The IPUS $\mathcal{C}.2$ Sound Understanding Testbed Acoustic Modeling Framework and Sound Library*

Frank Klassner† Ramamurthy Mani‡

CMPSCI Technical Report 95-65
July 1995

†Computer Science Department
University of Massachusetts
Amherst, Massachusetts 01003
klassner@cs.umass.edu

‡ECS Department
Boston University
Boston, MA 02125
rmani@bu.edu

Abstract

This report describes the models and features used to represent sounds in the IPUS (*Integrated Processing and Understanding of Signals*) Sound Understanding Testbed. It also catalogues the forty-five sounds in the acoustic library used to generate scenarios for testing the $\mathcal{C}.2$ configuration of the testbed.

*This work was supported by the Office of Naval Research under contract N00014-92-J-1450. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

1 Introduction

The $\mathcal{C}.2$ configuration of the IPUS SUT(*Integrated Processing and Understanding of Signals* Sound Understanding Testbed) is intended to serve as a research vehicle to explore (1) the use of approximate processing techniques within the basic IPUS framework [3] and (2) the scalability of the framework in handling moderately large sound libraries and complicated acoustic scenarios. The first part of this report describes the models and features used to represent sounds in the $\mathcal{C}.2$ SUT, while the second part catalogues the forty-five sounds in the acoustic library used to generate scenarios for testing the $\mathcal{C}.2$ testbed.

2 SUT Acoustic Modeling Framework

The SUT is designed to interpret acoustic scenarios containing sounds with a variety of frequency behaviors. Table 1 summarizes the sound categories for which we developed our acoustic modeling framework.

CATEGORY	PROPERTIES	EXAMPLES
chirp	the source has relatively smooth time-dependent frequency shifts	owl hoot, door creak
harmonic	the sound has a set of frequencies $\{f_1, \ldots, f_n\}$, such that they all can be represented as integer multiples of a fundamental frequency f_0. Note that some multiples can have zero energy.	fire alarm, car horn
impulsive	the source's acoustic energy is concentrated in a short period of time. Such sounds tend to have significant energy throughout the spectrum (wideband energy).	door knock, pistol shot
repetitive	the sound exhibits repetitive behavior in time; this repetition need not have a precise period.	footsteps, telephone ring
transient	the source exhibits well-defined attack (signal onset) or decay (signal turn-off) behaviors that show energy changes in frequency tracks found during the sound's steady-state behavior.	bell toll reverberation, hairdryer turning on

Table 1: *Sound categories represented in the IPUS acoustic testbed library.*

The $\mathcal{C}.2$ SUT's modeling framework for interpreting sounds from the above acoustic categories uses thirteen partially-ordered evidence representations. The representations are implemented through thirteen levels on the testbed's hypothesis blackboard. Figure 1 illustrates the support relationships among the representations, while the following discussion highlights the representations' content:

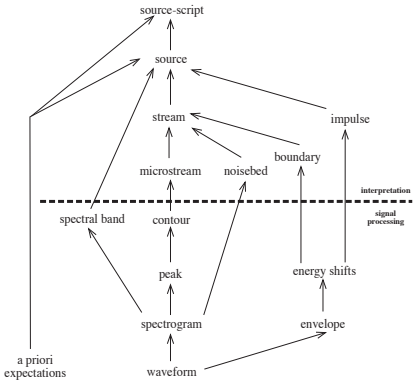

Figure 1: *The acoustic abstraction hierarchy for the IPUS sound understanding testbed.*

WAVEFORM: the raw waveform data. This representation is maintained since the testbed architecture will sometimes need to reprocess data. To conserve memory, only the last 3 seconds of waveform data are kept on the testbed's blackboard.

ENVELOPE: the envelope, or shape, of the time-domain signal. This level also maintain statistics such as zero-crossing density and average energy for each block of signal data.

SPECTROGRAM: spectral hypotheses derived for each waveform segment through algorithms such as the Short-Time Fourier Transform (STFT) and the Quantized Short-Time Fourier Transform (QSTFT) [4], an approximate SPA.

PEAK: peak spectral energy regions in each time-slice in a spectrogram. These indicate narrow-band features in a signal's spectral representation.

SHIFT: sudden energy changes in the time-domain envelope.

EVENT: time-domain events, which group shifts into boundaries (i.e. a step-up or step-down in time domain energy indicating the possible start or end of some sound) and impulses (i.e. sudden spikes in the signal).

CONTOUR: groups of peaks whose time indices, frequencies, and amplitudes represent a contour in the time-frequency-energy space with uniform frequency and energy behavior.

SPECTRAL BAND: regions of spectral activity (i.e. clusters of peaks) in a spectrogram. Our introduction of this abstraction level represents a knowledge approximation technique that avoids over-reliance on strict narrowband descriptions of sounds by mapping rough clusters of spectral activity in a spectrogram to only those sounds in the sound library that overlap those frequency regions.

MICROSTREAM: a track. A microstream represents a sequence of contours that has an energy pattern consisting of an attack region (signal onset), a steady region, and a decay (signal fadeout) region. The attack and decay regions can have frequency chirps in addition to their energy change. The steady region of "long-term" tracks have frequency and amplitude entropy measures indicating the frequency and energy variability of the peaks in in the track.

NOISEBED: the wideband component of impulsive areas within a sound source's acoustic signature. Microstreams (tracks) often form "ridges" on top of noisebed "plateaux," but not every noisebed has well-defined microstreams associated with it.

STREAM: we apply perceptual streaming criteria developed in the psychoacoustic research community [2] to group microstreams and noisebeds as support for stream hypotheses, or entities to be recognized as sound-sources. Specifically, our testbed knowledge sources group microstreams together when they have similar fates (e.g. synchronized onset- and end-times, synchronized chirp behavior), or when they share a harmonic relationship. Noisebeds are predicted and searched for only after a stream has been identified as a particular source's signature.

SOURCE: stream hypotheses, with their durations supported by boundaries, are interpreted as sound-source hypotheses according to how closely they match source-models available in the testbed library. Partial matches (e.g. a stream missing a microstream, or a stream with duration shorter than expected for a particular source) are accepted, but hypotheses resulting from these matches will later cause the testbed to attempt to account for the missing or ill-formed evidence (e.g. microstreams or noisebeds) as artifacts of inappropriate front-end procssing.

SCRIPT: temporal streaming of a sequence of sources into a single unit (e.g. a periodic source such as footsteps being composed of a sequence of footfalls, or the combination of cuckoo-chirps and bell-tones in a cuckoo-clock chime).

The following discussion elaborates the information about microstream entropies and noisebeds features that is maintained in the acoustic database's sound-source models.

2.1 Microstream Entropy Values

Microstream entropies are intended to express the variability in frequency and amplitude exhibited by the steady-state peaks in a sound's spectrogram. They are calculated only for sounds whose microstreams' steady-state behavior lasts longer than 0.75 seconds, and only for peaks generated from a 2048-point short-time Fourier Transform with a 256-point decimation and a 1024-point rectangular analysis window. Since both frequency and amplitude entropies are calculated in the same manner, we describe only the amplitude entropy calculation in detail.

The first step in calculating a microstream's amplitude entropy is to calculate the *segment amplitude entropy*,

$$\nu_a^k = \frac{\mu_k^a}{\sigma_k^a}$$

for each segment in the steady region of the microstream, where μ_a^k and σ_a^k are the mean and standard deviation, respectively, of the amplitudes of the M peaks in the k'th segment in the track. In our library a segment has $M = 30$ peaks, which covers 0.5 seconds in a 16KHz-sampled signal's spectrogram. Note that segments overlap; the k'th segment includes the k'th through $k + M - 1$'th consecutive peaks. By considering tracks of length at least 0.75-seconds, this gives us at least 15 segments per track. A microstream's *amplitude entropy* is defined by the values μ and σ, which are the mean and standard deviation of the segment entropies, respectively.

2.2 Noisebed Models

Our noisebed model is generated from an impulsive sound's spectrogram as follows. The spectrogram is produced from a 128-point short-time Fourier Transform with 32-point decimation and a 64-point rectangular analysis window. A 0.2-second region in the spectrogram is divided into an 8×10 (frequency × time) grid, starting at a point just before the sound reaches maximum energy. The grid's energy values are normalized with the maximum grid tile set to 1.0. From this grid four noisebed features are calculated: (1) the spectral center of gravity at the time the impulse is at its maximum energy, (2) the difference between the energy of the maximum-energy grid tile and that of the tile one frequency-slice *above* it at the same time, (3) the difference between the energy of the maximum-energy grid tile and that of the tile one frequency-slice *below* it at the same time, and (4) the least-squares exponential decay rate fitted to the energy values in the tiles in the same frequency-slice as the maximum-energy tile, but at time-slices after and including the maximum tile.

3 The Acoustic Library

This section describes the sound library that was used to generate scenarios for testing the $\mathcal{C}.2$ configuration of the IPUS acoustic testbed. The sound models in this library were derived from at least five instances for each sound; in the case of impulsive sounds the number of instances is often more. Each sound instance was captured in a signal stream at most 5 seconds long and sampled at 16 KHz.

3.1 Library Summary

The following is an alphabetical list of the sounds in the library. With the exception of the impulsive sounds "Knock," and "Clap," all sounds were extracted from a commercial tape provided by Auditec of St. Louis, Incorporated [1]. We indicate the cases where the name used for a sound in the IPUS library is different from the name used in the Auditec tape index.

1. **alarm clock 1** an old-fashioned bell+ringer.
2. **alarm clock 2** an electric clock alarm.
3. **bell1 toll** the "bell tolling" Auditec tape sound; lower-pitched, quick tolling succession.
4. **bell2 toll** the "country church bell" Auditec tape sound; higher-pitched, widely-spaced tolls.
5. **bicycle bell**
6. **bugle** the "coach horn" Auditec tape sound.
7. **burglar alarm**
8. **car running** background sound of car interior at 60 mph.
9. **car horn** horn is "beeped."
10. **chicken** the clucks of a chicken.
11. **chime** notes on an orchestra chime
12. **clap** a hand-clap
13. **clock chime**
14. **clock ticks**
15. **cuckoo clock** really has two simultaneous sounds: cuckoo & chimes.
16. **doorbell chimes**
17. **door creak**
18. **firengine bell**
19. **firehouse alarm**
20. **foghorn**
21. **footsteps** the "man walking" sound from Auditec tape.
22. **glass clink** the sound of glasses in a crate.
23. **gong**
24. **hairdryer** the turn-on, steady-state, and turn-off of a hairdryer
25. **knock** knocks on a wooden door.

26. **oven buzzer**

27. **owl** the hoots of a tawny-owl, to be precise.

28. **pistol shot**

29. **police car siren** a european-style police siren.

30. **razor** an electric razor turning on, running, then turning off.

31. **rooster** the crows of a rooster.

32. **seagull** the cries of a seagull.

33. **smoke alarm 1** steady whine.

34. **smoke alarm 2** staccato beeping.

35. **telephone dial** rotary phone's dialer turning.

36. **telephone ring** rings of a telephone bell.

37. **telephone tone** sound heard in phone earpiece that indicates "callee's" phone is ringing.

38. **triangle** tones on an orchestra triangle.

39. **truck motor** the sound of a truck engine idling.

40. **vending machine hum**

41. **viola**

42. **violin plain**

43. **violin vibrato** vibrato version of **violin plain**.

44. **wind** the "eerie wind" sound from Auditec tape.

To show the relative variety of acoustic behaviors represented by the C.2 SUT's library, we categorize the library's sources according to Table 1's entries. Note that the categories are not mutually exclusive, since sounds can exhibit more than one behavior over time.

chirp: door creak, hairdryer, owl, seagull

harmonic: alarm clock 1, alarm clock 2, bell1 toll, bell2 toll, bicycle bell, burglar alarm, car horn, car running chicken, chime, clock chimes, bugle, cuckoo clock, doorbell chimes, firengine bell, firehouse alarm, foghorn, gong, hairdryer, owl, oven buzzer, police car siren, rooster, razor, seagull, smoke alarm 1, smoke alarm 2, telephone dial, telephone ring, telephone tone, triangle, vending machine hum, viola, violin plain, violin vibrato

impulsive: bell1 toll, bell2 toll, burglar alarm, chime, clap, clock ticks, cuckoo clock, firehouse alarm, footsteps, glass clink, knock, telephone dial, triangle, pistol shot

repetitive: chicken, clock chimes, clock ticks, cuckoo clock, firehouse alarm, footsteps, owl, police car siren, rooster, seagull, smoke alarm 2, telephone ring, telephone tone, telephone dial

transients: bell1 toll, bell2 toll, chime, clap, clock chimes, door creak, gong, hairdryer, knock, razor, triangle

wide-band: clap, clock ticks, footsteps, knock, pistol shot, razor, truck motor, vending machine hum, wind

Figure 2 shows a histogram of the number of overlapping narrowband (e.g. $\leq$ 100-Hz wide) sound tracks in the library as an indication of the potential for interactions among sounds randomly selected from the library and placed in scenarios with random start times. Note that the higher the number of overlapping tracks there are in a spectral region, the greater the processing work that must be done to decide (1) whether in fact overlapping tracks are present in a scenario, and (2) which subset of the tracks that could be in the region of overlap are actually present.

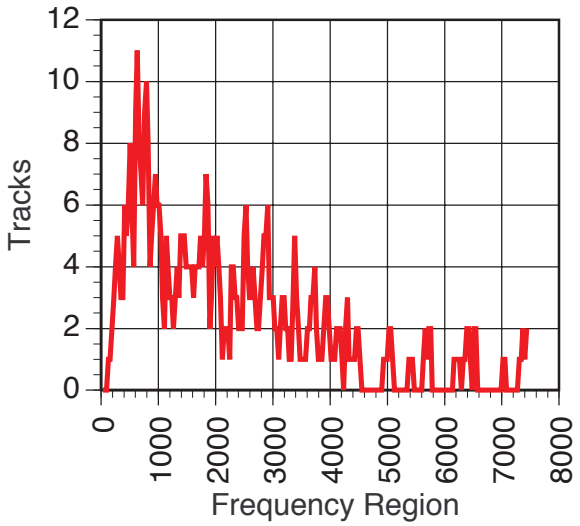

Figure 2: *Histogram showing the spectral distribution of the sources in the library used to evaluate IPUS-SUT. With the exception of three tracks (Track 5 and Track 6 of Firehouse Alarm and Track 1 of Alarm Clock 1), each sound's frequency tracks overlap at least one other sound's tracks. The average number of tracks per sound is 3.9, with the least number of tracks being 1 and the greatest number of tracks being 7.*

The sound-source information in the rest of this report is presented with the following conventions:

MICROSTREAM TIMES: Unless otherwise noted, a sound's microstreams (tracks) are assumed to start at the time the source starts. If tracks appear after the onset of a source, we note this with a "start" label that indicates how many seconds after the source begins does the track begin. Unless otherwise noted, tracks are assumed to last for the entire duration of the source. All time-lengths or time-points are in seconds.

ENTROPY VALUES: Entropies are represented by $\{\mu, \sigma\}$ pairs, where μ is the mean of all segment-entropy values, and σ is the standard deviation of all segment-entropy values. To aid in comparison, the entropy values have been normalized to a [0,100] range, with the telephone ring's 617 Hz track's maximum frequency entropy and the violin (vibrato)'s 887 Hz track's maximum amplitude entropy selected as the maxima.

MICROSTREAM FREQUENCIES: Track frequency ranges represent the total range of frequencies that the track will vary over from instance to instance; range bounds should **not** necessarily be taken as the actual width of the track. All tracks' frequencies were generated from a 2048-point short-time Fourier Transform with a 256-point decimation and a 1024-point rectangular analysis window.

MICROSTREAM RELATIVE ENERGIES: A track's relative energy is the range of ratios that its steady-state median energy can have with respect to the maximum-energy track in the sound. This information is provided only for those sounds whose ratio ranges are no wider than 0.5.

NOISEBED FEATURES: The grid-feature entries in the mean-value vectors and the covariance matrices are indexed left-to-right in the following order: (1) the spectral center

of gravity at the time the impulse is at its maximum energy, (2) the difference between the energy of the maximum-energy grid tile and that of the tile one frequency-slice *above* it at the same time, (3) the difference between the energy of the maximum-energy grid tile and that of the tile one frequency-slice *below* it at the same time, and (4) the least-squares exponential decay rate fitted to the energy values in the tiles in the same frequency-slice as the maximum-energy tile, but at time-slices after and including the maximum tile.

