Taylor & Francis
Taylor & Francis Group

Assessing the effect of sound file compression and background noise on measures of acoustic signal structure

Marcelo Araya-Salas[a,b], Grace Smith-Vidaurre[c] and Michael Webster[a]

[a]Laboratory of Ornithology, Cornell University, Ithaca, NY, USA; [b]Escuela de Biología, Universidad de Costa Rica, San José, CA, USA; [c]Department of Biology, New Mexico State University, Las Cruces, NM, USA

ABSTRACT
The study of animal acoustic signals is a central tool for many fields in ecology and evolution, but the diversity of analytical methods and sources of animal sound recordings poses important challenges for carrying out robust acoustic analyses. Sound file compression and background noise may both affect acoustic analysis, although little attention has been paid to their respective effects. We evaluated the effect of these factors by assessing the systematic deviation (i.e. bias) and measurement error (i.e. precision) that they generate on spectrographic parameters and two (dis)similarity methods (dynamic time warping on frequency contours and cross-correlation), which represent the most common methods currently used for quantitative characterization of acoustic signals. Measurements were taken across a wide range of signals from a diverse group of bird species, and compared between uncompressed files and decompressed files obtained from mp3-encoded files generated using the two most common mp3 encoders (Fraunhofer and LAME). Measurements were also compared across a range of synthetically-generated background noise levels. Compression did not significantly bias any of the acoustic or similarity measurements. However, the precision of acoustic parameters representing single extreme values (e.g. peak frequency) as well as dynamic time warping distances, was strongly affected by compression. High background noise biased most energy distribution-related parameters (e.g. spectral entropy) and affected the precision of most acoustic parameters and dynamic time warping. Overall, compression and background noise did have considerable effects on acoustic analyses. We provide recommendations to avoid potential pitfalls and maximize the information that can be reliably obtained.

ARTICLE HISTORY
Received 21 July 2017
Accepted 1 October 2017

KEYWORDS
Sound file compression; mp3; acoustic analysis; dynamic-time warping; cross-correlation; spectrographic parameters

Introduction

Animal acoustic signals have been an important study system for a wide variety of fields in ecology and evolution (Bradbury and Vehrencamp 2011). The particular usefulness of acoustic analyses resides in our ability to register acoustic signals with high fidelity and conduct very detailed and increasingly elaborated measures to characterize and compare

CONTACT Marcelo Araya-Salas ✉ marceloa27@gmail.com, araya-salas@cornell.edu

The supplemental data for this article can be accessed at https://doi.org/10.1080/09524622.2017.1396498.

their structure (Sueur, Aubin, and Simonis 2008; Tchernichovski et al. 2000; Lachlan 2007; Charif et al. 2010; Araya-Salas and Smith-Vidaurre 2017). In addition, the growing availability of recordings in acoustic libraries provides an unprecedented opportunity to study animal acoustic signals at large temporal, geographic and taxonomic scales (Araya-Salas 2012; Depraetere et al. 2012; Medina-García et al. 2015; Mason et al. 2016).

The diversity of both analytical methods and sources of animal sound recordings, however, poses potential challenges for carrying out robust acoustic analyses. The quality of recordings can be highly variable among sources. For instance, songs registered by an automatic recording device would typically be noisier than those taken by a field recorder from a focal individual. Even within focal individual recordings, the context of the recording, such as distance to target, habitat structure and the proficiency of the recordist, can affect the amplitude at which signals are registered relative to the background noise (Zollinger et al. 2012). Environmental noise can mask acoustic signals, potentially affecting the information obtained by receivers; it is considered a fundamental evolutionary force shaping animal acoustic signals (Morton 1975; Boncoraglio and Saino 2007). Despite being a ubiquitous and critical factor for acoustic communication, we are just starting to understand how noise can affect the ways in which acoustic signal structure is measured (Rempel et al. 2005; Ríos-Chelén et al. 2016, 2017; Brumm et al. 2017).

Acoustic analyses might also be affected by the methods used for recording and storing animal vocalizations. For example, recordings are sometimes collected or preserved in compressed formats, leading to an irreversible information loss (i.e. 'lossy' compression). Nonetheless, the use of compressed sound files is widespread in bioacoustic research (Botero et al. 2009; Weir et al. 2012; Doolittle and Brumm 2013; Gonzalez-Voyer et al. 2013; Mason et al. 2014; Medina-García et al. 2015; Pegan et al. 2015; Araya-Salas and Smith-Vidaurre 2017; Kaluthota et al. 2016). Sound files are typically compressed in the MPEG Audio Layer III lossy format (so called 'mp3'). This format was designed to reduce the size of sound files by removing information on acoustic features that were unlikely to be detected by the human listeners, based on human acoustic perceptual biases (Sterne 2012). Indeed, mp3 files can significantly shrink the size of audio recording file (Sterne 2012). Surprisingly, despite the wide use of mp3-compressed recordings, how the loss of information affects acoustic measurements has not been formally evaluated, although indirect evidence suggests they could have a considerable effect (Medina-García et al. 2015; Towsey et al. 2016).

In this study, we evaluated the effect of two potential confounding factors in bioacoustics research: sound file compression and background noise. We did so by assessing systematic deviations (i.e. bias) and measurement variability (i.e. precision) of commonly used acoustic measures across a wide range of signals from a diverse group of bird species. Acoustic measurements were conducted on original (uncompressed) .wav files and on .wav files derived from mp3-compressed files (hereafter uncompressed and compressed files, respectively), as well as across a range of synthetically-generated background noise levels. Our goal was to provide guidance for future work, helping researchers to avoid potential pitfalls and maximize the information that can be extracted from field recordings.

Methods

We obtained 840 recordings from sound libraries and personal collections (Table S1). The recordings represented 245 bird species, from 31 families and 11 orders, although most recordings came from hummingbirds (63%), parrots (27%), and songbirds (7%). Original sound files were recorded at a sample rate of either 44.1 or 48 kHz, and a bit depth of 16 or 24 bits in mono or stereo format. All files were converted to 44.1 kHz, 16 bits, single channel (mono) wave files prior to analyses using the PCM WAVE format in the software Audacity 2.1.2 (Audacity-Team 2014).

We used either visual inspection (i.e. directly clicking on the spectrograms) or automatic detection to determine the start and end time. Visual inspection was also used for defining frequency ranges. Automatic detections were made based on amplitude thresholds and were visually checked and corrected when necessary. We also removed the lowest quality signals by excluding those with a signal-to-noise ratio (hereafter 'SNR') lower than 2 dB, as those signals were difficult to distinguish from the background in the spectrograms (Figure 1). SNR was calculated as follows: SNR = (rms(S) − rms(N))/ rms(N), where S is the amplitude envelope of the signal, N is the amplitude envelope of the background noise 50 ms immediately before and after the signal and *rms* is the root mean square. SNR was then converted to decibels (SNR_{dB} = 20 * log10(SNR)). The final data-set contained 2642 signals from the 840 recordings described above (an average of 3.1 signals per recording).

