Discriminating Sound Textures

Jayson Lynch
MIT CSAIL
32 Vassar Street
Cambridge, MA 02139
jaysonl@mit.edu

Eric Mannes
MIT CSAIL
32 Vassar Street
Cambridge, MA 02139
mannes@mit.edu

ABSTRACT

[**xxx write abstract**]

Keywords: Machine Learning, Sound Textures, Human Audition

1. INTRODUCTION

This paper explores a new feature set for audio classification. Specifically, we look at a feature ensemble related to compressibility, sparsity, and temporal homogeneity in the problem of classifying audio signals as *sound textures*. Sound textures, such as rain or crackling fire, are the result of many similar acoustic events. These sounds seem to form a cohesive auditory category which may be processed by the brain in a distinct fashion[6, 5]. Their study has partially been motivated by the analogous visual textures which comprise a significant body of literature. Tomita and Tsuji provide a book on the computational analysis of visual textures[14]. The ideas that visual textures may be characterized by simple statistical features[2] or atomistic units[3] is carried over into some of the theory of sound textures. Recent work by McDermott and Simoncelli show many sound textures are well characterized by their time-averaged statistical properties[6, 5]. Saint-Arnaud's Master's Thesis attempts to extract the sound atoms that make up sound textures and use this as a basis for a classifier[11]. The thesis also discusses human perception of sound textures and the difficulties surrounding a good definition. A significant amount of the machine learning work in sound textures has been around synthesis often using wavelet hierarchies[4] or sound atoms[12]. Schwarz provides a general overview of methods as well as a classification scheme of the synthesis methods used[13].

McDermott and Simoncelli work shows that for many sound textures, the salient features that cause humans to classify these sounds are controlled by a number of time-averaged statistics of the sounds[6, 5]. This leads to an interesting question of why some sounds seem to be identified by time-averaged properties, when many other, such as speech or music, depend strongly on their temporal structure. McDermott et. al. conjecture that the 'sparsity' or 'compressibility' are related to the sounds that are perceived in this way. In addition, they mention that sound textures are 'temporally homogeneous' without giving formal definitions of these terms. This paper seeks to build upon these ideas both to investigate new features for audio classification as well as to lend evidence toward the previous conjectures about sound texture perception. We provide several formal definitions of audio compressibility, sparsity, and temporal homogeneity in Section 2. These are then extracted as a feature set and are used in a machine learning algorithm for sound textures, described in Section 3.

2. DEFINITIONS

We provide working definitions for examples of three main descriptors: sparsity, compressibility, and temporal homogeneity. Since these characteristics can be captured in different ways, which are not always equivalent, we choose to work with multiple formal definitions.

Compressibility was the simplest to work out. We chose a standard lossy and lossless compression algorithm and measured the compression ratio for the .wav flies. In this case we used [**xxx what audio compression did we use?**]

How to define sparsity was slightly less clear, since it usually refers to the quantity of zero entries with respect to some basis. To overcome this, we decided to use several natural representations. We looked at sparsity in the time domain with respect to the actual time series of the audio sample. We also looked at the sparsity in the frequency domain under both uniform and log-scale transforms. To be more precise, we took short-term Fourier transforms, constant q transforms, and Mel transforms of the time series (using the implementations in the Python library Librosa[7]) and counted the ratio of entries near zero. Due to noise and precision error, we assumed an entry was empty if the magnitude of the value was less than 10^{-5}. There are also algorithms for extracting low rank matrices assuming one has sampled from noisy data[8]. We would have liked to have implemented and run one of these algorithms, using the derived rank as a measure of sparsity, but we did not have time to extract this feature.

Finally temporal homogeneity refers to consistency in the signal over time. Obviously we can't have everything be exactly the same, so one must pick specific features with which to check consistency. A paper on visual textures provides a precise definition: X is homogeneous with respect

to the function $\Phi : \mathbb{R}^L \to \mathbb{R}$ with tolerance ε and probability p if the average of Φ over a sample $x \in X$ is a good approximation to the average of Φ over all of X [10].

$$P_x(\mathrm{E}(\Phi(x)) - \mathrm{E}(\Phi(X)) < \varepsilon) \geq p$$

In practice, this definition is slightly cumbersome to use, and requires picking either our probability or threshold arbitrarily. In the same spirit, we instead decide to use the variance of a given statistic over a sample. Thus we calculate the inhomogeneity of X with respect to Φ as

$$\Sigma_{x \in X}(\Phi(x) - \mathrm{E}(\Phi(X))^2$$

We additionally note that the original definition of homogeneity was given with respect to translations over a two dimensional space, whereas we take samples over a single time dimension.