3.2 Library Sounds

3.2.1 Alarm Clock 1

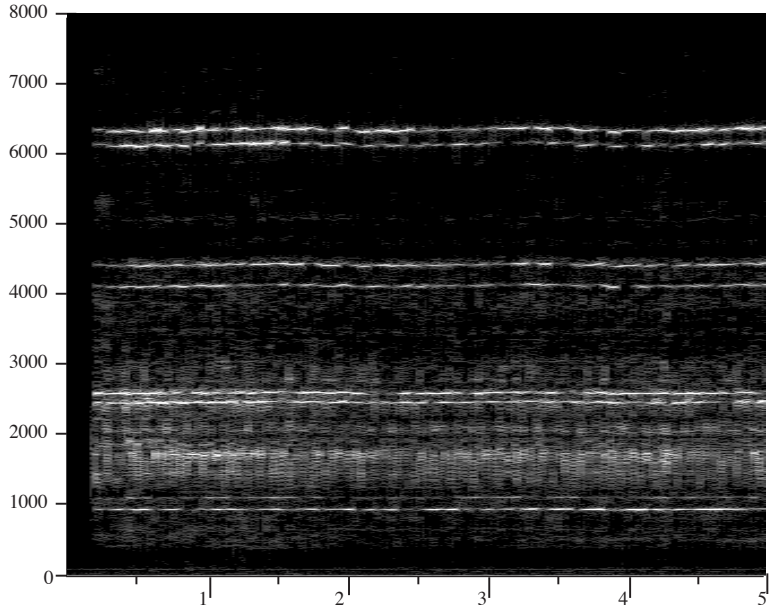

Figure 3: Lighter regions have higher energy.

track 1: 6297—6367 Hz, ampl. entropy: {64.9, 13.5}, freq. entropy: {4.7, 1.4}

track 2: 6090—6154 Hz, ampl. entropy: {78.1, 17.8}, freq. entropy: {4.4, 1.3}

track 3: 4398—4438 Hz, ampl. entropy: {68.3, 23.5}, freq. entropy: {4.5, 1.1}

track 4: 4093—4141 Hz, ampl. entropy: {65.9, 14.3}, freq. entropy: {4.6, 0.9}

track 5: 2586—2602 Hz, ampl. entropy: {81.1, 21.2}, freq. entropy: {4.8, 1.1}

track 6: 2453—2477 Hz, ampl. entropy: {58.8, 9.31}, freq. entropy: {5.7, 1.2}

track 7: 929— 953 Hz, ampl. entropy: {71.6, 11.9}, freq. entropy: {11.0, 2.1}

notes: Since this is an analog, bell+ringer clock there is a lot of variability in the strikes of the ringer on the bell. This leads to too much variability in track energy levels to permit a relative-energy characterization. Despite this observation, it should be noted that `Track 1` and `Track 2` are consistently the two highest-energy components. `Track 5` and `Track 6` are often not much higher in energy than the noisebed in that region. The sound's range of durations is arbitrarily set to [1.5, 10.0] seconds.

3.2.2 Alarm Clock 2

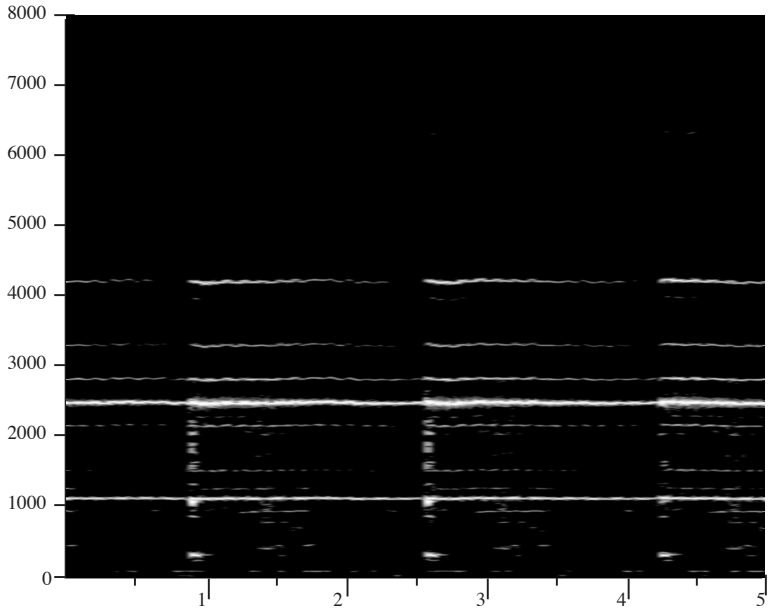

Figure 4: Lighter regions have higher energy.

track 1: 2450—2510 Hz, rel. energy = [1.00, 1.00]

track 2: 4190—4240 Hz, rel. energy = [0.22, 0.45]

track 3: 2800—2850 Hz, rel. energy = [0.18, 0.32]

track 4: 1110—1140 Hz, rel. energy = [0.09, 0.35]

track 5: 3300—3330 Hz, rel. energy = [0.13, 0.20]

track 6: 1510—1530 Hz, rel. energy = [0.01, 0.10]

track 7: 3950—3990 Hz, rel. energy = [0.06, 0.09]

notes: This is an electronic alarm clock with each "ring" being a 1.6-second 7-track stream. Energies are higher at the start of the stream, and drop 50% (linearly) by the end of the stream. Except for `Track 2`, higher relative energy ratios occur toward the end of the stream. The sound's range of durations is arbitrarily set to [3.2, 10] seconds. In the IPUS library this sound is represented as a script containing a sequence of rings abutting each other in time.

3.2.3 Bell-1 Toll

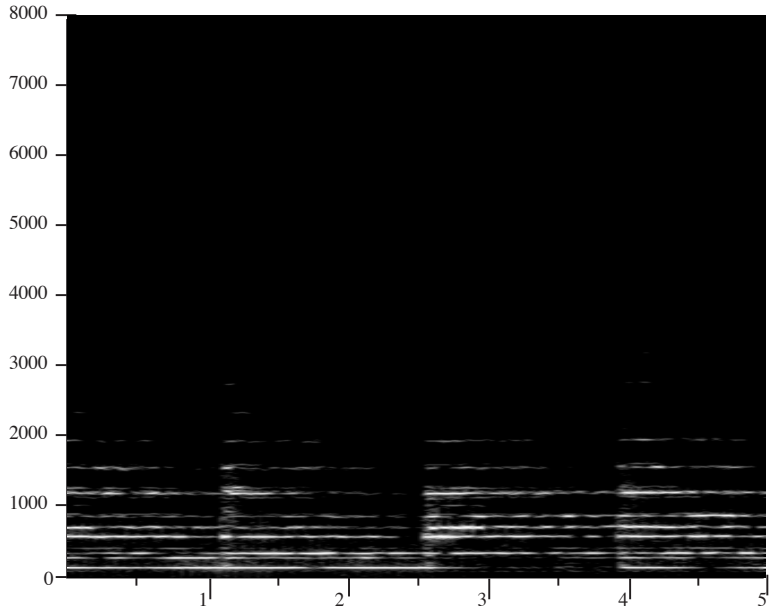

Figure 5: Lighter regions have higher energy.

track 1: 144— 150 Hz

track 2: 288— 295 Hz

track 3: 578— 585 Hz

track 4: 725— 732 Hz

track 5: 870— 877 Hz

track 6: 1160—1170 Hz

track 7: 1450—1460 Hz

track 8: 1935—1945 Hz

track 9: 2320—2335 Hz

notes: Each toll stream is a harmonic set with $f_0 = 48.5$ Hz. The tracks listed are the most prominent harmonics. In order of increasing track number, the harmonics represented here are 3, 6, 12, 15, 18, 24, 30, 40, and 48. Each toll is 1.375 seconds long, measured from strike to strike. There was too much energy variation among the tracks to determine meaningful relative energies.

3.2.4 Bell-2 Toll

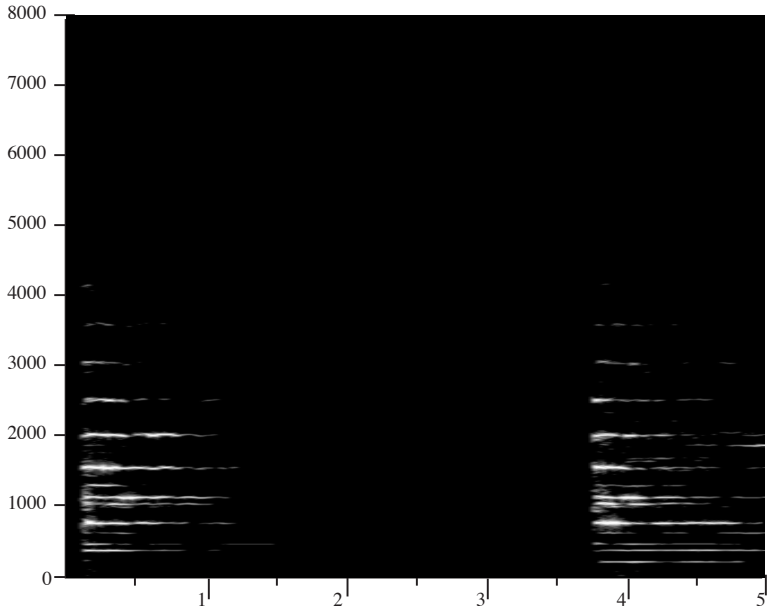

Figure 6: Lighter regions have higher energy.

track 1: 1555—1575 Hz, rel. energy = [1.00 1.00], start = 0.0, end = 0.7

track 2: 1120—1150 Hz, rel. energy = [0.16 5.00], start = 0.0, end = 0.8

track 3: 765— 790 Hz, rel. energy = [0.05 0.40], start = 0.0, end = 0.35

track 4: 2010—2040 Hz, rel. energy = [0.05 0.15], start = 0.0, end = 0.7

track 5: 1555—1575 Hz, rel. energy = [0.05 0.15], start = 0.1, end = 0.25

notes: Each toll stream is a harmonic set with $f_0 \approx 112$ Hz. The tracks listed are the most prominent harmonics. In order of decreasing energy, the harmonics represented here are 14, 10, 7, 18, and 22. Each toll is [0.8 1.0] seconds long, measured from strike time to when signal energy decays to the background energy level just prior to the strike. The *start* and *end* times for each track are relative to the start of a bell toll. Most of the time `Track 1` has the maximum energy of all tracks; however, around 0.2 seconds after the start of a toll `Track 2`'s energy grows to 5 times tk1's energy, and then decays similarly to all other tracks.

3.2.5 Bicycle Bell

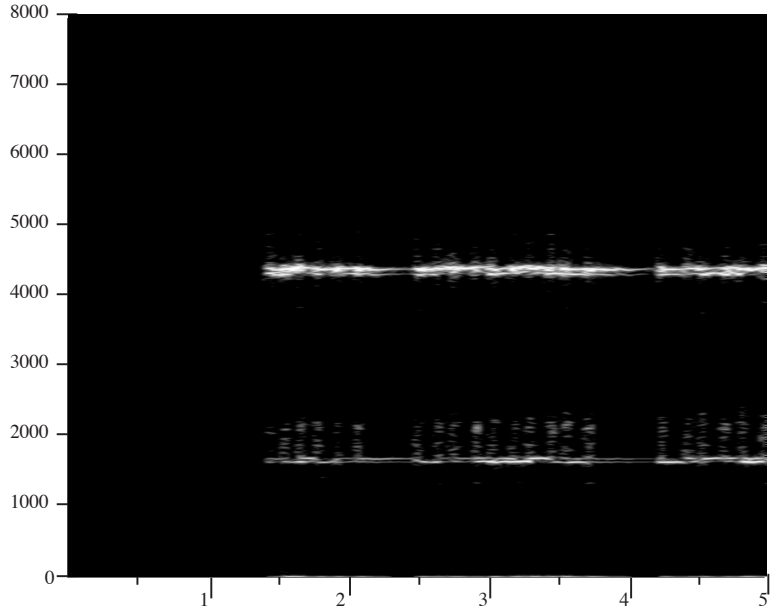

Figure 7: Lighter regions have higher energy.

track 1: 4310—4390 Hz, rel. energy = [1.00, 1.00]

track 2: 1640—1690 Hz, rel. energy = [0.06, 0.16], [0.32, 0.43]*

notes: There are two phases in this source. The first phase, called the "active" phase, covers the period over which the bell is being struck, and can last an indeterminate period of time. From wideband analysis, each strike (vertical bar in the spectrogram) is 0.08 second long and is separated from the next strike by a mnimum of 0.05 seconds. The second phase in the source is the "decay" phase, where the reverberations of the bell exponentially decay to background energy levels. This phase's length depends on the intensity of the last strike, but in the collected data this length has been found to be approximately 1.0 second.

Occasionally the two tracks each can be further resolved into two subtracks, giving a total of 4 tracks.

*The second energy range shows the rel. energy between the higher peak of each pair when the tracks split.

3.2.6 Bugle

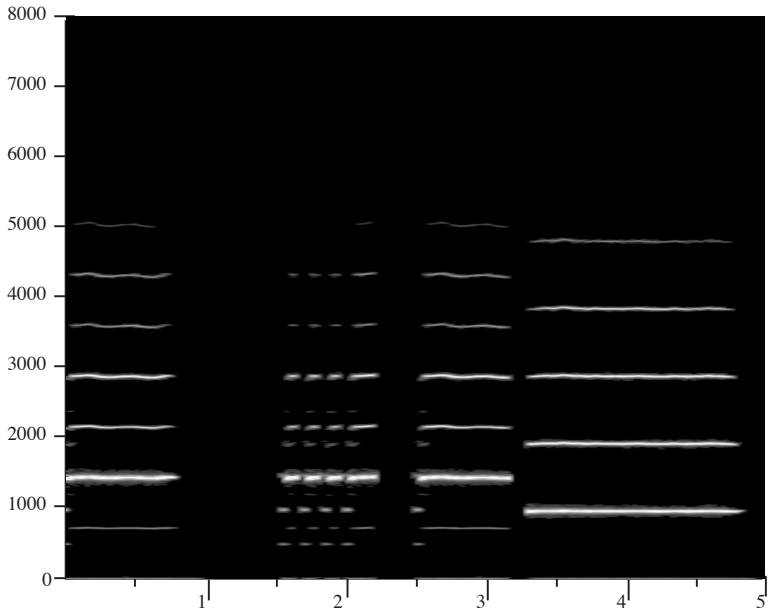

Figure 8: Lighter regions have higher energy.

Stream1 (time 0.3–1.3, 2.5–3.15)

track 1: 1430—1446 Hz, ampl. entropy: {18.1, 7.4}, freq. entropy: {6.5, 0.5}

track 2: 2148—2164 Hz, ampl. entropy: {37.4, 13.7}, freq. entropy: {5.2, 0.6}

track 3: 2859—2875 Hz, ampl. entropy: {82.2, 21.7}, freq. entropy: {3.1, 0.2}

Stream2 (time 3.3–4.9)

track 1: 929— 977 Hz, ampl. entropy: {10.8, 2.2}, freq. entropy: {3.4, 3.2}

track 2: 1906—1930 Hz, ampl. entropy: {10.6, 2.6}, freq. entropy: {4.8, 0.3}

track 3: 2867—2883 Hz, ampl. entropy: {32.1, 23.6}, freq. entropy: {1.6, 0.8}

track 4: 3828—3844 Hz, ampl. entropy: {32.0, 21.4}, freq. entropy: {2.0, 0.6}

notes: These figures are for the highest-energy tracks in the two longest notes in the sequence pictured above. Other tracks shown in the figure have only 1% of the energy of those for which data is given.

