IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 12, 2017, accepted December 9, 2017, date of publication December 13, 2017,
date of current version February 28, 2018.

Digital Object Identifier 10.1109/ACCESS.2017.2782778

Automatic and Efficient Denoising of Bioacoustics Recordings Using MMSE STSA

ALEXANDER BROWN, SAURABH GARG, AND JAMES MONTGOMERY
School of Engineering and ICT, University of Tasmania, Hobart, TAS 7001, Australia
Corresponding author: Saurabh Garg (saurabh.garg@utas.edu.au)

ABSTRACT Automatic recording and analysis of bird calls is becoming an important way to understand changes in bird populations and assess environmental health. An issue currently proving problematic with the automatic analysis of bird recordings is interference from noise that can mask vocalizations of interest. As such, noise reduction can greatly increase the accuracy of automatic analyses and reduce processing work for subsequent steps in bioacoustics analyses. However, only limited work has been done in the context of bird recordings. Most semiautomatic methods either manually apply sound enhancement methods available in audio processing systems such as SoX and Audacity or apply preliminary filters such as low- and high-pass filters. These methods are insufficient both in terms of how generically they can be applied and their integration with automatic systems that need to process large amounts of data. Some other work applied more sophisticated denoising methods or combinations of different methods such as minimum mean square error short-time spectral amplitude estimator (MMSE STSA) and spectral subtraction for other species such as anurans. However, their effectiveness is not tested on bird recordings. In this paper, we analyze the applicability of the MMSE STSA algorithm to remove noise from environmental recordings containing bird sounds, particularly focusing on its quality and processing time. The experimental evaluation using real data clearly shows that MMSE STSA can reduce noise with similar effectiveness [using objective metrics such as predicted signal quality (SIG)] to a previously recommended wavelet-transform-based denoising technique while executing between approximately 5–300 times faster depending on the audio files tested.

INDEX TERMS Noise removal, bioacoustics, big data.

I. INTRODUCTION

Human expansion and climate change have led to drastic changes in ecological balance, which has accelerated in recent years. This necessitates close monitoring of different species, particularly birds, which are very good indicators of environmental health. Traditionally, to monitor birds, experts needed to be present in the region of interest [1]. This is time consuming and expensive. Animals make distinct vocalisations that can be picked up using sound recorders, which can be later heard by experts to recognize different species present in certain ecosystems. However, with the large amount of recording data necessary to monitor an ecosystem, it is impractical for humans to listen to and manually label recordings [2]. Consequently, monitoring surveys are conducted based on selecting samples of recorded audio. However, this methodology can introduce bias and incompleteness. Hence, researchers have turned to automated techniques to process these environmental recordings.

The approach of automatically processing environmental recordings has recently seen significant research interest [3]–[5] because of its range of applications, including tracking bird migration [6], monitoring biodiversity [7], and detecting endangered species [8]. An issue currently proving problematic when processing environmental recordings is that interference from noise can mask vocalisations of interest and make them difficult to recognise [1], [9]. Sources of noise might be generated by geophony (environmental sounds such as wind and rain), anthrophony (noise generated by humans, though sources such as traffic and machines), and biophony (sounds from animals that are not of interest) [1]. In the context of bird acoustics, any sound other than birds is considered noise. In this paper, we focus on the automatic removal of background environmental noise that is present in recordings with bird vocalisations. Denoising speech signals is not a new topic [10], [11]; most research work in the area of bird acoustics adapts noise reduction techniques. In particular, some researchers apply low and high-pass filters [12]–[14], which attenuate audio in frequency regions known to not contain signals of interest. However, because bird vocalisations are often in the same frequency range as

interfering noise, a lot of noise remains in the recordings. The method for automatically removing noise from recordings with bird vocalisations should be sufficiently generic that it can be utilised in different contexts such as noise from different wind speeds, and different rain intensities. It should also not distort bird vocalisations. Moreover, the amount of acoustic data collected from multiple locations is sometimes so large that the time efficiency of the chosen denoising method becomes an important factor for consideration. Other research into denoising methods such as wavelet packet decomposition [9] does not consider time efficiency. Finally, not all types of denoising methods are applicable for bird acoustics, as environmental recordings generally have sounds of interest that are non-stationary as well as noise that is stationary. In other words, some noise is approximately constant within short time durations, and other noise is not [15]. Moreover, the background noise is from uncorrelated sources and additive in nature.