In this paper we decided to use the first through fourth moments of the time series and various sub-bands of the audio signal. In particular we compute the short-term Fourier transforms, constant q transforms, and Mel transforms of the time series. We further computed the amplitude envelopes of these waveforms and took the corresponding short-term Fourier transforms, constant q transforms, or Mel transforms of the resulting amplitude envelope following the auditory model in [6, 5]. We would have also liked to look at the temporal homogeneity of the cross-correlation of the bands, capturing all of the statistics used in that paper.

[xxx Give formal and informal definitions of the features we used]

3. METHODS

All of our models were written in Python with the help of a number of external libraries. NumPy[15] and Librosa[7] were used for audio processing; SciPy[1] provided a number of useful statistical methods; and scikit-learn[9] was used for our regression and classification.

Our dataset comes from the audio samples used in McDermott and Simoncelli's paper[6] as a basis for their synthetic sounds. The dataset contains 175 sound samples, all 7 seconds long.

[xxx Describe how we set up our feature extraction. How we set up our learning. What our dataset is. Any other important process things.]

4. RESULTS

[xxx What correlated and what didn't? What feature (ensemble) lead to a good classifier?]

5. CONCLUSION

[xxx Speculate if these could be useful features in audio classification. Speculate about the implication to human audition. Give future directions.]

There are a number of future directions left open by this paper. First, the space of reasonable features has not been fully explored and may lead to yet better classification methods or more insight into the nature of sound textures. These features include: other standard compression algorithms, the entropy of the audio signals, the Kolmogorov complexity of the audio signal, rank estimation of the audio signal, temporal homogeneity of the cross-band correlations, and different sparsity estimation criteria. Another obvious next direction is the integration of these features with existing audio classification methods to attempt to improve performance in more complex settings. Third, there is still much to be understood about the nature of these features and what they can tell us about the audio files. The authors would have been interested to look at the correlation between features, as many of them should be highly related. Understanding when things like sparsity and compressibility differer might hold new insight into signal characteristics. Further, the audio dataset was limited and constrained in a number of ways, and it would be interesting to see how these features vary with over a wider set of sounds. Finally, there still remain important questions about the nature of human perception of sound textures. We showed that intuition about what makes up a sound texture is partially captured by some formal measures, but given none appeared to be necessary or sufficient, it seems that this issue may be more complex than originally stated. However, we also did not fully explore what might reasonably be meant by sparse and temporally homogeneous audio signals, so it is still possible that another definition will better capture what causes this qualitatively different auditory category.

Acknowledgments

We thank Prof Josh McDermott for his support in answering our questions about their research and providing us with their dataset. We also thank the course staff of 6.867 for their instruction and support this semester. In particular, we appreciate Marzyeh Ghassemi advice and guidance in shaping our project plan.

References

[1] E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. [Online; accessed December 10, 2015].

[2] B. Julesz. Visual pattern discrimination. *Information Theory, IRE Transactions on*, 8(2):84–92, 1962.

[3] B. Julesz. Textons, the elements of texture perception, and their interactions. *Nature*, 290(5802):91–97, 1981.

[4] S. Kersten and H. Purwins. Sound texture synthesis with hidden markov tree models in the wavelet domain. In *Sound and Music Computing Conference*, 2010.

[5] J. H. McDermott, M. Schemitsch, and E. P. Simoncelli. Summary statistics in auditory perception. *Nature neuroscience*, 16(4):493–498, 2013.

[6] J. H. McDermott and E. P. Simoncelli. Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron*, 71(5):926 – 940, 2011.

[7] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto. librosa: Audio and music signal analysis in python. 2015.

[8] S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, pages 1069–1097, 2011.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos,

D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[10] J. Portilla and E. P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40(1):49–70, 2000.

[11] N. Saint-Arnaud. *Classification of sound textures*. PhD thesis, Massachusetts Institute of Technology, 1995.

[12] N. Saint-Arnaud and K. Popat. Analysis and synthesis of sound textures. In *in Readings in Computational Auditory Scene Analysis*. Citeseer, 1995.

[13] D. Schwarz. State of the art in sound texture synthesis. In *Proc. Digital Audio Effects (DAFx)*, pages 221–231, 2011.

[14] F. Tomita and S. Tsuji. *Computer analysis of visual textures*, volume 102. Springer Science & Business Media, 2013.

[15] S. van der Walt, S. Colbert, and G. Varoquaux. The numpy array: A structure for efficient numerical computation. *Computing in Science Engineering*, 13(2):22–30, March 2011.