3.2.7 Burglar Alarm

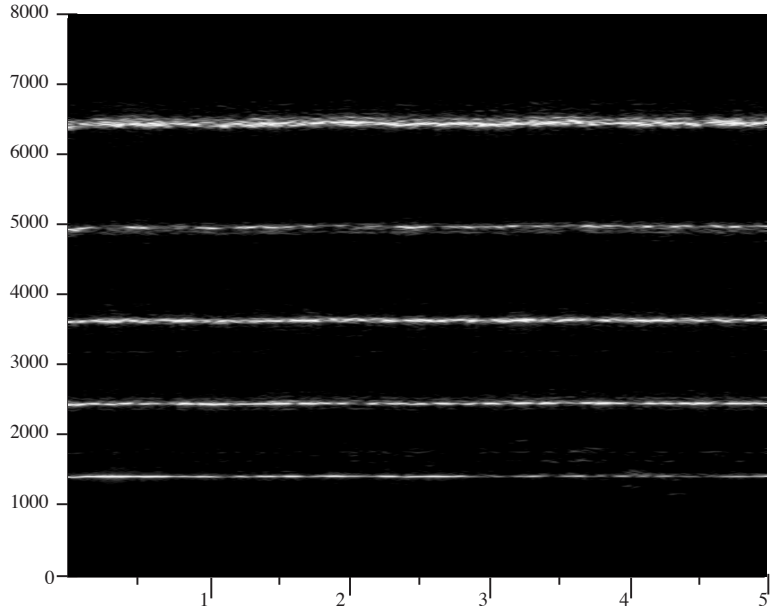

Figure 9: Lighter regions have higher energy.

track 1: 6413—6468 Hz, ampl. entropy: {49.9, 9.4}, freq. entropy: {4.8, 0.9}

track 2: 4875—5023 Hz, ampl. entropy: {74.8, 15.0}, freq. entropy: {7.1, 2.3 }

track 3: 3624—3665 Hz, ampl. entropy: {67.6, 12.1}, freq. entropy: {5.4, 1.1}

track 4: 2445—2485 Hz, ampl. entropy: {80.4, 12.8}, freq. entropy: {5.2, 1.2}

track 5: 1414—1446 Hz, ampl. entropy: {69.3, 15.6}, freq. entropy: {6.3, 3.0}

notes: `Track 1` is often the highest-energy track. At many time-points in its frequency region the track often contains a high central peak and one or two peaks on either side with [10%, 90%] of the central peak's energy. `Track 5` is often the lowest-energy track. The tracks' energies fluctuated too widely for any meaningful relative energy ratios to be determined.

3.2.8 Car Running

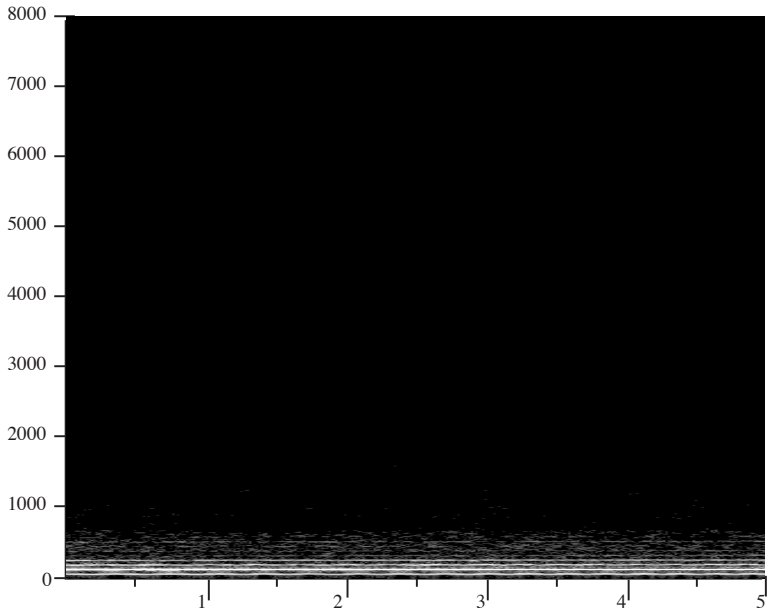

Figure 10: Lighter regions have higher energy.

track 1: 250— 281 Hz, ampl. entropy: {33.1, 5.2}, freq. entropy: {30.9, 7.18}

track 2: 188— 219 Hz, ampl. entropy: {31.5, 4.4}, freq. entropy: {45.0, 10.4}

notes: `Track 1` is often the higher-energy track. However, the tracks' energies fluctuated too widely for any meaningful relative energy ratios to be determined.

3.2.9 Car Horn

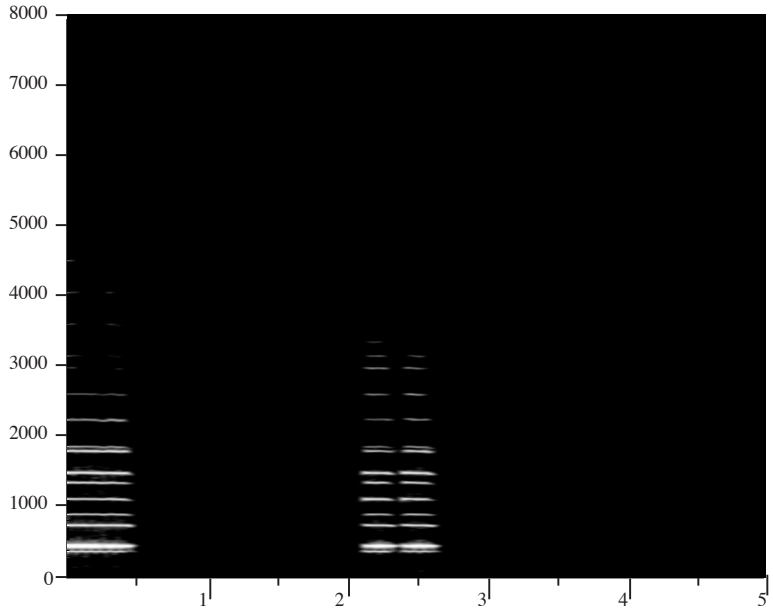

Figure 11: Lighter regions have higher energy.

track 1: 445— 460 Hz (f_0^1), rel. energy = [1.00, 1.00]

track 2: 1780—1805 Hz ($4f_0^1$), rel. energy = [0.10, 0.35]

track 3: 735— 750 Hz ($2f_0^2$), rel. energy = [0.02, 0.10]

track 4: 1100—1125 Hz ($3f_0^2$), rel. energy = [0.02, 0.10]

track 5: 360— 382 Hz (f_0^2), rel. energy = [0.08, 0.25]

track 6: 1470—1500 Hz ($4f_0^2$), rel. energy = [0.08, 0.16]

notes: The observed beep durations fall in the range [0.2, 0.50] seconds. The source has approximately 20 significant tracks that come from 2 harmonic sets. The first set has $f_0^1 \approx 450$ Hz, and contributes its first and fourth harmonics to the source. The second set has $f_0^2 \approx 370$ Hz, and contributes its first 4 harmonics to the source.

3.2.10 Chicken

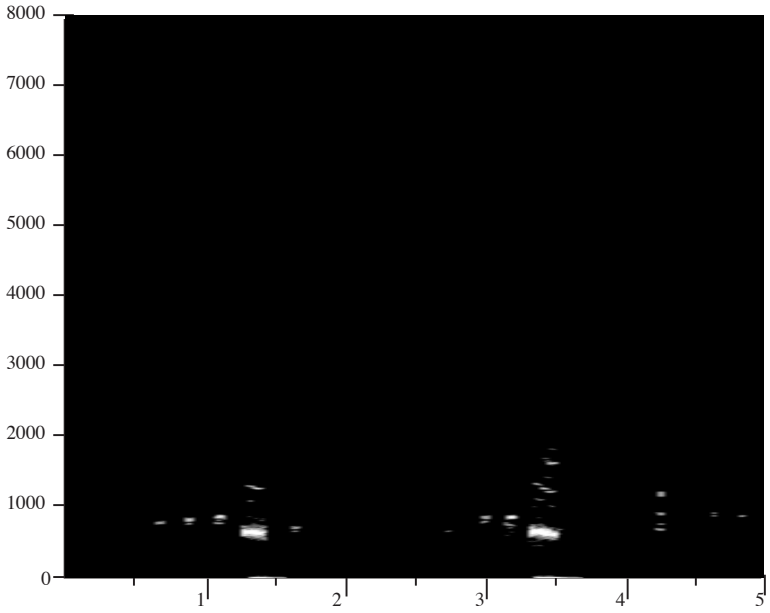

Figure 12: Lighter regions have higher energy.

track 1: a chirp from 670 to 610 Hz, rel. energy = [1.00, 1.00]

track 2: a chirp from 1350 to 1220 Hz, rel rel. energy = [0.10, 0.30]

notes: The three chicken clucks used to generate this model have durations lying in the range [0.13, 0.20] seconds.

3.2.11 Chime

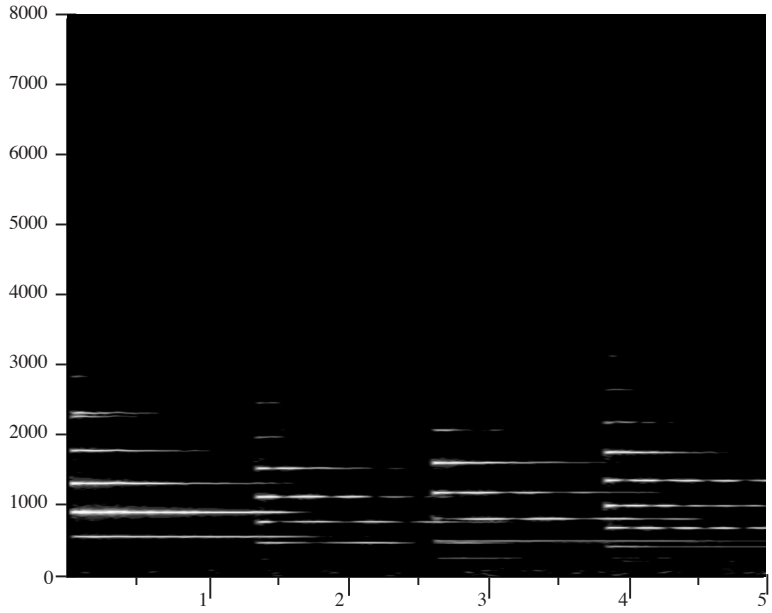

Figure 13: Lighter regions have higher energy.

track 1.1: 560— 590 Hz, rel. energy = [0.01, 0.10], duration = [1.4, 1.6] seconds

track 1.2: 900— 935 Hz, rel. energy = [1.00, 1.00], duration = [1.4, 1.6] seconds

track 1.3: 1300—1340 Hz, rel. energy = [0.05, 0.50], duration = [0.9, 1.0] seconds

track 1.4: 1780—1820 Hz, rel. energy = [0.01, 0.10], duration = [0.5, 0.6] seconds

track 1.5: 2275—2324 Hz, rel. energy = [0.01, 0.10], duration = [0.2, 0.25] seconds

track 2.1: 460— 500 Hz

track 2.2: 760— 795 Hz

track 2.3: 1115—1155 Hz

track 2.4: 1520—1565 Hz

track 3.1: 490— 535 Hz

track 3.2: 800— 840 Hz

track 3.3: 1175—1210 Hz

track 3.4: 1605—1645 Hz

track 3.5: 2065—2100 Hz

track 4.1: 410— 440 Hz

track 4.2: 675— 720 Hz

track 4.3: 1000—1040 Hz

track 4.4: 1350—1395 Hz

track 4.5: 1740—1785 Hz

track 4.6: 2155—2200 Hz

notes: There are 4 chime notes in the above spectrogram of an orchestra chime. Each has between 4 and 6 tracks. Each chime note is the result of a single strike followed by [1.2 2.5] seconds of reverberation. It was possible to determine relative energy ratios only for tracks in the first chime; the others had unpredictable energy variations. A rough estimate of the fundamental for each chime's harmonic set is:

- chime1—$f_0 = 51$Hz
- chime2—$f_0 = 95$Hz
- chime3—$f_0 = 64$Hz
- chime4—$f_0 = 85$Hz

3.2.12 Clap

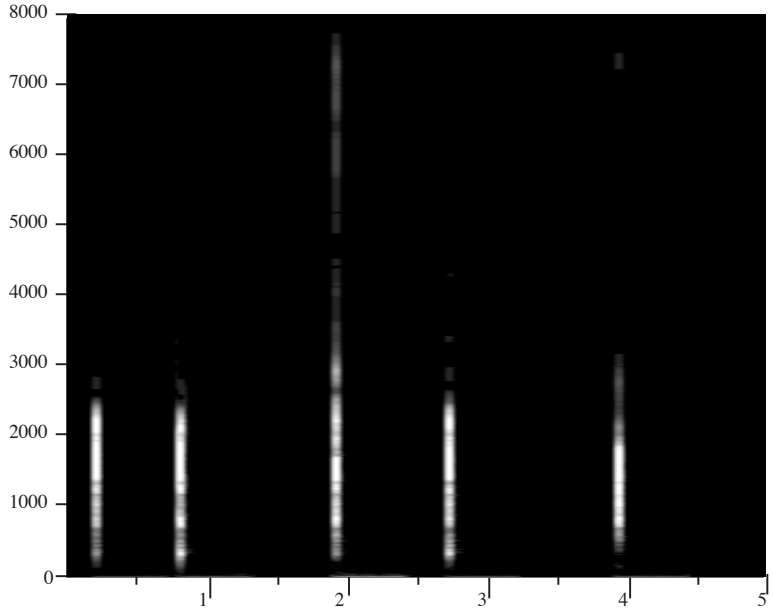

Figure 14: Lighter regions have higher energy.

noisebed 1: mean: [2.1933, 0.4440, 0.2067, 0.2268]

cov:
$$\begin{bmatrix} 7.5131s-2 & -3.4630s-3 & -1.0287s-2 & -6.3474s-3 \\ -3.4630s-3 & 8.9997s-4 & 4.2013s-4 & 3.6960s-4 \\ -1.0287s-2 & 4.2013s-4 & 4.5016s-3 & 2.7282s-3 \\ -6.3474s-3 & 3.6960s-4 & 2.7282s-3 & 2.0249s-3 \end{bmatrix}$$

notes: The feature values are listed in the order they are described at the beginning of this section. Ten isolated claps were used to generate these feature values.

3.2.13 Clock Chime

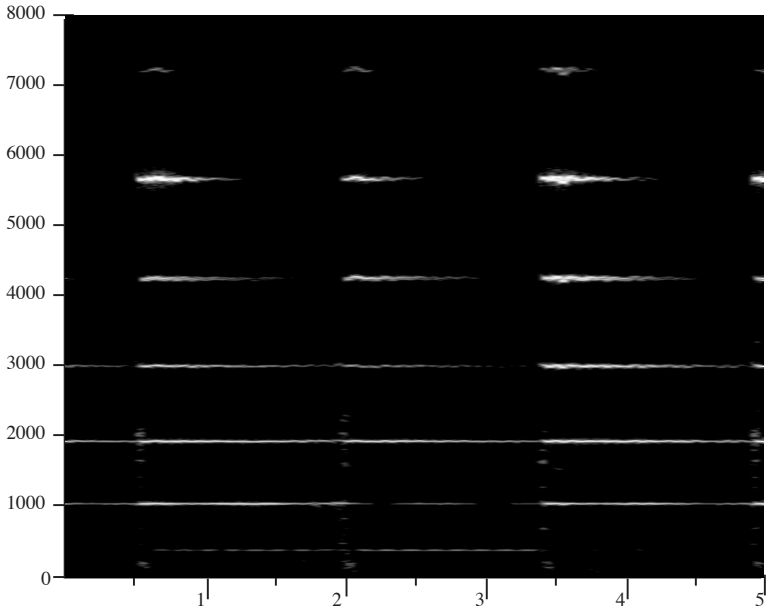

Figure 15: Lighter regions have higher energy.

track 1: 5650—5685 Hz, rel. energy = [1.00, 1.00], duration = [0.5, 0.55]

track 2: 1920—1950 Hz, rel. energy = [0.03, 0.25], duration = [1.9, 2.0]

track 3: 1040—1065 Hz, rel. energy = [0.03, 0.25], duration = [1.4, 1.5]

track 4: 2290—3025 Hz, rel. energy = [0.05, 0.25], duration = [0.9, 1.0]

track 5: 4240—4280 Hz, rel. energy = [0.08, 0.25], duration = [0.4, 0.5]

notes: Each isolated chime's reverberations last approximately 2.0 seconds, with `Track 2` lasting the longest of the harmonic set. The track with greatest energy is `Track 1`; however, the other tracks last longer and have slight rises in energy as `Track 1` decays.

