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

In this paper we investigate a feature set for the classification of sound textures based on signal compressibility, sparsity, and temporal homogeneity. The feature set was examined and used in a classifier for sound textures. We show that for sound samples with normalized amplitude, these features have very little predictive power with respect to whether a sound is textural.

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 might be processed by the brain in a distinct fashion[7, 8]. 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[17]. The ideas that visual textures may be characterized by simple statistical features[4] or atomistic units[5] 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[8, 7]. 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[14]. 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[6] or sound atoms[15]. Schwarz provides a general overview of methods as well as a classification scheme of the synthesis methods used[16].

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[8, 7]. 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.

2.1 Compression Rate

Compressibility was the simplest to work out. We chose a standard lossy compression algorithm and measured the compression ratio for the .wav flies. In this case we used the Ogg Vorbis variable-bitrate compression format.

Definition Given a compression scheme, the *compression rate* of a file is equal to

$$\frac{\text{size of uncompressed file (B)}}{\text{size of compressed file (B)}}.$$

2.2 Sparsity

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[9]) and counted the ratio of entries near zero to the total number of entries. Due to noise and precision error, we assumed an entry was empty if the magnitude of the value was less than 10^{-5}. In the samples we inspected closely this value was sufficiently small that all apparent signals were much larger than this value.

The Short Term Fourier Transform attempts to compute the Discrete Fourier Transform at every point for a changing signal by computing the DFT over a small sliding window. The STFT at a time n with a window size ω can be expressed as

$$STFT\,(x, \omega) = \Sigma_{m=-\infty}^{\infty} x[m]w[n-m]e^{-j\omega n}$$

where $x[i]$ and $w[i]$ are points in our signal[2].

The Constant Q and Mel Transforms are analogs to the Fourier Transform, but end up with a logarithmic spacing in the frequency domain, making them behave more like human auditory signal processing. We refer the reader to Judith Brown's paper[1] for details on the Constant Q Transform and to Saha and Sahidullah's paper[13] for details on the Mel Transform, as well as the Librosa source code[1] for their specific implementation details.

Another measure of sparsity is the rank of the matrix needed to fully specify a transform. There are also algorithms for extracting low rank matrices assuming one has sampled from noisy data[10]. 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 additional feature.

2.3 Temporal Homogeneity

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[12].

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

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

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 a related value we call the temporal inhomogeneity.

Definition The inhomogeneity of X with respect to Φ is

$$\Xi\,(X) = \Sigma_{x \in X}\left(\Phi(x) - \mathrm{E}\left(\Phi(X)\right)\right)^2$$

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. The k^{th} moment of a continuious function $f(x)$ is $\mu_k = \int_{-\infty}^{\infty} x^k f(x)dx$. As with Fourier Transforms, we need a discrete analogue to calculate this for our signal, so instead we use the sample moment $\sigma_k = \frac{1}{n}\Sigma_{i=1}^{n} s_i^k$ where s_i is our i^{th} point in our signal.

We still haven't fully specified our statistics. We decided to compute the temporal inhomogeneity of the first four moments of 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 inspired by the auditory model in [8, 7]. Additionally, we computed the temporal inhomogeneity of the root mean square energy of the signal. 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.

3. METHODS

All of our models were written in Python with the help of a number of external libraries. NumPy[18] and Librosa[9] were used for audio processing; SciPy[3] provided a number of useful statistical methods; and scikit-learn[11] was used for our regression and classification.

Our dataset comes from the audio samples used in McDermott and Simoncelli's paper[8] as a basis for their synthetic sounds. The dataset contains 175 sound samples, all 7 seconds long in a .wav format. Of these, 28 were discard due to naming ambiguities which prevented them from being identified with the values in the paper. These samples included many examples of sound textures, such as four different wind samples and multiple different rain samples, as well as some audio files that are obviously not sound textures, such as person speaking English and various rhythmic drumbeats. The spectrograms of several samples can be seen in Figure 1 and Figure 2 A full list can be found in Appendix A. In the prior work, these samples were used to generate synthetic sounds with similar statistical properties and these synthetic sounds. An experiment was run where these synthetic sounds were given realism ratings by people on a scale of 1 to 7. Since [7] showed that many sound textures seemed to be well characterized by these same sets of statistics, we use this realism score as a proxy for the 'texturality' of the audio samples.

