Giovany Vega Viera

Review Rebuttal Letter

Below, we address (in bold) the comments and observations made by the reviewers of the manuscript “Audio segmentation using flattened local trimmed range for ecological acoustic space analysis”.

Reviewer 1 (Anonymous)

Basic reporting

My only "Basic Reporting" criticism given the guidelines is that the data has not been made clearly available to readers, which seems to go against PeerJ guidelines. The subsection "Data" in "DATA AND METHODOLOGY" presents no indication of where others can find the data sets used.

The data are published in figshare. In order for it to be properly referenced, I have cited the datasets used for the article in the “Data” subsection, along with bibliographic entries for each dataset.

Experimental design

No comments

Validity of the findings

No comments

Comments for the author

I found this to be a well written and interesting paper where the authors propose a new way to screen for generic "audio events" in acoustic recordings. Sometimes I felt that words were somewhat arbitrarily used, and provide further comments on this below.

While screening “recordings” for “audio events” is an interesting question, and one likely to become more and more important as recordings become more and more common, it is at odds with the perhaps more usual task found in the literature (that I am most familiar with), where one is trying to identify specific signals (e.g. a coqui sound, a bird song, a whale sound) in recordings. Therefore, I believe that the authors must bring forward that distinction and in particular must find a precise way to define what is an “acoustic event”. If that definition is unclear, how can we judge a given algorithm is useful or efficient? One needs to unambiguously define what are the “audio events” (i.e. the signal) one needs to detect in order to quantify true positives and false positives, etc. This must be done from the start, around line 43 or before when the term is first used.

A definition of an acoustic event has been provided in the article. Acoustic event is now used throughout all the document, rather than using both acoustic and audio event in the previous version.

When one reads the abstract sentence “Our goal is to develop an algorithm that is not sensitive to noise, does not need any prior training data and works with any type of audio event.” or the introduction sentence, even more optimistic, “What is needed is an algorithm that works for any recording, is not targeted to a specific type of audio event, does not need any prior training data, is not sensitive to noise, is fast and requires as little user intervention as possible., one has to wonder, what is the catch? One usually says there are no free lunches in statistics. So what is the price to pay? And the price you pay is specificity. When you choose to detect “audio events” without clearly defining them a priori, then you actually lose the ability of doing many of the analysis that you might do when you target a specific signal. This is fine if is intended, but that must be explicitly stated so, and eventually discuss clearly the notion that this is a precursor step in other more elaborate and fine-tuned procedures.

We agree and the non-specificity of the method was stated explicitly in the Introduction. We also add that the method’s intent is to work as a data reduction procedure (i.e. it selects the interesting parts in the recordings).

This is to me an important point: I thought that all of these discussions regarding the distribution of variables when events were detected is not really possible unless we know the distribution of the variables available for the entire duration of the recordings was available. It would mean very different things to have many events at 18:00 hours if most of the recordings were at 18:00 hours versus if few recordings were at 18:00s! (see e.g. discussions around line 292-294).

The data in the sites dataset was sampled so that it is evenly distributed over the whole day. We have added a sentence in the “Data” subsection of “Data and Methodology” to clarify this point.

More detailed/specific comments that might be used to improve the paper follow below.

Line 10 – 11 – I would say that “thousands of recordings” is wording to be avoided. Recording is a meaningless unit. It should be easy to replace by something that is unambiguous to readers. Same on line 48, “recordings” again used as a unit. This is not a sensible unit. Please check the entire text for the use of the word, as sometimes it makes sense, other not. In particular when you describe the data, again “recordings” is not a useful description. I know you are thinking about Arbimon, but this must be general. How long is a recording say?

Arbimon recordings are usually one minute long. we have verified the usage of phrases with the word “recording”. In cases where recording is used as a unit of audio we have either rewritten it as minutes of recorded audio or qualified it as one-minute recordings. In other cases, such as when recording is used as a bearer of audio data, it has been left as is.

Line 46-47 – the “easily draw a boundary around any audio event” is strictly not true, especially at low signal to noise ratios, or when a given frequency band is saturated and so many overlapping events occur continuously for a long time period. Please reword.

You’re correct. we have reworded it as “draw a boundary around an acoustic event”.