3.2.14 Clock Ticks

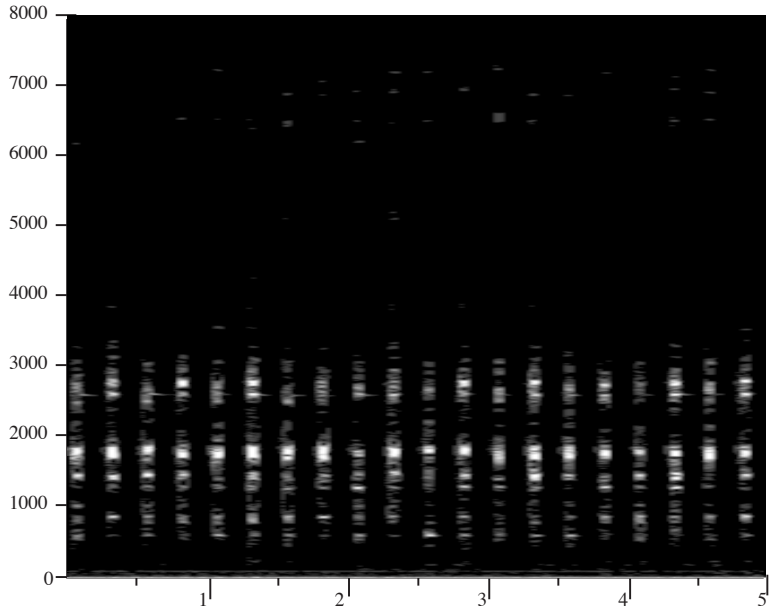

Figure 16: Lighter regions have higher energy.

track 1: 1730—1770 Hz

noisebed 1: mean: [2.3224, 0.3689, 0.1056, 0.1299]

cov:
$$\begin{bmatrix} 3.3037s-2 & -1.5382s-3 & -4.7619s-3 & -2.5159s-3 \\ -1.5382s-3 & 2.1002s-3 & -3.6221s-5 & 2.5113s-4 \\ -4.7619s-3 & -3.6221s-5 & 3.0816s-3 & 2.8253s-3 \\ -2.5159s-3 & 2.5113s-4 & 2.8253s-3 & 3.1981s-3 \end{bmatrix}$$

notes: Each tick is 0.1 seconds long (practically no variation), and there is 0.15 seconds between ticks. There is a significant noisebed throughout the narrow time-slice that each tick lasts. However, over all examined ticks, there was a consistent peak at the "track" frequencies noted above. Twenty clock ticks were used to generate the noisebed feature values.

3.2.15 Cuckoo Clock (Cuckoo + Hour Chime)

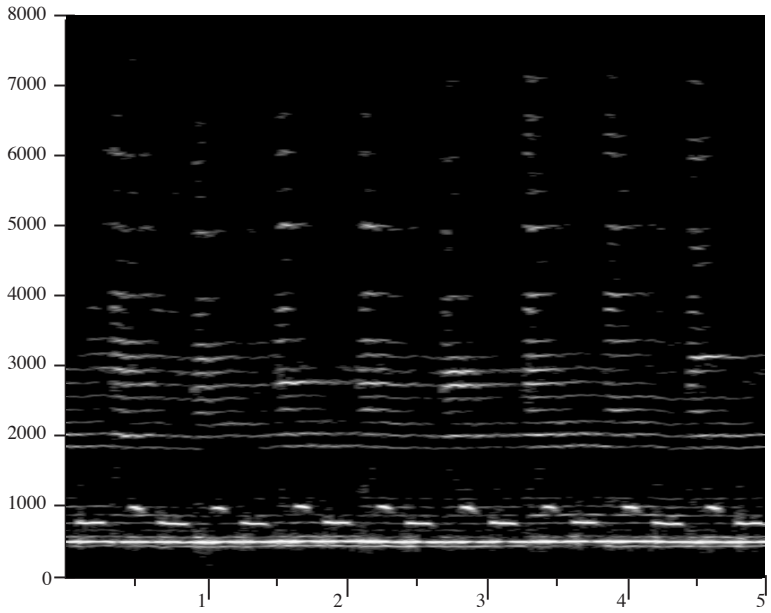

Figure 17: Lighter regions have higher energy.

track 1: 960—1020 Hz, rel. energy = [0.05, 0.35]

track 2: 750— 800 Hz, rel. energy = [0.25, 0.75]

track 3: 495— 525 Hz, rel. energy = [1.00, 1.00]

track 4: 2000—2050 Hz, rel. energy = [0.02, 0.10]

track 5: 2740—2790 Hz, rel. energy = [0.01, 0.15]

notes: This sound is actually the result of two simultaneous sources. `Track 1` and `Track 2` are produced by the cuckoo-call, while `Track 3`, `Track 4`, and `Track 5` are produced by a clock-chime. Each "cuckoo-chime" combination lasts [0.56, 0.63] seconds (i.e. time between successive strikes to the clock-chime). The final chime's reverberations last 0.3 seconds beyond the 0.6-second chime-length. This model was developed from 12 "cuckoo-chime" instances.

3.2.16 Doorbell Chimes

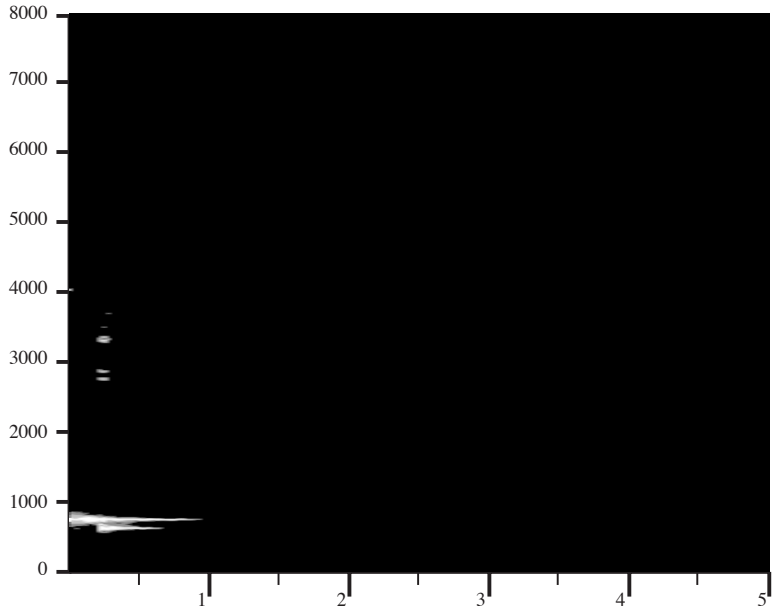

Figure 18: Lighter regions have higher energy.

track 1: 740— 760 Hz, rel. energy = [1.00, 1.00]

track 2: 615— 635 Hz, rel. energy = [0.20, 0.90], start time:[0.20, 0.25]

notes: The short "tracks" above `Track 1`'s frequency have extremely low energy and appear prominent only because of the image-enhancement process used to generate the spectrogram shown here. They are not included in the sound model.

3.2.17 Door Creak

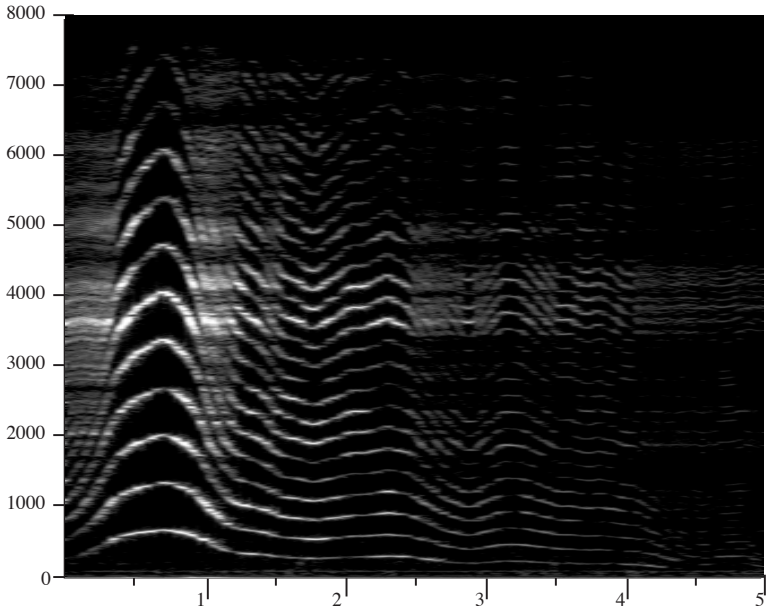

Figure 19: Lighter regions have higher energy.

track 1: 300— 680— 330 Hz, rel. energy = [0.80,1.20]

track 2: 625—1350— 670 Hz, rel. energy = [0.80,1.20]

track 3: 870—2030—1000 Hz, rel. energy = [0.80,1.20]

track 4: 1100—2600—1400 Hz, rel. energy = [0.80,1.20]

track 5: 1400—3375—1800 Hz, rel. energy = [0.80,1.20]

track 6: 1600—4050—2000 Hz, rel. energy = [1.00,1.00]

notes: All attack lengths fall in the range [0.7,0.8] seconds, while all decay lengths fall into the range [0.5,0.6] seconds. The frequencies shown are derived from the "knee" in the first 1.5 seconds of spectrogram data.

3.2.18 Firengine Bell

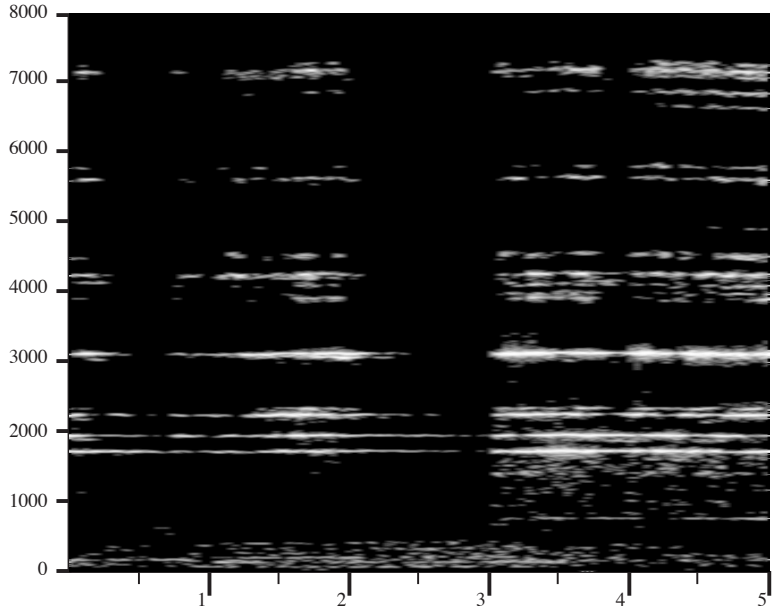

Figure 20: Lighter regions have higher energy.

track 1: 3070—3148 Hz, ampl. entropy: {80.1, 22.9}, freq. entropy: {5.0, 1.5}

track 2: 2242—2266 Hz, ampl. entropy: {80.7, 16.3}, freq. entropy: {6.4, 0.6}

track 3: 1945—1970 Hz, ampl. entropy: {84.9, 25.6}, freq. entropy: {5.7, 0.8}

track 4: 1725—1750 Hz, ampl. entropy: {72.2, 23.1}, freq. entropy: {6.6, 0.8}

track 5: 3304—3327 Hz, ampl. entropy: {92.0, 14.4}, freq. entropy: {6.5, 0.6}

track 6: 4211—4234 Hz, ampl. entropy: {84.9, 18.3}, freq. entropy: {4.9, 0.5}

track 7: 2820—2843 Hz, ampl. entropy: {74.0, 13.5}, freq. entropy: {8.0, 0.5}

notes: This sound is produced by a bell being struck. `Track 1` is the most energetic track. There is a large amount of variability in relative track energies; the track ordering is a rough indication of track ranking by average energy. The source has no natural duration bounds. For the purposes of this database the bounds on active duration are set to be [1.5, 6.0] seconds. The reverberation decay ranges from 0.5 to 1.0 seconds.

3.2.19 Firehouse Alarm

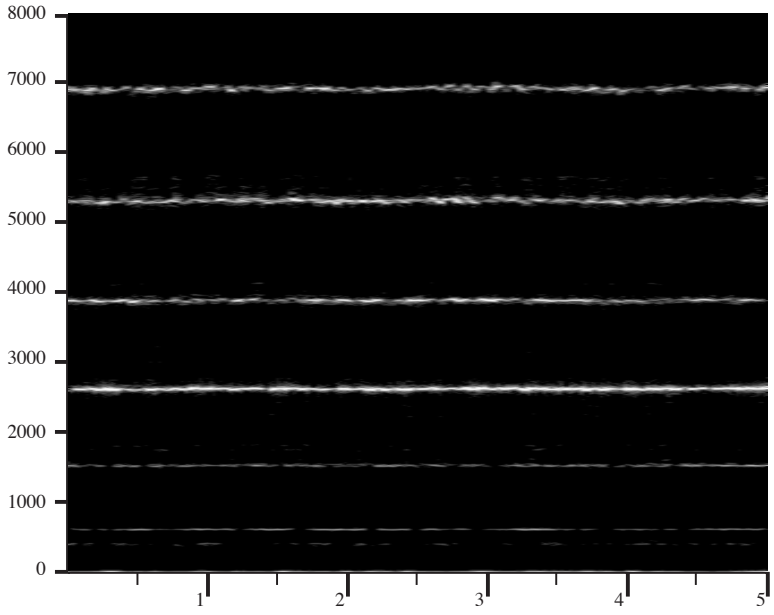

Figure 21: Lighter regions have higher energy.

track 1: 600— 616 Hz, ampl. entropy: {78.4, 12.5}, freq. entropy: {15.8, 2.8}

track 2: 1515—1539 Hz, ampl. entropy: {65.6, 12.5}, freq. entropy: {11.7, 2.4}

track 3: 2617—2633 Hz, ampl. entropy: {75.7, 15.9}, freq. entropy: {5.1, 0.8}

track 4: 3883—3906 Hz, ampl. entropy: {78.7, 17.8}, freq. entropy: {4.5, 0.9}

track 5: 5312—5342 Hz, ampl. entropy: {65.6, 13.3}, freq. entropy: {4.8, 1.0}

track 6: 6914—6954 Hz, ampl. entropy: {75.2, 23.3}, freq. entropy: {4.0, 1.1}

notes: The source's tracks have too much variability in energy to determine useful rel. energy ratios. `Track 1` generally is the lowest-energy track. The tracks form the following harmonics from a harmonic set with f_0 =[186, 188] Hz: 3, 8, 14, 21, 28, and 37. The sound's range of durations is nominally set to [3.0, 10.0] seconds.

