IEEE AASP SCENE CLASSIFICATION CHALLENGE USING HIDDEN MARKOV MODELS AND FRAME BASED CLASSIFICATION

May Chum, Ariel Habshush, Abrar Rahman, Christopher Sang

The Cooper Union
Electrical Engineering Department
41 Cooper Square
New York, NY 10003

ABSTRACT

The IEEE AASP Challenge involves the detection and classification of acoustic scenes and events. The scene classification (SC) challenge consists of 10 different scenes of 10 audio files or length 30 seconds each, totaling a number of 100 audio clips. The list of scenes is: busy street, quiet street, park, open-air market, bus, subway-train, restaurant, shop/supermarket, office, and subway station. The goal is to test on a development set that is composed of audio clips of the same scenes as the training set and determine what scene the audio clips originated from. One of the algorithms presented in this paper to discriminate between these different scenes include the use of hidden Markov models (HMMs) and Gaussian mixture models (GMMs). The features that were used include the following: short time Fourier transform, loudness, and spectral sparsity. Using these features yielded 72% correct classification with 10 fold crossvalidation. The other algorithm implemented uses the same features as before plus temporal sparsity to classify individual frames of an audio clip, then vote on the class. This algorithm achieved 62% accuracy.

Index Terms— hidden Markov models (HMMs), Gaussian mixture models (GMMs), spectral sparsity

1. INTRODUCTION

As humans, we find ourselves each day using our different senses to recognize our surrounding environments. However, our senses play different roles depending on the type of environment we are situated in. In this paper we focus on the role of our sense of hearing and environments that can be characterized solely based on their sounds. On the surface, our ability to characterize an acoustic environment may seem like a simple task, but in fact our brains perform complex calculations to arrive at these conclusions. This complexity shows itself when we try to automate this classification task.

Over the last few years, scene classification has been gradually receiving attention in the field of audio signal processing. Much of this research has been spent developing feature extraction algorithms specific for environment classification. These features are composed of temporal domain as well as frequency domain features. In fact, many of these algorithms extract combined temporal-frequency (TF) features. For example, [1] and [2] use the matching pursuit (MP) algorithm to obtain TF features. Due to the diversity of sound sources, [3] introduces the concept of using both short-term and long-term features. In this paper, we implemented some of these feature extraction techniques in order to obtain defining features for each type of environment.

2. FRAME-BASED CLASSIFICATION

One approach to classification is to use a frame-based technique. The input binaural audio signal is converted to a single channel by averaging and then is downsampled by 4 to reduce computational requirements for the classifier. Then, the signal is divided into overlapping short term and long term frames using an approach similar to that in [3]: Short term frames are 512 samples or approximately 45 ms in length. The overlap is set to 256 samples. Long term frames are 16384 samples or approximately 1.5 s. The overlap is set to 16128 samples so that for every short term frame there is a corresponding long term frame. A Hamming window is used for windowing the frames.

2.1. Features

For each short term frame, the following features are computed:

- Magnitude response, defined as the magnitude of the first 256 coefficients of a 512-point FFT of the frame, in decibels.
- Loudness, defined as

$$x = 20\log_{10}(rms + \alpha) \tag{1}$$

  where rms is the root mean square level of the frame and α is a smoothing parameter to prevent zeros from being input into the logarithm.
- Spectral sparsity,

$$x = \frac{\max\{|X(1)|, \ldots, |X(M)|\}}{\sum_{m=1}^{M} |X(m)|} \tag{2}$$

  where $X(m)$ is the m-th coefficient of the M-point FFT previously computed [3].

In addition, for each long term frame, we compute the following feature:

- Temporal sparsity,

$$x = \frac{\max\{rms_1, \ldots, rms_K\}}{\sum_{k=1}^{K} rms_k} \tag{3}$$

where rms_k is rms value of the k-th short term subframe of the long term frame [3].

2.2. Classifier

For each class, we set up a support vector machine (SVM) with Gaussian radial basis function kernel. Each SVM is a binary classifier that detects whether or not a particular frame belongs to the class. An M-ary classifier is formed by running the SVMs in parallel and choosing the class with the highest likelihood.

For each frame of an input signal, we concatenate the features into a single feature vector. Then, we use the collection of feature vectors from all signals and frames to train the M-ary classifier. To classify an unknown audio signal, first all of the frames are classified. If certain frames cannot be classified at a sufficient confidence level, they are skipped. At the end, the final result is the class that the most frames are identified as belonging to.

