TECHNICAL REPORT

Title: The calculation of acoustic indices derived from long-duration recordings of the natural environment.

Author: Michael Towsey

Original Deposit: August 2017
History: Edited March 2018
Edited April 2018

Institution: **QUT Ecoacoustics Research Group** | www.ecosounds.org
Science and Engineering Faculty,
Queensland University of Technology,
Brisbane, Australia

Keywords: Acoustic environment, acoustic indices, summary indices, spectral indices.

1. Introduction

The work described in this technical report is part of ongoing research within the QUT Ecoacoustics Research Group to build practical tools for the manipulation, analysis and visualisation of long-duration recordings of the natural environment. This report describes methods we use to calculate *spectral* and *summary acoustic indices* derived from recordings of the natural environment. Typically, we break long-duration recordings (anything from 30 minutes to 24 hours into one-minute segments and calculate indices independently for each segment. However, we also calculate high-resolution indices (up to 100ms resolution) for special purposes, usually as acoustic features for species recognizers. In our terminology, a spectral index is a vector of *N* values, one for each frequency bin of the spectrogram. Typically, our frame size is 512, so the spectral index is a 256-element vector. A summary index is a scalar value, in some cases derived directly from the waveform envelope and in other cases from the 256 values of the corresponding spectral index. Note that the term 'acoustic index' as used in the literature, usually refers only to summary indices.

An important step in the calculation of many indices is the prior removal of 'background noise' from the waveform envelope or from the spectrogram. It is important to note that, in the context of audio recordings of the environment, 'noise' can have several interpretations. Noise does not mean just electronic or microphone noise as engineers understand it. Much of the 'noise' in environmental recordings is of physical origin due to wind, rain, leaf rustle, etc. (referred to as *geophony* in the ecoacoustics literature) and of biological origin (also called *biophony*, due to cicada, cricket and other persistent animal vocalisations). In this work, we define 'noise' as that acoustic energy which persists throughout the duration of a recording segment, usually one minute. Thus, it is possible that the same acoustic source will contribute to both 'noise' and specific events (signal) within a recording. For example, if we assume that crickets are evenly distributed in the landscape around a sensor, there will be a background 'murmur' of crickets (noise) but the chirps of those crickets closest to the microphone will register as specific acoustic events (signal) within the background. Likewise, wind gusts will stand out as specific acoustic events within a constant background of noise generated by moving air and leaves. There is, of course, a third sense in which the term 'noise' is used: any acoustic event that is not of interest and/or obstructs the detection of those events that are of interest. But this definition is not relevant to the calculation of the acoustic indices described in this document.

The remainder of this report is divided into six sections:

Signal acquisition and pre-processing
Noise removal
The calculation of spectral indices
The calculation of summary indices
The clustering of spectra
Ridge Indices

2. Signal acquisition and pre-processing

2a. Signal Processing

Since publication of our earlier reports (e.g. http://eprints.qut.edu.au/61399/), our lab has primarily acquired signals from two brands of commercial sensors, the BAR (Frontier Labs, Brisbane, Australia) and Song Meters, SM2 and SM4 (Wildlife Acoustics, Massachusetts, USA). Regardless of the recording sample rate (usually 22.05 kHz, but anywhere between 16 kHz and 96 kHz), we currently down-sample to 22.05 kHz before analysis and visualisation. The primary justification for this is storage constraint. It does mean however, that we lose frequencies above 11,025 Hz where some insects, higher bird harmonics and of course bat calls are located. We always use a bit depth of 16. The audio format is either WAVE or WAC (Wildlife Acoustics Audio Compression), but at the time of writing this report, we are also exploring other compression options. We no longer record in MP3 because it generates spectrogram artefacts that greatly affect the values of some acoustic indices, most especially the Acoustic Complexity Index.

Typically we process a long recording by splitting it into one-minute segments of audio. If a signal is re-sampled at 22,050 samples per second and divided into non-overlapping frames of 512 samples each, there will be 2584 frames per one minute of recording, each frame having a duration ~23.2ms. If a final fractional frame occurs, it is discarded.

2b. Production of a wave envelope

A wave envelope is derived from the signal waveform by taking the maximum absolute value in each frame. The number of values in the wave envelope of a one-minute audio segment will therefore equal the number of complete frames. Absolute amplitude values are converted to decibels (dB) using: $dB = 20.\log_{10}(A)$. To protect against zero signal values (which can occur due to equipment failure) the minimum dB value is set to -90 dB. This value is comparatively high but suitable for acoustic recordings of the environment using our hardware. This minimum value should be adjusted according to the quality of your microphones, bit rate etc.

2c. Spectrograms

For the preparation of spectrograms, signals are typically framed using a window of 512 or 1024 samples and 0% or 50% overlap. A Hamming window function is applied to each frame prior to performing a Fast Fourier Transform (FFT). The spectra are smoothed with a moving average window of width three. Depending on the subsequent analysis to be performed, the spectral amplitude values (Fourier coefficients, A) may be converted to spectral energy/power (A^2) or to decibels (dB) using: $dB = 20\times\log_{10}(A)$, with minimum value set to -90 dB. Note that decibels being a ratio, the dB values at this stage are with respect to a hypothetical signal having unit amplitude in each frequency bin. Where the frame width = 512 samples, the frequency bin width = ~43.1 Hz.

