EVENT DETECTION AND CLASSIFICATION

Sameer Chauhan, Sharang Phadke, Christian Sherland

Cooper Union for the Advancement of Science and Art
Electrical Engineering
41 Cooper Square
New York, NY 10003

ABSTRACT

The IEEE AASP Challenge addresses the problem of acoustic event detection and classification in an office environment. Our system performs segmentation and event classification on a continuous stream of acoustic activity in an office using basic feature extraction techniques and a two layer classifier (we haven't implemented two layers yet).

Index Terms—

1. INTRODUCTION

The task of event detection and classification is fundamentally important in computational auditory scene analysis (CASA). Detecting events such as words in a sentence of speech, the arrival of a bus or train, or any other common occurrence in an audio stream is the first step of any speech or audio processing system that needs to make use of this information. Many such systems then need to classify the detected events in order to take appropriate actions.

In this project submission, the problem of event detection and classification, as applied to an office environment, was approached from a pattern recognition perspective. Our system consists of a segmentation stage, a feature extraction stage, and two classification stages. The first stage detects the onset and offset of events in a live recording, the second extracts features from each event that can be used to classify events, and the final stages classify each detected event using two pre-trained classifiers. The classification system initially classifies the detected event as one of a few groups of events. The system then classifies the events in each group as a specific type of event using a set of features that extracts the most discriminatory information from the events within each group.

2. FEATURE EXTRACTION

In order to classify events, our system extracts a set of features from the training data. This set includes the spectral centroid, spectral flux, spectral sparsity,temporal sparsity, loudness, short time energy and Mel-frequency cepstrum coefficients (MFCC). The MFCCs are based on frequency bands equally spaced on the mel scale which approximates the human auditory system's reposnse. Each feature highlights different acoustic properties of the signal. Spectral sparsity is expected to be very large for pure sine tones or bells and smaller for sounds with significant "noise" characteristics that imply a wide frequency spectrum. Temporal sparsity is large for sounds such as footsteps in relative silence and is useful for indexing and retrieving these types of sounds. Short time energy is analogous to the volume of the event and the spectral centroid is essentially the center of mass of the spectrum. They are both expected to be reliable indicators of silence. The spectral flux, also called spectral variation, measures how quickly the power spectrum changes. It can be used to determine the timbre of the audio signal. Different permutations of these features will be tested and the best combination will be used for the system training and classification stages of our system.

Our feature set is computed within a 40ms window with no overlap, which forces our frames to be 40ms long. This limits our precision but increases our accuracy. In order to increase the performance of the system, we will choose features to be calculated over a long period (1s) . This 1s window will slide every 20ms and overlap will be included with the features obtained within the 40ms window to help obtain a more precise feature set for the events.

3. TRAINING

The classification stage of our system employs two levels of classification. As input the classifier takes in an event signal and first classifies what type of event it is. Then based upon the type of event the signal is passed through a type-specific classifier which determines the exact event contained in the signal. In the second stage each type has its own classifier.

Each level of classification in the system was trained with features extracted from the provided training data. The first stage classifier was trained using the full set of extracted features for all events. The second stage classifiers were each trained using features extracted only from training signals of the correct type. Additionally, the features used for training were chosen to best describe the type of event being classified.

4. SEGMENTATION

Before events within a continuous input signal can be classified, the segments of the input which contain events must be extracted. A speech segmenter [**?**] accomplishes this by examining the short time energy and spectral centroid features. The segmenter examines peaks of the features and checks them against a threshold to determine the onset and offset of events. The onset and offset times are

Thanks to the Cooper Union

then used to extract the portions of the input signal which contain an event and pass them on to the classification stage.

A continuous sound file is input into the segmenter which applies a Chebyshev Type II low pass filter. This step was added to the original segmenter to reduce the amount of background noise and increase onset and offset detection. The segmentation process requires three parameters for feature extraction: window length, step time, and the weighting facter that will be used to compare the signal to a threshold. The segmenter was exhaustively tested against the development data to determine the parameters that provided optimal segmentation. Each run of the segmentation test was checked for the number of events it detected. The runs were narrowed down to those that returned the same number of events as the development data indicated. Then the mean square error was calcuated using onset and offset times and used to determine the optimal choice of parameters.The following is the cost function used to measure the segmenter, where $\tau_{(on)i}$ is the estimated onset time and $tau_{(off)i}$ is the estimated offset time and t_{on} and t_{off} are the truth values.

$$MSE = \min\left(\tau_{(on)i} - t_{on}\right)^2 + \left(\tau_{(off)i} - t_{off}\right)^2 \qquad (1)$$

The values for the window length, step time, and threshold weight that results in the minimum MSE were 0.05s, 0.04s and 3 respectively. During the exhaustive process, it was also determined that the threshold weight had to be an integer value.

