer technology can help to provide such supervision at lower cost and with fewer privacy concerns than other approaches. As part of a broader collaboration known as the ASSIST project, the UMass Computer Vision Lab is working to develop systems for monitoring and assisting elderly people living alone, in part by detecting abnormal behavior patterns which might be signs of trouble requiring intervention. This research explores the use of microphones and audio information to classify and understand the everyday sounds produced in a kitchen, with an application towards tracking behavior patterns.

Approach

Hidden Markov Models (HMMs) are a machine learning technique commonly used to classify data which changes over time. HMMs assume that the data are generated by a Markov process: a state machine in which the probability of transitioning to a given state depends only on the current state, and in which the output produced depends on which state the process is in. When modeling real-world data, the states of this process are "hidden" – we don't necessarily know anything about the internal structure of the object making a particular sound – so we must guess that there are some particular number of states, and use the statistical properties of the data to estimate the most likely set of probabilities for transitioning between them.

In our case, the data are recorded sounds. To extract the most relevant features of these sounds, we divide each sound into overlapping 25-millisecond windows and compute the **Mel Frequency Cepstral Coefficients (MFCCs)** of each window. MFCCs are commonly used in speech and environmental sound recognition applications, and are intended to produce a compact representation of the most perceptually important qualities of an audio signal. Applying the MFCC transformation condenses the relatively large number of audio sample points in a 25ms sound clip into a single small vector which represents higher-order features.

Using this processed data, we train an HMM for each class of sounds we want to distinguish. Within the HMM, the output of each state is modeled as a Gaussian mixture – a sum of multiple Gaussian distributions in the multidimensional feature-space – and the training process works by choosing the parameters of those mixtures, as well as the state-transition probabilities of the HMM, to maximize the likelihood that the model could have produced the data observed. Once a set of models is trained, unknown sounds can be classified by calculating the likelihood that the sound could have been produced by each model, and choosing the class whose model gives the highest likelihood.

We implemented a software system to perform this process using MATLAB along with Kevin Murphy's Hidden Markov Model Toolbox and the MIR Audio Toolbox developed by the University of Jyväskylä.

Data Collection

To evaluate the effectiveness of the HMM/MFCC approach to sound classification, we collected data representing examples of common kitchen sounds recorded in three different kitchens in relatively compact apartment settings. Sounds were recorded using two omnidirectional microphones and mixed together into a monophonic signal. Each sound was recorded using the native equipment (appliances, cookware, and so on) of the kitchen it was created in, and manually trimmed to minimize extraneous noise.