Although several mp3 encoders have been developed, two are currently the most widely used (Berman 2015): LAME (free software, http://lame.sourceforge.net) and Fraunhofer (proprietary software, Fraunhofer Institut Integrierte Schaltungen, Germany; http://www.iis.fraunhofer.de). These two encoders are found in the most popular software packages for audio file manipulation (e.g. Adobe Audition, Audacity, iTunes, Windows Media player, SoX), and therefore the most likely to be used by recordists converting to mp3 formats. Both encoders can produce mp3 files using two encoding methods: constant bit rate and variable bit rate (hereafter CBR and VBR, respectively). As their names suggest, CBR uses the same bit rate across the entire sound file while VBR increases the rate during more acoustically complex sections of sound files. We used both encoders (LAME and Fraunhofer) and encoding methods (CBR and VBR, for a total of four treatments) to generate mp3-compressed files. That is, we converted the original wave files to 44.1 kHz, 16 bit and 128 kbps mp3 format in both CBR and VBR using both LAME and Fraunhofer encoders, and then

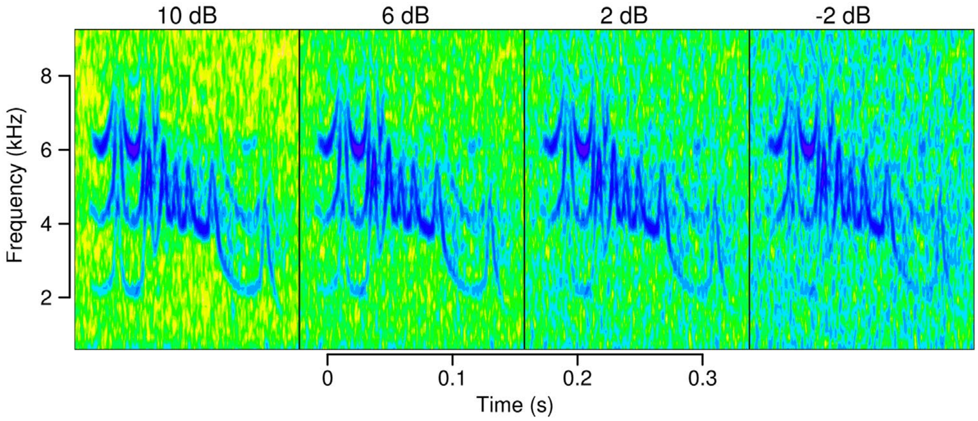

Figure 1. Spectrograms of a long-billed hermit (*Phaethornis longirostris*) song at decreasing values of signal-to-noise ratio (10 to −2 dB, labels on top).
Note: Background noise was synthetically generated using the R package 'Seewave' (Sueur, Aubin, and Simonis 2008).

decompressed them back into 44.1 kHz 16 bit wave format (through the same encoder used for compressing them) for analyses, because acoustic analysis software typically requires uncompressed sound files. We used the most recent versions of both encoders. The LAME encoder (version 3.99.5) was used through the UNIX command-line interface, and the Fraunhofer encoder was used through Adobe Audition CS6 (version 5.0, Adobe Systems Inc., San Jose, CA, USA).

We measured 12 acoustic parameters on both uncompressed and compressed files: mean frequency, mean dominant frequency, minimum dominant frequency, maximum dominant frequency, peak frequency, frequency range, modulation index, 1st quartile frequency, 3rd quartile frequency, interquartile range, skewness and spectral entropy (Figure 2). A full description of the acoustic parameters is provided in Table 1. Parameters were measured using the 'specan' function from the R package 'warbleR' (Araya-Salas and Smith-Vidaurre 2017). Although several other parameters can be measured by this function, we choose a subset representing the most commonly used acoustic measurements: parameters calculated across the entire signal (e.g. mean dominant frequency, spectral entropy), representing a single extreme value (e.g. maximum frequency, peak frequency) or derived from those (e.g. dominant frequency range, modulation index). Dominant frequencies were measured as the highest amplitude value within the predefined frequency range for every time window in the spectrogram. The selected parameters are also measured by the most common bioacoustic analysis software, making our findings relevant to most researchers working in this field. The analyses were conducted under a 512-point fast-Fourier transformation window length (for a 11.3 ms time resolution), 90% window overlap with a 'hanning' window function, and a 10% amplitude threshold for detecting dominant frequencies (relative to the maximum amplitude in the signal).

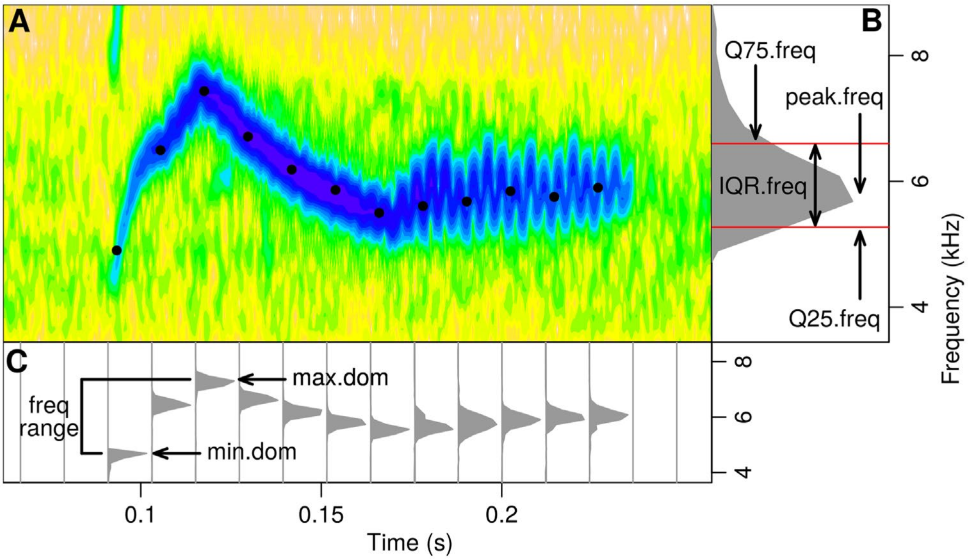

Figure 2. Schematic diagram for seven acoustic parameters (out of 12) measured in this study.
Notes: (A) Spectrogram of a White-chinned Sapphire (*Hylocharis cyanus*) vocalization. The black dots represent the dominant frequency measured across the signal. (B) Frequency spectrum of the signal (i.e. relative amplitude of the frequencies; amplitude on the x-axis). (C) Frequency spectrum 'slices' across the signal, showing the dominant frequency for each time window. Dominant frequency is also shown as black dots on the spectrogram (A). Parameter names and descriptions are detailed in Table 1.

Table 1. Description of the acoustic parameters and pairwise similarity measurements evaluated in this study.

Parameter type	Parameter	Abbreviated name	Description	Equivalent parameters in other software
Average frequency	Mean frequency	Mean.freq	Weighted average of frequency by amplitude (in kHz)	Avisoft: mean spectrum of entire element
	Mean dominant frequency	Mean.dom	Average of regularly spaced dominant frequency measures across the signal (in kHz)	Avisoft: mean param. of entire element Raven: mean of peak frequency contour
Single value frequency	Minimum dominant frequency	Min.dom	Minimum of dominant frequency measured across the signal (in kHz)	Avisoft: min param. of entire element Raven: peak frequency contour min frequency
	Maximum dominant frequency	Max.dom	Maximum of dominant frequency measured across the signal (in kHz)	Avisoft: max param. of entire element Raven: peak frequency contour max frequency
	Peak frequency	Peak.freq	Frequency with the highest energy (in kHz)	Avisoft: peak frequency Raven: peak frequency
Derived frequency measurement	Dominant frequency range	Freq.range	Range of dominant frequency measured across the signal (in kHz)	Avisoft: difference between max param. of entire element and min param. of entire element Raven: difference between peak frequency contour max frequency and peak frequency contour m frequency
	Modulation index	Mod.indx	Ratio of the cumulative absolute differences between adjacent measurements of dominant frequencies to the dominant frequency range (1 to ∞)	Avisoft: analogous to relative stddev of entire element
Amplitude distribution	Fist quartile frequency	Q25.freq	The frequency at which the signal is divided in two frequency intervals of 25% and 75% energy respectively (in kHz)	Avisoft: lower quartile (25%) Raven: 1st quartile frequency
	Third quartile frequency	Q75.freq	The frequency at which the signal is divided in two frequency intervals of 75% and 25% energy, respectively (in kHz)	Avisoft: upper quartile (75%) Raven: 3rd quartile frequency
	Interquartile frequency	IQR.freq	Frequency range between 'freq.Q25' and 'freq.Q75' (in kHz)	Raven: IQR bandwith
	Spectral skewness	Skewness	Asymmetry of the spectrum (index: −1 to 1)	
	Spectral entropy	Sp.entropy	Energy distribution of the frequency spectrum. Pure tone ~0, noisy ~1 (index: 0 to 1)	Avisoft: Wiener entropy Raven: aggregated entropy
Pairwise similarity	Spectrographic cross-correlation	Cross.corr	Maximum pairwise correlation of amplitude matrices in the time-frequency space (index: −1 to 1)	Avisoft: template cross-correlation Raven: spectrogram correlation
	Dynamic time warping distance	DTW.dist	Pairwise dynamic time warping distance of dominant frequency contours (0 to ∞)	Luscinia: time warping analysis