While I understand the purpose of the last sentence in the introduction, I believe it includes several bits of information that should belong in the methods. It makes sense to end the discussion with laying the paper that lies ahead, but one should avoid technical details like “2051 manually labelled audio events” or “20 recordings”. One gets confused also because in the text you say this is the workflow of the article, the figure states it’s the workflow of the AED methodology. These can be naturally closely related, but are not one and the same.

We have removed technical details such as the number of manually labeled audio events or the number of recordings on each step. We have also rewritten the caption in Figure 1 so it reflects better the workflow in the article.

Figure 2 legend – to me Yen threshold means nothing, and legends should be self-explanatory. I suggest removing as it’s mostly diversionary here, this is in the text after line 153.

We have removed the “Yen threshold” parenthesis.

Line 89 – here you refer fig 2C, but fig 2B was not mentioned yet. It would flow better if the description matched the figure order.

We have moved the figure reference to the next sentence, so that it can also include fig 2B.

Property A1. – what is tau? And eta? I mean, I know, but rigorously it would be useful to define these.

We have included a formal definition for the usage of S(t, f), which defines tau and eta as the bounds of t and f respectively.

Line 118 – I believe the notation needs tweaking, as S_db(t,f) has no I which is what it needs to be summed over?

Corrected. summation index was t, not i.

Line 103-109 – I think you cannot extend this to infinity. Would the algorithm not break apart if an entire recording had a given frequency band saturated? Discuss, or change wording, please.

Depends really on the nature of the recording. If what gets recorded is essentially stationary (pretty much the same noises most of the time), then it wouldn't necessarily break. However, as soon as any of the assumptions change then it would definitely degrade.
If a band gets saturated, then the b(f) estimation part would degrade, since some of the saturation (the part that's less than the band median) would be confused with a constant noise source and removed. This is discussed in the Discussion section.

Line 124 – "valuation"?

This has been removed, and the sentence has been rewritten.

Line 132 – Surely this is over a given time frame in practice? What is the time frame considered? Will you discuss the sensitivity to different time frames considered?

In practice, the timeframe is the whole recording. In theory it should be enough time to guarantee that epsilon is stationary.

Line 133 There's something missing as like an "at least" before "1-\rho(f) proportion", right?

\rho(f) is the proportion of samples that have a localized energy process within them and1-\rho(f) exactly is the opposite. We have added the interpretation of \rho(f) so it is clearer.

Lines 141-142, I find this confusing… would it not be clearer if you delete just ”by estimating it as” and replace by “:”

We moved the equation to the next sentence and removed the “by estimating it as” part.

Just before line 146 – he “r>0” is this in time, frequency, both?

We have specified that it is in time and frequency.

Line 146 – Equations should read just as text. Therefore, the “Where” must be “where” and not indented, and there’s no dot after the previous equation. Same in line 201. Check remaining instances.

We have revised all equations and added the necessary punctuations.

Line 148 – reword the arbitrary “the estimator should have a small response”. What are you referring to? What is an estimator response?

The response of the estimator is just the values that the estimator produces for the given regions. The phrase has been reworded to “the estimator would give small values”.

Line 155 – I understand what you mean, but clarify “image values”, since there’s strictly no images here.

The phrase “image values” has been changed to “spectrogram values”.

Line 156 – “Th” should be “T”?

Th and T are not the same variable, T is an independent variable in the entropic correlation TC(T), while Th is the value that maximizes it (i.e. the threshold). This subsection has been reworded to make the distinction more explicit, also the variable

names have changed to t being the independent variable in TC(t) and t-hat being the selected threshold.

Line 157 – the contiguous here refers to both time and frequency?

Yes, we have rephrased it as “adjacent (t, f) coordinates” to make it clearer.

Line 168 – not sure why you need the descriptor “the sites dataset”? The first data set was also collected on sites…

While the first dataset is collected from different sites, the purpose is to validate the method and thus, the site distinction is not needed. We named the second dataset the sites dataset because the recordings are processed site-wise.

Line 175 – as I said, here’s a good example. You set to find “events” here. What is an event? Is this a circular definition to some extent, since events are sounds you are able to classify as events??? In particular, how these relate e.g. to the 3 types of sound you mention in the first paragraph of the introduction?