3. Noise removal

This section on noise removal supersedes a previous report titled “Noise removal from waveforms and spectrograms derived from long-duration recordings of the natural environment”, M. Towsey (2013 September), http://eprints.qut.edu.au/61399/. Three noise subtraction methods are used in the literature for noise removal: subtraction of the mean, subtraction of the median and subtraction of the mode. We find that for environmental recordings containing normal sounds due to biophony and geophony, there is not much difference in the visual appearance of spectrograms prepared using these three means of noise removal (see Figure 1. Modal noise subtraction assumes an additive noise model (see later). Median noise subtraction is most often used because it makes no assumptions about a noise model.

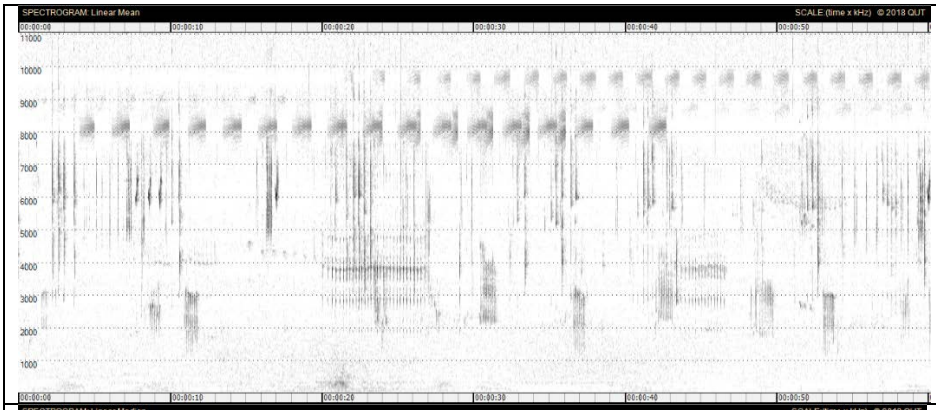

Figure 1a: Spectrogram after mean noise subtraction. Background noise in frequency bin 100 = -90.7 dB.

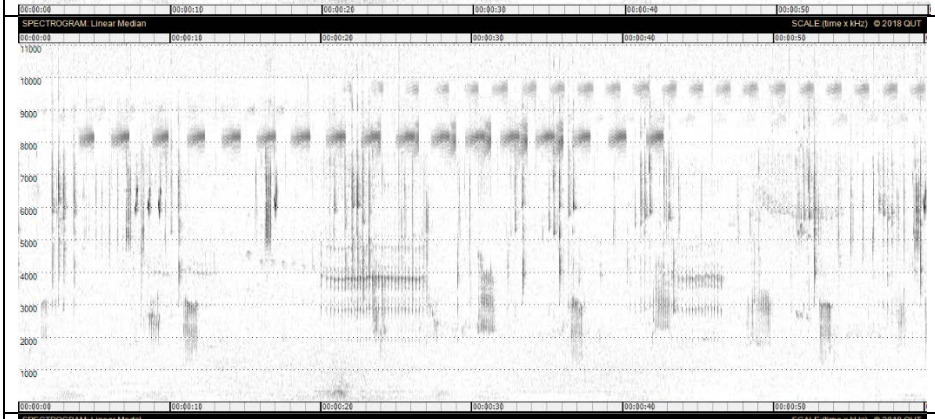

Figure 1b: Spectrogram after median noise subtraction. Background noise in frequency bin 100 = -91.5 dB.

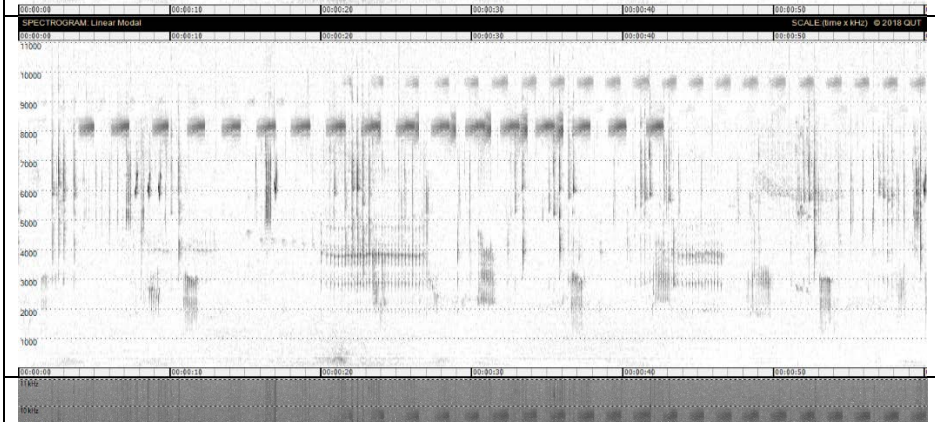

Figure 1c: Spectrogram after modal noise subtraction. Background noise in frequency bin 100 = -91.6 dB.