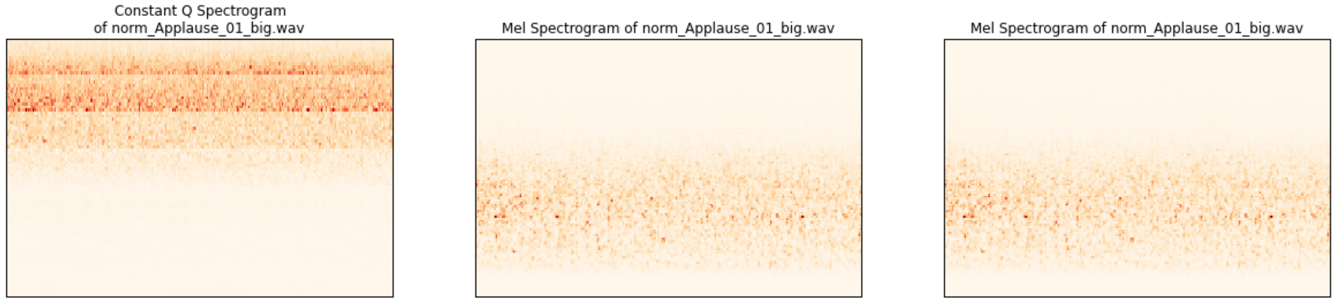
We performed a linear regression with L_1 regularization with all of our features. Half of the data was randomly selected to be used as training data, one quarter was reserved for a validation step to select our regularization constant, and a quarter was reserved to determine our final r-squared values. **[xxx talk about feature normalization]**

[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

When we ran LASSO regressions, we consistently found that two features were of the most predictive value: the RMS energy of a sound and the time homogeneity of the RMS

[1] https://github.com/bmcfee/librosa



(a) Constant Q spectrogram of an applause sound sample.

(b) Mel spectrogram of an applause sound sample.

(c) Short Term Fourier Transform spectrogram of an applause sound sample.

Figure 1: Three different spectrograms of the same sound sample of applause.

energy. A linear regression with either one of these features had R^2 of, on average, .45 on the test set.

While this early success seemed promising, we realized that it was, in fact, a problem. The RMS energy of a sound is essentially its loudness. As seen in Figure 3a, the louder the recording, the more like a sound texture it was.

Because the degree to which a sound is a sound texture is not based on how loud it is played, we assume that this is because of an artifact in how the sounds were recorded. Many things that are clearly sound textures are quite loud (e.g., applause) or were likely recorded with a microphone close to the source of the sound (e.g., running water). On the other hand, sounds that are definitely not sound textures, like speech or church bells, may have been recorded at more of a distance.

5. CONCLUSION

Out of the large number of features investigated, we've shown that none of them have predictive power when we normalized the amplitudes of our sound samples to fall in the same range. Although we have only ruled out their usefulness in predicting sound textures, it seems unlikely that they will be generally useful in other types of audio classification. Further, this negative result has implications in the understanding of sound textures and human audition. It suggests that one informal conjecture about the nature of sound textures is incorrect. This will hopefully spur a refinement of ideas and definitions surrounding sound textures and provide a useful piece of evidence when considering how to move forward. One might now believe the decision to throw away temporal information about an audio signal to be more nuanced than previously thought and hopefully further work on features such as temporal coincidence, reverberation, natural harmonics, and temporal regularity will now be increasingly motivated.

There are a number of future directions left open by this paper. First, the space of reasonable features has not been fully explored and could still lead to useful features for classification 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. Although these features are all related, they are certainly not the same and the authors know of no way to rule them out without direct investigation. Another possible direction is the integration of these features with existing audio classification methods to attempt to improve performance in more complex settings. We do not believe this is likely to be a fruitful search; however, the domain of attempting to classify the texturalality of sound is very narrow compared to the entirety of computer sound classification. Third, there is still much to be understood about the nature of these features and what they do tell us about an audio file. The authors would have been interested to look at the correlation between features and examples where these notions do and do not line up. Understanding when things like sparsity and compressibility differer might hold new insight into signal characteristics. Finally, there still remain important questions about the nature of human perception of sound textures. Now that we've ruled out a number of characterizing features, the question of what makes something a sound texture is even more tantalizing.

Acknowledgments

We thank Professor Josh McDermott for his support in answering our questions about his and Simoncelli's research and for 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's advice and guidance in shaping our project plan.

References

[1] J. C. Brown. Calculation of a constant q spectral transform. *The Journal of the Acoustical Society of America*, 89(1):425–434, 1991.

