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 spectrotemporal 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 commonalties 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 a number of technologies including smart robots, surveillance and security systems, target tracking in sensor networks as well as adaptive communication aids for the sensory-impaired.

Unlike visual scenes, the difficulty of parsing auditory scenes stems from challenges of segmenting and separating the different components given the complex temporal dynamics that different sound events have, as well as the time-varying nature of their spectral details. Efforts towards classification of auditory scenes have focused on extracting informative features from the acoustic waveforms, that are then exploited to learn generative or discriminative statistical models of the sound classes of interest. Such efforts have led to notable successes in recognizing different acoustic events [2, 3, 4]. Most approaches rely on a short-time analysis of the signal and derive time-varying spectral information, mostly based on Mel Frequency Cepstral Coefficients (MFCC) and their related statistics. As for the statistical analysis of features, various discriminative approaches such as support vector machines [4, 5], multilayered perceptron [2] and generative approaches such as Gaussian Mixture Models (GMM) [6] have been proposed. It has further been suggested that discriminative approaches outperform the generative approaches [6].

Unfortunately, the applicability of these approaches is hindered by the usefulness of features such as MFCC for a task like scene classification. By nature, cepstral coefficients capture only the global spectral details of the signal and fail to analyze the detailed and subtle changes in the spectrum as it changes over time. Studies on mammalian auditory processing suggest that neurons at the level of primary auditory cortex are more directed at analyzing the local spectral and temporal modulations in the signal; hence capturing both details of spectral profile, as well as its changing dynamics over time [7]. In this study, we explore the use of such detailed feature analysis in parsing informative characteristics of auditory scenes. We propose a simplified system motivated by processing in the mammalian auditory system that can perform scene classification in isolation, 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 $\mu(t; \tau) = e^{-t/\tau}u(t)$ and $\tau = 2ms$ mimics the loss of phase-locking at the level of the midbrain. Given an input signal $s(t)$,

This work was partly supported by grants from NSF (IIS-0846112), AFOSR (FA9550-09-1-0234), NIH (1R01AG036424) and ONR (N000141010278). We would like to thank Malcolm Slaney for valuable discussions on topics of scene analysis and the role of attention.

the output of the first stage $y(t, f)$ is derived as follows:

$$\begin{aligned} y_{coch}(t,f) &= s(t) \otimes_t h(t;f) \\ y_{lin}(t,f) &= max[\partial_f y_{coch}(t,f), 0] \\ y_{mid}(t,f) &= y_{lin} \otimes_t \mu(t;\tau) \\ y(t,f) &= (y_{mid}(t,f))^{\frac{1}{3}} \end{aligned} \qquad (1)$$

where $\otimes_t$ represents convolution with respect to time.

The output $y(t, f)$ is then further analyzed to capture the detailed spectral modulations (or scales $\mathfrak{s}$ in cycles/octave) and temporal modulations (or rates $\mathfrak{r}$ in Hz) along the frequency and time axes, respectively. This analysis is done using a bank of modulation-tuned filters (MF) defined as

$$MF(f,t;\mathfrak{s},\mathfrak{r}) = \frac{1}{2\pi\sigma_t\sigma_f} e^{-\frac{1}{2}\left(\frac{t^2}{\sigma_t^2}+\frac{f^2}{\sigma_f^2}\right)} e^{2\pi i(\mathfrak{r}t+\mathfrak{s}f)} \qquad (2)$$

where σ_t, σ_f denote the spread of the filter. The filters are a linear approximation of auditory cortex neurons [12, 13]. The resulting representation $\mathfrak{R}$ is a high-dimensional tensor, parameterized by time t, frequency f, scale $\mathfrak{s}$ and rate $\mathfrak{r}$; given by:

$$\mathfrak{R}(f,t;\mathfrak{s},\mathfrak{r}) = |y(f,t) \otimes_{f,t} MF(f,t;\mathfrak{s},\mathfrak{r})| \qquad (3)$$

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

Finally, we integrate $\mathfrak{R}$ 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 C0 energy component of the MFCCs are ignored to yield a 12 dimensional vector [our analysis shows that C0 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.

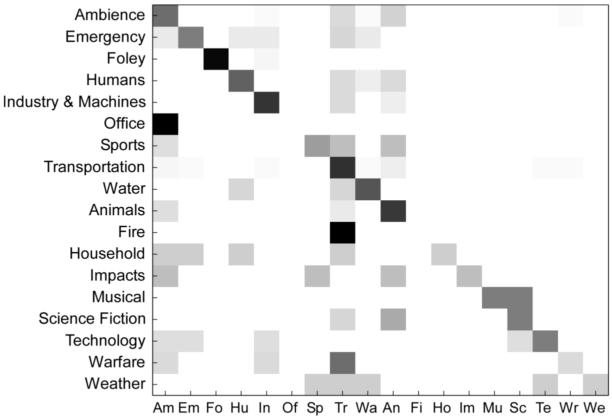


Figure 1: The confusion matrix of the classifier with the true labels on the vertical axis and predicted labels on the horizontal axis.