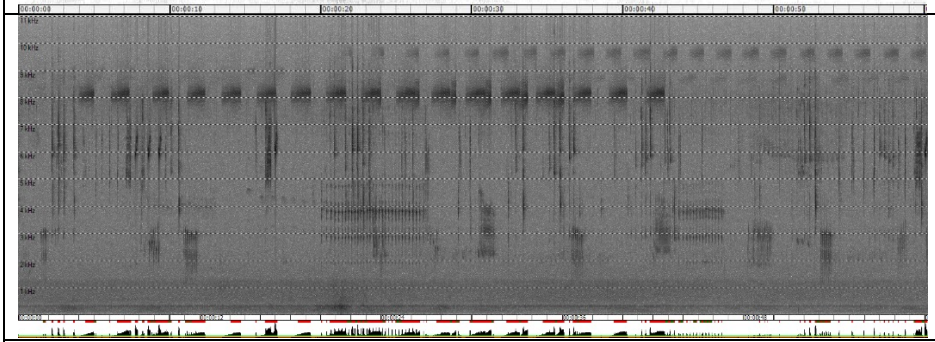

Figure 1d: The original recording

Figure 1. “Noise” subtraction from a one-minute recording of the natural environment

Our ‘noise’ removal algorithm uses modal noise subtraction. It assumes an additive model where the energy of discrete acoustic events is added to a continuous acoustic background whose energy is assumed to have a Gaussian distribution. It is further assumed that the recording contains sufficient ‘silence’ (absence of significant acoustic events) that the mean and standard deviation of the background noise can be estimated reliably. Typically, the recordings from which we remove noise have a duration of 30 seconds to 2 minutes and over that duration, the condition of sufficient silence is usually met.

First, we describe noise removal from the signal waveform and then noise removal from the spectrogram.

3a. Noise removal from the waveform

We remove noise from a one-minute waveform using the *adaptive level equalisation* algorithm originally used for end-point detection in speech recordings (Lamel, L.F. et al., 1981). Lamel’s method assumes that the wave envelope values are in decibels. It also assumes an additive noise model where the noise has a Gaussian distribution whose average does not exceed 10 dB above the minimum value of the wave envelope. The application in Lamel et al is signal transmission over a telephone line. In our implementation, we ignore the 10-dB constraint.

1. Find the minimum and maximum values of the waveform. Note that in our implementation the minimum will never be less than -90dB.
2. Compute a 100-bin histogram of the decibel intensity values. The minimum and maximum bins match the minimum and maximum of the waveform.
3. Smooth the histogram using a moving average filter (window = 5 or 7 works well).
4. The mode of the histogram distribution (the histogram bin having maximum count) corresponds to the mean of the noise, assuming an additive noise model.
5. If required, one can calculate the standard deviation of the (assumed Gaussian) noise distribution by summing counts in bins below the mode until 68% of total counts below the mode are accumulated.
6. The value of background noise in the recording waveform is given by the modal intensity $\pm N$ times the standard deviation. Higher values of N will remove more energy from the waveform. In practice, we usually follow Lamel et al. in setting $N = 0$. $N > 0$ removes excessive signal. For this reason, we have sometimes used $N = -0.1$.
7. The noise-reduced signal waveform is calculated by subtracting the value found in Step 6 above from each waveform value and truncating negative values to zero. Adaptive level equalisation has the effect that during silence, the noise-reduced waveform power is close to zero.

Recalling the above ‘working’ definition of noise as that acoustic energy which persists through the duration of a one-minute recording, it is more accurate to state that we set the background ‘noise’ value equal to the mode of the distribution of energy values in the signal waveform.

3b. Noise removal from spectrograms

The contribution of noise to recordings of the environment typically declines with increasing frequency. However, we do not assume a standard pink noise model. Rather, we estimate the modal value independently for each of the frequency bins in the spectrogram of each one-minute recording. We use a modified version of the same *adaptive level equalisation* algorithm described above, with N equal zero. Note that this modified version can also be applied to amplitude and power spectrograms where values have not been converted to decibels. Having calculated a ‘background’ intensity value for each frequency bin, we subtract it from each value in that bin (with truncation of negative values to zero).

In more detail:
A: For each frequency bin (or row of spectrogram values):

1. Compute a 100-bin histogram of the intensity values. The minimum and maximum histogram bins match the minimum and maximum values respectively in the current frequency bin.
2. Smooth the histogram using a moving average filter (window = 5).
3. The modal intensity value corresponds to the bin containing maximum count. If the modal bin is in 96-100, set the modal intensity value to the 95th bin.
4. The value of 'background noise' in the current frequency bin is given by the intensity value corresponding to the bin obtained in Step A3.

B: Step A produces a 'noise profile', a vector of 'background noise' values, one value for each frequency bin.

1. Smooth the noise profile using a moving average filter (window = 5). Smoothing the noise profile eliminates possible "banding" in the noise reduced spectrograms.
2. Subtract the resulting background noise values from the values in each frequency bin. Truncate negative values to zero.
3. As noted above, the alternative to subtracting the modal value from each frequency bin is to subtract the mean or the median. For a 'normal' recording of the environment the resulting spectrograms do not look different (see Figure 1).