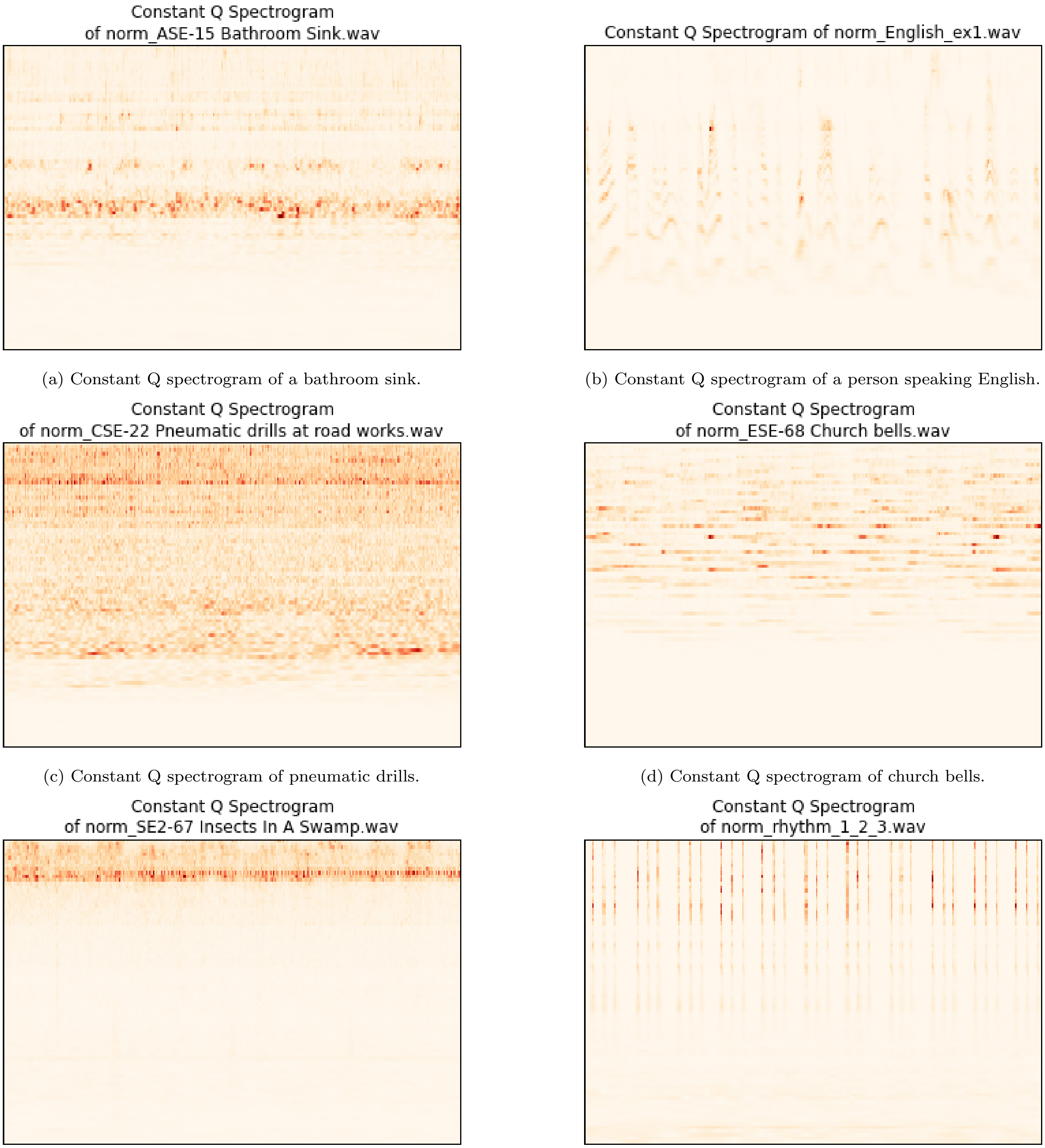
[2] R. Gutierrez-Osuna. Lecture notes 6 for introduction to signal processing, 2001–. [Online; accessed December 10, 2015].

[3] E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python.

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

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

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



(a) Constant Q spectrogram of a bathroom sink.

(b) Constant Q spectrogram of a person speaking English.

(c) Constant Q spectrogram of pneumatic drills.

(d) Constant Q spectrogram of church bells.

(e) Constant Q spectrogram of insects in a swamp.