An acoustic event is, basically, anything that stands out from an audio recording. The sources of these events are those three types of sounds. Defining what “stands out” is the tricky part, since our most basic of reference is the ability to detect changes in either the raw audio or any other transform we apply to it.

Line 179 – 21 by 21 and alpha=5. These are fundamental details. Are these values optimal in any way? Why? What is the sensitivity of the method to changing them? What are recommendations for users?

We have included a paragraph about the parameter choice in the discussion. Essentially they are ad-hoc parameters that follow certain structure.

Line 188 and 191 (2 times) – I think the word “automatically” needs to be added for clarity, e.g. “were detected over the total count” is “were automatically detected over the total count”

Added the word “automatically” to these phrases.

Line 187-192 – I’d introduce the wording true and false positives here, and note explicitly this is a typical confusion 2 by 2 matrix but one of the cells is absent (i.e. there are no true negatives).

The true and false positive wording has been included, along with formulas for sensitivity and positive predictive value.

Line after 200 – you need a more rigorous wording. “To measure the degree of separation of each variable on the audio event density”… there is no such characteristic as “variable separation”… what do you mean exactly?

We have rewritten this, emphasizing the information content of the variable.

Line 201 i-th should be i^{th} (superscript)

Corrected i-th to i^{th}

Line 203 – “The 2-variable marginal distributions” reword to bivariate?

Changed all instances of “2-variable marginal” to “bivariate” in the text

Line 205 – why this one chosen? Just as an example? If so say so.

Yes, it is an example of what can be done with the method. we have said so explicitly now.

Line 208 – words missing or plural vs singular mistake

The text has been thoroughly edited to remove these mistakes.

Table 1 legend – last sentence – you say “arbitrary number of false negative examples”. Don’t you mean “true negatives”?

Corrected the term to “true negative”.

Figure 3 legend – need to label the columns, since each column is a different type of image, right?

We labeled the columns with the respective image type.

Figure 4,5,6 – you need at the very least to state explicitly that dark is less and white is more, but a scale would be helpful!

A color bar has been added to the side of the plot matrix.

Figure 6 – plot y_max by tod. What is the reason for the weird vertical bands? It's hard for me to imagine what would create these tod "discontinuities"?

The "discontinuities" are produced by aggregating the recordings into 24-1 hour bins.

Line 228 – so does "cov", in fact with a larger H value than _max. Why do you not mention it

"**Cov" Is now mentioned in the results.**

Line 236-237 – explanation? Is this a feature or an artefact? If you raised the question you can't leave it unanswered.

This has been added to the discussion.

Line 243 – seems inconsistent to me to do this "separation" (and I do not like the word as I said above) only visually in 2d but with an H statistic in 1d, why this choice?

Separation has been replaced with information content using joint entropy

Line 247 and figure 7 – 6? Which? I can't see this. Please mark them all in figure 7.

All the areas have been marked with a rectangle. The Area of Interest is pointed to by an arrow.

Figure 7 legend – "Close-up on the first area"… why the first.

The word "first" has now been removed.

Line 251 and several others – all latin names must be italicized, including in references.

All species names have been italicized.

Line 262-264 – Discuss problems with intense chorus on a given noise band.

In recordings with an intense chorus, rho(f) would be >= .5, and thus the theorem would not be true. Regardless, energy peaks within the chorus would still be preserved. Depending on their size, they could appear in the flattened spectrogram, and still be possibly detected. Added a paragraph with this reasoning in the discussion.

Line 281 –use of word recordings here is inconsistent (cf. with say line 168-171), you say 6, these are presumably what you referred as dataset?

The six recording referred to in line 281 are the recordings from the validation dataset as part of the discussion of the FLTR validation results. The recordings referred to in lines 168-171 are the recordings from the sites dataset, which is used in the FLTR application step.

Line 282 – remove the two instances of the word “any”, not useful.. but this needs added clarification, not sure what is meant here?

The two instances of “any” have been removed. The paragraph is saying that although our validation dataset is small, some statistical bias in the results is to be expected, but that we used recordings from different environments in the test to reduce this bias.

Lines 300-303 – I’d like to see comments regarding whether these have also been missed sometimes or not?