3.2.20 Foghorn

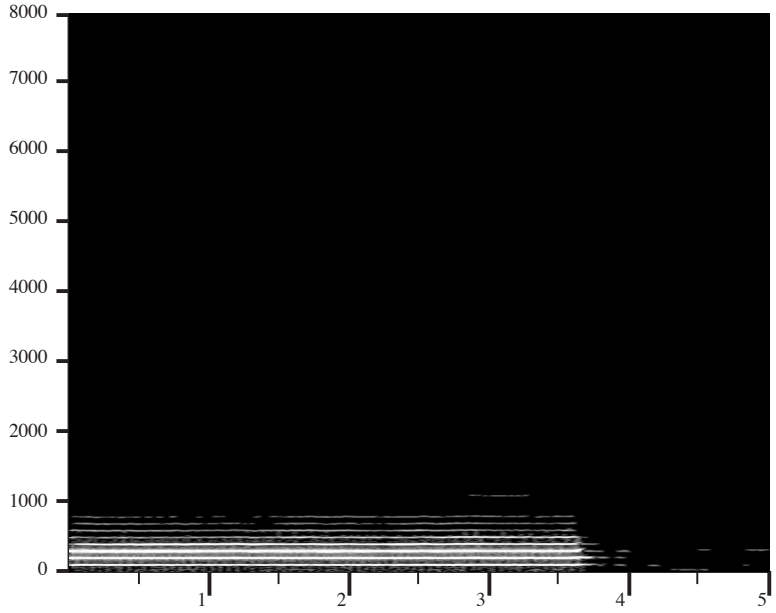

Figure 22: Lighter regions have higher energy.

track 1: 257— 336 Hz, rel. energy = [1.00, 1.00], ampl. entropy: {12.4, 6.7}, freq. entropy: {9.7, 13.2}

track 2: 171— 235 Hz, rel. energy = [0.50, 0.60], ampl. entropy: {13.6, 7.5}, freq. entropy: {5.3, 12}

track 3: 492— 516 Hz, rel. energy = [0.10, 0.30], ampl. entropy: {32.1, 9.4}, freq. entropy: {14.3, 5.5}

track 4: 593— 610 Hz, rel. energy = [0.20, 0.40], ampl. entropy: {32.8, 8.1}, freq. entropy: {14.9, 2.4}

track 5: 687— 711 Hz, rel. energy = [0.03, 1.00], ampl. entropy: {46.2, 24.7}, freq. entropy: {13.8, 1.6}

track 6: 789— 805 Hz, rel. energy = [0.03, 1.00], ampl. entropy: {32.9, 6.2}, freq. entropy: {11.9, 2.0}

notes: Each horn blast lasts 3.8 seconds. A horn blast is a harmonic stream with $f_0 = 97$ Hz. The energy ratios are quite stable; there is a slow rise in energy in `Track 5` and `Track 6` as the horn finishes sounding, which accounts for their wide energy ranges. The tracks indicated here represent the following harmonics: 2, 3, 5, 6, 7, and 8.

3.2.21 Glass Clink

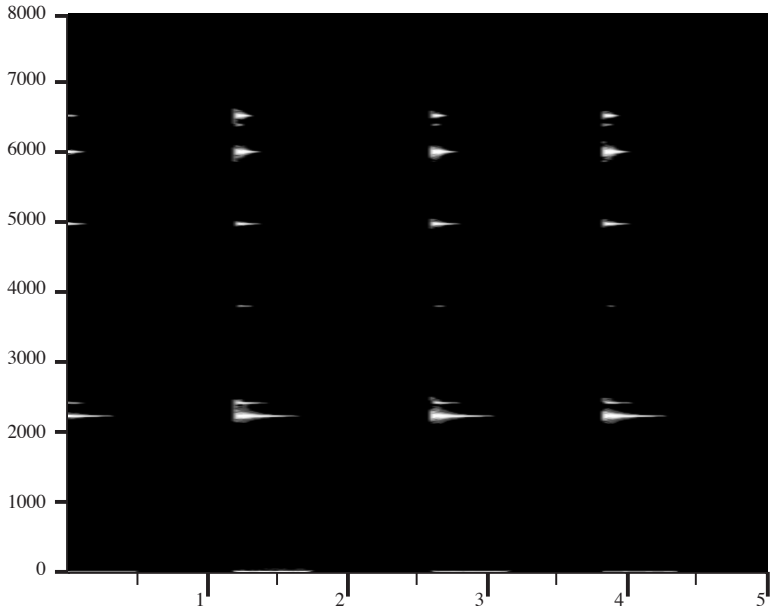

Figure 23: Lighter regions have higher energy.

track 1: 2225—2243 Hz, rel. energy = [1.0, 1.0]

track 2: 4960—4990 Hz, rel. energy = [0.5, 1.0]

noisebed 1: **mean:** [4.9140, 0.06986, 0.03265, 0.02894]

cov:
$$\begin{bmatrix} 1.7730\times10^{-1} & 9.7419\times10^{-3} & 4.6897\times10^{-4} & -1.2236\times10^{-3} \\ 9.7419\times10^{-3} & 7.9156\times10^{-4} & -2.0913\times10^{-5} & -1.1002\times10^{-4} \\ 4.6897\times10^{-4} & -2.0913\times10^{-5} & 3.1226\times10^{-5} & 1.8969\times10^{-5} \\ -1.2236\times10^{-3} & -1.1002\times10^{-4} & 1.8969\times10^{-5} & 3.0866\times10^{-5} \end{bmatrix}$$

notes: Each glass-clink is [0.15, 0.2] seconds in duration. Ten isolated glass clinks were used to generate the feature values. The spectrogram tracks not listed had low and highly variable energies.

3.2.22 Footsteps

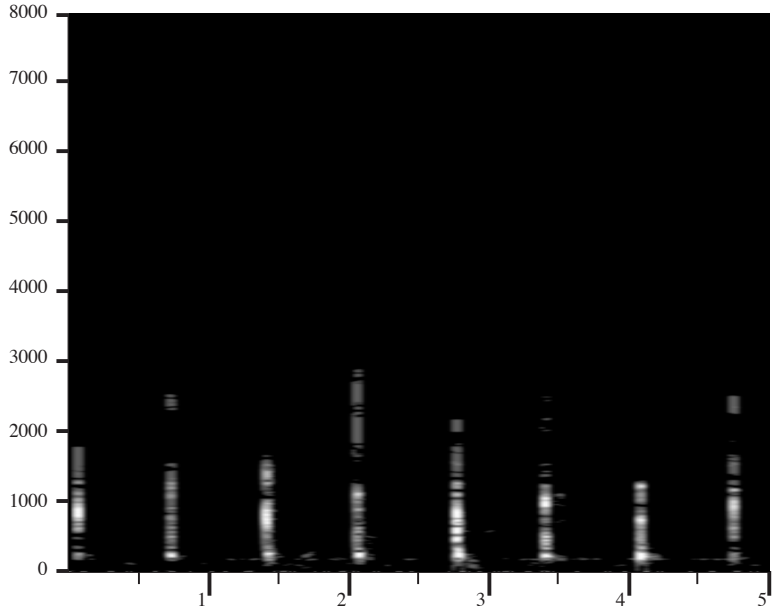

Figure 24: Lighter regions have higher energy.

noisebed 1: mean: [1.2753, 0.2762, 0.1619, 0.2129]

cov:
$$\begin{bmatrix} 4.5466\times10^{-2} & -6.0995\times10^{-3} & -1.0380\times10^{-2} & -3.3756\times10^{-3} \\ -6.0995\times10^{-3} & 5.1067\times10^{-3} & 2.6788\times10^{-3} & 1.0639\times10^{-3} \\ -1.0380\times10^{-2} & 2.6788\times10^{-3} & 6.1067\times10^{-3} & 4.5021\times10^{-3} \\ -3.3756\times10^{-3} & 1.0639\times10^{-3} & 4.5021\times10^{-3} & 6.1556\times10^{-3} \end{bmatrix}$$

notes: Each footstep is at most 0.3 seconds in duration. In the IPUS sound library we use the source model "footfall" to represent a single footstep. The library uses the acoustic script "footsteps" to represent the concept of a sequence of footfalls. Eight footfalls were used to generate the noisebed feature values.

3.2.23 Gong

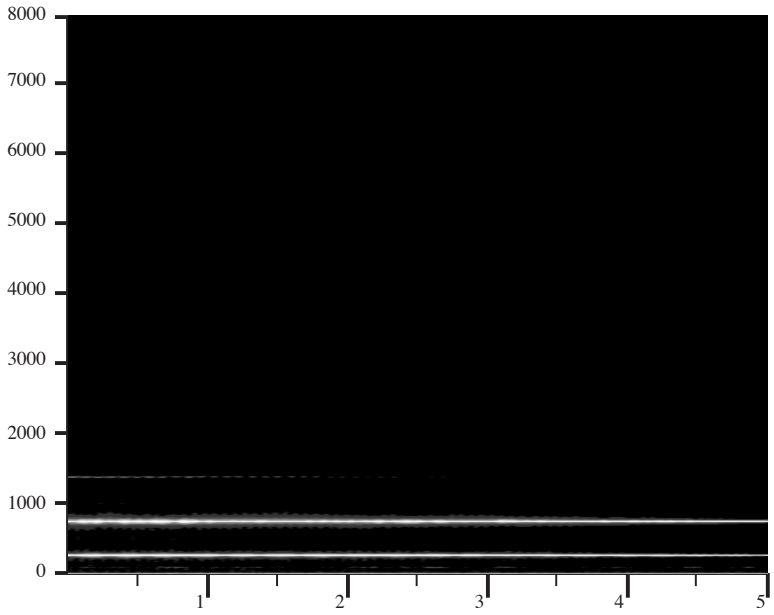

Figure 25: Lighter regions have higher energy.

track 1: 743— 757 Hz, rel. energy = [1.00, 1.00]

track 2: 258— 273 Hz, rel. energy = [0.30, 0.40]

track 3: 1375—1390 Hz, rel. energy = [0.005, 0.01], duration=[1.2, 1.4] seconds

notes: The majority of the sound's spectral energy is in the first two tracks listed. However, `Track 3` is included in the model because its energy is strong enough to interfere with the energy measurement of other sounds' tracks that overlap its frequency region. The sound's duration range is [6.3, 7.1] seconds.

3.2.24 Hairdryer

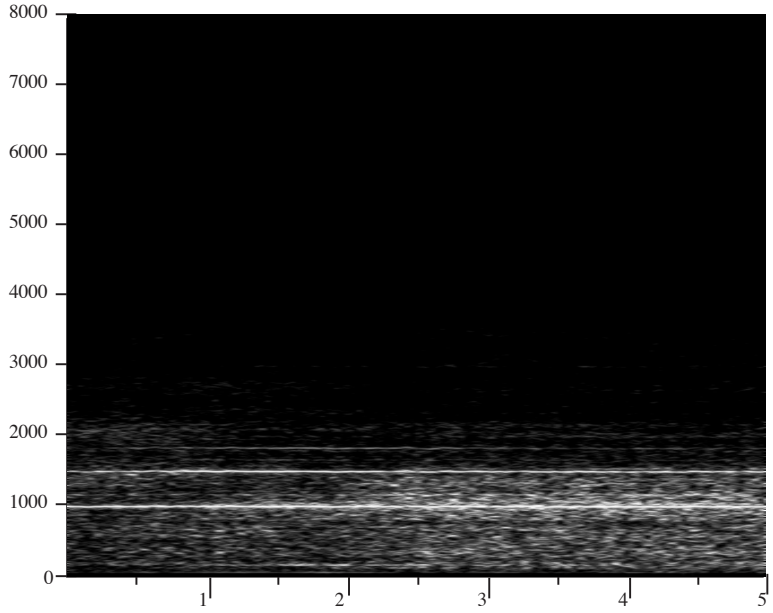

Figure 26: Lighter regions have higher energy.

track 1: 980—1015 Hz, rel. energy = [1.00, 1.00]

track 2: 1475—1510 Hz, rel. energy = [0.33, 0.54]

track 3: 1810—1840 Hz, rel. energy = [0.07, 0.18]

notes: This is the steady-state behavior of a hairdryer. In the IPUS database this sound's *overall* range of durations (encompassing attack, steady, and decay behaviors) is set to [3.0, 10.0] seconds.

3.2.25 Hairdryer Off

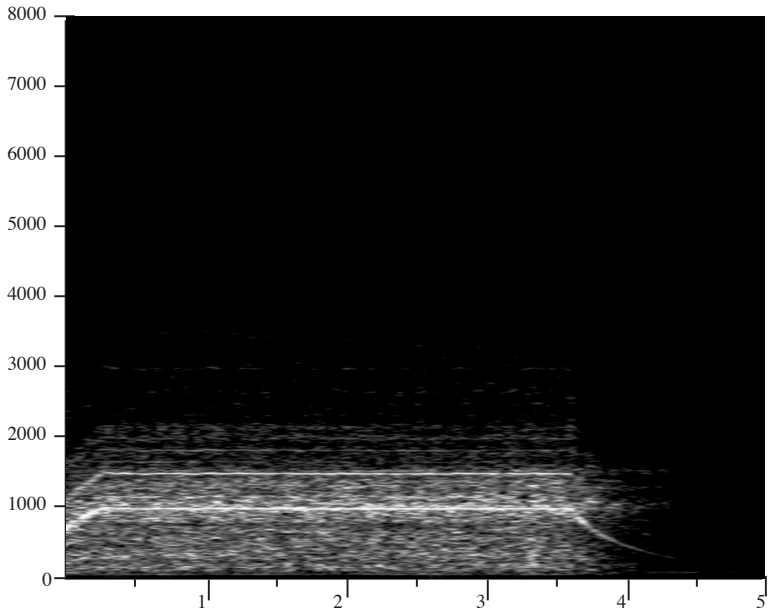

Figure 27: Lighter regions have higher energy.

track 1: A decaying exponential chirp from 1015 to 210 Hz.

notes: This is the transient behavior of a hairdryer as it is turned off. The chirp lasts approximately 0.5 seconds.

3.2.26 Hairdryer On

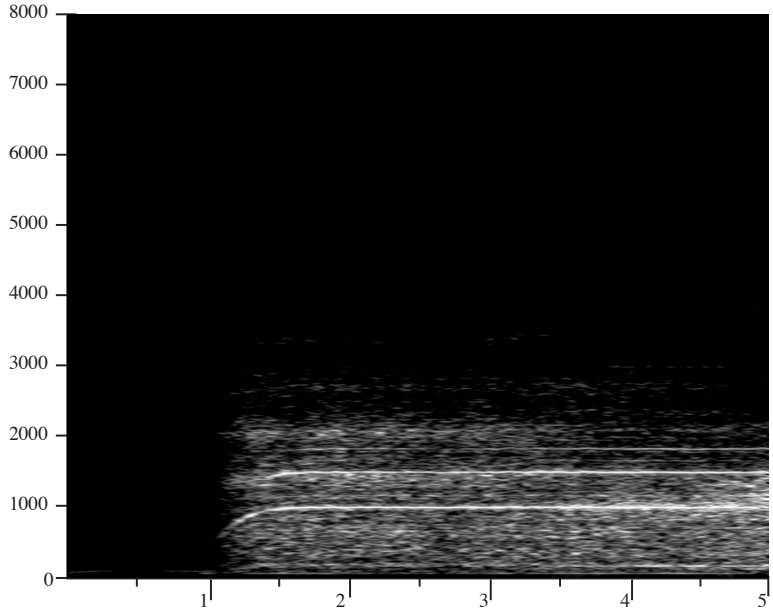

Figure 28: Lighter regions have higher energy.

track 1: A chirp from 210 to 1015 Hz.

notes: This is the transient behavior of a hairdryer as it is turned on. The chirp lasts approximately 0.5 seconds.

3.2.27 Knock

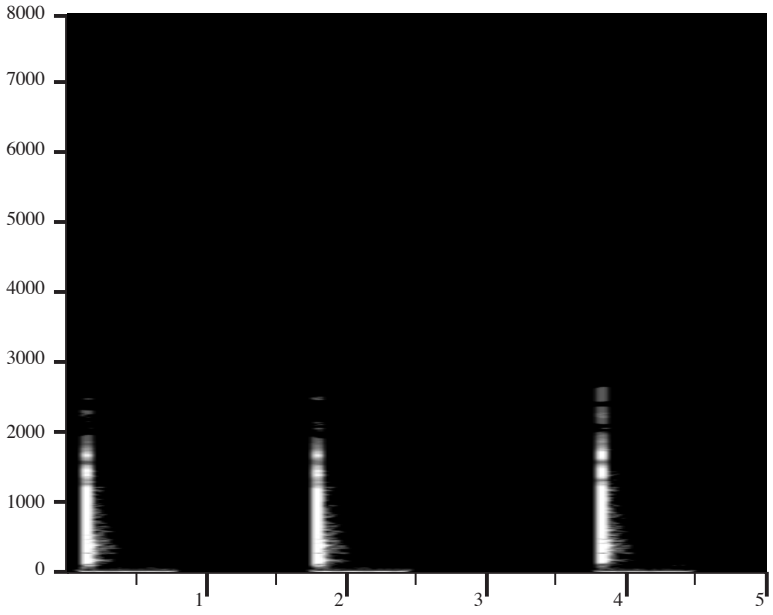

Figure 29: Lighter regions have higher energy.

noisebed 1: mean: [1.2257, 0.2551, 0.2364, 0.3248]

cov:
$$\begin{bmatrix} 1.1641\times10^{-3} & -5.0455\times10^{-5} & -1.0276\times10^{-3} & -6.2427\times10^{-4} \\ -5.0455\times10^{-5} & 2.8732\times10^{-4} & 5.8748\times10^{-6} & -4.4741\times10^{-5} \\ -1.0276\times10^{-3} & 5.8748\times10^{-6} & 9.3234\times10^{-4} & 5.8904\times10^{-4} \\ -6.2427\times10^{-4} & -4.4741\times10^{-5} & 5.8904\times10^{-4} & 4.0269\times10^{-4} \end{bmatrix}$$

notes: Ten instances of the knock sound were used to generate the noisebed feature values.