C: We implement an additional noise removal step which endeavours to preserve the structural integrity of complex acoustic events (e.g. bird calls) but removes noise from background locations further removed from those events.

5. For each neighbourhood (3 frames × 9 frequency bins) centred on any element/pixel in the spectrogram, calculate the average spectrogram value, $\bar{a}$.
6. If $\bar{a}$ is less than a user determined threshold, θ, set the value of the central element/pixel equal to the minimum in the neighbourhood. Typically, the minimum will be zero because this step is carried out after Steps A and B. We set θ = 0.015 or 2–4 dB depending on the units of the spectrogram.

Adaptive level equalisation has the effect that during silence, the power in every frequency bin fluctuates around 0 dB but during an acoustic event it is considerably higher. Thus it becomes possible to define a single absolute threshold for the detection of an acoustic event that spans multiple frequency bins.

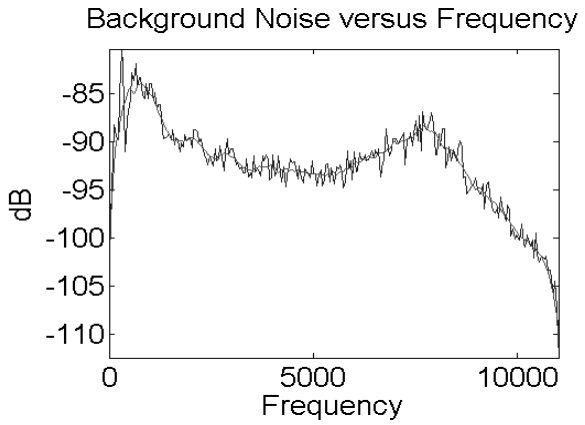

Figure 2. Noise intensity versus frequency for a typical spectrogram derived from a mobile phone recording; original and smoothed values shown. This background noise profile is typical when a mobile phone is used as a sensor. The power drop-off below 500 Hz and above 8-9 kHz is due to a filter applied in the phone. For this reason, we no longer use mobile phones for long-duration recordings of the environment.

Our approach assumes that each frequency bin derived from a one-minute audio segment contains sufficient frames without signal to estimate the noise intensity for that bin. In practice, we find that most eco-acoustic recordings permit accurate estimates of the modal background noise. However, counter-intuitive results can occur when insects (especially cicadas and orthoptera) call continually in a narrow frequency band. (For example, the peak at 8 kHz shown in Figure 2 is due to a chorus of crickets.) This noise reduction technique removes continuous insect tracks so that they do not appear in the noise reduced spectrogram.

4. **The calculation of spectral indices**

We calculate the following spectral indices. A three-letter code is used for ease of identifying each index. The subscript suffix 'sp' identifies the index as a spectral index as opposed to a scalar summary index.

a. **Background Noise (BGNsp):** This is the noise profile calculated as described above in Section 3b, 'Noise removal from spectrograms', step B1.
b. **Power Minus Noise (PMNsp):** The difference between the maximum decibel value in each frequency bin and the decibel BGN value for the corresponding bins.
c. **Activity (ACTsp):** The fraction of cells in each noise-reduced frequency bin whose value exceeds the threshold, $\theta = 3$ dB.
d. **Events (EVNsp):** The number of acoustic events per minute in each noise-reduced frequency bin. An event is counted each time the decibel value in a bin crosses the 3-dB threshold from lower to higher values.
e. **Temporal Entropy (ENT_{sp}):** A measure of the acoustic energy 'concentration' in each frequency bin, *f* (Towsey et al., 2014). The squared amplitude values in each frequency bin (indexed *f*) are normalized to unit area and treated as a probability mass function (*pmf*). The entropy of the normalized values is a measure of the energy 'dispersal' through time and is calculated as:

$$H_t[f] = -(\sum_i pmf_{if} \times \log_2(pmf_{if})) / \log_2 N \qquad (1)$$

where *i* is an index over frames and *N* is the number of frames. To obtain a more 'intuitive' index, we convert $H_t(f)$ to 'energy concentration' as follows:

$$ENT[f] = 1 - H_t[f] \qquad (2)$$

f. **Acoustic Complexity Index (ACI_{sp}):** This index is a 256-element vector that quantifies the relative change in acoustic intensity (*A*) in each bin (*f*) of the amplitude spectrogram:

$$ACI[f] = \sum_i |A_{if} - A_{i-1,f}| / \sum_i A_{if} \qquad (3)$$

where *i* is an index over frames and *f* is an index over frequency bins (Farina et al., 2014; Farina et al., 2013; Pieretti et al., 2011). It is widely used as a measure of biophony in environmental recordings. However, it is also highly sensitive to some non-biological sound sources, such as rain. Normal practice in ecological studies is to manually exclude recordings containing rain and wind. However, in this study, the presence of wind and rain in a soundscape is also of interest. Note that we modify the published method by dividing ACI[f] by the number of frames, *N*, in the spectrogram. This converts the sum in (3) to an average and normalises for different segment durations and frame sizes.