Notes: Abbreviated names correspond to the ones used in figures. Parameters were measured using the R packages 'seewave' (Sueur, Aubin and Simonis, 2008) and 'warbleR' (Araya-Salas and Smith-Vidaurre 2017). Dominant frequencies were measured as the highest amplitude value within the predefined frequency range for every time window in the spectrogram (see Figure 2). Equivalent parameters from Raven Pro 1.5 (Charif et al. 2010), Avisoft 5.2 (Specht 2004) and Luscinia 2.16 (Lachlan 2007) are also listed.

We also evaluated the effect of compression on two pairwise acoustic (dis)similarity methods: dynamic time warping and spectrographic cross-correlation (hereafter 'cross-correlation'). Briefly, dynamic time warping measures the alignment of two numeric sequences as a unitless distance (e.g. dissimilarity) that penalizes mismatches between the sequences. This method has been successfully used to compare frequency contours of bird songs (e.g. the dominant frequency values across signals, Kogan and Margoliash 1998). Cross-correlation, on the other hand, compares the whole matrix of amplitude values in the bi-dimensional time-frequency space of the signals. The method 'slides' one spectrogram over the other calculating a correlation of the amplitude values at each step (Clark et al. 1987). The peak correlation coefficient is taken as a measure of similarity between the signals. This method also has been widely used in bioacoustic research (Cortopassi and Bradbury 2000; Cramer 2013).

For acoustic (dis)similarity analyses, pairwise comparisons were conducted on a subset of long-billed hermit (*Phaethornis longirostris*) recordings. This is a lekking species in which individual males have a song repertoire of only one song, songs are constantly repeated in a singing bout with little variation among renditions, song types are shared by subgroups of males within leks, and several song types can be found in a lek (Stiles and Wolf 1979; Araya-Salas and Wright 2013). Songs are typically composed of frequency-modulated pure tones with moderate harmonic structure, sometimes combined with broadband sounds (Figure 1). Variation in song structure in this species can be easily distinguished by visual inspection of spectrograms (Araya-Salas and Wright 2013). Thus, song similarity within and between singing neighbourhoods is expected to cover a wide range of similarities (from very different to very similar) on a set of signals with similar durations and frequency ranges, in which the performance of acoustic similarity methods can be properly assessed. Songs were recorded from individuals at seven leks at La Selva Biological Station between 2008 and 2016 (Araya-Salas and Wright 2013). We paired songs from the same lek to ensure that comparisons contained songs from both the same and different song types. Each song was used in a single comparison to avoid pseudoreplication. Dynamic time warping was conducted using the 'dfDTW' function from the R package 'warbleR' (Araya-Salas and Smith-Vidaurre 2017), which extracts the dominant frequency contours as time series and returns a pairwise distance matrix. This function uses a smoothing spline to interpolate frequency values which is used to obtained frequency contours of equal length for all signals, regardless of their duration. The function also interpolates the frequency value in signal sections in which the dominant frequency did not exceed the amplitude threshold (10% in our case). We set the function to obtained 30 regularly spaced dominant frequency measures across signals. Frequency contours were z-transformed to 'remove' differences in absolute frequency and focus the comparison on contour shapes. Cross-correlation was conducted with the 'xcorr' function from 'warbleR', using Pearson product-moment coefficient to measure signal similarity. Both dynamic time warping and cross-correlation were conducted using a 300-point fast-Fourier transformation window length (6.8 ms time resolution, 147 Hz frequency resolution), and 90% window overlap with a 'hanning' window function. The R packages 'tuneR' (Ligges, Krey, Mersmann, and Schnackenberg 2014), and 'Seewave' (Sueur, Aubin, and Simonis 2008) were used for importing sound files in the R environment and generating spectrograms, respectively.

The effect of compression on acoustic structure metrics can be described in terms of the agreement between the same metrics measured in original uncompressed files and compressed files. Several statistical tools have been used to evaluate measurement agreement. However, the Bland-Altman analysis (Bland and Altman 1986) is generally regarded as the most reliable quantitative estimation for measurement agreement (Hanneman 2010). Under this analysis the accuracy is estimated as the overall mean difference between its measurements and those of an established method (so-called 'method bias'). In addition, the variation around mean differences (e.g. confidence intervals) provides information about the precision of the method. In our case the bias on the measurements derived from compressed sound files was calculated as the difference between the acoustic parameters of original uncompressed files and those of compressed files. The precision was estimated as the 95% confidence intervals (1.96 * SD, 'limits of agreement' in Bland-Altman terminology).

The significance of a method bias should be interpreted relative to the expected measurement error and the magnitude of the differences that have been detected in previous studies for the given acoustic parameters (Hanneman 2010). Biases within the expected error range should be dismissed as these may arise as an artefact of measurement uncertainty. On the other hand, if bias surpasses commonly reported differences, such deviations could impair our ability to detect significant effects. In our case, the error of a frequency measurement is a function of the precision in the frequency domain, which is defined by the ratio of the sampling rate to the window length (Beecher 1988). For a 44.1 kHz sound file with a 512-point window length (as in this study) the associated error is 86 Hz. Variation in song frequencies between populations from different habitats are usually reported when evaluating predictions of the acoustic adaption hypothesis (i.e. the degree to which the structure of animal sounds has evolved to optimize transmission in particular habitats; Morton 1975) and thus provide an a priori threshold for unacceptable frequency biases or confidence intervals. The most recent review on this topic gives average effect sizes for frequency shifts due to habitat selection (Boncoraglio and Saino 2007), and reported the lowest average effect size of 160 Hz (Table 1, maximum frequency). We used this value as a conservative threshold for detecting frequency biases of possible concern. Note that the lack of similar estimates for non-frequency parameters (to the best of our knowledge) prevents the use of bias and precision thresholds for skewness, spectral entropy, modulation index and pairwise similarity measures such as dynamic time warping and cross-correlation. Hence, bias and precision for these parameters are shown as a percentage of the overall observed parameter range for the original uncompressed files to provide some estimate of the effect of compression.

We also estimated the agreement between measurements on uncompressed and compressed files, using the intraclass correlation coefficient. This test evaluates the degree to which measurements on uncompressed and compressed sound files provide the same results (i.e. repeatability). The test statistic provides information on measurement agreement (0 ~ no agreement, 1 ~ total agreement) similar to the confidence intervals of the Bland-Altman bias, although as mentioned above, the latter provides a more robust estimator of method agreement (Bland and Altman 1986; Müller and Büttner 1994; Hanneman 2010). Nevertheless, we included repeatability as a complementary measure of agreement given that it can be directly compared across methods with different units (or unitless), including pairwise similarity measures. Repeatability was measured using the R package 'ICC' (Wolak et al. 2012).