3.2.28 Oven Buzzer

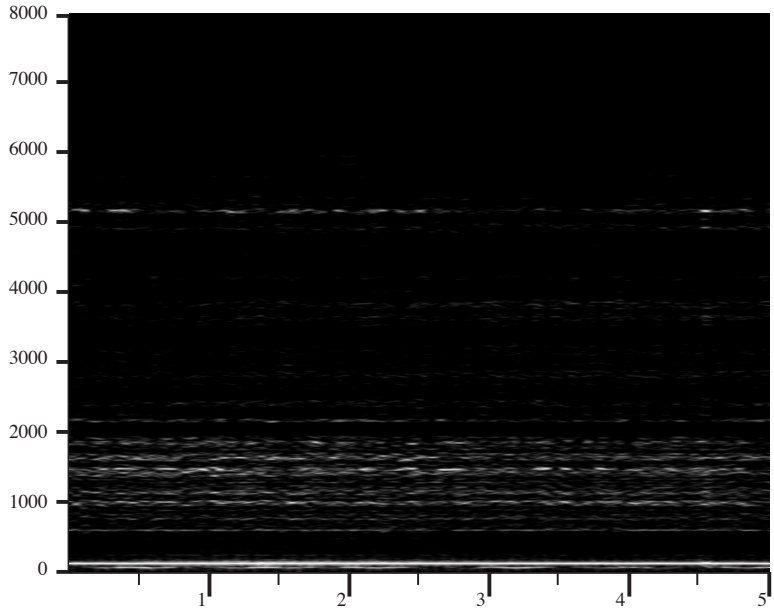

Figure 30: Lighter regions have higher energy.

track 1: 1000—1016 Hz, ampl. entropy: {55.8, 11.4}, freq. entropy: {14.8, 1.9}

track 2: 594— 602 Hz, ampl. entropy: {44.5, 11.8}, freq. entropy: {10.5, 3.4}

track 3: 117— 133 Hz, ampl. entropy: {9.0, 3.3}, freq. entropy: {0.1, 0.1}

notes: This sound's expected duration range in the IPUS library is arbitrarily set to [3.0, 10.0].

3.2.29 Owl

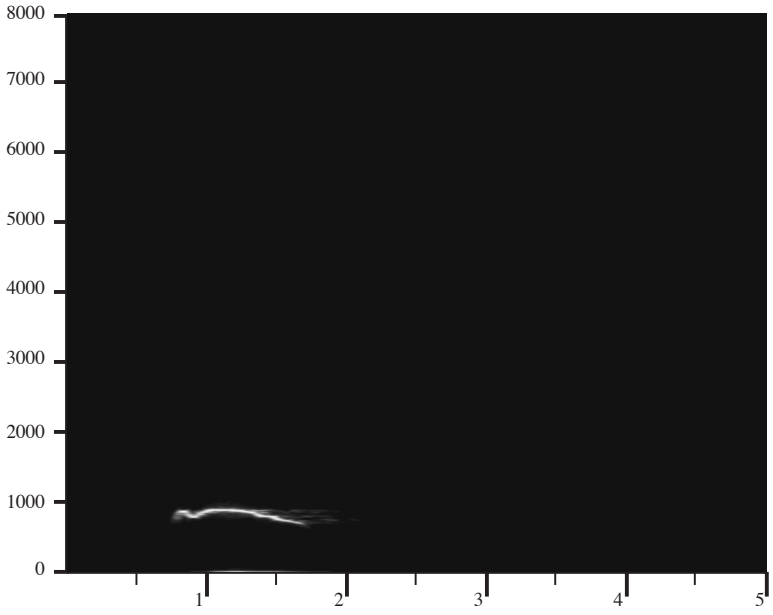

Figure 31: Lighter regions have higher energy.

track 1: 700— 880 Hz, rel. energy = [1.00, 1.00], duration = [0.9, 1.0]

track 2: 1350—1760 Hz, rel. energy = [0.05, 0.09], duration = [0.9, 1.0]

track 3: 2250—2660 Hz, rel. energy = [0.04, 0.20], duration = [0.7, 0.8]

track 4: 2000—2200 Hz, rel. energy = [0.06, 0.15], start = [0.20, 0.25], duration = [0.25, 0.3]

track 5: 1800—2000 Hz, rel. energy = [0.06, 0.15], start = [0.65, 0.70], duration = [0.50, 0.6]

notes: `Track 4` and `Track 5` are both very diffuse in frequency.

3.2.30 Pistol Shot

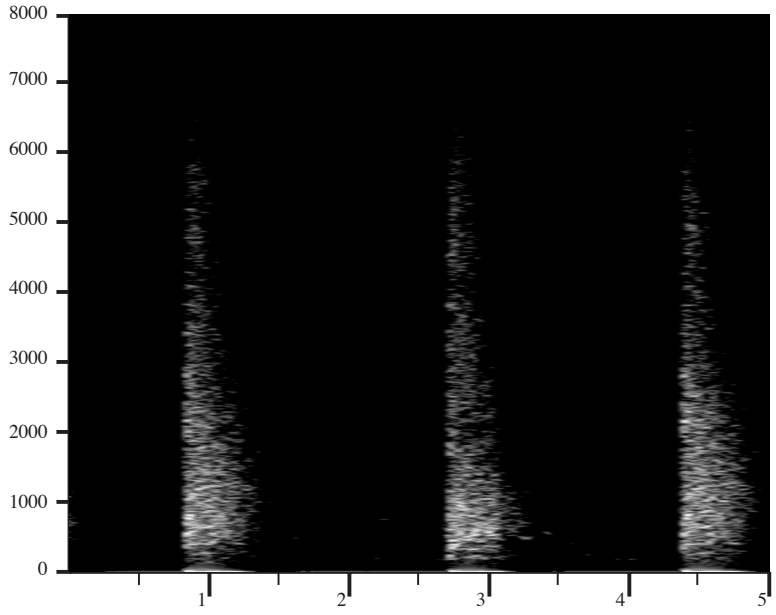

Figure 32: Lighter regions have higher energy.

noisebed 1: mean: [1.6869, 0.5350, 0.04049, 0.04070]

cov:
$$\begin{bmatrix} 3.7361\times10^{-2} & 3.0124\times10^{-3} & -1.0742\times10^{-3} & -3.0123\times10^{-3} \\ 3.0124\times10^{-3} & 1.6050\times10^{-1} & 3.7407\times10^{-3} & 1.2623\times10^{-3} \\ -1.0742\times10^{-3} & 3.7407\times10^{-3} & 1.6393\times10^{-4} & 1.3829\times10^{-4} \\ -3.0123\times10^{-3} & 1.2623\times10^{-3} & 1.3829\times10^{-4} & 2.8300\times10^{-4} \end{bmatrix}$$

notes: This sound has extremely wide-band energy, with each shot stream including a frequency decay of approximately 0.3 seconds. Five isolated pistol-shot instances were used to generate the noisebed feature values shown here.

3.2.31 Police Car Siren

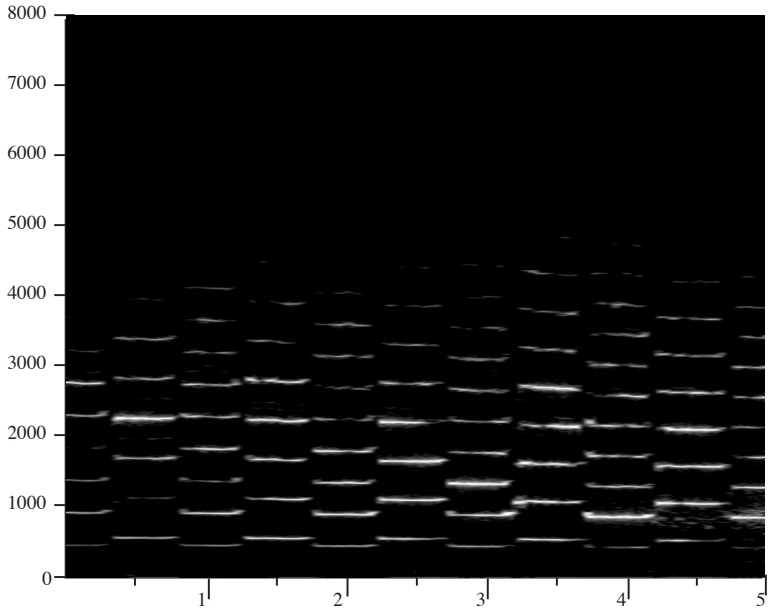

Figure 33: Lighter regions have higher energy.

Stream1 (duration 0.45—0.5)

track 1: 555— 570 Hz

track 2: 1110—1132 Hz

track 3: 1657—1690 Hz

track 4: 2227—2257 Hz

track 5: 2774—2812 Hz

track 6: 3338—3375 Hz

track 7: 3868—3898 Hz

Stream2 (duration 0.45—0.5)

track 1: 446— 468 Hz

track 2: 907— 947 Hz

track 3: 1360—1406 Hz

track 4: 1821—1843 Hz

track 5: 2274—2304 Hz

track 6: 3180—3210 Hz

track 7: 3634—3664 Hz

notes: This source has two alternating streams. One is defined by the harmonic set with $f_0 \approx 460$ Hz, and the other is defined by a harmonic set with $f_0 \approx 560$ Hz. In both streams there is too much variation in energy in the first 4 harmonics to clearly determine a maximum-energy track and, therefore, relative-energy ratios cannot be determined.

3.2.32 Razor On

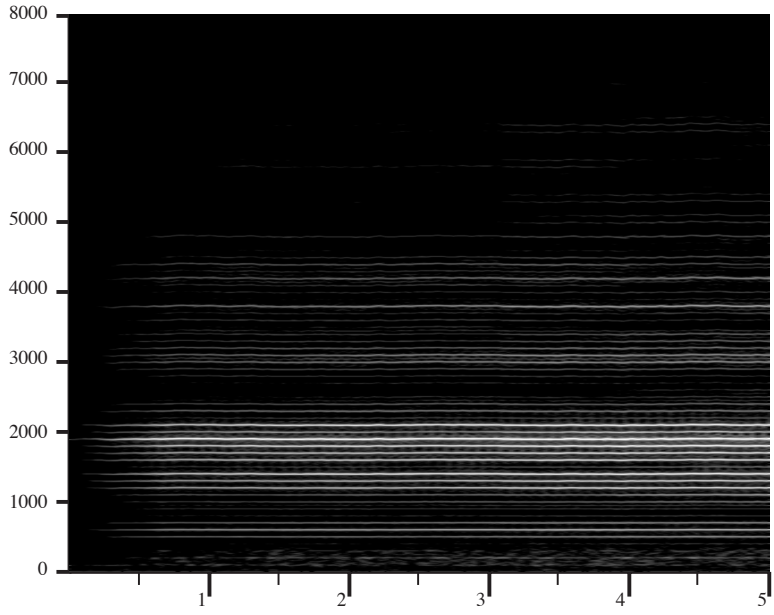

Figure 34: Lighter regions have higher energy.

track 1: 1890—1910 Hz, rel. energy $= [1.00, 1.00]$, ampl. entropy: $\{27.0, 20.3\}$, freq. entropy: $\{4.8, 0.6\}$

track 2: 2090—2110 Hz, rel. energy $= [0.60, 0.64]$, ampl. entropy: $\{35.6, 21.9\}$, freq. entropy: $\{2.3, 1.8\}$

track 3: 1790—1810 Hz, rel. energy $= [0.56, 0.60]$, ampl. entropy: $\{30.8, 21.2\}$, freq. entropy: $\{5.1, 0.4\}$

track 4: 1690—1710 Hz, rel. energy $= [0.42, 0.46]$, ampl. entropy: $\{32.4, 17.6\}$, freq. entropy: $\{4.3, 1.1\}$

track 5: 1590—1610 Hz, rel. energy $= [0.38, 0.42]$, ampl. entropy: $\{32.0, 20.1\}$, freq. entropy: $\{2.2, 2.2\}$

track 6: 1390—1410 Hz, rel. energy $= [0.38, 0.42]$, ampl. entropy: $\{22.6, 19.3\}$, freq. entropy: $\{5.6, 1.8\}$

notes: The attack time is [0.5, 0.6] seconds long. The signal is best described as a harmonic set with $f_0 = 100$ Hz. The signal has *very* steady relative energy ratios among its tracks. However, wideband (short-time) analysis reveals that the tracks all have sinusoidal amplitude modulation. The range of durations for the overall length of a razor sound is arbitrarily set to [5.0, 11.0] seconds in the IPUS database.

3.2.33 Razor Off

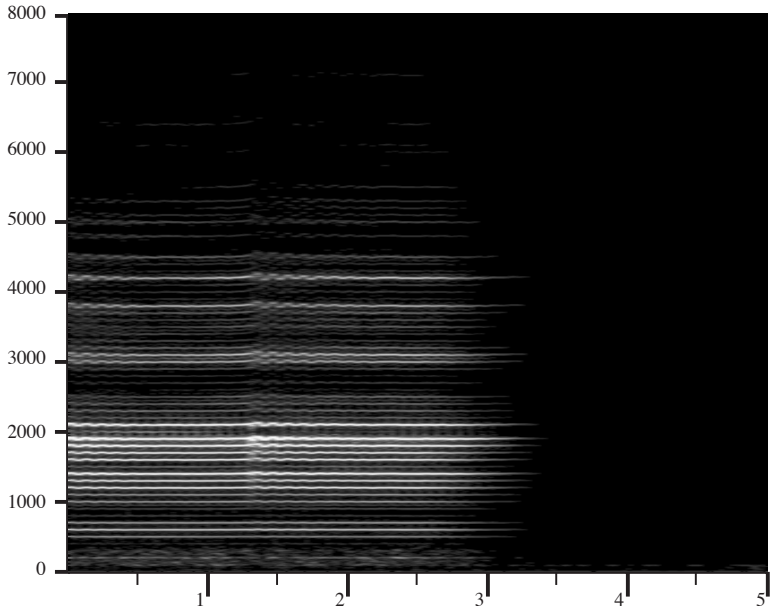

Figure 35: Lighter regions have higher energy.

track 1: 1890—1910 Hz, rel. energy = [1.00, 1.00], ampl. entropy: {27.0, 20.3}, freq. entropy: {4.8, 0.6}

track 2: 2090—2110 Hz, rel. energy = [0.60, 0.64], ampl. entropy: {35.6, 21.9}, freq. entropy: {2.3, 1.8}

track 3: 1790—1810 Hz, rel. energy = [0.56, 0.60], ampl. entropy: {30.8, 21.2}, freq. entropy: {5.1,0.4 }

track 4: 1690—1710 Hz, rel. energy = [0.42, 0.46], ampl. entropy: {32.4, 17.6}, freq. entropy: {4.3, 1.1}

track 5: 1590—1610 Hz, rel. energy = [0.38, 0.42], ampl. entropy: {32.0, 20.1}, freq. entropy: {2.2, 2.2}

track 6: 1390—1410 Hz, rel. energy = [0.38, 0.42], ampl. entropy: {22.6, 19.3}, freq. entropy: {5.6, 1.8}

notes: The decay time is [0.5, 0.6] seconds long. The signal is best described as a harmonic set with $f_0 = 100$ Hz. The signal has *very* steady relative energy ratios among its tracks. However, wideband (short-time) analysis reveals that the tracks all have sinusoidal amplitude modulation. The range of durations for the overall length of a razor sound is arbitrarily set to [5.0, 11.0] seconds in the IPUS database.