g. **Ridge Indices (RHZ_{sp}, RPS_{sp}, RVT_{sp}, RNG_{sp}, $R3D_{sp}$)**: These indices are derived from the noise-reduced decibel spectrogram. Many animal sounds, particularly bird songs, have a harmonic structure resulting in one to many formants. We attempt to "detect" these using 5×5 ridge masks. We calculate four ridge indices corresponding to the four directions of ridge slope: Ridge Horizontal (RHZ_{sp}), Ridge Vertical (RVT_{sp}), Ridge Positive having an upward slope (RPS_{sp}) and Ridge Negative having downward slope (RNG_{sp}). These are 256-element vectors, each element of which is the average decibel value of ridge cells identified

within the corresponding frequency bin (*f*). See Section 7 for more detail. We calculate a fifth ridge index, $R3D_{sp}$, which equals the maximum of {RHZ_{sp}, RPS_{sp}, RNG_{sp}}. Formants in the mid-band are typically due to bird song. Vertical ridges on the other hand are frequently not due to biophony but rather to rain drops, electronic clicks, etc.

h. **Spectral Peak Tracks (SPT_{sp})**: A measure of the presence of spectral peak tracks in a spectrogram.

Step One: Detect the location of local maxima in the spectra, one frame at a time. A spectral cell counts as a local maximum if its decibel value exceeds a threshold (we use 6 dB) and if its value is greater than the values in the two frequency bins on either side.

Step Two: Sum the decibel values of all those cells identified as spectral peaks within a frequency bin.

Step Three: The SPT vector consists of the sums obtained in step 2 divided by the number of frames in a frequency bin.

5. The calculation of summary indices

Summary indices fall into four categories. The first category consists of four indices derived from the signal's waveform envelope converted to decibels. They correspond to the first four spectral indices defined above.

1. **Background Noise (BGN):** An estimate of the background noise in each one-minute recording. It is set equal to the mode of the energy distribution in the waveform envelope as described in Section 3a.
2. **Signal to Noise Ratio (SNR):** The difference between the maximum decibel value in the decibel envelope and the decibel value of BGN.
3. **Activity (ACT):** The fraction of values in the noise-reduced decibel envelope that exceed the threshold, $\theta = 3$ dB.
4. **Events per Second (EVN):** A measure of the number of acoustic events per second, averaged over the same noise-reduced one-minute segment. An acoustic event is defined as starting when the decibel envelope crosses a threshold, θ, from below to above, where $\theta = 3$ dB.

The following three summary indices (5, 6, and 7) are derived from the noise-reduced decibel spectrogram. They compare acoustic activity in the low, middle and high frequency bands. The mid-band bounds (1000-8000 Hz) were chosen to capture most of the bird vocalisations but to avoid much of the anthropophony which predominates at low frequencies. The decibel spectrogram was noise-reduced by applying the steps in section 3b above.

5. **Low-frequency Cover (LFC):** The fraction of noise-reduced spectrogram cells that exceed 3 dB in the low-frequency band (1-1000 Hz).
6. **Mid-frequency Cover (MFC):** As for LFC but in the mid-frequency band (1000-8000 Hz).
7. **High-frequency Cover (HFC):** As for LFC but in the high-frequency band (8000–11025 Hz).

The following four indices (8, 9, 10 and 11) describe different 'entropy' measures of the distribution of acoustic energy within a recording. Index 8 (temporal entropy) is derived from the waveform envelope. Indices 9, 10 and 11 are different measures of the distribution of acoustic energy in the mid-frequency band (1000-8000 Hz) of each noise-reduced, amplitude spectrogram. Note that for indices 9, 10 and 11, the noise-reduced spectrograms were derived from the amplitude spectrogram and not from the decibel spectrogram. Amplitude values were squared to give an energy value. These indices are similar in purpose to the Gini index used in [Briggs et al., 2014] to describe energy distribution within acoustic events. Entropy values give a measure of the *flatness* of a distribution. To obtain a more intuitive index, we subtract each entropy value from 1.0, to obtain a measure of energy 'concentration'.

8. **Temporal entropy (ENT):** Entropy of the energy (squared amplitude) values of the signal waveform (henceforth *temporal entropy*). The squared amplitude values were normalised to unit area and treated as a probability mass function (*pmf*). The entropy (*H*) of the signal was calculated as:

$$H_t = -(\Sigma_i \, pmf_i \times \log_2(pmf_i)) / \log_2 N,$$

where *i* is an index over all *N* values in the signal envelope [Depraetere et al., 2012]. Finally, a measure of energy 'concentration' is calculated as ENT = $1 - H_t$.