In this paper, we analyse and adapt the MMSE STSA filter Ephraim and Malah [16] to remove stationary and uncorrelated noise from environmental recordings with differing characteristics. We investigate the effectiveness of different parameter settings to identify those that should be used for automatic denoising. We compare the accuracy and time efficiency of our proposed MMSE STSA denoising method for bird recordings with a recent recommended wavelet decomposition based method [9]. The contributions of this paper are:

1) Analysis of the applicability of MMSE STSA for automatic denoising of large scale environmental recordings containing bird vocalisations. The algorithm is tested with six different categories of noisy recordings.
2) Analysis of different settings of the MMSE STSA estimator for denoising environmental recordings containing bird vocalisations.

We discuss related works in the next section, followed by an introduction to the MMSE STSA algorithm in Section III. We present our experimental methodology and evaluation in Sections IV and V, respectively, followed by conclusions and future directions in Section VI.

II. RELATED WORK

In recent decades, several noise removal and sound enhancement methods have been proposed for processing human speech signals. Recently, there has been increasing interest in finding ways to automatically recognise bird species in environmental recordings. Noise interference has been a significant problem in this research area as it can potentially decrease the accuracy of bird recognition.

The simplest approaches to reducing background noise in audio recordings are low and high-pass filtering. These filters attenuate frequencies in regions of audio known not to contain any signal. In the context of bioacoustics, the calls of animals of interest are often known to be in a certain frequency range, so anything not in this frequency range can be eliminated. Birds typically do not make sounds above 12 kHz or below 1 kHz [17], so sounds outside of this region can be ignored. Neal *et al.* [12] uses a 1 kHz high-pass filter as part of an effort to segment bird sounds. Baker and Logue [13] use the same approach as part of a technique to compare differences between chickadee sounds across populations. Similarly, Bardeli *et al.* [14] use low-pass filtering to help detect two endangered bird species. However, as recordings usually have noise in frequency regions that also contain bird calls [17], these filters on their own cannot remove all noise from bioacoustics recordings. However, because they aggressively remove any sound from target frequency regions, they can be used in combination with other techniques to improve noise reduction [9].

Another common approach for noise reduction is spectral subtraction, as derived by Boll [10]. This was one of the first algorithms developed to reduce general noise in audio. This approach collects a 'noise profile', which is a sample of audio only containing noise. It then analyses the noisy component of the signal, breaking it down into its component frequencies. It then subtracts these noise components from the entire audio file, theoretically leaving only the signal. A problem with this process is that it is prone to introducing processing artifacts that can sound like musical tones [18]. Patti and Williamson [19] used spectral subtraction as a preprocessing step in a bird species classification problem, but did not test the effectiveness of the noise reduction itself.

Noise gating is similar approach that utilises a noise profile for estimating the intensity of noise and reduces the volume of any part of the recording which is below a noise threshold [15]. Bedoya *et al.* [20] adapted this methodology to aid in the detection of anuran (frog-like) species. While the noise reduction itself was not tested for its effectiveness, the overall system proved to be successful in classifying 13 anuran species, achieving an accuracy of 99.38%–100%, which compares favourably to similar studies. Due to their effectiveness, spectral subtraction and noise gating are employed by the widely-used audio editors SoX [21] and Audacity [22], respectively. However, as these methods require estimation of noise from a noise profile, their applicability is limited to the context of developing an automated system for recognising bird sounds from diverse environmental recordings, because researchers need to collect noise profiles that cover all different types of background noise featured in the recordings. Figure 1 illustrates this problem. In this example, identical audio files are processed by the same spectral noise gating approach, but one (Figure 1b) uses a noise profile selected from a portion of audio from a different time of the recording, while the other (Figure 1c) uses a noise profile selected from an uneventful part of the recording. When using a general noise profile from another time in the recording, the noise filter removes much less noise.

The Wiener filter approaches noise filtering in a similar way to spectral subtraction, in that it assumes a signal is made up of a desired component and a noisy component, but it approaches the estimation of these components differently. This filter aims to optimise the minimum mean square error