To examine the effects of background noise, we built an R routine to adjust the background noise level on (uncompressed) sound files using the R package 'seewave'

(Sueur, Aubin, and Simonis 2008). The routine gradually added uniform noise (i.e. uniformly distributed across frequencies, aka. 'white noise') to the signals until reaching a target SNR (± 0.1 dB). Other noise types (with different power spectrum) might be a better approximation to some natural acoustic environments, particularly pink noise in which the power spectrum decades in logarithmic scale. However, pink noise would not affect all frequencies bands equally. Hence, the effect of noise would also depend on the signal frequency range and frequency spectrum, making inference across different signals difficult. We ran this routine on the subset of signals with the highest SNR (SNR > 10 dB). Acoustic parameters and pairwise similarity measures were then calculated on noise-adjusted signals at 10 SNR levels (from 1 to 10 dB; Figure 1). Again, dynamic time warping and cross-correlation were calculated only on the subset of long-billed hermit songs. Mean bias, 95% confidence intervals, and repeatability were estimated at each SNR level compared to the highest SNR level (SNR = 10 dB). All analyses for non-pairwise similarity measures were run on the complete data-set and on the subset of long and short duration signals. Pairwise similarity measures were not split by duration categories as they were only conducted on long-billed hermit songs, which show little variation in song length.

Results

Compression

The mean bias of all frequency parameters from compressed files remained within the range of the frequency measurement error (86 Hz) regardless of the encoder or encoding method (*n* = 2642 signals, Figure 3). The most biased parameter (peak frequency) showed a mean positive deviation of only 14 Hz. However, 95% confidence intervals were much broader for most measurements, surpassing the a priori defined acceptable difference (160 Hz) in 4 of the 9 frequency parameters: minimum dominant frequency, maximum dominant frequency, peak frequency and frequency range (Figure 3). Interestingly, CBR produced narrower confidence intervals than VBR for LAME encoded files (Figure 3). Modulation index, skewness and spectral entropy showed a similar pattern; low mean biases but higher variation (particularly modulation index), and within those, greater variation on VBR-LAME encoded files (Figure 4). Dynamic time warping showed little bias but significant variation for all four treatments (~13% of the overall range of dynamic time warping distances). Compression produced slightly negatively biased cross-correlation coefficients (~−1.5% of the overall range) for most treatments except for VBR-LAME compression, which showed the largest bias (−3.7%) and also considerable variation. Long signals seem to be more affected by compression than short signals, except for modulation index (Figures 3 and 4).

Most acoustic parameters (both frequency and non-frequency parameters) produced high repeatability values despite compression, although VBR-LAME encoded files showed lower repeatability, particularly for modulation index, peak frequency and frequency range (Figure 5). Dynamic time warping distances were less repeatable than cross-correlation (or most acoustic parameters) regardless of encoder or encoding method (*n* = 242 independent signal pairs, Figure 5). In both pairwise similarity methods VBR-LAME encoded files showed the lowest repeatability. Again, except for modulation index, long signals are more affected by compression than short signals (Figures 3 and 4).

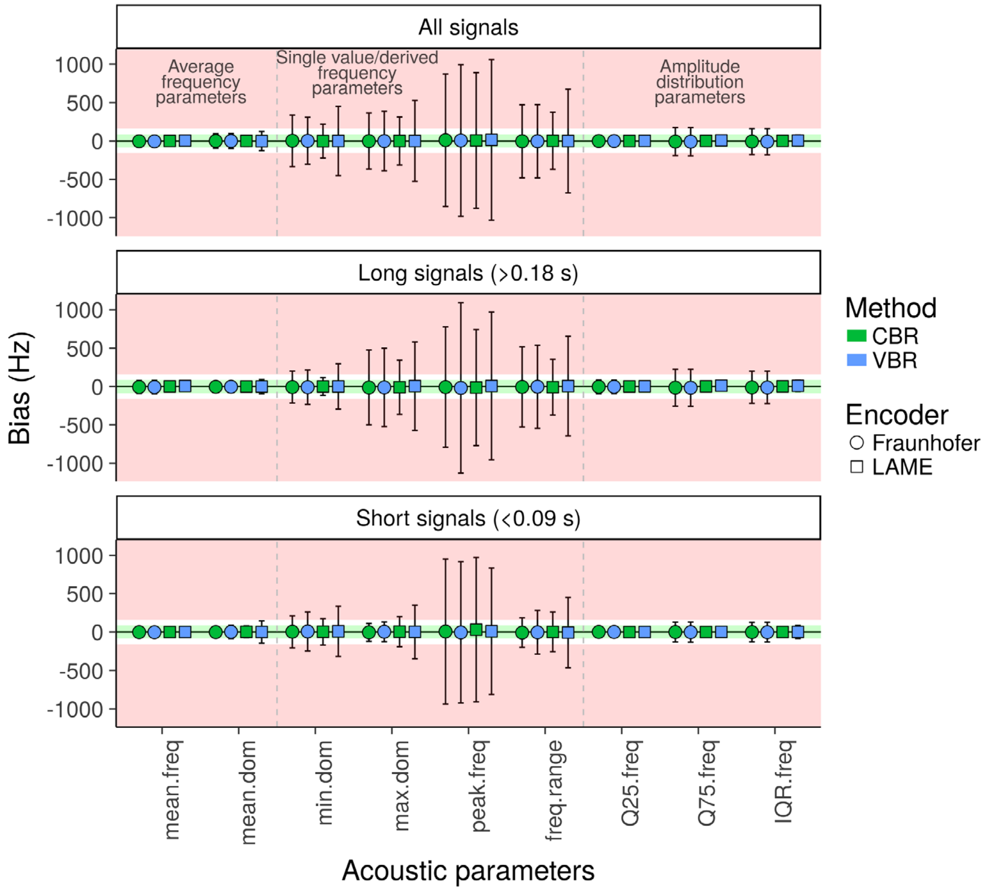

Figure 3. Bland-Altman bias and 95% confidence intervals for frequency parameters on mp3 compressed sound files compared to the same measurements on uncompressed files (all signals *n* = 2642; short signals *n* = 659, long signals *n* = 659).

Notes: Four biases are shown for each parameter, which correspond to each of the encoders (LAME and Fraunhofer) under the two encoding methods (CBR: constant bit rate, VBR: variable bit rate). The measurement error range is highlighted in light green. The light red area highlights values of possible concern according to a predefined threshold. Parameter names and descriptions are detailed in Table 1.

Background noise

Frequency parameter biases remained low across the SNR range, except for 3rd quartile and interquartile frequencies at a SNR below 2 dB (n = 1689 signals for all parameters, Figure 6). However, confidence intervals extended above the acceptable threshold with lower SNR (i.e. precision decreased with higher background noise) for several frequency parameters: mean frequency, mean dominant frequency, minimum dominant frequency, 1st quartile, 3rd quartile and interquartile frequencies. Some of these parameters displayed decreased precision even at intermediate background noise levels. The remaining frequency parameters showed high levels of variation (above the defined threshold) across the entire SNR range (Figure 6). From the non-frequency acoustic parameters, spectral entropy showed both important positive bias and high variation at low SNR values (Figure 7). Skewness showed a similar pattern, but to a much smaller extent compared to its overall range. Modulation index remains unbiased but with high variation across the entire SNR range.