We implement all of these mechanisms in MATLAB using the Pattern Recognition Toolbox [4].

3. HIDDEN MARKOV MODELS

The other classification algorithm used was based on the Hidden Markov Model(HMM). The scenes were assumed to have states it transitioned between. Each of these states were modeled to be a Gaussian Mixture Model(GMM), in other words, a sum of Gaussian probabilities. Procedures for an HMM learn the transition probabilities from state to state. These transition probabilities follow the Markov property where the next state is only dependent on the current state. An HMM learned all the possible transitions between states as a transition matrix. This transition matrix was learned based on the aforementioned features: magnitude response, loudness, and spectral sparsity. A transition matrix was learned for each of the classes. For each of the unlabeled audio files, features are computed and processed so they could be compared to the transition matrices of each class. The HMM computes a likelihood that the features came from a specific class. The class that gave the highest likelihood was the most likely to be the class of the unlabeled scene.

4. RESULTS

The results for the frame-based classification are shown in Figure 1. With this method, we achieved a total classification accuracy of 62%. We obtained a classification accuracy of 80% or more in the first four classes: bus, busy-street, office, and open-air-market. Notice, however, that park had been misclassifed as office 40% of the time. Similarly, supermarket was misclassified as open-air-market 40% of the time. Perhaps the reason for this misclassification is that the park and open-air-market audio files didn't seem to have distinct audible characteristics.

The results of the HMM-based classifier are shown in Figure 2. With this method, we achieved a total classification accuracy of 72%. Notice that we classified bus, busy-street, and quiet-streey with 100% accuracy. We did not misclassify any other files as bus or busy-street; however, we did misclassify park and office as quiet-street which leads to some suspicion in the quiet-street results. As opposed to the frame-base classification, with HMMs we acheived a classification rate to 50% or more for the last four classes. However, the most drastic result with HMMs is that the park classification rate went down to 0%.

5. ACKNOWLEDGMENT

Many thanks to Professor Sam Keene for his guidance and advice.

6. REFERENCES

[1] G. Wichern, J. Xue, H. Thornburg, B. Mechtley, and A. Spanias, "Segmentation, indexing, and retrieval for environmental and natural sounds," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 3, pp. 688–707, 2010.

[2] B. Ghoraani and S. Krishnan, "Time-frequency matrix feature extraction and classification of environmental audio signals," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 2197–2209, 2011.

[3] S. Chu, S. Narayanan, and C.-C. Kuo, "Environmental sound recognition with time-frequency audio features," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 6, pp. 1142–1158, 2009.

[4] New Folder Consulting L.L.C. Pattern Recognition Toolbox. [Online]. Available: http://www.newfolderconsulting.com/prtdoc/.