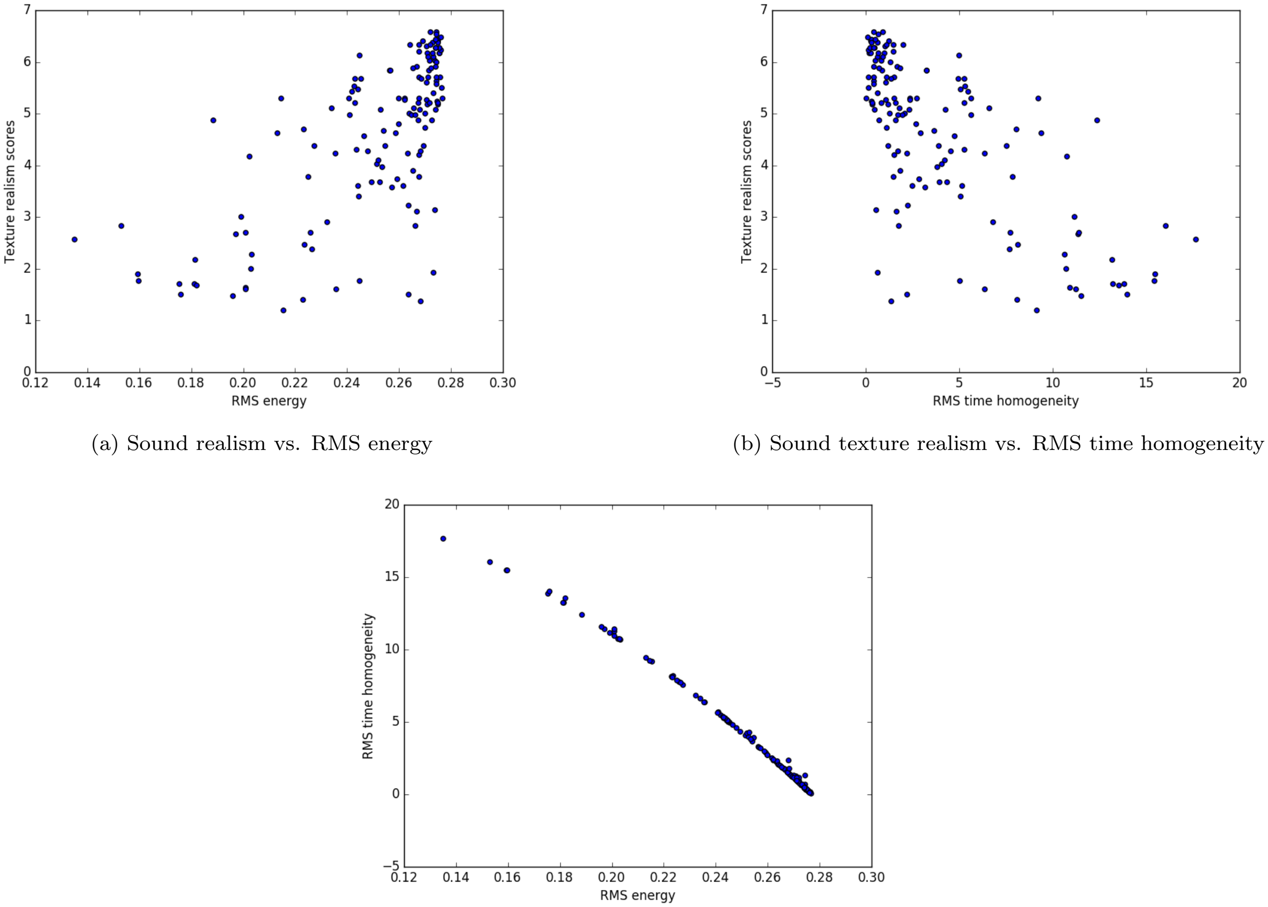
(f) Constant Q spectrogram of a tapped rhythm.

Figure 2: Constant Q spectragrams of a variety of different sounds. The left column contains sound textures and the right does not.

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

[8] 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.

[9] 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.



(a) Sound realism vs. RMS energy

(b) Sound texture realism vs. RMS time homogeneity

(c) RMS time homogeneity vs. RMS energy

[10] 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.

[11] 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.

[12] 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.

[13] M. Sahidullah and G. Saha. Design, analysis and experimental evaluation of block based transformation in {MFCC} computation for speaker recognition. *Speech Communication*, 54(4):543 – 565, 2012.

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

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

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

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

[18] 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.

APPENDIX

A. DATASET

Dataset with realism ratings from [8].