The segmenting function returns a two column vector where the first column indicates each onset, and the second column indicates each event offset. The signal between the onset and offset times is used in the classification stage of the system.

5. CLASSIFICATION

After a set of event signals has been segmented, the classifier extracts a full set of features for each event signal. For each event, these features are passed into the first level classifier. Based upon the output of the first stage classifier, each event is labeled with its subgroup.

Each event signal was then passed through the classifier that corresponds to its subgroup label and reclassified based upon the subset of features that best describes its type of event.

Types were chosen such that members of a given type generally have many similar features. The set of features that is used to determine the specific event that occurred in stage two is the set which varies most among the members.

6. RESULTS

In order to evaluate the performance of our system we examined the percentage of frames correctly classified with our segmentation and with perfect segmentation. The perfect segmentation and ground truth label for each frame were determined using the annotation to the development data.

The percentage of frames correct given perfect segmentation gives a sense of how well the classification stage of our system performs. Currently, given perfect segmentation, our system correctly classifies 48% of frames.

The percentage of frames correct with segmentation determined by our system when compared to the other metric gives a good sense of how segmentation effects classification. With our segmentation, the system correctly classified 25% of frames.

The confusion matricies of classifier stages of the perfectly segmented events, as well as the events segmented using our algorithm can be seen in Figure 1 and Figure 2 respectively. (We need to make the figure bigger...)

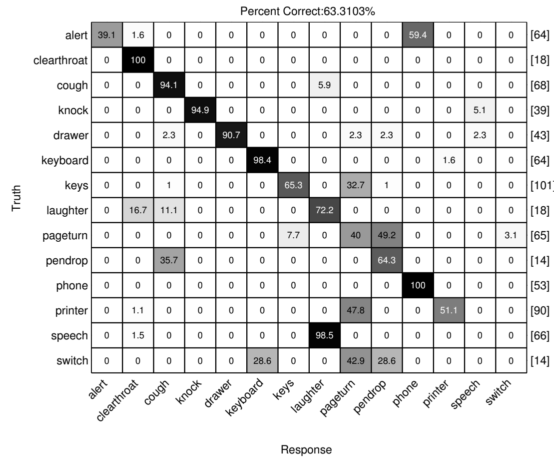

Figure 1: Example of a figure with experimental results.

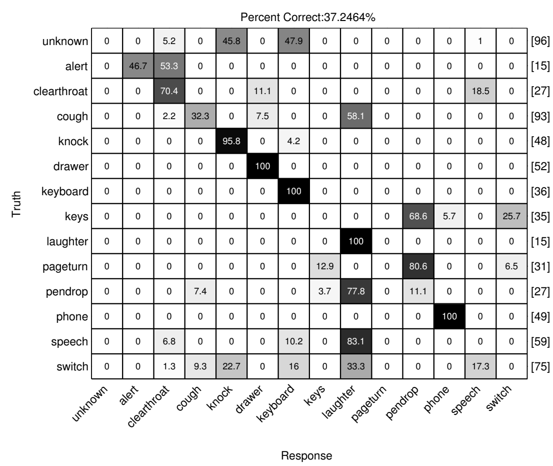

Figure 2: Example of a figure with experimental results.

7. CONCLUSION

8. REFERENCES

List and number all bibliographical references at the end of the paper. The references should be numbered in order of appearance in the document. When referring to them in the text, type the corresponding reference number in square brackets as shown at the end of this sentence [?], [?]. For LaTeX users, the use of the BibTeX style file IEEEtran.bst is recommended, which is included in the LaTeX paper kit available from the workshop website [?].

9. ACKNOWLEDGMENT

The preferred spelling of the word acknowledgment in America is without an “e” after the “g.” Try to avoid the stilted expression, “One of us (R. B. G.) thanks ...” Instead, try “R.B.G. thanks ...” Put sponsor acknowledgments in the unnumbered footnote on the first page.