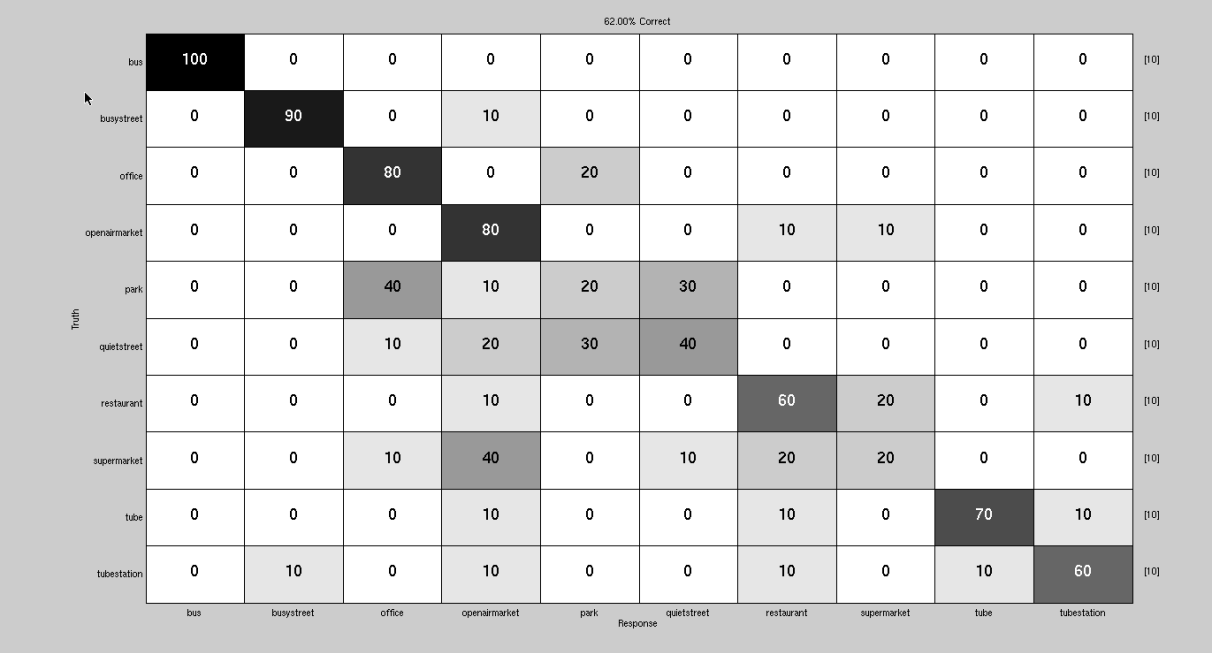

Figure 1: Frame-based classification results.

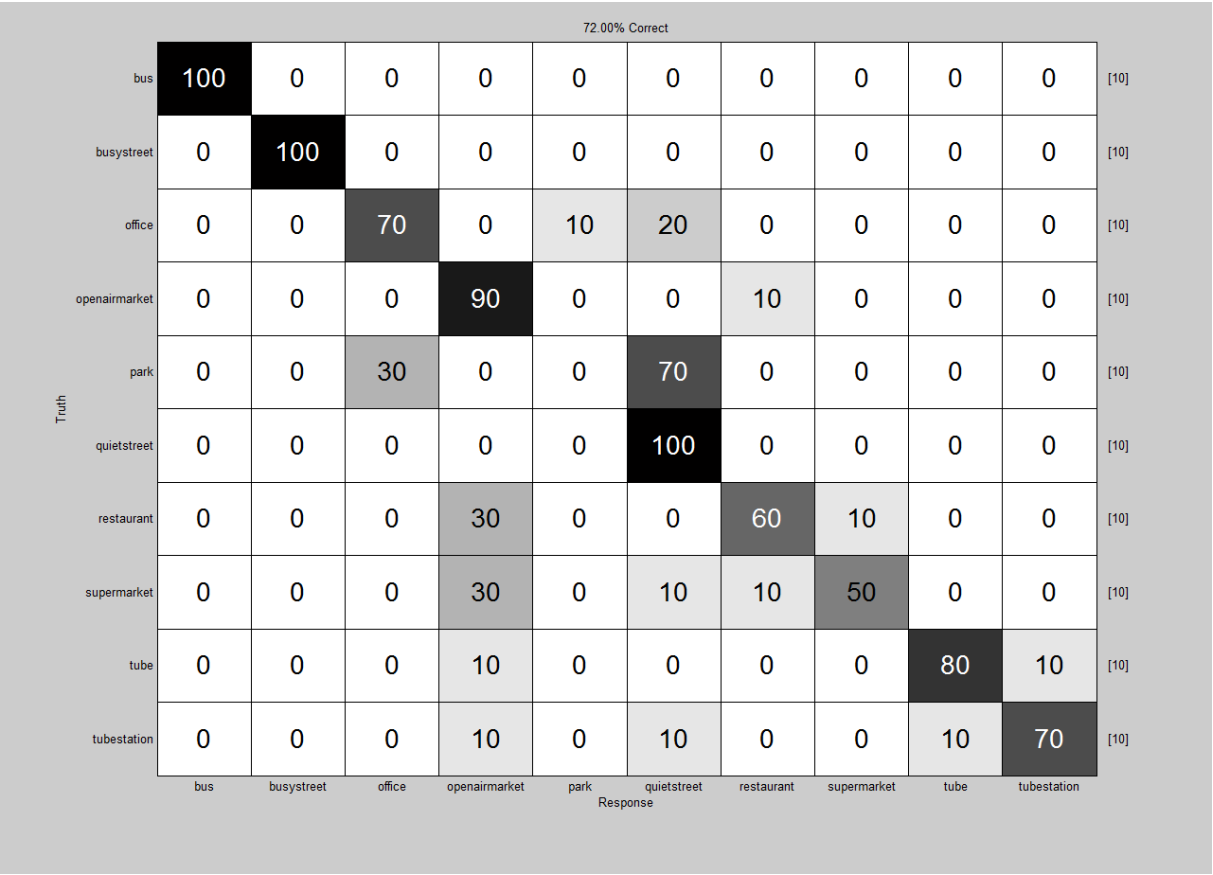

Figure 2: HMM classification results.