3.2.34 Rooster

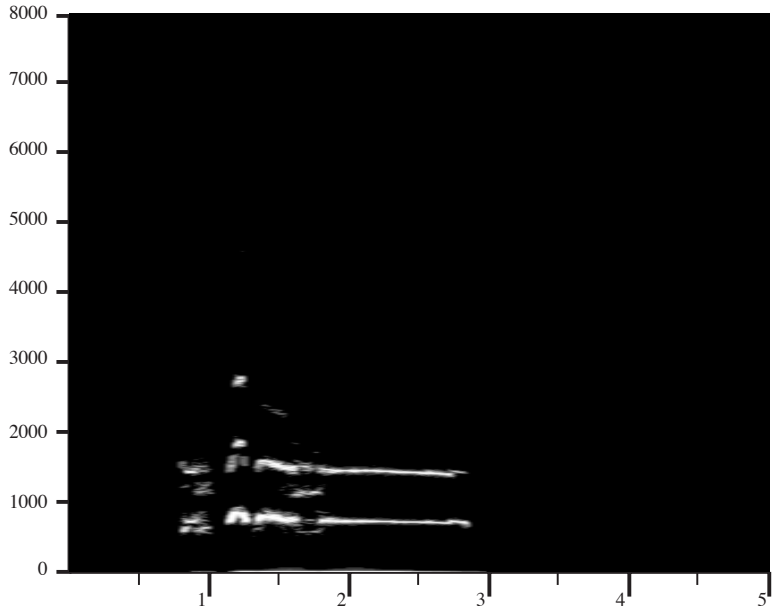

Figure 36: Lighter regions have higher energy.

track 1: 1415—1515, 1500—1860—1500, 1500—1590—1392 Hz, rel. energy = [0.30, 0.90]

track 2: 627— 730, 730— 830— 730, 730— 800— 690 Hz, rel. energy = [1.00, 1.00]

notes: Each "track" has three chirp phases, with the corners indicated above. The first phase lasts approximately 0.3 seconds; the second phase's chirps evenly split a 0.2-second period. The last phase's initial chirp lasts 0.08 seconds, while its decay chirp lasts 1.5 seconds. The entire rooster crow lasts approximately [2.1, 2.3] seconds.

3.2.35 Seagull

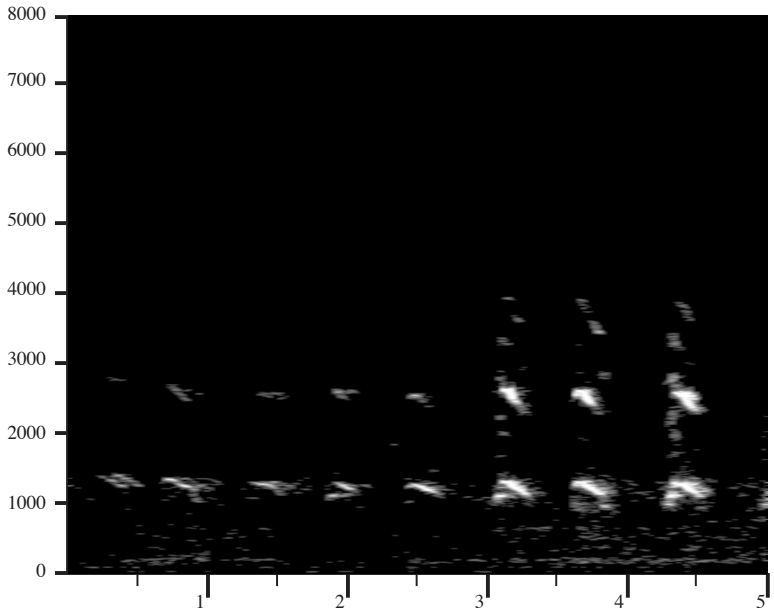

Figure 37: Lighter regions have higher energy.

track 1: 1280—1310—1150 Hz, rel. energy = [0.40, 2.26]

track 2: 2560—2620—2300 Hz, rel. energy = [1.00, 1.00]

track 3: 3200—3275—2875 Hz, rel. energy = [0.06, 0.45]

track 4: 3840—3930—3450 Hz, rel. energy = [0.16, 2.82]

track 5: 4480—4585—4025 Hz, rel. energy = [0.03, 0.40]

track 6: 5120—5240—4600 Hz, rel. energy = [0.02, 0.08]

notes: All prominent tracks appear to be members of a harmonic chirp set, with the "corners" of the chirps having the following fundamentals:

- start: 640 Hz
- max peak: 655 Hz
- end: 575 Hz

In order of track number, the above tracks represent the following harmonics: 2, 4, 5, 6, 7, and 8. Each cry is 0.25 seconds long. The initial positive-slope chirp lasts approximately 0.05 seconds. `Track 5` and `Track 6` are discernable only when the gull cries loudly. Most of the time `Track 2` has the highest energy. At the end of the cry, however, `Track 1` and `Track 4` surge in energy while `Track 2` drops. `Track 3`, `Track 5`, and `Track 6` remain at relatively constant energy levels throughout each cry.

3.2.36 Smoke Alarm 1

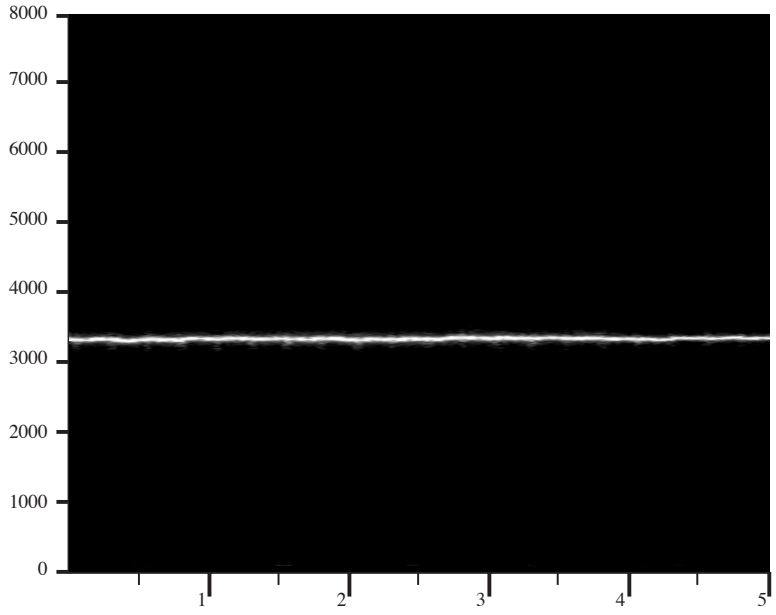

Figure 38: Lighter regions have higher energy.

track 1: 3336—3352 Hz, rel. energy = [1.00, 1.00], ampl. entropy: {39.6, 23.7}, freq. entropy: {3.5, 1.2}

notes: `Track 1`'s narrow peak widens in sideband energy every 0.25 seconds. At the same time, `Track 2`'s energy dips to a local minimum. This sound's duration range is nominally set to [3.0, 10.0] seconds.

3.2.37 Smoke Alarm 2

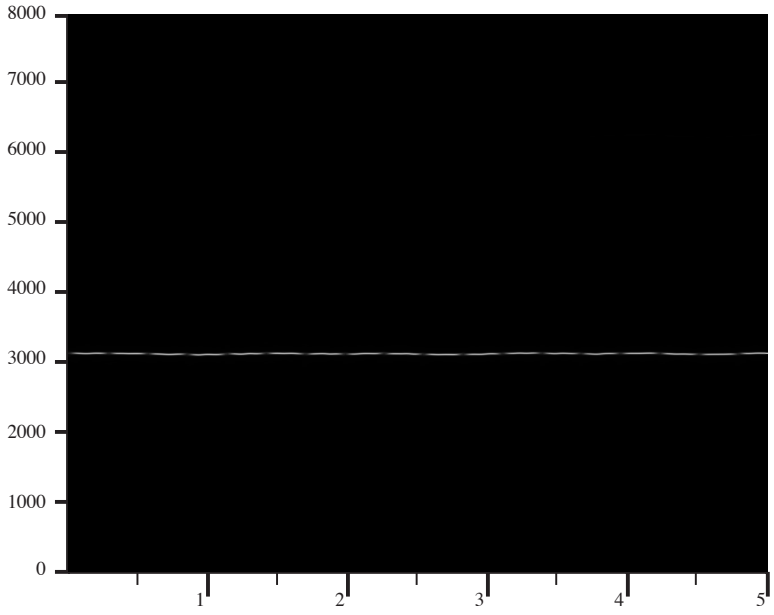

Figure 39: Lighter regions have higher energy.

track 1: 3125—3133 Hz, rel. energy = [1.00, 1.00], ampl. entropy: {58.9, 6.4}, freq. entropy: {2.3, 0.8}

track 2: 6250—6280 Hz, rel. energy = [0.02, 0.20]

notes: According to short-time analysis, every 0.7 seconds, the total signal energy drops. At these times, `Track 1`'s energy drops by a factor of 10, while `Track 2`'s energy drops by a factor of 2. The high relative energy of `Track 2` occurs at this point. The low-energy dip lasts for 0.05 seconds.

This sound's duration range is nominally set to [3.0, 10.0] seconds.

3.2.38 Telephone Dial

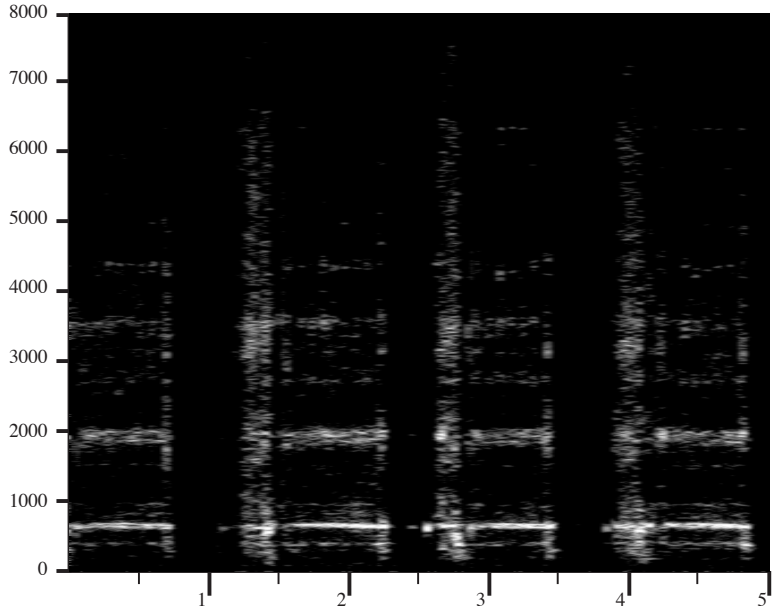

Figure 40: Lighter regions have higher energy.

track 1: 640— 690 Hz, rel. energy = [1.00, 1.00]

track 2: 1930—2000 Hz, rel. energy = [0.05, 0.20]

noisebed 1: start-time = [0.25, 0.3],

mean: [1.5481, 0.2088, 0.06939, 0.07352]

covariance:
$$\begin{bmatrix} 1.9684\times10^{-2} & 1.2455\times10^{-2} & -1.0637\times10^{-3} & -1.0697\times10^{-3} \\ 1.2455\times10^{-2} & 2.9754\times10^{-2} & -9.3511\times10^{-4} & -9.5440\times10^{-4} \\ -1.0637\times10^{-3} & -9.3511\times10^{-4} & 2.2955\times10^{-4} & 2.4747\times10^{-4} \\ -1.0697\times10^{-3} & -9.5440\times10^{-4} & 2.4747\times10^{-4} & 2.6835\times10^{-4} \end{bmatrix}$$

noisebed 2: start-time = 0.1 seconds before the end of the dial,

mean: [1.2605, 0.2114, 0.08285, 0.08847]

covariance:
$$\begin{bmatrix} 3.3743\times10^{-2} & -6.1750\times10^{-4} & -1.5244\times10^{-3} & -1.0827\times10^{-3} \\ -6.1750\times10^{-4} & 1.0062\times10^{-3} & 9.8642\times10^{-5} & 7.7836\times10^{-5} \\ -1.5244\times10^{-3} & 9.8642\times10^{-5} & 2.1811\times10^{-4} & 2.0461\times10^{-4} \\ -1.0827\times10^{-3} & 7.7836\times10^{-5} & 2.0461\times10^{-4} & 1.9815\times10^{-4} \end{bmatrix}$$

notes: This is the sound of a rotary phone being dialed. The sound has two impulses: the first representing the impact of the finger and the stopper at the end of the clockwise dial rotation, and the second at the end of the sound, representing the end of the counterclockwise dial rotation. The dial sounds range in duration from 0.5 seconds for dials from the digit 1 to 2.0 seconds for dials from the digit 0.

3.2.39 Telephone Ring

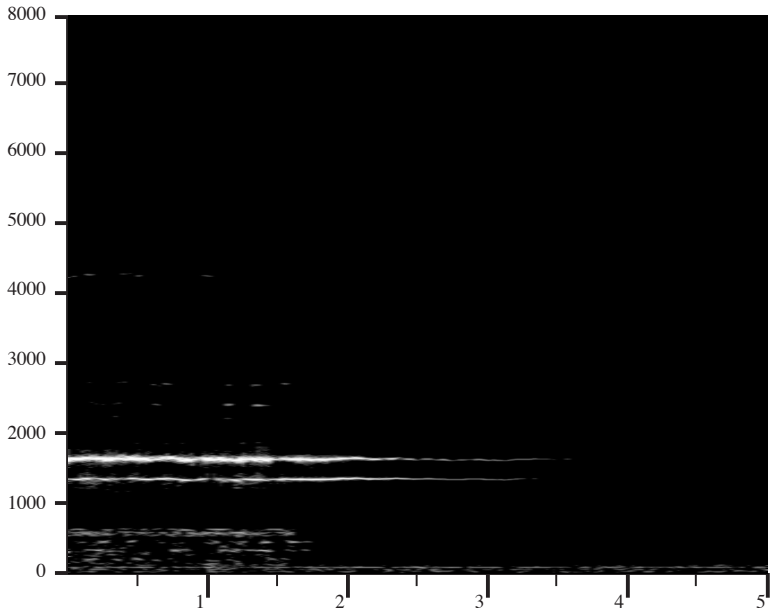

Figure 41: Lighter regions have higher energy.

track 1: 1630—1655 Hz, ampl. entropy: {76.5, 15.6}, freq. entropy: {8.2, 1.5}

track 2: 1345—1365 Hz, ampl. entropy: {50.1, 13.0}, freq. entropy: {8.7, 2.1}

track 3: 550— 620 Hz, ampl. entropy: {44.9, 7.5}, freq. entropy: {89.5, 6.7}

notes: For each ring, the striker actively hits the bell for 1.7 seconds; reverberations last 3.3 seconds after that. `Track 1` is the highest-energy track. `Track 3` has energy diffused throughout its range. There is too much variability in track energies to determine meaningful relative energy ratios.

3.2.40 Telephone Tone

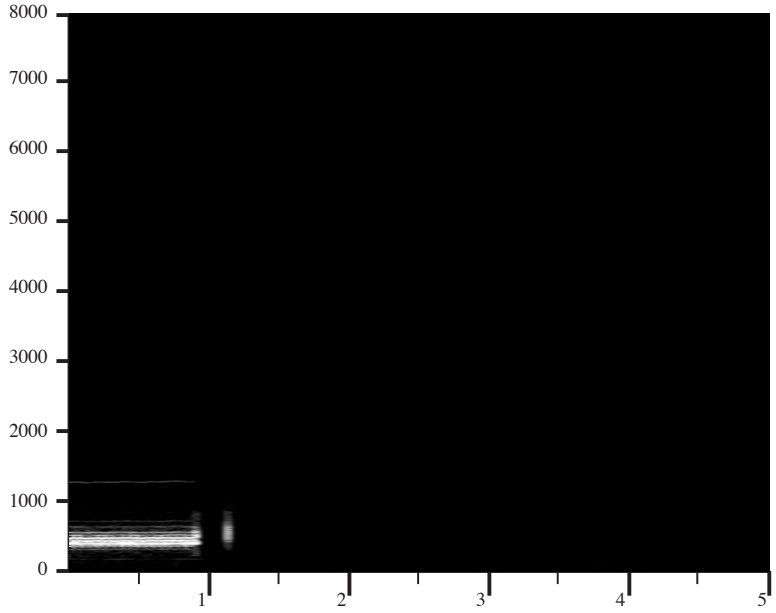

Figure 42: Lighter regions have higher energy.

track 1: 398— 415 Hz, rel. energy = [1.00, 1.00], ampl. entropy: {13.6, 0.8}, freq. entropy: {17.4, 4.7}

track 2: 1270—1290 Hz, rel. energy = [0.20, 0.30]

notes: Wideband analysis shows that the tracks are actually the result of 0.01-second-long spikes that occur every 0.03 seconds. The sound's range of durations is [0.8, 2.0] seconds.