9. **Entropy of the Spectral Peaks (EPS):** A measure of the 'concentration' of spectral maxima in the mid-band.
    i. For each frame, determine the bin within the mid-band having the maximum amplitude, the spectral maximum.
    ii. Prepare a histogram (having bin count equal to the number of frequency bins) of the count of spectral maxima occurring in each frequency bin.
    iii. Calculate the entropy of the distribution, H_p, as for H_t, except that *N* = number of frequency bins in the mid-band. Finally, EPS = $1 - H_p$.
10. **Entropy of the Average Spectrum (EAS):** A measure of the 'concentration' of mean energy within the mid-band of the mean-energy spectrum. The entropy of the distribution, H_a, is calculated as for H_t. EAS = $1 - H_a$.
11. **Entropy of the Spectrum of Coefficients of Variation (ECV):** Derived in similar manner to EAS except that the mid-band spectrum is derived from the variance divided by the mean of the energy values in each frequency bin. ECV = $1 - H_c$, where H_c is the entropy of the distribution of mid-band spectrum composed of coefficients of variation.

Indices 12, 13 and 14 are 'ecological indices' which attempt an acoustic measure of species richness.

12. **Acoustic Complexity Index (ACI):** This summary index is derived from the spectral ACI index calculated as described above. It is the average of the mid-band ACI[f] values only. It is worth mentioning that this index is of decreasing usefulness as the duration of tbe recording segment is reduced because its value becomes increasingly subject to random effects which mask the contribution of biophony. In practice, we do not utilise ACI values for recording segments shorter than about 15-20 seconds.
13. **Cluster Count (CLS):** The number of distinct spectral clusters in the mid-frequency band of a one-minute segment of recording. Calculated as described in Section 6 below, "The Clustering of Spectra". This index is an attempt to measure the degree of internal acoustic structure, or spectral diversity, within the mid-band where bird calls predominate. It is expected that more bird species will generate greater vocal diversity which will increase the spectral cluster count.
14. **Spectral Peak Density (SPD)**: A measure of the number of cells in the mid-frequency band of a one-minute spectrogram that are identified as being local maxima. This index is calculated using the initial steps described for the spectral index, SPT.

    Step One: Detect the location of local maxima in the spectra, one frame at a time. A spectral cell is deemed a local maximum if its decibel value exceeds a threshold (we use 6 dB) and if its value is greater than the values in the two frequency bins on either side.

    Step Two: Count the total of spectrogram cells that are a local spectral maximum.

    Step Three: The SPD index equals the value obtained in step 2 divided by the number of frequency bins. As a warning, note that this index is not properly normalised to be independent of frame size and frame overlap.

Note that five of the above fourteen indices are derived from the waveform. Seven are derived from the mid-band of the spectrograms where bird calls are expected to predominate. This reflects the importance of the contribution of bird calls to the biophony at most recording sites of

ecological interest. If frogs are of particular interest, then the frequency bounds of the mid-band will need to be adjusted accordingly.