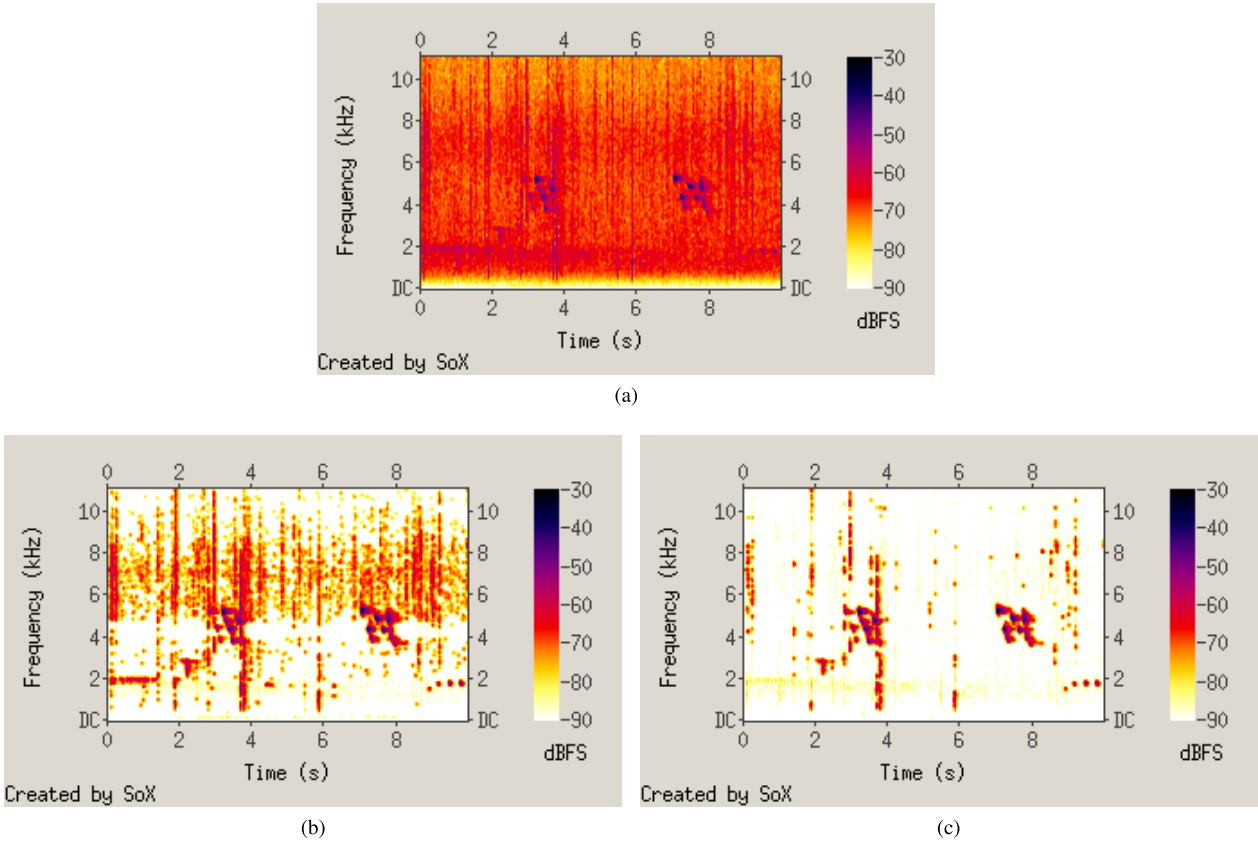


FIGURE 1. Comparison between usage of noise profiles. (a) Raw Recording. (b) General Noise Profile. (c) Specific Noise Profile.

between the target signal and the predicted signal. An issue with this technique is that it assumes that noise and signal are both stationary [9], which is not necessarily true in long duration environmental recordings.

The **MMSE STSA** estimator derived by Ephraim and Malah [16] is another noise reduction approach. It derives a near-optimal short time spectral amplitude estimator to significantly improve noise reduction with a reduction in artifacts compared to spectral subtraction and Wiener filtering. An improvement to this method uses log spectra, which was found by Ephraim and Malah [23]. While this approach is intended as a speech enhancement technique [24], it was used successfully in a bioacoustics scenario by Alonso *et al.* [1] as part of a semi-automated segmentation technique for auran identifications. However, they did not analyse the general effectiveness of MMSE STSA for different environmental recordings, in particular for bird identifications. In our paper, we extensively study the applicability of MMSE STSTA for denoising environmental recordings for bird identification application.