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 $\mathbf{66.3}\%$, compared to $\mathbf{45.5}\%$ using MSS features and $\mathbf{39.43}\%$ 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 in 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.

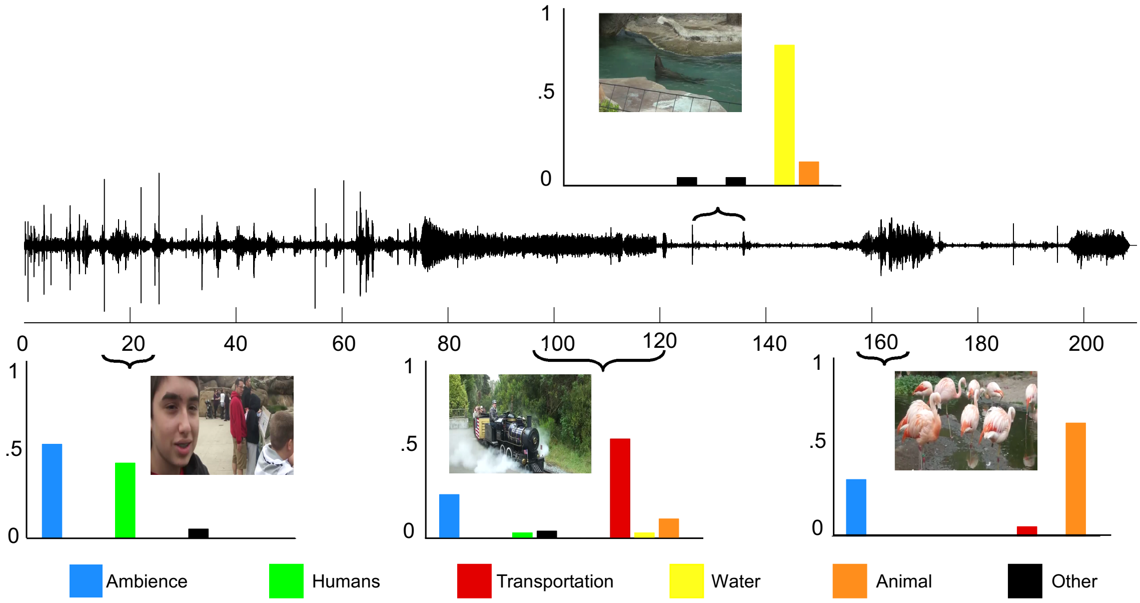


Figure 2: An audio recording collected from youtube as a function of time(in seconds) is shown here. Examples of histograms of labels derived from the classifier over different segments of time are also shown. The duration between 17s to 22s has human speech along with a background of other speakers. The sound of a train from 94s to 120s and water sounds from 128s to 134s is heard along with the babble of other speakers. Similarly, animal sounds occur from 158s to 164s.

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 occurence 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 $\mathfrak{R}(f,t;\mathfrak{s},\mathfrak{r})$ 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 $W(f) = \sum_{t,\mathfrak{s},\mathfrak{r}} M(f,t;\mathfrak{s},\mathfrak{r})$ where M indicates the mean of $\mathfrak{R}$ 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).

$$\mathfrak{R}_w(f,t;\mathfrak{s},\mathfrak{r}) = |y(f,t) \otimes_{f,t} MF(f,t;\mathfrak{s},\mathfrak{r})| * W(f) \quad (4)$$

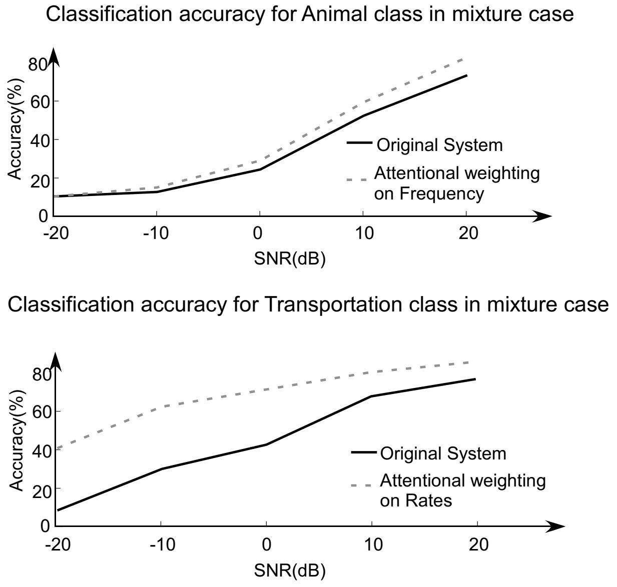
where $\mathfrak{R}_w$ is the representation upon applying the attention weighting W.

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].


Classification accuracy for Animal class in mixture case

Classification accuracy for Transportation class in mixture case

Figure 3: The accuracy of the classifier for animal class and transportation class when mixed with other classes, with and without weighting on the frequency and rate axis respectively is shown as a function of SNR.

5. References

[1] A. Bregman, *Auditory scene analysis: the perceptual organization of sound*. MIT Press, 1990.