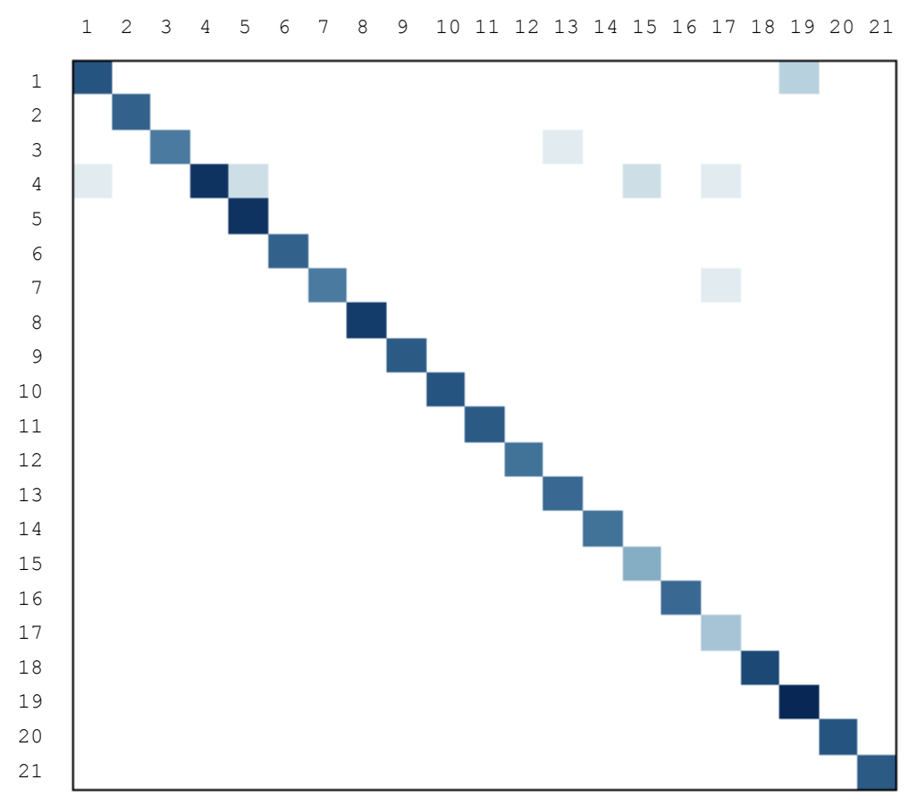
Class	Description	Total time (s)	Total examples	Mean time/example (s)
bag-rustle	plastic bag rustling	150.16	42	3.58
boil	water at full boil in a pot on the stove	160	30	5.33
cabinet	cabinet door being closed (not slammed)	19.65	36	0.55
chop	an onion being chopped on a cutting sheet on the counter	26.25	66	0.40
dishes	ceramic dishes clinking against each other	50.55	48	1.05
fridgeclose	refrigerator door being closed	11.15	31	0.36
fridgeopen	refrigerator door being opened	13.45	32	0.42
mwavebeep	button being pressed on the microwave	9.7	42	0.23
mwaveclose	microwave door closing	15.32	34	0.45
mwaveopen	microwave door opening	15.47	34	0.46
mwaverun	microwave running on high	86.5	31	2.79
ovclose	oven door being closed	22.45	30	0.75
ovopen	oven door being opened	17.15	30	0.57
phone	nearby phone ringing	51.2	26	1.97
pourwater	water being poured into a glass	68.47	20	3.42
silence	pure background noise	80	28	2.86
sinkfillglass	faucet turned on medium, into a cup or glass	54	11	4.91
sinkrun	faucet turned on high, into an empty sink basin	106	38	2.79
sizzle	onions frying on the stove	270	52	5.19
speech	single human speaker	94.4	35	2.70
stove-pot	metal pot full of water being set down on a stove burner	17.27	37	0.47

Results

We began by combining the sounds from all three kitchens into one large set. Half of the sounds (selected uniformly across kitchens and classes) were used for training, and the remainder were held out for testing. Classification accuracy on the test set was 95.1%, with the true class falling within the top three predicted classes 99.5% of the time.

Mixed-kitchen results by class: recall and precision

Class	Recall	Precision
1. bag-rustle	1.00	0.84
2. boil	1.00	1.00
3. cabinet	0.67	1.00
4. chop	0.82	1.00
5. dishes	0.96	0.89
6. fridgeclose	1.00	1.00
7. fridgeopen	1.00	0.89
8. mwavebeep	1.00	1.00
9. mwaveclose	1.00	0.89
10. mwaveopen	1.00	0.89
11. mwaverun	1.00	1.00
12. ovclose	0.93	0.88
13. ovopen	1.00	1.00
14. phone	1.00	1.00
14. pourwater	1.00	0.90
15. silence	1.00	0.93
17. sinkfillglass	0.83	1.00
18. sinkrun	1.00	1.00
19. sizzle	1.00	1.00
20. speech	1.00	1.00
21. stove-pot	0.84	0.94
AVERAGE	0.95	0.96