Ren *et al.* [18] apply a similar noise removal model to wavelet transforms, rather than Short Time Fourier Transforms (STFTs) as used by other techniques. This is designed to avoid the problem of choosing window sizes for STFTs, which have a trade-off between time resolution and frequency resolution, depending on the window size. Instead, wavelet transforms implicitly use different window sizes for different frequency components, which reduces the problems presented by this trade-off. Ren *et al.* [18] tested this wavelet based method against other noise reducing techniques, such as spectral subtraction and MMSE STSA (with 32 ms windows, 75% overlap, and smoothing factor $\alpha = 0.98$). They modified clean audio recordings of different animals by adding white noise and environmental noise. They found that their approach increased the signal to noise ratio and segmental signal to noise ratio (which considers the signal to noise ratio for smaller segments of audio) more than other approaches where the signal to noise ratio of the original audio is lower. Once the signal to noise ratio of the audio became closer to 0 dB, the standard MMSE STSA approach started to reduce the noise slightly more effectively. Other approaches, such as spectral subtraction, did not perform well.

Priyadarshani *et al.* [9] similarly use wavelet transforms to remove noise in a bird sound recording. They use Shannon entropy to determine noise thresholds. The intuition is that noise will have a higher entropy than signal, and this can be used as a basis to remove noise information. They combine this with band-pass filters which remove frequencies outside of the range of the signals. They evaluate their

technique using noisy environmental recordings, as opposed to Ren *et al.* [18], who artificially added noise to their recordings. They define a 'success ratio', which compares the initial noise level to the noise level after denoising, and a modified peak signal to noise ratio, which considers the ratio between the maximum value and mean squared error of the signal. They found large improvements in all metrics compared to applying simple band-pass filtering. However, the use of signal to noise ratio to evaluate filter quality is problematic, as this cannot be used to determine how well the original signal has been preserved. As such, our evaluation uses measures that can evaluate noise removal and (retained) signal quality. We use these measures to compare results with those of Priyadarshani *et al.*'s [9] wavelet transform method.

In summary, most work in bird identification from environmental recording applies simple denoising methods such as low and high-pass filters. Most researchers use off the shelf methods without fully analysing the wide applicability of the denoising or signal enhancement techniques. Denoising methods such as spectral subtraction and noise gating require estimation of noise using a noise profile, and thus are not applicable for automatic denoising of diverse environmental recordings. According to Priayadarshani *et al.* [9], Wiener filters are not applicable in the context of environmental recordings. MMSE STSA and wavelet transform based methods appear to be viable solutions based on previous methods. However, to the best of our knowledge, no existing research considers the time efficiency of the denoising algorithms, which is becoming increasingly important given the very large and increasing amount of environmental recordings that are being collected every day. In this paper, we propose using the MMSE STSA method with band-pass filters for generalised automated denoising of environmental recording for bird identification. We compare our proposed method with the wavelet transform based denoising method proposed by Priyadarshani *et al.* [9].

III. MMSE STSA ALGORITHM

As discussed, environmental recordings generally contain noise which interferes with the actual signal, making identifying bird sounds more difficult, particularly for automated processes. Any developed denoising method should be applicable to a wide range of recordings and should be able to be integrated with automated systems for processing large amounts of recordings to identify birds. In environmental recordings, noise may vary over long durations. For example, changing wind speed can vary the amount of background noise. It also might not always be possible to cancel out the noise completely, particularly if it has non-stationary components. We aim to reduce the effects of the noise on the signal's average spectral amplitude. The Minimum Mean-Square Error (MMSE) Short-time Spectral Amplitude (STSA) estimator designed by Ephraim and Malah [16] gives a theoretically optimal estimation of the clean spectral amplitude and possesses significant advantages over other spectral based methods when dealing with non-stationary noise, which is the context of environmental recordings. This approach is based on modeling speech and noise spectral components as statistically independent Gaussian random (i.e. normally distributed) variables. The signal to noise ratio (SNR) of the audio is estimated *a priori*, and the filter adapts based on how high the SNR is (it is more aggressive when the SNR is low). We present an overview here for the reader's convenience.

Let $Y(k), N(k)$, and $X(k)$ be the Short Time discrete Fourier Transform (STFT) of original noisy signal, noise signal and clean signal, respectively, and integer k represent the frequency index. In the frequency domain, the noisy signal can be represented as:

$$Y(k) = X(k) + N(k) \tag{1}$$

which is defined for each **frequency index** k as

$$Y_k \exp^{j\theta_{Y_k}} = X_k \exp^{j\theta_{X_k}} + N_k \exp^{j\theta_{N_k}} \tag{2}$$

where Y_k, X_k, N_k and $\theta_{\{\cdot\}}$ are the magnitudes and phase of the frequency spectrum. The MMSE STSA filter is summarised using the equation for the minimum mean squared estimate of spectral amplitude of the clean signal($\hat{X}$):