Insects in swamp 6.57
Heavy rain on hard surface 6.53
Frogs 6.53
Rain 6.47
Applause
big room 6.43
Radio static 6.43
Stream 6.43
Jungle rain 6.40
Air conditioner 6.40
Stream near small waterfall 6.37
Frogs 6.37
Frogs and insects 6.37
Frying eggs 6.33
Frogs 6.33
Wind
blowing 6.33
Wind
whistling 6.33
Insects during day in South 6.30
Radio static 6.30
Frogs 6.30
Heavy rain falling and dripping 6.27
Applause
large crowd 6.27
River running over shallows 6.27
Construction site ambience 6.23
Waterfall 6.20
Sparrows
large excited group 6.17
Pneumatic drills 6.17
Small river 6.17
Fast running river 6.17
Rain in woods 6.13
Water trickling into pool 6.10
Bathroom sink 6.10
Water running into sink 6.03
Frying bacon 6.03
Rain in ihe woods 6.00
Fire
forest inferno 5.97
Birds in forest 5.90
Linotype 5.90
Bee swarm 5.90
Applause 5.90
Bath being drawn 5.90
Rustling paper 5.87
Train speeding down railroad tracks
steam 5.87
Rattlesnake rattle 5.83
Fire
burning room 5.83
Bubbling water 5.83
Fire
burning room 5.83
Thunder and rain 5.73
Fire 5.70
Wind
moaning 5.70
Bulldozer 5.70
Babble 5.70
Fire 5.70
Wind
spooky 5.70
Water lapping gently 5.67
Shaking coins 5.67
Helicopter 5.67
Seagulls 5.63
Crunching cellophane 5.63
Sander 5.60
Radio static 5.60
Teletype
city room 5.57
Steam shovel 5.53
Pigeons cooing 5.50
Metal lathe 5.47
Bee swarm 5.47
Lapping waves 5.43
Geese cackling 5.40
Train speeding down railroad tracks
Diesel 5.30
Lake shore 5.30
Sanding by hand 5.30
Blender 5.30
Teletype 5.30
Birds in tropical forest 5.27
Drumroll 5.27
Surf hitting beach 5.23
Industrial machinery 5.20
Crowd noise 5.20
Rolling coin 5.20
Ducks quacking 5.20
WWII bomber plane 5.17
Applause 5.17
Idling boat 5.17
Jackhammer 5.10
Brushing teeth 5.10
Horse trotting on cobblestones 5.07
Scratching beard 5.07
Printing press 5.07
Writing with pen on paper 5.00
Train locomotive
steam engine 5.00
Helicopter fly by 4.97
Pouring coins 4.97

Motorcycle idling 4.97
Fire 4.93
Crumpling paper 4.87
Ship anchor being raised 4.87
Jingling keys 4.87
Electric adding machine 4.80
Horse walking in snow 4.73
Cymbals shaking 4.70
Fire
in chimney 4.67
Tambourine shaking 4.67
Pouring coins 4.63
Rhythmic applause 4.63
Cat lapping milk 4.57
Seaside waves 4.43
Rustling paper 4.37
Horse pulling wagon 4.37
Vacuum cleaner 4.37
Horse and carriage 4.30
Power saw 4.30
Tire rolling on gravel 4.27
Horse and buggy 4.27
Steam engine 4.23
Cement mixer 4.23
Power saw 4.23
Castanets 4.23
Ox cart 4.20
Battle explosions 4.17
Chickens squawking 4.10
Rubbing cloth 4.03
Rain beating against window panes 3.97
Typewriter
IBM electric 3.90
Lawn mower 3.77
Gargling 3.77
Horse gallop on soft ground 3.73
Applause
foreground clapper 3.67
Sawing by hand 3.67
Crumpling paper 3.60
Wolves howling 3.60
Fast breathing 3.57
Dogs 3.40
Out of breath 3.23
Windshield wipers 3.20
Pile driver 3.13
Silly mouth noise 3.10
Large diner 3.00
Filing metal 2.90
Typewriter
manual 2.83
Fire alarm bell 2.83
Knife sharpening 2.83
Typewriter
old 2.70
Pile driver 2.70
Clock ticking 2.67
Jogging on gravel 2.67
Castanets 2.57
Hammering copper 2.47
Laughter 2.47
Tapping rhythm 2.37
Running up stairs 2.27
Typewriter
IBM selectric 2.17
Men marching together 2.00
Tapping on hard surface 1.93
Railroad crossing 1.90
Tapping 1
2 1.77
Wind chimes 1.77
Corkscrew against desk edge 1.70
Reverse drum beats
snare 1.70
Tapping 1
2
3 1.67
Snare drum beats 1.63
Walking on gravel 1.60
Snare rimshot sequence 1.60
Music
Apache drum break 1.50
Music
mambo 1.50
Bongo loop 1.47
Firecrackers 1.40
Person speaking French 1.37
Church bells 1.20
Person speaking Engli