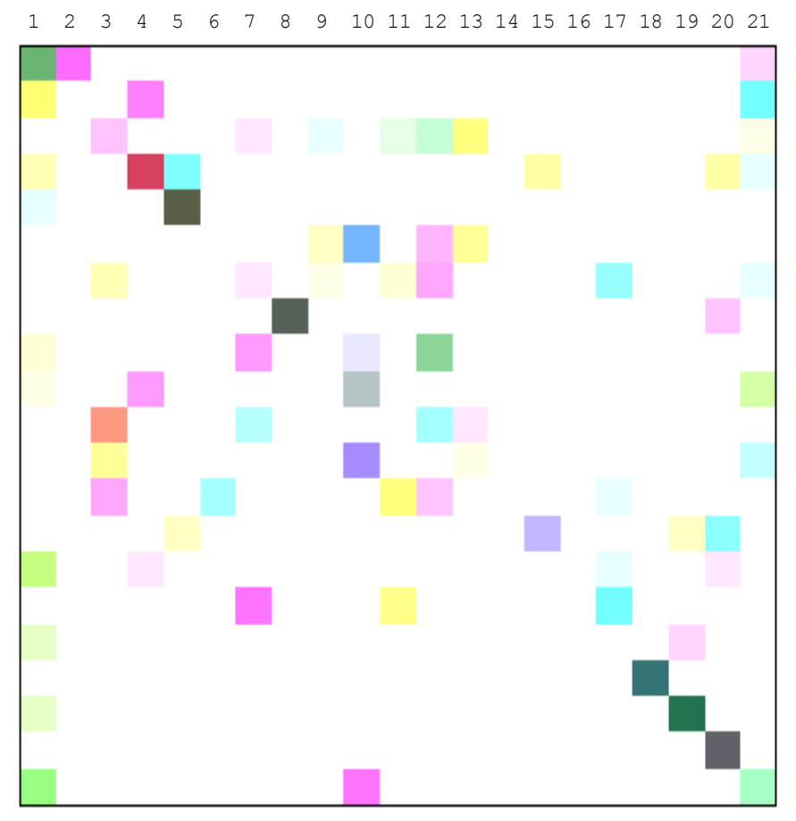
The confusion matrix above visualizes the relationship between true class on the vertical axis and predicted class on the horizontal axis. The darkness of the shade indicates the number of sound samples which fit into a particular cell, so classes with fewer overall samples tend to have lighter shades.

To evaluate generalization performance, we used all of the sounds from two of the three kitchens to train a classifier, and used it to classify all of the sounds from the third kitchen. This process was repeated using each of the three kitchens as the test kitchen, and the results averaged. Mean classification accuracy was 38.6%, with 60.4% of sounds having the true class as one of the top three predicted classes.

Results (continued)

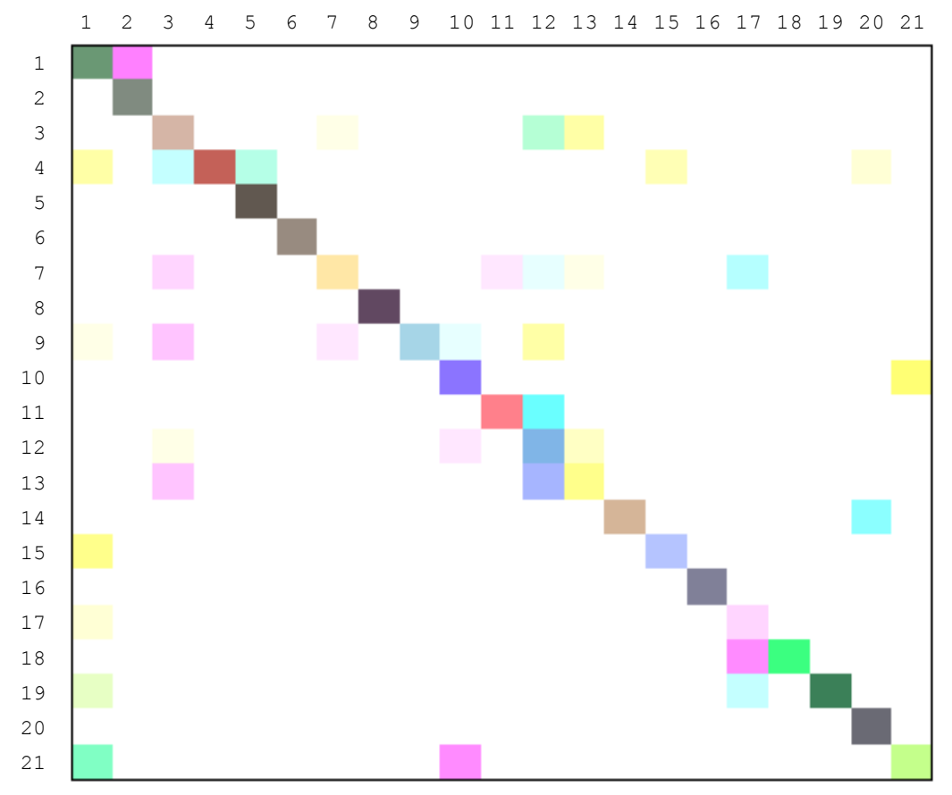
Average cross-kitchen results by class: recall and precision