$$\hat{X} = G_{MMSE}(k) Y_k \tag{3}$$

where G_{MMSE} is the spectral gain factor, given by:

$$G_{MMSE}(k) = \frac{\sqrt{\pi v_k}}{2} \exp^{\frac{-v_k}{2}} \left[(1 + v_k) I_0 \left(\frac{v_k}{2} \right) + v_k I_1 \left(\frac{v_k}{2} \right) \right] \tag{4}$$

where $I_0(\cdot)$ and $I_1(\cdot)$ are modified Bessel functions of the zeroth and first order, respectively, and v_k is defined as:

$$v_k = \frac{\xi_k}{(1 + \xi_k)} \gamma_k \tag{5}$$

where ξ_k and γ_k are estimated *a priori* and *a posteriori* signal to noise ratios for each spectral components. The *a posteriori* signal to noise ratio obtained is defined as the ratio of the actual noisy signal to the variance of noise power (σ_n):

$$\gamma_k = \frac{Y_k^2}{\sigma_n^2(k)} \tag{6}$$

Ephraim and Malah proposed a decision-directed method to iteratively compute the *a priori* and *a posteriori* SNR. Initially the variance of noise ($\sigma_n^2(k)$) is computed based on silence regions and then the *a posteriori* SNR is obtained on a frame by frame basis. Generally, estimation is based on the first few frames in the recording. The initial value of the *a priori* SNR $\xi_k(0)$ is given by

$$\xi_k(0) = \alpha + (1 - \alpha) P[\gamma_k(0) - 1] \tag{7}$$

where $P[\cdot]$ is a rectification function to ensure the STSA estimator is positive even, and α is the smoothing constant with typical value of 0.98.

For each frame m, we update the *a priori* SNR estimate using

$$\xi_k = \alpha \gamma_k(m-1) G_{MMSE_k}^2(m-1) + (1 - \alpha) P[\gamma_k(m) - 1], \quad 0 < \alpha < 1 \tag{8}$$

Algorithm 1 MMSE STSA Implementation

Data: Input: Audio File = *af*; Window Size = *WSize*; Noise Threshold = *Thresh*
Result: Denoised Audio File *paf*;
Apply 1 kHz high-pass filter;
Apply Hamming Window, 50% overlap, Window Size=*WSize*. This splits into frames of size WSize. Let *Frames*[*k*] be these frames, where k is a frame ID. **for** *i in* (*Frames*) **do**
    Coeffs[*i*, *j*]= FFT(*i*);
    `// Where FFT is the Fast Fourier Transform. j is the frequency index of the coefficients (which is of size Wsize)`
    Let *Magnitude*[*i*, *j*] = mod(*Coeffs*[*i*, *j*]);
    `// The volume of each frame and each frequency index`
end
Let *InitialFrames*[*i*, *j*] be the *Magnitude* of the frames in the segment used to initialise mean noise level and mean noise variance (approximately 0.1 seconds);
Let *OtherFrames*[*i*, *j*] be the *Magnitude* all other frames;
Let *NoiseMean*[*j*] be the mean noise level for each frequency index *j*. Initialise this to be the mean of all *InitialFrames*;
Let *NoiseVar*[*j*] be the mean noise variance for each frequency index *j*. Initialise this to be the mean volume squared of all *InitialFrames*;
for *k in* (*OtherFrames*) **do**
    for *m in length*(*WSize*) **do**
        Apply VAD;
        `// Voice Activity Detection. Detects if a sample contains animal sound.`
        if *OtherFrames*[*k*, *m*] *has no signal* **then**
            Update *NoiseMean*[*m*] and *NoiseVar*[*m*];
        end
        Calculate spectral gain factor for *OtherFrames*[*k*, *m*]. Set this to *SpectralGain*[*k*, *m*];
        Apply Spectral Gain (*otherFrames*[*k*, *m*] = *otherFrames*[*k*, *m*] × *SpectralGain*[*k*, *m*]);
    end
end
paf = Inverse FFT(*Frames*);
`// Where Frames contains InitialFrames followed by the newly processed otherFrames`

A. DENOISING ALGORITHM IMPLEMENTATION

The actual implementation is summarised in Algorithm 1. This consists of the following steps:

Firstly, audio files are converted to 22.05 kHz mono wave files. This is chiefly to reduce computation time in later analysis steps. A 1 kHz high-pass filter is then applied to these files. This attenuates the sound below 1 kHz, which can be done without loss of signal as no birds make sound below 1 kHz [17]. For each audio file, the MMSE-STSA algorithm is applied, where each file is divided into predefined window frames. In the experimental evaluation, we utilise a native Java implementation[1] of the MMSE STSA estimator as described by Ephraim and Malah [16].