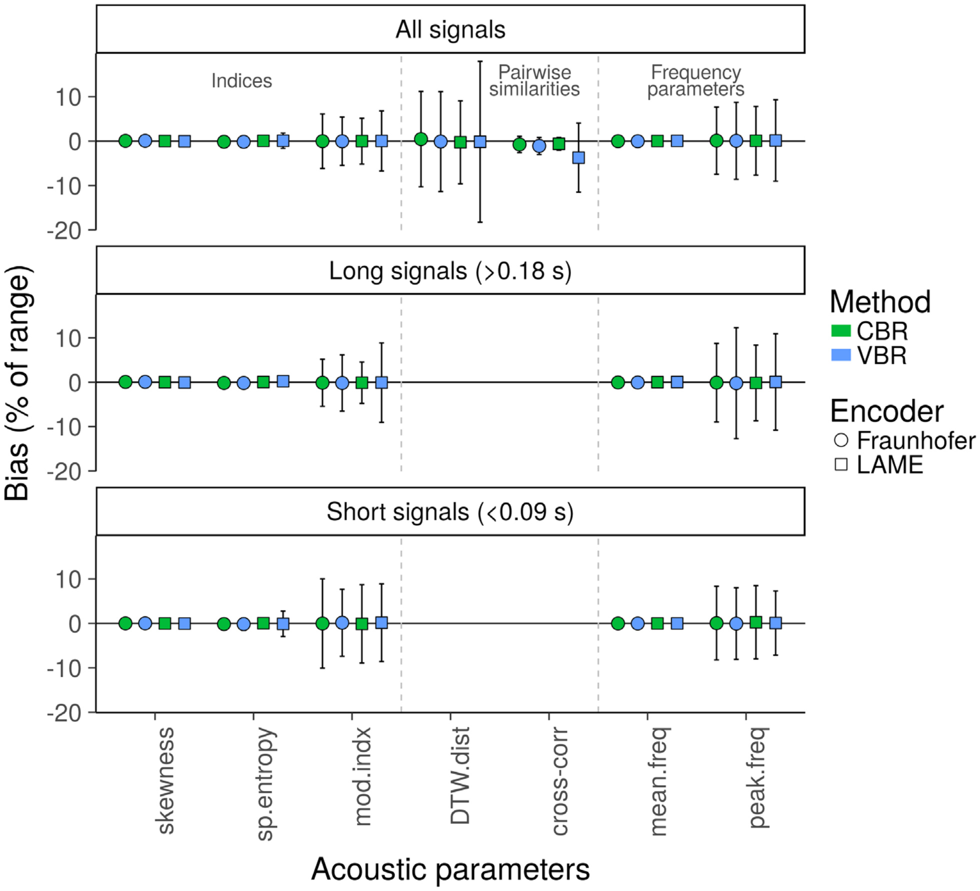

Figure 4. Bland-Altman bias and 95% confidence intervals for non-frequency parameters (modindx, skewness, sp.entropy; all signals $n = 2642$; short and long signals $n = 659$ each) and acoustic similarity methods (DTW.dist, cross.corr; $n = 242$ signal pairs; not split by duration due to little variation across long-billed hermit songs) on mp3 compressed sound files compared to the same measurements on uncompressed files.

Notes: Two parameters included in Figure 1 (mean.freq and peak.freq) are also shown to facilitate comparison to biases in frequency parameters. Four biases are shown for each parameter, which correspond to each of the encoders (LAME and Fraunhofer) under the two encoding methods (CBR: constant bit rate, VBR: variable bit rate). Biases are presented as the percentage of the overall parameter range in uncompressed files.

Both pairwise similarity methods showed little biases due to increasing background noise, although confidence intervals were much broader for dynamic time warping across the SNR range ($n = 214$ independent signal pairs, Figure 7). Signal duration had no apparent effect on parameter bias although showed a small effect on measurement precision. Minimum/maximum frequency and frequency range show slightly wider confidence intervals for long signals while confidence intervals were wider in short signals for peak frequency and amplitude distribution parameters.

The repeatability of acoustic parameters remained high across the SNR range, except for spectral entropy, which was clearly affected at SNR lower than four (Figure 8). Cross-correlation seemed to be highly repeatable regardless of the background noise level. Dynamic time warping was more affected by background noise, with low repeatability

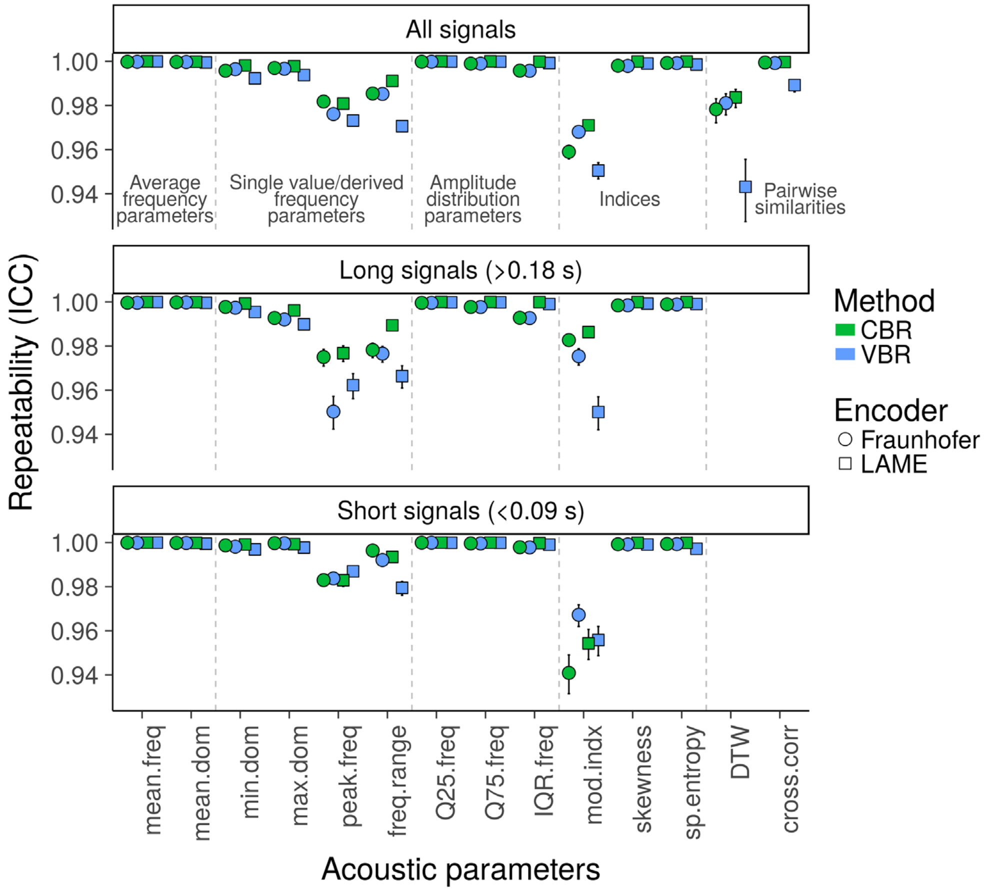

Figure 5. Repeatability and 95% confidence intervals of acoustic parameters (all signals $n = 2642$; short signals [duration < 0.09 s] $n = 659$, long signals [duration > 0.18 s] $n = 659$) and pairwise similarity measures (DTW.dist and cross.corr, $n = 242$ signal pairs; not split by duration due to little variation across long-billed hermit songs) measured on mp3 compressed sound files and uncompressed files.