3.2.41 Triangle

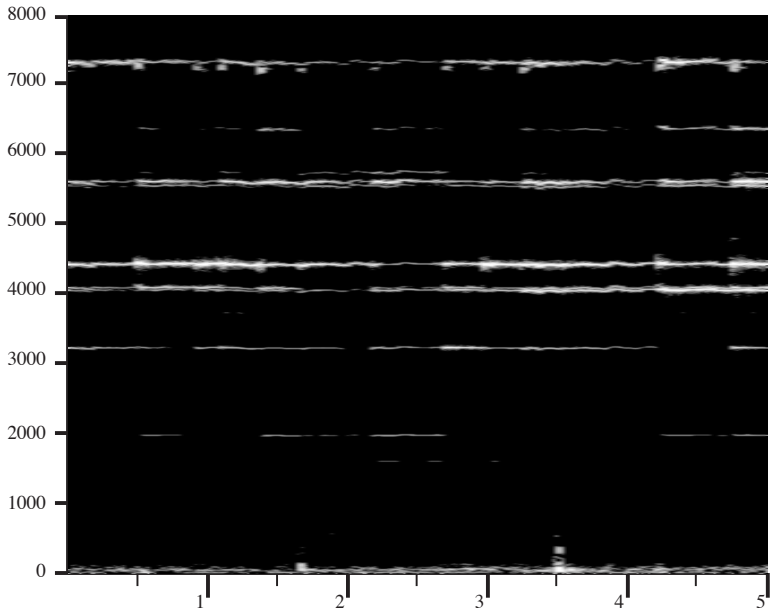

Figure 43: Lighter regions have higher energy.

track 1: 4415—4440 Hz, rel. energy = [1.00, 1.00]

track 2: 7297—7320 Hz, rel. energy = [0.10, 0.20]

track 3: 4047—4085 Hz, rel. energy = [0.20, 0.40]

track 4: 5571—5609 Hz, rel. energy = [0.15, 0.40]

notes: The above model was generated from five isolated triangle-strike instances. The durations all fall in the range [1.0,1.2] seconds.

3.2.42 Truck Motor

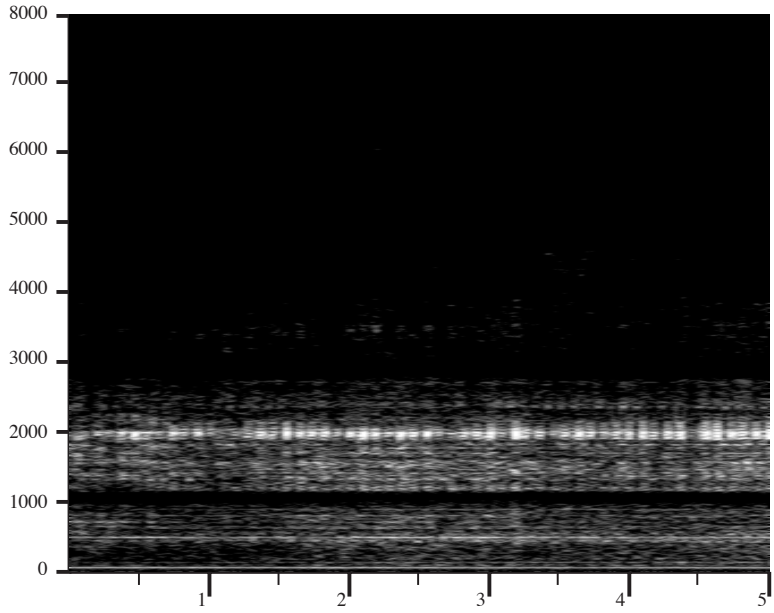

Figure 44: Lighter regions have higher energy.

track 1: 1970—2010 Hz, rel. energy = [1.00, 1.00]

track 2: 475— 507 Hz, rel. energy = [0.03, 0.35]

notes: `Track 1` is really a series of discontinuous frequency spikes. Wideband analysis shows that each burst on the spectrogram shown here is a double spike with 0.02 seconds between each spike in the pair, and with each spike lasting approximately 0.01 seconds. The distance between the last spike of one pair and the first spike of the next is 0.05 seconds. The three time values yield a pair period of approximately 0.1 seconds. The sound's range of durations is arbitrarily set to [5.0, 30.0] seconds.

3.2.43 Vending Machine Hum

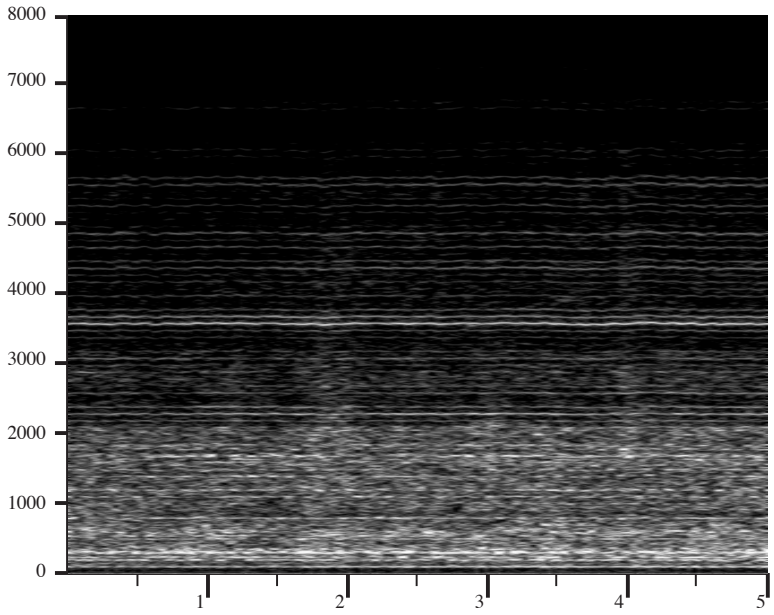

Figure 45: Lighter regions have higher energy.

track 1: 3565—3595 Hz, rel. energy = [1.00, 1.00]

track 2: 3665—3695 Hz, rel. energy = [0.55, 1.00]

track 3: 5540—5580 Hz, rel. energy = [0.40, 0.75]

track 4: 5640—5680 Hz, rel. energy = [0.30, 0.65]

notes: Although the narrowband tracks are well-defined, note that this source has a significant noisebed from 0 to 8000 Hz. The sound's range of durations is arbitrarily set to [5.0, 30.0] seconds.

3.2.44 Viola

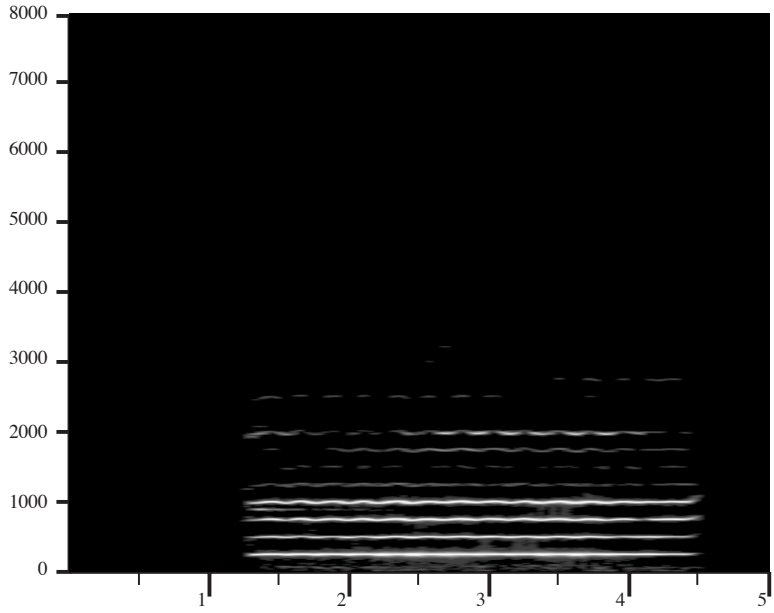

Figure 46: Lighter regions have higher energy.

track 1: 992—1032 Hz, ampl. entropy: {26.6, 7.7}, freq. entropy: {19.3, 1.4}

track 2: 750— 782 Hz, ampl. entropy: {51.0, 6.7}, freq. entropy: {18.5, 2.5}

track 3: 492— 514 Hz, ampl. entropy: {52.9, 14.8}, freq. entropy: {22.5, 3.5}

track 4: 242— 274 Hz, ampl. entropy: {25.9, 3.9}, freq. entropy: {14.6, 12.0}

notes: The signal has the harmonic set $f_0 \approx 260$ Hz. Its duration range in the IPUS database is [2.4, 3.0] seconds.

3.2.45 Violin Plain

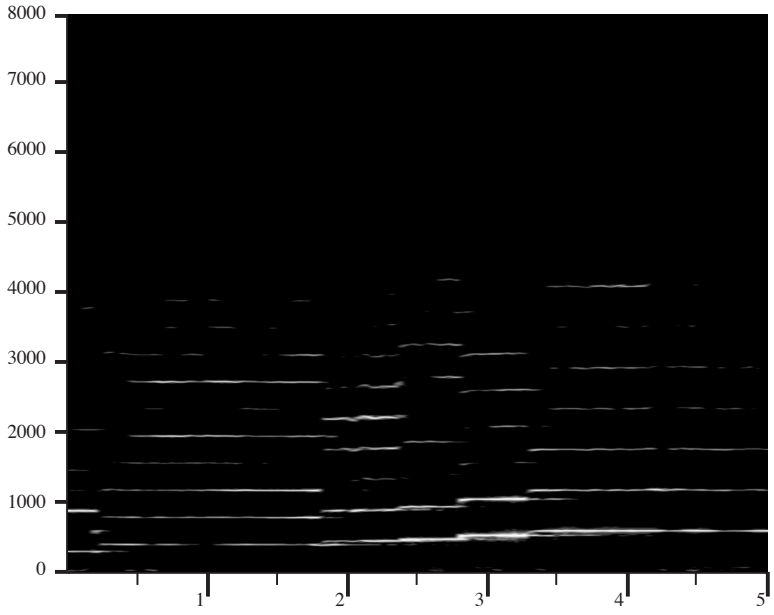

Figure 47: Lighter regions have higher energy.

track 1: 1172—1188 Hz, ampl. entropy: {43.6, 17.6}, freq. entropy: {8.8, 1.6}

track 2: 875— 898 Hz, ampl. entropy: {72.9, 8.3}, freq. entropy: {16.2, 3.6}

track 3: 586— 609 Hz, ampl. entropy: {53.3, 30.9}, freq. entropy: {0.4, 1.7}

track 4: 445— 485 Hz, ampl. entropy: {68.9, 34.8}, freq. entropy: {42.7, 18.0}

track 5: 390— 414 Hz, ampl. entropy: {70.8, 19.9}, freq. entropy: {19.2, 12.5}

notes: This is the same note as in *Violin Vibrato*, played without vibrato style. Note the lower frequency entropy values listed for the tracks. The sound's range of durations is [2.4, 3.0] seconds.

3.2.46 Violin Vibrato

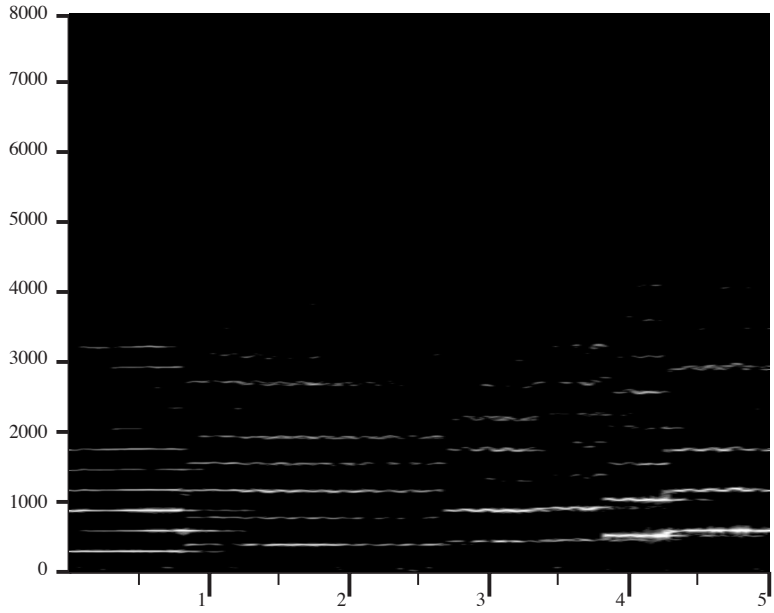

Figure 48: Lighter regions have higher energy.

track 1: 1172—1188 Hz, ampl. entropy: {58.7, 21.8}, freq. entropy: {7.2, 1.1}

track 2: 875— 898 Hz, ampl. entropy: {99.2, 41.2}, freq. entropy: {13.5, 6.8}

track 3: 586— 609 Hz, ampl. entropy: {81.6, 11.4}, freq. entropy: {15.7, 6.9}

track 4: 445— 485 Hz, ampl. entropy: {93.0, 40.7}, freq. entropy: {44.2, 15.6}

track 5: 390— 414 Hz, ampl. entropy: {74.8, 25.9}, freq. entropy: {25.5, 3.4}

notes: This is the same note as in *Violin Plain*, but played vibrato style. Note the higher frequency entropy values listed for the tracks. The sound's duration range is [2.4, 3.0] seconds.

3.2.47 Wind

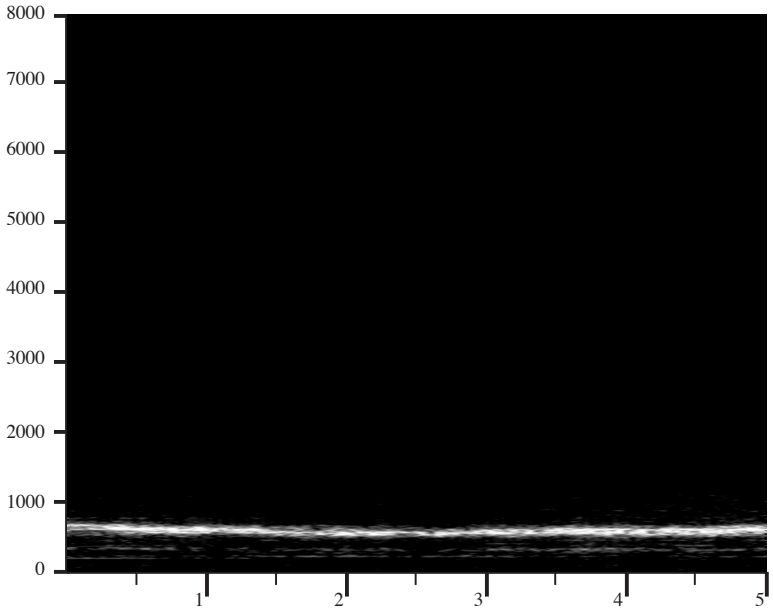

Figure 49: Lighter regions have higher energy.

track 1: 625— 766 Hz, ampl. entropy: {71.8, 14.6}, freq. entropy: {46.9, 12.6}

notes: The spectral energy in `Track 1` is diffused throughout the track's frequency region. The range of durations for the overall length of this sound is arbitrarily set to [5.0, 30.0] seconds in the IPUS database.

References

[1] *"Sound Effects" Tape*, Auditec of St. Louis, Inc., 330 Selma Avenue, St. Louis, MO, 63119.

[2] Bregman, A. "Auditory Scene Analysis: The Perceptual Organization of Sound," MIT Press, 1990.

[3] Lesser, V., Nawab, H., Klassner, F., "IPUS: An Architecture for the Integrated Processing and Understanding of Signals," *Artificial Intelligence Journal*, vol. 77, no. 1, August/September 1995.

[4] Nawab, S. H. and Dorken, E., "Efficient STFT Computation Using a Quantization and Differencing Method," *The Proceedings of the 1993 IEEE Conference on Acoustics, Speech and Signal Processing*, vol. 3, pp. 587–590, Minneapolis, Minnesota, April 1993.