Goal-Oriented Auditory Scene Recognition

Kailash Patil, Mounya Elhilali

Center for Speech and Language Processing, Department of Electrical and Computer Engineering
Johns Hopkins University, Baltimore, MD, USA.
kailash@jhu.edu, mounya@jhu.edu

Abstract

How do we understand and interpret complex auditory environments in a way that may depend on some stated goals or intentions? Here, we propose a framework that provides a detailed analysis of the spectrottemporal modulations in the acoustic signal, augmented with a discriminative classifier using multilayer perceptrons. We show that such representation is successful at capturing the non-trivial commonalities within a sound class and differences between different classes. It not only surpasses performance of current systems in the literature by about 21%, but proves quite robust for processing multi-source cases. In addition, we test the role of feature re-weighting in improving feature selectivity and signal-to-noise ratio in the direction of a sound class of interest.

Index Terms: Scene understanding, Acoustic Event Recognition, Attention, Bottom-up, Top-down

1. Introduction

One of the most remarkable feats that humans are able to perform rapidly and reliably is to recognize and understand the complex acoustic world that surrounds them. This process, referred to as ‘auditory scene analysis’ [1] is a multi-faceted problem which encompasses various aspects of auditory perception. It encompasses the ability to detect, identify and classify sound objects; to robustly represent and identify these objects in multi-source environments; and to guide actions and behaviors in line with complex goals and shifting acoustic soundscapes. Such capability can provide much needed robustness and flexibility to auditory processing in an online setting as well as in presence of other sources. The proposed model is described in Sec. 2.

In addition to ‘passive’ scene understanding, we also explore the role of goal-directed attention in biasing the processing of complex scenes. It is believed that goal-directed attention modulates sensory processing of relevant features in a scene in order to improve selectivity of the most informative sensory inputs [8]. This process involves a complex interplay between bottom-up, stimulus-driven analysis of the sound features and top-down, feedback processes that can adaptively modulate sensory features, either via scaling, re-tuning or scanning prioritization [9]. Here, we test the role of re-weighting in improving goal-directed scene understanding in a complex setting when multiple sources are present. Details of the proposed architecture are presented in Sec. 3.2.

2. Modulation-based System

We employ a model inspired from mammalian processing of acoustic stimuli [10, 11]. The first stage models analysis from the ear up to the level of the auditory midbrain. The cochlear processing is implemented as 128 asymmetric filters ($h(t; f)$), equally spaced on a logarithmic axis over 5.3 octaves starting from 180Hz. Next, a first-order derivative along the frequency axis followed by a half wave rectifier provides additional spectral sharpening. Finally a short-term integration with $\mu(t; \tau) = e^{-\tau} u(t)$ and $\tau = 2ms$ mimics the loss of phase-locking at the level of the midbrain. Given an input signal $s(t)$,
the output of the first stage \( y(t, f) \) is derived as follows:
\[
\begin{align*}
y_{\text{coch}}(t, f) &= s(t) \otimes h(t; f) \\
y_{\text{lin}}(t, f) &= \text{max}[\partial_f y_{\text{coch}}(t, f), 0] \\
y_{\text{mid}}(t, f) &= y_{\text{lin}} \otimes \mu(t; \tau) \\
y(t, f) &= \left(y_{\text{mid}}(t, f)\right)^{1/2}
\end{align*}
\] (1)

where \( \otimes \) represents convolution with respect to time.

The output \( y(t, f) \) is then further analyzed to capture the detailed spectral modulations (or scales \( s \) in cycles/octave) and temporal modulations (or rates \( r \) in Hz) along the frequency and time axes, respectively. This analysis is done using a bank of modulation-tuned filters \( MF \) defined as
\[
MF(f, t; s, r) = \frac{1}{2\pi \sigma_f \sigma_t} e^{-\frac{1}{2} \left( \frac{s^2}{\sigma_t^2} + \frac{r^2}{\sigma_f^2} \right)} e^{2\pi i (st + fr)}
\] (2)

where \( \sigma_t, \sigma_f \) denote the spread of the filter. The filters are a linear approximation of auditory cortex neurons [12, 13]. The resulting representation \( \mathcal{R} \) is a high-dimensional tensor, parameterized by time \( t \), frequency \( f \), scale \( s \) and rate \( r \); given by:
\[
\mathcal{R}(f, t; s, r) = |y(f, t) \otimes_{f,t} MF(f, t; s, r)|
\] (3)

where \( \otimes_{f,t} \) indicates convolution in time and frequency.

Finally, we integrate \( \mathcal{R} \) over the duration of the audio segment considered. We employ 220 \( MF \) filters, at 11 scales (0.25 cycles/octave to 8 cycles/octave), 10 positive and 10 negative rates (between 2 and 250Hz), equally spaced on a logarithmic scale axis. The final representation is then reduced via Tensor Singular Value Decomposition[14], keeping 336 dimensions which maintain 99% of the variance in the data.

Each recording is segmented into non-overlapping 1s segments, and features are extracted for each non-silent segment. A discriminative classifier is then trained to learn the boundaries between classes using multilayer perceptron algorithm. It contains one hidden layer with 1500 hidden nodes with sigmoid non linearity. The output layer has soft-max non linearity which yields the estimated posterior probabilities of the classes. 10% of the training data is used as cross validation during the training of the classifier. The posterior probabilities are then averaged over the entire duration of the recording. A simple K Nearest Neighbor classifier (with \( K=7 \)) then predicts the class label of any given test recording using the cosine distance between these average posteriors.

The scene classification task is performed on data from the BBC Sound Effects Library [15]. It has 2400 recordings in total, amounting to 68 hours of audio data. The recordings are organized into 18 classes for example Ambience, Machine, Animal etc. Each recording is resampled to 16kHz sampling rate and pre-processed through a pre-emphasis filter with filter coefficients [1 -0.97]. 90% (random selection) of the recordings are used for training and 10% for testing.