This begins by applying a Hamming Window with 50% overlap is applied to each chunk. The window size is specified as an input parameter. A Fast Fourier Transform (FFT) is applied to each frame. The amplitude of the audio at a given frequency is given by the modulus of the resulting complex coefficients.

An initial segment of audio is used to estimate the mean noise level and variance of the audio for each frequency given by the FFT. The length of the segment chosen is set to be approximately 0.1 seconds. The number of frames varies depending on the windows size and sample rate. For a sample rate of 22.05 kHz, this is equivalent to 7 frames for a 512 windows size, 16 frames for 256 samples, and 33 frames for 128 samples, i.e. for frames with 50% overlap.

At this point, Voice Activity Detection (VAD) is applied on each of the other windows in the audio. This begins by calculating the volume difference (in dB) between the current signal and the mean noise level for each frequency. Any negative values are truncated to zero. The mean of the noise differences over all frequencies is computed. The mean noise level is calculated using

$$\bar{n} = \frac{l * \bar{n}_{old} + Y_k}{l + 1} \tag{9}$$

where $\bar{n}$ is the mean noise level, Y_k is the magnitude of the frequency at the frequency index k for the given frame, and l is the noise length.

In the existing implementation, the noise length is set to be constant, although we observed greater success initialising it to 0 and incrementing by 1 each time the noise profile is updated, so as to accurately calculate the noise mean. Keeping this static in the implementation is likely done to make newer frames detected as noise having a higher weighting in the noise mean. The same principle applies with the noise variance.

If this mean difference is below a noise threshold, it is classified as noise. This noise threshold is specified as an input parameter. If there has been a pre-specified number of consecutive frames of noise (called the 'frame reset' by the implementation), then the sample is flagged as not containing speech (i.e., a bird call) and the noise mean and variance is updated to include the current frame. Otherwise, it is said to contain speech. The frame reset is an input parameter, but early experimentation found that varying this in the range [1, 20] did not have a noticeable effect on the audio, so its value is left at the default of 8.

[1] Available at https://github.com/alexanderchiu/AudioProcessor

Window Size	128			256			512		
Noise Threshold	2	6	10	2	6	10	2	6	10
Category									
1	3.23	3.21	3.20	3.30	3.30	3.30	3.27	3.37	3.29
2	3.13	3.21	3.12	3.60	3.41	3.78	3.55	3.29	3.23
3	2.31	2.21	2.22	2.53	2.40	2.41	2.71	2.56	2.56
4	2.81	2.65	2.64	2.89	2.77	2.79	2.96	2.91	2.90
5	3.22	3.11	3.08	2.92	2.86	2.81	3.00	2.91	2.97
6	2.96	2.83	2.83	3.03	3.00	3.00	3.11	3.10	3.24
Average	2.94	2.87	2.85	3.05	2.96	3.02	3.10	3.02	3.10
Std. Dev.	0.35	0.39	0.37	0.37	0.37	0.48	0.29	0.30	0.28

(a)

Window Size	128			256			512		
Noise Threshold	2	6	10	2	6	10	2	6	10
Category									
1	2.03	2.04	2.04	2.07	2.08	2.07	2.01	2.10	2.04
2	2.06	2.17	2.10	2.40	2.27	2.56	2.32	2.12	2.07
3	2.09	2.05	2.05	2.10	2.06	2.07	2.10	2.06	2.06
4	1.90	1.93	1.95	1.92	1.97	1.99	1.91	2.01	1.98
5	1.89	1.95	1.96	1.66	1.66	1.66	1.68	1.68	1.68
6	1.50	1.51	1.51	1.51	1.51	1.51	1.54	1.54	1.54
Average	1.91	1.94	1.93	1.94	1.92	1.98	1.93	1.92	1.90
Std. Dev.	0.22	0.23	0.22	0.32	0.28	0.37	0.28	0.24	0.23

(b)

Window Size	128			256			512		
Noise Threshold	2	6	10	2	6	10	2	6	10
Category									
1	2.21	2.21	2.20	2.30	2.30	2.30	2.23	2.36	2.26
2	2.14	2.30	2.18	2.75	2.51	3.00	2.64	2.30	2.22
3	1.86	1.78	1.79	1.98	1.90	1.90	2.07	1.97	1.97
4	2.00	1.91	1.91	2.05	1.99	2.01	2.08	2.07	2.06
5	2.43	2.40	2.39	2.00	1.97	1.94	2.05	2.00	2.03
6	2.04	1.98	1.98	2.08	2.07	2.07	2.13	2.13	2.13
Average	2.11	2.09	2.07	2.19	2.12	2.20	2.20	2.14	2.11
Std. Dev.	0.19	0.24	0.22	0.30	0.23	0.42	0.22	0.16	0.11

