MULTIRESOLUTION AUDITORY REPRESENTATIONS FOR SCENE CLASSIFICATION

Kailash Patil and Mounya Elhilali

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

ABSTRACT

Here, we propose a framework that provides a detailed analysis of the spectrotemporal modulations in the acoustic signal, augmented with a discriminative classifier using support vector machines. We have seen that such representation is successful at capturing the non-trivial commonalties within a sound class and differences between different classes[1, 2, 3].

***Index Terms*—** Multiresolution analysis, Auditory representation, Modulation domain.

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' [4] 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 [5, 6, 7]. 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 [7, 8], multi-layered perceptron [5] and generative approaches such as Gaussian Mixture Models (GMM) [9] have been proposed. It has further been suggested that discriminative approaches outperform the generative approaches [9].

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 [10]. 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. The proposed model is described in Sec. 2 and the results on the giving training data is described in Sec. 3.

2. METHODS

The proposed model is divided into Sensory Processing, Object Representation modules. Each of these modules and the experimental setup is described below.

2.1. Sensory Processing

The incoming sound is processed to extract informative features using techniques that mimic the behavior of the mammalian auditory system. This can be further divided into two steps - the subcortical stage and the cortical processing stage. In the subcortical stage, the waveform is passed through a set of 128 asymmetric filters $h(t;f)$ placed uniformly on a logarithmic axis covering 5.3 octaves starting from $180Hz$. This is similar to the frequency-space transformation of the cochlear membrane. This is followed by a spectral derivative and a half wave rectification stage, which models the lateral inhibition networks in the cochlear nucleus, sharpening the frequency resolution of these filters. The mid brain processing is implemented as a short term integration with window $\mu(t;\tau) = e^{-t/\tau}u(t)$ and $\tau = 2ms$ followed by cubic root compression. These subcortical transformations can be collectively written as in Eq. 1 and the details of implementation can be found in [11].

$$y(t,f) = (max(\partial_f(s(t) \otimes_t h(t;f)), 0) \otimes_t \mu(t;\tau))^{\frac{1}{3}} \quad (1)$$

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

This resulting time-frequency representation is referred to as the auditory spectrogram. In the cortical stage, this spectrogram is analyzed locally for joint spectrotemporal modulations using a bank of modulation tuned filters. These filters as defined in Eq.

2, are shaped like 2D Gabors, which are known to be a linear approximation to the receptive field shapes of auditory cortex neurons [12, 13]. The temporal modulation rate and spectral modulation rate are denoted by $\mathfrak{r}$ and $\mathfrak{s}$ respectively. The filtering operation can then be written as simple two dimensional convolution as in Eq. 3 which yields a four dimensional tensor representation.

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

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

The MF filters are tuned to 10 upward rates and 10 downward rates $\{\mathfrak{r}$ = 2, 3.4, 5.7, 9.5, 16, 26.9, 45.3, 76.1, 128, 215.3 Hz$\}$ and 11 scales $\{\mathfrak{s}$ = 0.25, 0.35, 0.5, 0.71, 1, 1.41, 2, 2.83, 4, 5.66, 8 cycles/octave$\}$, resulting in a total of 220 filters.

2.2. Object Representation

Each audio recording is windowed into $1s$ segments with an overlap of $0.5s$. We integrate the cortical representation $\mathfrak{R}$ over the time duration of each window. To facilitate the machine learning module we reduce the number of dimensions via Tensor Singular Value Decomposition [14] to keep 99% of the variance resulting in a 96 dimensional feature vector. We learn the boundaries between classes using one vs one SVM framework with RBF kernel. To classify an unknown test recording, the distance from boundaries is converted to a probability estimate for each class. The probabilities from each window are weighed by the energy present in the window and finally averaged over the top 90% energetic frames.

3. RESULTS

On the training data provided we ran a 5-fold cross validation and achieved an average accuracy of 73% with a standard deviation of 13%.

4. REFERENCES

[1] K. Patil, D. Pressnitzer, S. Shamma, and M. Elhilali, “Music in our ears: The biological bases of musical timbre perception,” *PLoS Comput. Biol.*, vol. 8, no. 11, p. e1002759, 11 2012. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pcbi.1002759

[2] K. Patil and M. Elhilali, “Goal-oriented auditory scene recognition,” in *Proceedings of INTERSPEECH 2012*, Portland, USA, September 2012.

[3] ——, “Task-driven attentional mechanisms for auditory scene recognition,” in *Proceedings of ICASSP 2013*, Vancouver, Canada, May 2013.

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

[5] 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.

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

[7] 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.

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

[9] 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.

[10] 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.

[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, “Spectro-temporal 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.