Results from the proposed system are compared to standard features for acoustic event classification, based on statistics of Mel Frequency Cepstral Coefficients(MFCC) [2, 3]. We compute 13 dimensional MFCC feature for every 10ms with a hamming window length of 25ms. The CO energy component of the MFCCs are ignored to yield a 12 dimensional vector [our analysis shows that CO exclusion did not affect performance]. We then derive 2 sets of features: a) MS: Mean and Standard Deviation of the MFCCs are computed over time to form a 24 dimensional vector b) MSS: Mean, Standard Deviation and Skew of the MFCCs are computed over time to form a 36 dimensional vector.

3. Results

We test the performance of modulation features and baseline features using the setup described in Sec. 2. The proposed features yield an accuracy of 66.3\%, compared to 45.5\% using MSS features and 39.43\% using MS features. The difference in performance is directly related to the information content captured by the features and not their dimensionality; this has indeed been confirmed elsewhere [16]. This remarkable improvement in accuracy of the modulation-based features clearly demonstrates the need for localized analysis that can better capture the subtle nuances in sound as well as the co-presence of specific spectral profiles and time dynamics. It is worth noting that each class by itself is relatively variable and heterogeneous.

The confusion matrix of the classifier trained on modulation features is shown in Fig. 1. Some of the common misclassifications are that of Office with Ambience, Warfare with Transportation, and Music with Science Fiction. These errors are not unexpected as the classes are themselves ambiguously defined, for example Warfare class consists of examples recorded from vehicles used for transportation in war.

3.1. Parsing a natural scene

Next, we test the robustness of the proposed model in a more realistic uncontrolled setting using an example recording from the web (youtube), due to lack of annotated databases of this nature. In this analysis, we perform a continuous parsing of the audio into 1s windows, with an overlap of 0.75s. Each segment is then analyzed through the scene classification system and the class with maximum posterior probability is chosen as the label for the segment. Fig 2 shows the recording of a visit to a zoo [17] where the acoustic scene is generally a mixture of different sources and these scenes change over time. Examples of time durations where there is predominance of human, vehicle, animal or water sounds are depicted. The histograms of labels for these time durations clearly shows that the model is able to capture the sources in the changing acoustic scenes remarkably well even though the classifier was never trained on the these kinds of examples. This indicates that the modulation features capture the inherent characteristics of different acoustic scenes.
3.2. Goal-oriented scene understanding

As is often the case, sound sources never occur in isolation. Natural recordings are often a mixture of different sources, and cannot be given a simple class label. We therefore test the robustness of the proposed model in multi-source mixtures. Two samples from different classes in the BBC database are mixed together at one of the following SNRs: −20db, −10db, 0db, 10db, 20db. Fig. 3(a) shows examples of the model performance for recognizing transportation and animal classes as a function of SNR. The figure highlights the fact that the model performance degrades significantly for SNR values below 10db.

Next, we explore the role of attention in the process of scene analysis. Humans demonstrate the ability to pay attention to a given goal using some prior knowledge and do a better job at recognizing or identifying the occurrence of that goal [18]. Studies have shown that the gain of the auditory cortex neurons is enhanced when attention is being paid to a target [19, 20]. In a simplistic attempt to model this, we use the mean representation of $R(f, t; s, r)$ from Sec. 2 as the prior knowledge the system has about a class. Attention is modeled as a simple weighting along either the frequency, rate or scale axis.

For example the weights for the frequency are calculated as $W(f) = \sum_{i,s} M(f, t; s, r)$ where $M$ indicates the mean of $R$ over all training examples in the class. The weights $W(.)$ are then scaled and shifted to be in the range 0.8 to 1 to limit the mismatch in representation presented to the classifier. This attentional weighting is then applied as shown in (4).

$$R_w (f, t; s, r) = |y(f, t) \otimes_f M(f, t; s, r)| \ast W(f) \quad (4)$$

where $R_w$ is the representation upon applying the attention weighting $W$.

Mixing of natural scenes leads to a spreading of energy to a wider region. Applying the attention weighting as described would nullify this effect and enhance the regions of interest. The effect of applying attention weighting for the transportation and animal classes, along rate and frequency axis respectively, is shown in Fig. 3a). On average, over all SNRs, the model performance improved by 22.9% for transportation and 4.7% for animal class. Note that even though this simplistic approach has shown significant promise, it does not always yield an improvement in performance. For example weighting on the scale axis resulted in an improvement of 0.7% for transportation but decreased the accuracy of animals by 6.5%.

4. Conclusions

We proposed a novel approach to scene classification based on biologically inspired models of auditory processing. We show that the global analysis of modulations as done by MFCC is not sufficient for scene classification. Further we show that a more detailed local analysis of the spectral and temporal modulations in a joint manner is able to capture the identity of acoustic scenes more successfully. We also show the high degree of generalizability of this model to new unseen recordings in Sec. 3.1.

Attention is known to be a complex mechanism where the top-down prior knowledge is imposed on the bottom-up analysis. In the visual domain, top-down attention has been incorporated into computational models [21, 22, 23] to successfully enhance task performance. Attention in the auditory modality is even more challenging since an acoustic scene is non-stationary and dynamic, with multiple events occurring simultaneously. Nevertheless, the proposed attentional mechanism in Sec. 3.2 based on weighting the appropriate features for the given task improves the performance significantly. We are currently exploring alternative models of attention, guided by neurophysiological evidence that attention can not only change the gain of the auditory cortex receptive fields [19, 20] but also change its selectivity and tuning properties [24, 25, 26].
5. References