Class	Recall	Recall (Top 3)	Precision
1. bag-rustle	0.62	0.69	0.28
2. boil	0.00	0.20	0.00
3. cabinet	0.14	0.72	0.12
4. chop	0.47	0.77	0.60
5. dishes	0.83	0.94	0.73
6. fridgeclose	0.00	0.10	0.00
7. fridgeopen	0.13	0.50	0.12
8. mwavebeep	0.90	1.00	1.00
9. mwaveclose	0.03	0.26	0.10
10. mwaveopen	0.38	0.79	0.22
11. mwaverun	0.00	0.52	0.00
12. ovclose	0.00	0.17	0.00
13. ovopen	0.03	0.40	0.05
14. phone	0.00	0.00	0.00
15. pourwater	0.00	0.50	0.00
16. silence	0.00	0.46	0.00
17. sinkfillglass	0.00	0.09	0.00
18. sinkrun	1.00	1.00	0.97
19. sizzle	0.88	1.00	0.85
20. speech	1.00	1.00	0.61
21. stove-pot	0.27	0.59	0.22
AVERAGE	0.32	0.56	0.28



This confusion matrix uses the red, green and blue color channels to reflects tests on each of the three kitchens. As can be seen from the table at left, some classes (e.g. speech) appear to generalize quite well across kitchens, while others do not fare as well.

These results are not particularly impressive, but they are much better than random guessing. Accuracy does improve significantly if we are allowed to add one sample from each sound class in the test kitchen to our training set, in addition to the sounds from the other two kitchens (with the test set consisting of all sounds which were not used for training). Then 68.8% of sound samples are correctly classified, and the true class falls within the top three 87.5% of the time.



Confusion matrix for the case in which the training set is supplemented by a small number of samples from the test kitchen, again with the red, green, and blue channels corresponding to tests on each of the three kitchens.