[2] O. Kalini, S. Sundaram, and S. Narayanan, “Saliency-driven unstructured acoustic scene classification using latent perceptual indexing,” in *Proceedings of the IEEE International Workshop on Multimedia Signal Processing (MMSP), Rio de Janeiro, Brazil*, October 5-7, 2009.

[3] X. Zhuang, X.Zhou, T.Huang, and M.Hasegawa-Jhonson, “Feature analysis and selection for acoustic event detection,” in *Proceedings of ICAASP08*, 2008.

[4] J. Portelo, M. Bugalho, I. Trancoso, J. Neto, A. Abad, and A. Serralheiro, “Non-speech audio event detection,” in *Proceedings of ICAASP’09*, april 2009, pp. 1973 –1976.

[5] A. Temko and C. Nadeu, “Classification of acoustic events using svm-based clustering schemes,” *Pattern Recognition*, vol. 39, pp. 682–694, 2006.

[6] W. Chu, W. Cheng, J. Wu, and J. Y. jen Hsu, “A study of semantic context detection by using svm and gmm approaches,” in *ICME04*, 2004.

[7] L. Miller, M. Escabi, H. Read, and C. Schreiner, “Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex.” *J Neurophysiol*, vol. 87, no. 1, pp. 516–527, Jan 2002.

[8] S. Shamma, M. Elhilali, and C. Micheyl, “Temporal coherence and attention in auditory scene analysis,” *Trends in Neurosciences*, vol. 34, no. 3, pp. 114–123, 2011.

[9] L. Itti, G. Rees, and J. K. Tsotsos, Eds., *Neurobiology of Attention*. San Diego, CA: Elsevier, Jan 2005, the first encyclopedic volume on attention research, with 111 chapters from over 160 experts in the field.

[10] T. Chi, P. Ru, and S. Shamma, “Multiresolution spectrotemporal analysis of complex sounds,” *Journal of the Acoustical Society of America*, vol. 118(2), pp. 887–906, August 2005.

[11] X. Yang, K. Wang, and S. Shamma, “Auditory representations of acoustic signals,” *IEEE transactions on information theory*, vol. 38(2), pp. 824–839, March 1992.

[12] T. Ezzat, J. Bouvrie, and T. Poggio, “Spectrotemporal analysis of speech using 2-d gabor filters,” in *INTERSPEECH-2007*, 2007, pp. 506–509.

[13] F. E. Theunissen, K. Sen, and A. J. Doupe, “Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds,” *The Journal of Neuroscience*, vol. 20, no. 6, pp. 2315–2331, March 2000.

[14] L. D. Lathauwer, B. D. Moor, and J. Vandewalle, “A multilinear singular value decomposition,” *SIAM Journal on Matrix Analysis and Applications*, vol. 21(4), pp. 1253–1278, 2000.

[15] *The BBC Sound Effects Library Original Series*. http://www.soundideas.com, May 2006.

[16] K. Patil, D. Pressnitzer, S. Shamma, and M. Elhilali, “Music in our ears: The biological bases of musical timbre perception,” (Under Review).

[17] *http://www.youtube.com/watch?v=1dPYZIskA0c*, March,2012.

[18] N. L. Wood and N. Cowan, “The cocktail party phenomenon revisited: Attention and memory in the classic selective listening procedure of cherry (1953),” *Journal of Experimental Psychology: General*, vol. 124, no. 3, pp. 243–262, 1995.

[19] V. Poghosyan and A. Ioannides, “Attention modulates earliest responses in the primary auditory and visual cortices,” *Neuron*, vol. 58, pp. 802–813, 2008.

[20] C. Karns and R. Knight, “Intermodal auditory, visual, and tactile attention modulates early stages of neural processing,” *Journal of Cognitive Neuroscience*, vol. 21, pp. 669–683, 2008.

[21] R. J. Peters and L. Itti, “Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention.” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[22] S. Han and N. Vasconcelos, “Biologically plausible saliency mechanisms improve feedforward object recognition,” *Vision Research*, vol. 50, pp. 2295–2307, 2010.

[23] V. Navalpakkam and L. Itti, “An integrated model of top-down and bottom-up attention for optimizing detection speed,” in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2, 2006, pp. 2049 – 2056.

[24] J. Fritz, M. Elhilali, S. David, and S. Shamma, “Does attention play a role in dynamic receptive field adaptation to changing acoustic salience in a1?” *Hear Res*, vol. 229, pp. 186–203, 2007.

[25] J. Fritz, S. Shamma, M. Elhilali, and D. Klein, “Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex,” *Nature Neuroscience*, vol. 6, pp. 1216–1223, 2003.

[26] J. Ahveninen, M. Hmlinen, I. Jaaskelainen, S. Ahlfors, S. Huang, F.-H. Lin, T. Raij, M. Sams, C. Vasios, and J. Belliveau, “Attention-driven auditory cortex short-termplasticity helps segregate relevant sounds from noise,” *Proceedings of the National Academy of Sciences*, vol. 108, pp. 4182–4187, 2011.