As this was a sample of the detections from a region of interest in an application of the algorithm, we only focused on cataloguing the detected events. However we can estimate the missed events with the computed sensitivity and positive predictive value. We added a paragraph indicating this.

Line 334 – “University Press”

“university” was capitalized.

Line 344- incomplete ref?

A new reference is included.

Line 347 – “Conference on…”- incomplete

Conference name was included before the “Conference on” text. Corrected the sentence.

Line 355 – “pages”???

Not sure what the error is. Revised references to include complete information.

Line 366 – species name needs italics

***Eleutherodactylus* written in italics.**

Reviewer 2 (Michael Towsey)

Basic reporting

The English language is good.

The mathematical description has not helped the presentation. It took time to understand and was not adding much more than could have been said quickly in words. In particular, symbols are used which are not defined - e.g. eta, n, tau. The numeral '1' is used for the indicator function but it is not stated as an indicator function and is doubly confusing because bold or blackboard-bold font is not used.

The symbols were defined, capital I is now used instead of 1 for the indicator (support) function. The mathematical description was revised and simplified a bit.

T in the equation just above line 154 is not defined. I had to go on-line to see what it was. It is the threshold which is later referred using the letter 'Th". All this suggests that not a lot of care was put into proof reading.

T is the variable used in the entropic correlation TC(T), and Th is the value maximizing the correlation. we have changed the variables to t as the variable for TC(t), v as the index in the sums in TC(t), and t-hat for the threshold. we have also made explicit the relation between TC(t) and t-hat.

The images are just of sufficient resolution to support the text.

Figures 4, 5 and 6 are puzzling in that the images at top and bottom of the left column have different time scales. Also it appears from the top left image that more events happen at night than in the day. The reader needs a lot more help to understand these figures.

The second time scale has been removed, now all tod figures have same time scale. Also, additional explanation is included.

Experimental design

The method for subtracting base-line value is valid although it is based on an important assumption. The signal model is not that different from an additive noise model. The authors state that they do not make an assumption about the distribution of the noise, symbol epsilon in text, and this is a nice feature. (Although later the use of 5 percentile tails to establish cutoffs implies that something like gaussian noise is assumed).

The nicest feature of the method is that used to calculate the Range estimator and the use of entropic correlation. I am not aware of this being done elsewhere and is the main interesting result of the paper.

The authors report their accuracy based on overlap of observed rectangle with predicted rectangle. This criterion is far too liberal because even a slight overlap can lead to a correct prediction for the wrong reasons. I would suggest that at least a 50% overlap is required which is indeed the case in some of their images.

We added a second comparison test which is measured as the proportion of the area of the intersection over the area of the union. In this test, a hit is decided when this ratio is greater than .25

Validity of the findings

The authors imply that there are few thresholds or critical parameters that must be tuned for their system. However the use of the 21x21 window size is surely important and must have been determined by trial and error.

We agree that the selection of the window size was not addressed, we now include an acoustic interpretation to the window size in the Methodology / FLTR Validation section.

If an event was too large in area, this window would leave "holes" in the event due to the way they calculate the range estimator. In general I felt the authors were too uncritical of their method.

This method is basically finding the borders of an acoustic event. While this does leave holes for big events it is just a matter of filling the interior of the event. Text has been added in the Theory section (Thresholding subsection) to include this detail.

The important assumption of
rho(f) < 0.5 is hidden in extensive and not very helpful mathematics. It is a reasonable assumption - other methods have to make similar assumptions but the authors claim some superiority for their method.

The mathematics try to formalize a reason on why the method should work. The theorem and its proof have been simplified a bit. And also, the requirement for rho(f) < 0.5 has been restated as an extra assumption (A3 is now A4, and this is A3).

Comments for the author

This could be a nice paper. The method of using the Range estimator is very nice and something I am sure others will emulate when the paper is published. However the paper is inadequate in three respects:
1) The authors have been too uncritical in promoting the advantages of their method.

The document was reviewed and comments towards this issue were made.

2) The estimates of accuracy are based on a very easy success criterion.

A second (more vigorous) success criterion was added.

3) There is no comparison with another method. I totally agree that a fixed threshold technique is not useful but there are better techniques to compare their method with.

We did not analyze results for other methods that only detect acoustic events.