Discussion and Future Work

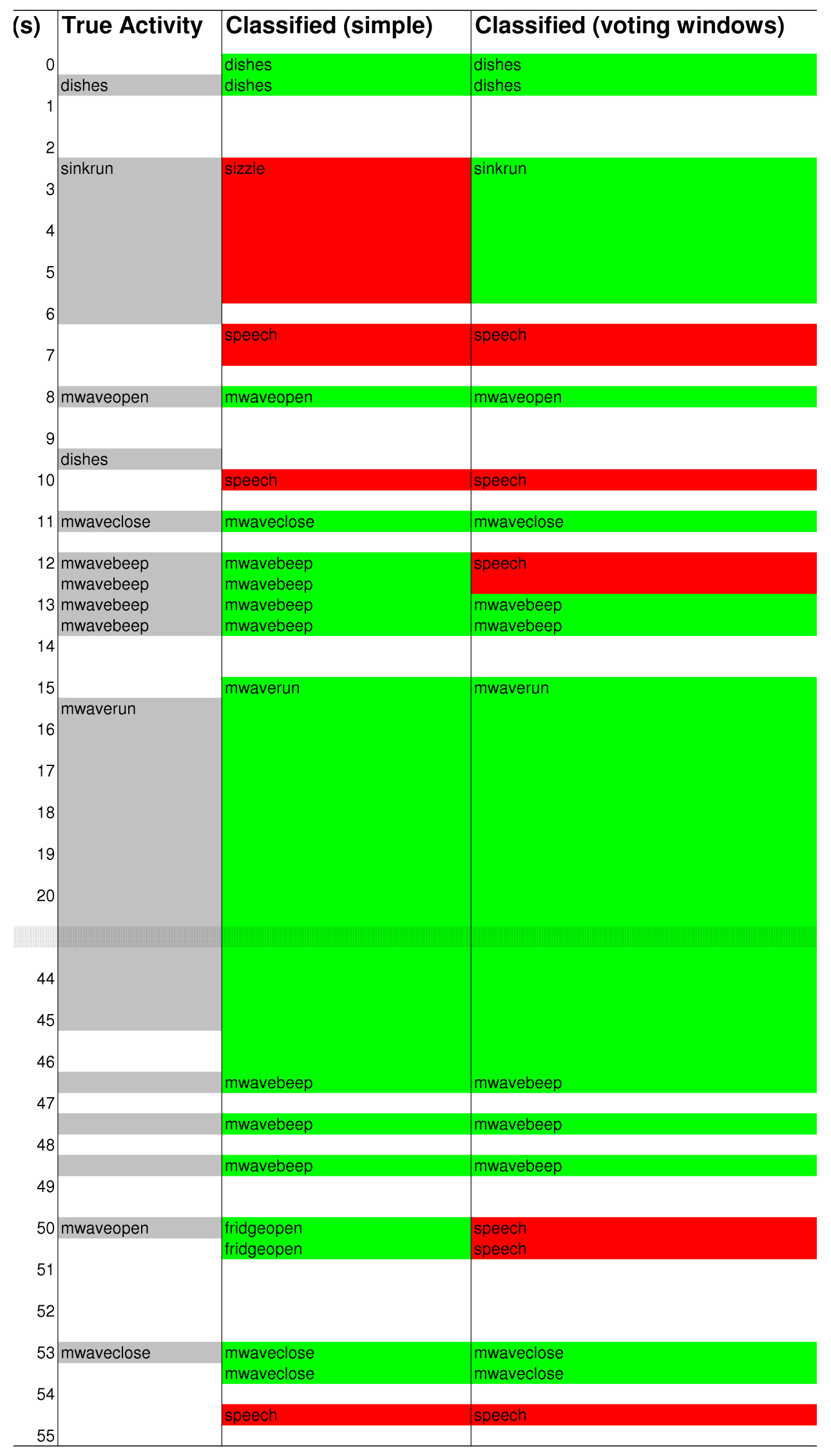
We have shown that a relatively simple, standard approach is capable of identifying kitchen sounds with good accuracy, given a well-labeled training set which includes sounds from the kitchen in which it is used. Performance degrades dramatically when the system is applied to a kitchen on which it has not been trained, although several classes (e.g. speech, sizzle, sinkrun, mwavebeep, and dishes) seem to maintain relatively high accuracy in such circumstances, so it may still be possible to recognize a limited subset of sounds in unfamiliar kitchens. Adding in even a small amount of training data from the kitchen being tested improves results significantly.

Initial tests on simple sequences of actions show promising results using naïve segmentation by silences, but a more sophisticated segmentation method will be necessary to handle real-world data in which events are not always cleanly separated by silence. In addition, since the most relevant sound classes are likely different in every kitchen, it might be productive to explore unsupervised clustering or semi-supervised learning methods which could alleviate the need for a labor-intensive manual training process and allow the system to automatically refine its responses based on experience.

Application

Classifying individual sounds is the first step towards being able to extract behavior information from recordings of kitchen activity. We recorded several sequences of actions in each kitchen, and have experimented with several approaches for classifying them.

We perform naïve segmentation using a volume threshold, assuming that our sounds of interest are separated by silence. We can then classify each sound holistically as before, or we can split it into smaller windows (200ms, in this case), classify each window, and use the most common result to label the entire sound. Both methods are shown in the chart on the right (green shading is correct; red denotes an error).



Acknowledgements

This work is partially supported by the Research Experience for Undergraduates Program of the National Science Foundation under NSF award number CCF-0755376. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the National Science Foundation.

References

Chen, J., Kam, A.H., Zhang, J., Liu, N. and Shue, L. (2005) Bathroom Activity Monitoring Based on Sound. Proceedings of the International Conference on Pervasive Computing (Pervasive 2005), 47-61.

F. Kraft, R. Malkin, T. Schaaf, and A. Waibel. Temporal ICA for classification of acoustic events in a kitchen environment. In Proceedings of ICSLP-Interspeech, 2005.

Martin F McKinney and Jeroen Breebaart, "Features for audio and music classification", in Proceedings of the Fourth International Conference on Music Information Retrieval (ISMIR) 2003.

Kevin Murphy (2005). Hidden Markov Mopdel Toolbox for Matlab. http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html.

Lartillot, O. & Toiviainen, P. (2007). A Matlab Toolbox for Musical Feature Extraction From Audio. International Conference on Digital Audio Effects, Bordeaux, 2007.