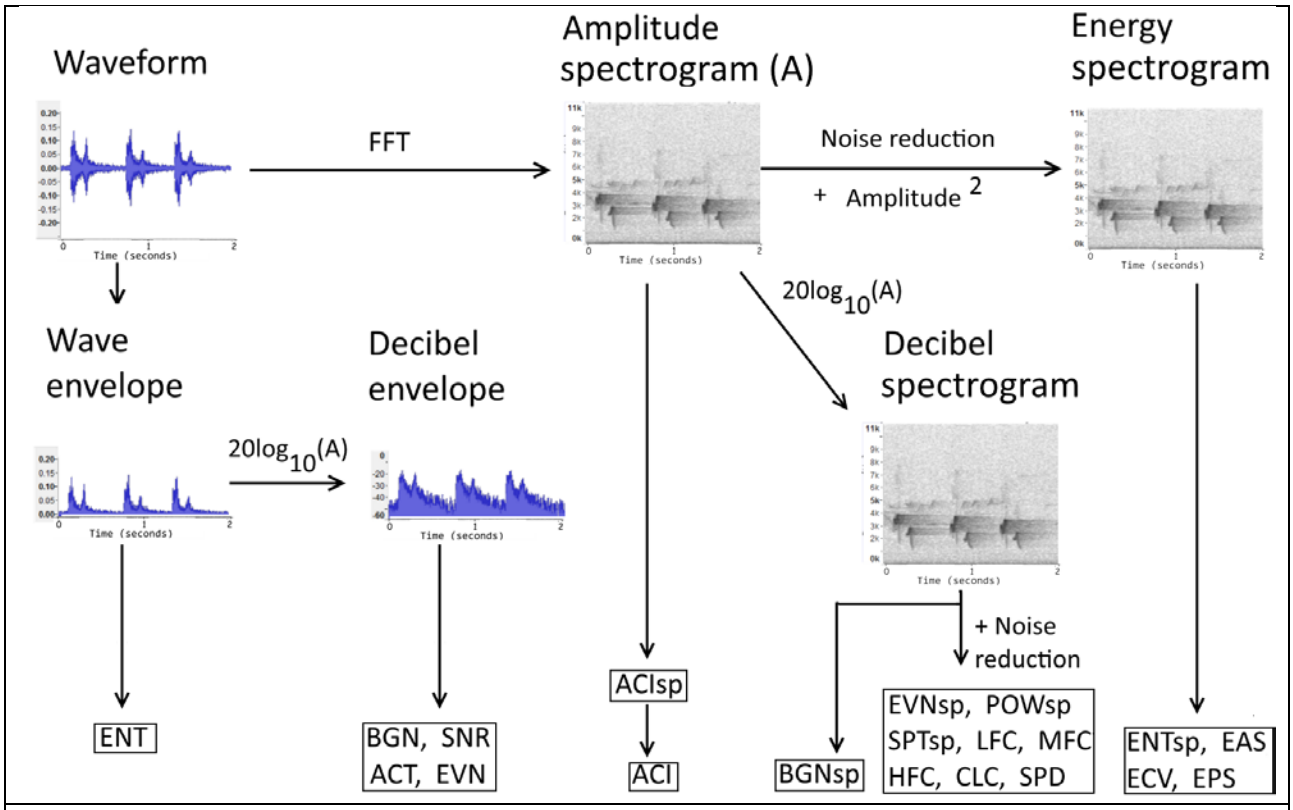

Figure 3: Overview of the sequence of calculating summary and spectral acoustic indices. This image was originally published in the PhD thesis of Yvonne Phillips, QUT 2018.

6. The Clustering of Spectra

The clustering algorithm used to obtain summary index CLS is a modification of ART1, an unsupervised iterative learning algorithm designed to cluster binary input vectors (Stephen Grossberg and Gail Carpenter, http://cns.bu.edu/Profiles/Grossberg/CarGro2003HBTNN2.pdf). This description of our implementation supersedes the previous QUT ePrints report: "An Algorithm to Cluster the Spectra in a Spectrogram", Michael Towsey, July 2013 located at http://eprints.qut.edu.au/61509/.

In our application, we cluster only the mid-frequency band (1000-8000 Hz), which typically contains much of the bird biophony. Given the above settings for sampling rate (22050), frame width (512) and frame overlap (zero), the mid-band contains 162 bins. Clustering is performed only after prior removal of background noise from the decibel spectrogram.

1. Reduce the length of each spectrum to one-third of its original length (from 162 to 54) by averaging values in consecutive, non-overlapping groups of three. This step is to reduce computational burden and to reduce spectral detail.
2. Convert each spectrum (length = 54) to a binary vector using a threshold = 6 dB.
3. Add the binary spectrum to a dataset of training spectra only if the sum of its non-zero elements exceeds one.
4. If the final number of training spectra is ≤ 8, return a cluster count = 0.

5. Execute the clustering algorithm on the remaining training set of spectra. To reduce computational burden, parameters are adjusted to achieve fast convergence.
    a. Initial count of seed clusters = 2. Selecting a low initial cluster count allows the cluster count to grow at each iteration. Setting a high number of initial clusters (> 10) results in a final cluster count that differs little from the initial cluster count.
    b. Initialise the seed clusters by random selection of one training instance for each cluster.
    c. Set a *vigilance* or *similarity* parameter, $v = 0.15$. This determines the minimum similarity a new instance must have with an existing cluster, if it is to be added to the cluster.
    d. Begin iterating the training data over the clusters.
        i. Calculate the similarity of every training spectrum to each cluster representative where:
        $Similarity = (\sum_i \text{logical-AND}[i]) / (\sum_i \text{Logical-OR}[i])$.
        The index *i* is over the elements in the two vectors whose similarity is being calculated. The logical-OR is a normalisation factor to prevent two input vectors with mostly zero values gaining a high similarity score. The similarity score will vary between 0.0 (having no non-zero elements in common) and 1.0 (perfect match).
        ii. Each training spectrum is assigned to the cluster to which it has greatest similarity. However, if the greatest similarity < *v*) then create a new cluster whose single representative is the new unmatched training spectrum. Increasing the vigilance, *v*, proliferates new clusters/categories. Note that the elements of the single vector representing a cluster never change. In other words, the first vector to create a new cluster remains that cluster representative unchanged. This is equivalent in the original Grossberg-Carpenter algorithm to setting the *momentum factor* = 1.0 which means weights never change.
        iii. At the end of each iteration (one pass of the training data), remove any cluster containing only one training instance. Because cluster representatives are selected from the training data, each cluster must have at least one member.
    e. The training data is iterated across the cluster representatives until *either* there is no change in the cluster-assignment of the training instances *or* the number of iterations = 20. In practice, this maximum number of iterations is seldom, if ever, reached with the above parameters values.
6. Prune the resulting list of spectral clusters by removing clusters that contain less than *N* members, where we set $N = 3$.

As is typical for clustering algorithms, the final cluster count and composition is sensitive to the choice of spectra that seed the clustering process and to other parameter choices. Nevertheless, it is generally indicative of the spectral diversity in a one-minute recording. The binary threshold of 6 dB above background noise limits detection to loud animal calls or, in the case of birds, calls relatively near to the microphone. It also reduces the number of resulting spectral clusters. In one particular recording sequence (obtained in a natural Australian bush setting), the maximum cluster count in any minute over a five-day period of continuous recording was 16. Reducing the dB threshold to (4 dB) increased the maximum cluster count to 50.