Notes: Parameter names and descriptions are detailed in Table 1. Repeatability was estimated using the intraclass correlation coefficient (ICC). Four repeatability values are shown for each parameter, which correspond to each of the encoders (LAME and Fraunhofer) under the two encoding methods (CBR: constant bit rate, VBR: variable bit rate). Note that most parameters showed very narrow confidence intervals that do not stand over mean value symbols. Parameter names and descriptions are detailed in Table 1.

and high variation in repeatability across the SNR range (Figure 8). Signal duration had little effect on repeatability. Only spectral entropy, modulation index and frequency range presented slight differences between the two duration categories, although with important overlap between repeatability estimations.

Discussion

We evaluated the extent to which sound file compression and increasing levels of background noise affect the performance of the most common measurements for quantifying acoustic signal structure. Overall, the accuracy of mp3 compression remained high across the different mp3 encoding algorithms, although low precision (i.e. broad confidence intervals) was

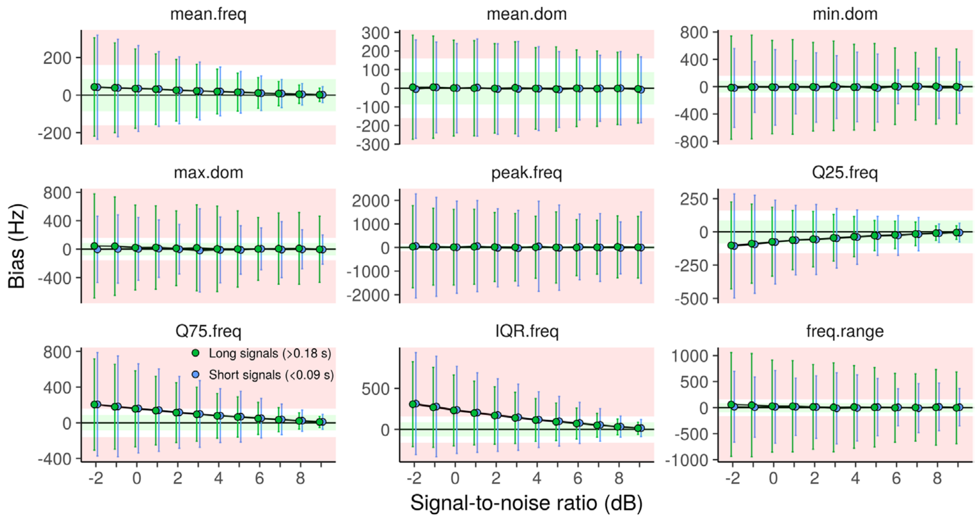

Figure 6. Bland-Altman bias and 95% confidence intervals for frequency parameters across the range of signalto-noise ratio values for short (duration < 0.09 s; *n* = 244) and long signals (duration > 0.18 s; *n* = 244).

Notes: Bias and confidence intervals for the complete data-set (*n* = 1689 signals) are not shown as these closely resemble those of the subsets included. The light red area highlights values of possible concern according to a predefined threshold and the light green area corresponds to the measurement error range. Parameter names and descriptions are detailed in Table 1.

found in four out of nine frequency measurements on compressed files (Figures 3 and 4). Cross-correlation seemed to be slightly negatively biased by file compression (i.e. decreased similarity when comparing compressed files). Dynamic time warping produced less biased similarity measures, but its precision was much more affected by mp3 compression than was the precision of cross-correlation, particularly for VBR-LAME encoded files (Figure 4). These patterns were confirmed by analyses of repeatability, in which measurements on VBR-LAME encoded files were consistently less repeatable, and dynamic time warping showed the lowest repeatability of all parameters and similarity methods.

Effects of background noise were more heterogeneous. Biases remain low for most parameters (Figures 6 and 7), except for 3rd quartile frequency and interquartile range, in which a high bias was found at high background noise levels (Figure 6). These results are in accordance with previous studies that have shown unbiased automatic frequency measures in noisy signals (Brumm et al. 2017; Ríos-Chelén et al. 2017). In a few cases, precision decreased with increasing background noise, but for most parameters precision was low (i.e. broad confidence intervals) regardless of the background noise level (Figure 6). In terms of repeatability, spectral entropy and peak frequency were most affected by background noise, particularly at low SNR levels. Both cross-correlation and dynamic time warping remained unbiased across background noise levels, although the precision of the latter was significantly more affected (Figures 6 and 7). Recent studies have shown that measuring frequency range using an amplitude threshold on a power spectrum is not prone to bias by noise (Brumm et al. 2017; Ríos-Chelén et al. 2017). Our results support the use of amplitude thresholds to obtained unbiased measures of frequency ranges when dealing with signals in variable noise levels. Nonetheless, results (particularly the lack of significant differences) should be interpreted cautiously given the decrease in precision.

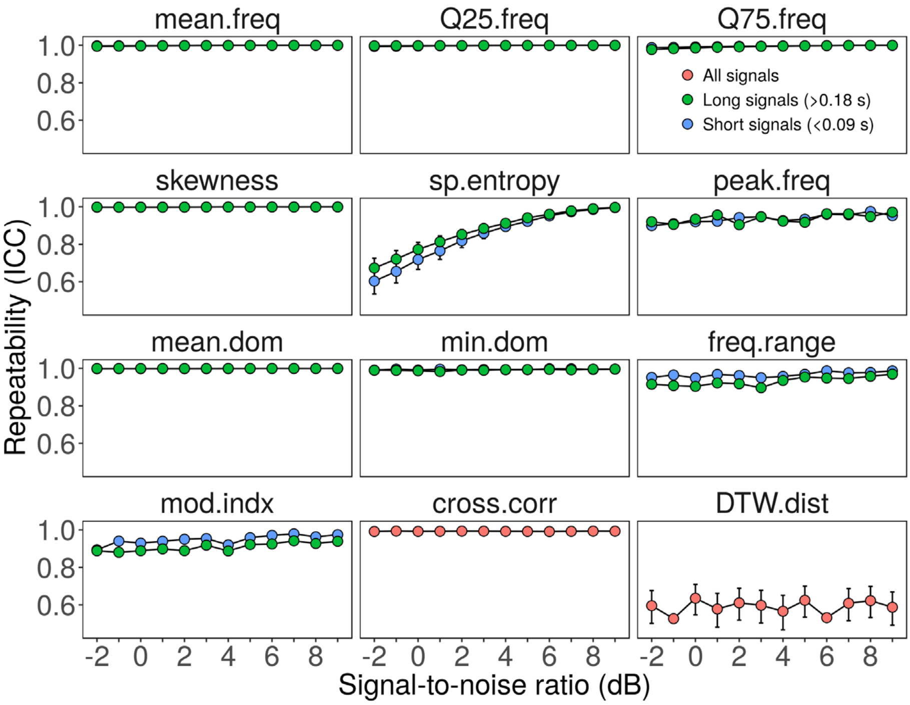

Figure 7. Bland-Altman bias and 95% confidence intervals for non-frequency parameters (skewness, sp.entropy, mod.indx; short and long signals *n* = 422 signals each) and pairwise similarity methods (cross.corr and DTW.dist; *n* = 214 signal pairs) across the range of signal-to-noise ratio values.

Notes: A parameter included in Figure 4 (IQR.freq) is also shown to facilitate comparison to biases in frequency parameters. Bias and confidence intervals for the complete data-set of non-frequency parameters (*n* = 1689 signals) is not shown as it closely resembles those of the subsets included. Pairwise similarity measures were not split by duration due to little variation across long-billed hermit songs. Parameter names and descriptions are detailed in Table 1. Biases are presented as percentage of the overall parameter range in uncompressed files.