(c)

FIGURE 2. **Comparison between the MMSE STSA configurations. (a) SIG. (b) BAK. (c) OVL.**

The spectral gain factor for each frequency index is computed by evaluating, in order, Equations 6, 8, 5 and 4, substituting variables computed in the previous equations. If the calculated gain is infinite, due to precision errors in the Java implementation the gain is instead set to

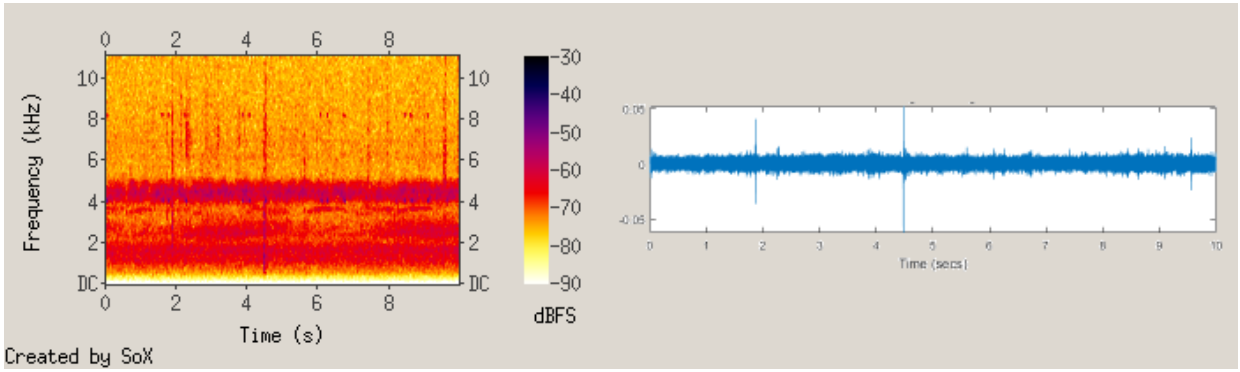
$$G_{MMSE}(k) = \frac{\xi_k}{1+\xi_k} \tag{10}$$

This occurs if the modified Bessel functions (see Equation 4) give very high values that are approximated to infinity in the implementation. This is an infrequent occurrence (it usually is not applied to any frames, and usually less than 100 frames out of 44000).

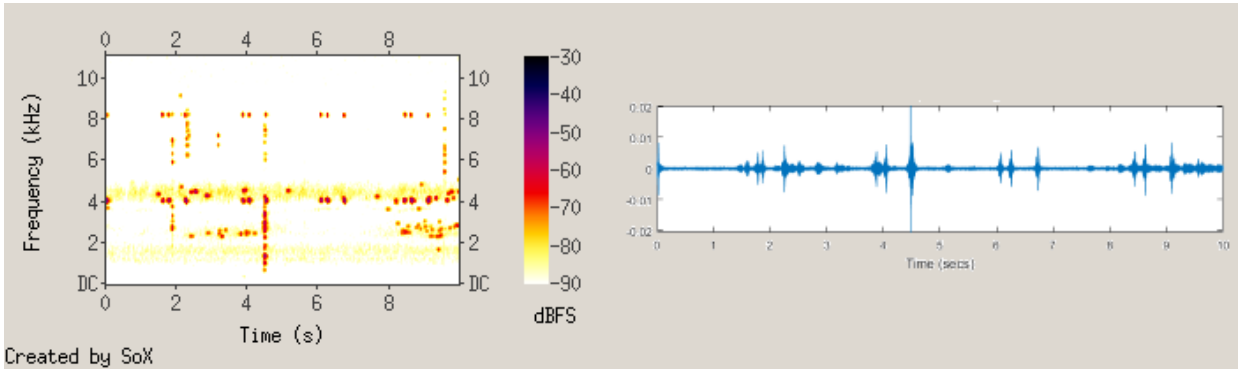
The magnitude of each discrete frequency component of the current window is multiplied by the computed gains for each of these components. The signal is converted back into the time domain using an Inverse Fast Fourier Transform. Windows are combined to form the processed signal, with overlapping components being added together. This signal is written to a new file.