Once a set of spectral clusters has been determined for a given spectrogram mid-band, it is possible to assign a cluster ID to every ‘active’ spectrogram frame and then derive other indices from the sequence of cluster IDs. As one example, we calculate an index called 3-*gram Count* using the following steps: **Step 1:** Calculate the spectral clusters as above. **Step 2:** Each ‘active’

frame (as determined when calculating the summary index, **ACT**) in a one-minute segment is assigned to its nearest spectral cluster. **Step 3:** Calculate the number of 3-gram sequences that occur more than once. As with many other acoustic indices we have derived, 3-gram count has not stood out to be particularly useful in our applications.

7. Ridge indices

We identify spectral ridge components in a noise-reduced decibel spectrogram by convolving it with 5×5 masks, one mask for each ridge direction. Note that a constant frequency whistle would show in the spectrogram as a ridge having direction 0 radians (RHZ_{sp}, using the above abbreviations) and a broadband click would show as a ridge having direction $\pi/2$ (RVT_{sp}). We originally investigated two sets of masks: a set of four masks for the directions 0, $\pi/4$, $\pi/2$, and $3\pi/4$ radians and a set of eight masks for the directions 0, $\pi/8$, $\pi/4$, $3\pi/8$, $\pi/2$, $5\pi/8$, $3\pi/4$, and $7\pi/8$. Experiments with birdcall retrieval (Dong, et al., 2015) established that four directional masks produced more consistent results. We also experimented with 3×3 and 7×7 masks but found 5×5 masks to produce more consistent results (Dong, et al., 2015). Ridge direction scores were obtained by taking the dot-product of a zero-sum mask with the decibel values in a 5×5 neighbourhood of the spectrogram. As an example, the mask to detect horizontal ridges was,

```
{
    { -0.1, -0.1, -0.1, -0.1, -0.1 },
    { -0.1, -0.1, -0.1, -0.1, -0.1 },
    {  0.4,  0.4,  0.4,  0.4,  0.4 },
    { -0.1, -0.1, -0.1, -0.1, -0.1 },
    { -0.1, -0.1, -0.1, -0.1, -0.1 },
};
```

The four masks yielded four dot-product scores, s_h, s_p, s_v, s_n. To determine the presence of a ridge and its direction, dot-product scores of less than 0.1 were first set equal to zero and then the four dot-product scores were summed: $S_r = \Sigma_d s_d = s_h + s_p + s_v + s_n$, where *d* is an index over the four directions. Five spectrogram cells constituted a ridge only if the largest of the four ridge direction scores, s_{max}, exceeded a threshold, that is, $s_{max} > S_r/3$. When a ridge satisfied this condition, its five cells were assigned the score, s_{max}. When convolving the masks with the entire spectrogram, the direction and score for a spectrogram cell was only updated when the new value of s_d exceeded any pre-existing value of s_d. After convolution, the elements of the four direction vectors, RHZ_{sp}, RPS_{sp}, RVT_{sp}, RNG_{sp}, are assigned values equal to the sum of the corresponding s_d scores divided by the frame count.

REFERENCES

F. Briggs, B. Lakshminarayanan, L. Neal, X. Z. Fern, R. Raich, S. J. K. Hadley, et al., "Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach," The Journal of the Acoustical Society of America, vol. 131, pp. 4640–4650, 2012.

Depraetere M, Pavoine S, Jiguet F, Gasc A, Duvail S, Sueur J. “Monitoring animal diversity using acoustic indices: Implementation in a temperate woodland”. *Ecological Indicators*. 2012; **13**(1):46–54.

Xueyan Dong, Michael Towsey, Anthony Truskinger, Mark Cottman-Fields, Jinglan Zhang, Paul Roe. (2015) “Similarity-based birdcall retrieval from environmental audio”, *Ecological Informatics*, **29** pp66–76

Farina, A., Buscaino, G., Ceraulo, M., & Pieretti, N. (2014). The soundscape approach for the assessment and conservation of mediterranean landscapes: Principles and case studies. *Journal of Landscape Ecology*, **7**(1), 10–22. doi: 10.2478/jlecol-2014-0007

Farina, A., Pieretti, N., & Morganti, N. (2013). "Acoustic patterns of an invasive species: the Red-billed Leiothrix (Leiothrix lutea Scopoli 1786) in a Mediterranean shrubland". *Bioacoustics: The International Journal of Animal Sound and its Recording*, 22(3), 175-194. doi:10.1016/j.ecolind.2010.11.005

Lamel L.F. et al. (1981) *An improved endpoint detector for isolated word recognition*. IEEE Trans. ASSP ASSP-**29**: 777-785.

Pieretti, N., Farina, A., & Morri, D. (2011). A new methodology to infer the singing activity of an avian community: the acoustic complexity index (ACI). *Ecological Indicators*, **11**(3), 868–873. doi: 10.1016/j.ecolind.2010.11.005

Towsey, M., Zhang, L., Cottman-Fields, M., Wimmer, J., Zhang, J., & Roe, R. (2014). Visualization of long-duration acoustic recordings of the environment. Procedia Computer Science, 29, 703-712. doi:10.1016/j.procs.2014.05.063