Overall, acoustic parameters calculated across the entire signal (e.g. mean dominant frequency, spectral entropy) showed better precision than parameters representing a single extreme value (e.g. maximum frequency, peak frequency) or parameters derived from those (e.g. dominant frequency range, modulation index). This effect seems to be stronger in long duration signals. In general, this applied to both compression (particularly using the VBR-LAME encoder) and background noise analyses. Dominant frequency range and modulation index have already been found to be affected by compression (Medina-García et al. 2015). The single exception was spectral entropy, which was strongly affected by background noise (entropy increases with increasing noise), likely because higher background noise levels would tend to homogenize amplitude values across the signal.

The lack of precision on parameters representing single extreme values could also explain the strong effect of both compression and background noise on dynamic time warping distances. Frequency contours are sequences of single frequency values. Hence, the lack of precision on each measurement may be 'accumulated', resulting in a stronger effect on the overall contour shape. Cross-correlation was robust to both compression and background noise. This matches with previous findings in which background noise levels showed small effects on cross-correlation performance (Cortopassi and Bradbury 2000).

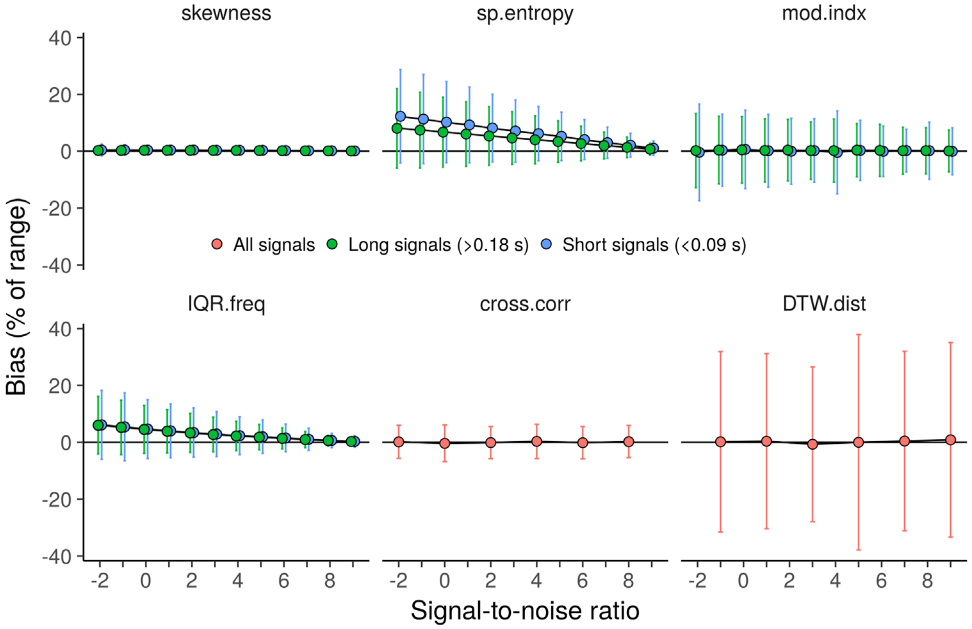

Figure 8. Repeatability and 95% confidence intervals of acoustic parameters (short and long signals *n* = 422 signals each) and pairwise similarity measures (cross.corr and DTW.dist *n* = 214 signal pairs) at different levels of signal-to-noise ratio.

Notes: Repeatability was estimated using the intraclass correlation coefficient (ICC). Bias and confidence intervals for the complete data-set of acoustic parameters (*n* = 1689 signals) are not shown as these closely resemble those of the subsets included. Pairwise similarity measures were not split by duration due to little variation across long-billed hermit songs. Parameter names and descriptions are detailed in Table 1. Repeatability of 'Q75.freq' and 'max.dom' (not shown) closely resembled those of 'Q25.freq' and 'min.dom', respectively. Most parameters showed very narrow confidence intervals that do not stand over mean value symbols.

Several conclusions can be drawn from our findings. First, mp3 compression (at least at 128 kbps or higher) does not generate a systematic deviation, on average, in the most commonly used acoustic measurements. However, single extreme value parameters or metrics derived from them appear to be less precise after file compression, particularly when using VBR-LAME encoded sound files. Second, cross-correlation should be chosen over dynamic time warping when comparing signals from mp3 compressed files or when recordings differ substantially with regard to background noise. Third, high background noise biases most acoustic parameters derived from the distribution of energy in the signals (e.g. interquartile range, spectral entropy) which can lead to spurious results when noise levels differ among treatments (particularly if noisy signals show higher entropy, 3rd quartile frequencies, and/or interquartile ranges). Finally, in general, background noise affects the precision of acoustic parameters and dynamic time warping, but has a smaller effect on cross-correlation.

These conclusions translate into specific recommendations and guidelines for dealing with mp3 compressed files or acoustic signals and/or those with variable levels of background noise. When aiming to quantify signal structure, uncompressed files are preferred over files that have undergone mp3 compression. VBR-LAME encoding should be avoided if file compression is required. Parameters measured across entire signals would be more reliable, whereas single-value acoustic parameters and metrics derived from them should be avoided (e.g. 'frequency excursion index', Podos et al. 2016). Spurious statistical differences are unlikely to arise as a result of compression, as acoustic measurements showed small biases. However, spurious differences could result when energy distribution parameters are used to characterize signals if background noise levels co-vary with a predictor factor (e.g. when looking at group-level acoustic signatures and groups inhabit areas with different background noise levels).

The low measurement precision generated by compression and background noise could result in failing to detect actual biological differences. Furthermore, the use of parameters prone to be biased by noise (as the energy distribution parameters) to compare signals with different noise levels could generate spurious statistical differences. New analytical tools allow users to precisely measure background noise levels for each signal (e.g. signal-to-noise ratio, Araya-Salas and Smith-Vidaurre 2017). This parameter can be used to statistically control for the effect of noise, to directly evaluate differences among treatments, or to identify the parameters most strongly affected by noise (and remove them). Any of these approaches will help to focus the analysis on the actual acoustic differences of the groups being compared, strengthen the validity of the results. The negative effects of compression and background noise could be less problematic when conducting comparing acoustic parameters across species, as the differences are usually stronger than at the within species level (e.g. Hall et al. 2013).

The study of animal acoustic signals remains an important tool for many fields in biological research. The rapid diversification of recording devices, analytical approaches and bioacoustic repositories can pose important challenges for achieving robust studies. Sound file compression and variation in background noise are two putative confounding factors for acoustic analysis, although little attention has been devoted to their respective effects on acoustic metrics. We have shown that mp3 compression and background noise can indeed influence parameters commonly used to quantify acoustic signals. However, careful consideration of these parameters, or the use of statistical tools to directly assess their effect, can help mitigate some negative effects of compression and background noise. It is also likely that the magnitude of some of the observed effects is related to the signal structure itself (e.g. precision of dynamic time warping on pure tone signals could be less affected by compression). More studies will be warranted to reach a more detailed understanding of the effects of these factors in acoustic signal quantification.

Authors' contributions

MAS, MSW and GSV conceived the ideas and designed methodology; MAS and GSV gathered the recordings and analysed the data; MAS led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Data accessibility

All recordings are publicly available at acoustic libraries. Libraries and catalogue numbers for all recordings are provided in supplementary Table S1.

