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 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 that best fit the observed data. 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 a relatively compact apartment setting. 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.

To produce a compact representation of the most perceptually important features of these sounds, we divided each sound into overlapping 25-millisecond windows and compute the relevant features of these sounds, we divide each sound into overlapping 25-millisecond windows and compute the Frequency Cepstral Coefficients (MFCCs) over each window.

Hidden Markov Models (HMMs) are a machine learning process: a state machine in which the probability of transitioning to one state depends 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.

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 this test set was 95.1%, with the true class falling within the top three predicted classes 99.9% of the time.

Mixed-kitchen results by class: recall and precision

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 was obtained. 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, steen, 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 scheme will be necessary to handle real-world data in which events are not always clearly 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.

Results

Average cross-kitchen results by class: recall and precision

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 was obtained. 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, steen, 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 scheme will be necessary to handle real-world data in which events are not always clearly 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.

Discussion and Future Work

To evaluate generalization performance, we used all of the sounds from the 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 36.6%, with 60.8% of sounds having the true class as one of the top three predicted classes.

HMMs are a machine learning process: a state machine in which the probability of transitioning to one state depends 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.

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 was obtained. 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, steen, 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 scheme will be necessary to handle real-world data in which events are not always clearly 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.

References


Acknowledgements

This work is partially supported by the Research Experience for Undergraduates Program of the National Science Foundation under NSF award number CCF-0755377. 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.