Acknowledgements

We thank William-Pier Vellinga, Russ Charif, Greg Budney and anonymous reviewers for their useful suggestions at different stages of the manuscript. We also thank the Macaulay Library, Laboratorio de

Bioacústica de La Universidad de Costa Rica, Instituto Humboldt, Florida Museum of Natural History, British Library and Borror Laboratory of Bioacoustics for providing access to bird song recordings.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Marcelo Araya-Salas http://orcid.org/0000-0003-3594-619X

References

Araya-Salas M. 2012. Is birdsong music? Evaluating harmonic intervals in songs of a Neotropical songbird. Anim Behav. 84:309–313.

Araya-Salas M, Smith-Vidaurre G. 2017. warbleR: an R package to streamline analysis of animal acoustic signals. Methods Ecol Evol. 8:184–191.

Araya-Salas M, Wright T. 2013. Open-ended song learning in a hummingbird. Biol Lett. 9:20130625.

Audacity-Team. 2014. Audacity: free audio editor and recorder. Version 2.0.0. [accessed 2014 April 20]. http://audacity.sourceforge.net/.

Beecher MD. 1988. Spectrographic analysis of animal vocalizations: implications of the 'uncertainty principle'. Bioacoustics. 1:187–208.

Berman J. 2015. Analysis of zero-level sample padding of various MP3 codecs. Denver: University of Colorado.

Bland JM, Altman DG. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 327:307–310.

Boncoraglio G, Saino N. 2007. Habitat structure and the evolution of bird song: a meta-analysis of the evidence for the acoustic adaptation hypothesis. Funct Ecol. 21:134–142.

Botero CA, Boogert NJ, Vehrencamp SL, Lovette IJ. 2009. Climatic patterns predict the elaboration of song displays in Mockingbirds. Curr Biol. 19:1151–1155.

Bradbury J, Vehrencamp S. 2011. Principles of animal communication. Sunderland (MA): Sinauer Associates.

Brumm H, Zollinger SA, Niemelä PT, Sprau P. 2017. Measurement artefacts lead to false positives in the study of birdsong in noise. Methods Ecol Evol. doi: 10.1111/2041-210X.12766

Charif R, Waack A, Strickman L. 2010. Raven Pro 14 user's manual. Ithaca (NY): The Cornell Lab of Ornithology.

Clark CW, Marler P, Beeman K. 1987. Quantitative analysis of animal vocal phonology- an application to swamp sparrow song. Ethology. 76:101–115.

Cortopassi KA, Bradbury JW. 2000. The comparison of harmonically rich sounds using spectrographic cross-correlation and principal coordinates analysis. Bioacoustics. 11:89–127.

Cramer ERA. 2013. Measuring consistency: spectrogram cross-correlation versus targeted acoustic parameters. Bioacoustics. 22:1–11.

Depraetere M, Pavoine S, Jiguet F, Gasc A, Duvail S, Sueur J. 2012. Monitoring animal diversity using acoustic indices: Implementation in a temperate woodland. Ecol Indic. 13:46–54.

Doolittle E, Brumm H. 2013. O Canto do Uirapuru: consonant intervals and patterns in the song of the musician wren. Journal of interdisciplinary music studies. 6:55–85.

Gonzalez-Voyer A, den Tex R-J, Castelló A, Leonard JA. 2013. Evolution of acoustic and visual signals in Asian barbets. J Evol Biol. 26:647–659.

Hall ML, Kingma SA, Peters A, Blevins W, Vanbroeckhoven C. 2013. Male songbird indicates body size with low-pitched advertising songs. PLoS ONE. 8:e56717.

Hanneman S. 2010. Design, analysis and interpretation of method-comparison studies. AACN Adv Crit Care. 19:223–234.

Kaluthota C, Brinkman BE, Santos EB, Rendall D. 2016. Transcontinental latitudinal variation in song performance and complexity in house wrens (Troglodytes aedon). Proc R Soc B Biol Sci. 283(1824):20152765.

Kogan J, Margoliash D. 1998. Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: a comparative study. J Acoust Soc Am. 103:2185–2196.

Lachlan RF. 2007. Luscinia: a bioacoustics analysis computer program. Version 1.0. Computer program.

Ligges U, Krey S, Mersmann O, Schnackenberg S. 2014. tuneR: analysis of music. R Packag version 121.

Mason N, Shultz A, Burns K. 2014. Elaborate visual and acoustic signals evolve independently in a large, phenotypically diverse radiation of songbirds. Proc R Soc B Biol Sci. 281:20140967.

Mason NA, Burns kJ, Tobias JA, Claramunt S, Seddon N, Derryberry EP. 2016. Song evolution, speciation, and vocal learning in passerine birds. Evolution. 1–27.

Medina-García A, Araya-Salas M, Wright TF. 2015. Does vocal learning accelerate acoustic diversification? Evolution of contact calls in Neotropical parrots. J Evol Biol. 28:1782–1792.

Morton E. 1975. Ecological sources of selection on avian sounds. Am Nat. 109:17–34.

Müller R, Büttner P. 1994. A critical discussion of intraclass correlation coefficients. Stat Med. 13:2465–2476.

Pegan TM, Rumelt RB, Dzielski SA, Ferraro MM, Flesher LE, Young N, Freeman AC, Freeman BG. 2015. Asymmetric response of costa rican white- breasted wood-wrens (henicorhina leucosticta) to vocalizations from allopatric populations. PLoS ONE. 10:1–16.

Podos J, Moseley D, Goodwin S, McClure J, Taft B, Strauss A, Rega-Brodsky C, Lahti D. 2016. A fine-scale, broadly-applicable index of vocal performance: frequency excursion. Anim Behav. 116:203–212.

Rempel RS, Hobson KA, Holborn G, Wilgenburg SL Van, Elliott J. 2005. Bioacoustic monitoring of forest songbirds: interpreter variability and effects of configuration and digital processing methods in the laboratory. J Ornith. 76:1–11.

Ríos-Chelén AA, Lee GC, Patricelli GL. 2016. A comparison between two ways to measure minimum frequency and an experimental test of vocal plasticity in red-winged blackbirds in response to noise. Behaviour. 153:1445–1472.

Ríos-Chelén AA, McDonald AN, Berger A, Perry AC, Krakauer AH, Patricelli GL. 2017. Do birds vocalize at higher pitch in noise, or is it a matter of measurement? Behav Ecol Sociobiol. 71:29.

Specht R. 2004. Avisoft-SASLab Pro. Berlin: Avisoft.

Sterne J. 2012. The meaning of a format MP3. Durham (NC): Duke University Press.

Stiles FG, Wolf LL. 1979. Ecology and evolution of lek mating behavior in the long-tailed Hermit hummingbird. Ornith Monogr. 27: iii–78.

Sueur J, Aubin T, Simonis C. 2008. Equipment review: seewave, a free modular tool for sound analysis and synthesis. Bioacoustics. 18(2):213–226.

Tchernichovski O, Nottebohm F, Ho C. 2000. A procedure for an automated measurement of song similarity. Anim Behav. 59:1167–1176.

Towsey MW, Truskinger AM, Roe P. 2016. The navigation and visualisation of environmental audio using zooming spectrograms. Proceedings – 15th IEEE International Conference on Data Mining Workshop, ICDMW 2015. p. 788–797.

Weir JT, Wheatcroft DJ, Price TD. 2012. The role of ecological constraint in driving the evolution of avian song frequency across a latitudinal gradient. Evolution. 66:2773–2783.

Wolak ME, Fairbairn DJ, Paulsen YR. 2012. Guidelines for estimating repeatability. Methods Ecol Evol. 3:129–137.

Zollinger SA, Podos J, Nemeth E, Goller F, Brumm H. 2012. On the relationship between, and measurement of, amplitude and frequency in birdsong. Anim Behav. 84:e1–e9.