IV. EXPERIMENTAL METHODOLOGY

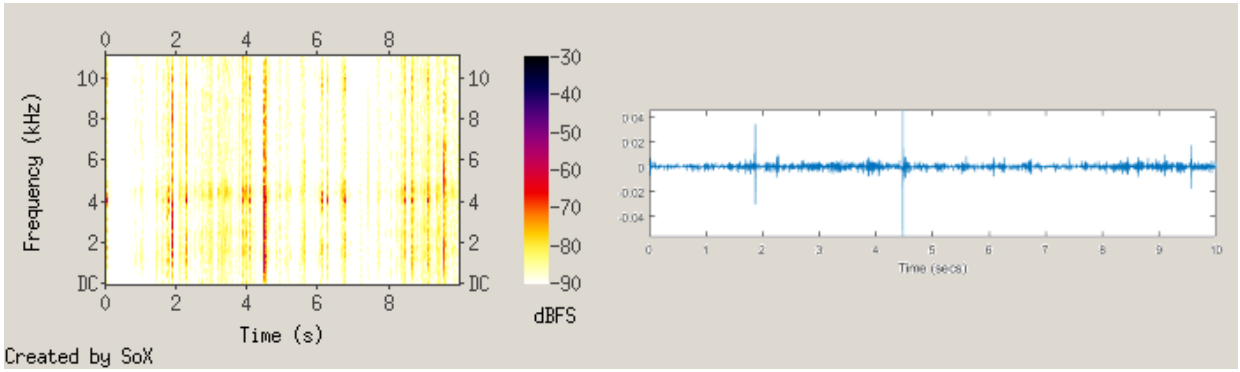
The aim of this research is to present a denoising method which can be use in automatic bird identification systems. In other words, the denoising method should be generally applicable in different situations and it should have low execution time. Accordingly, experiments are designed such that these features of our proposed denoising method can be evaluated. For experiments, we utilise real data collected from four different locations recorded by the Samford Ecological



(a)



(b)



(c)

FIGURE 3. Effect of high Shannon entropy on the Wavelet Transform algorithm in a Category 1 (Low SNR) recording in terms of spectrograms (left) and waveforms (right). (a) Raw. (b) Clean. (c) Wavelet Transform (Entropy = 4.5).

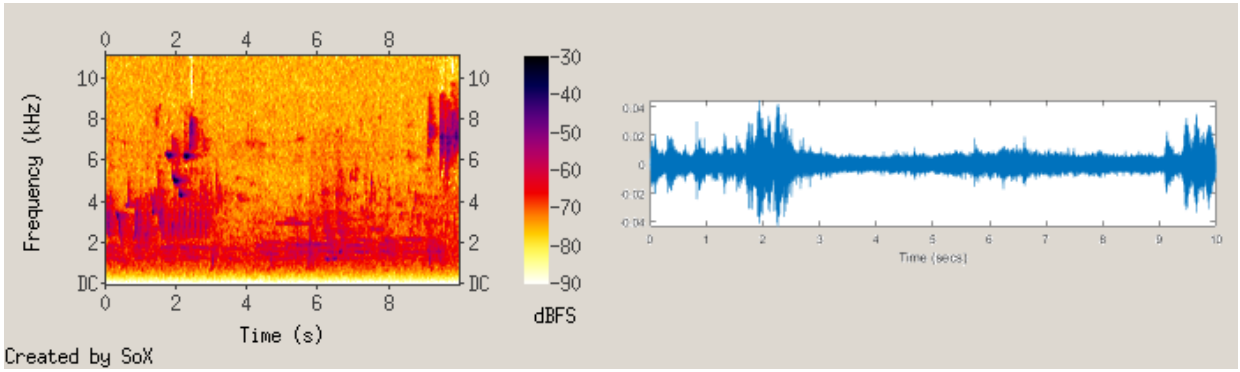
Research Facility (SERF), operated by the Queensland University of Technology. The SERF recordings were taken over five days between October 12 2010 and October 16 2010. Recordings from this group have been used in several research papers in the field [4], [25]. We randomly chose audio samples from these four locations from one day of this recording for evaluation. We conducted two types of experiments:

1) a sensitivity analysis of the algorithm to identify the most appropriate parameter values to effectively reduce noise from bioacoustics recordings without degrading the signal and minimising distortion; and
2) a comparison of the performance of the proposed method against that of the wavelet transform based method by Priyadarshani *et al.* [9].

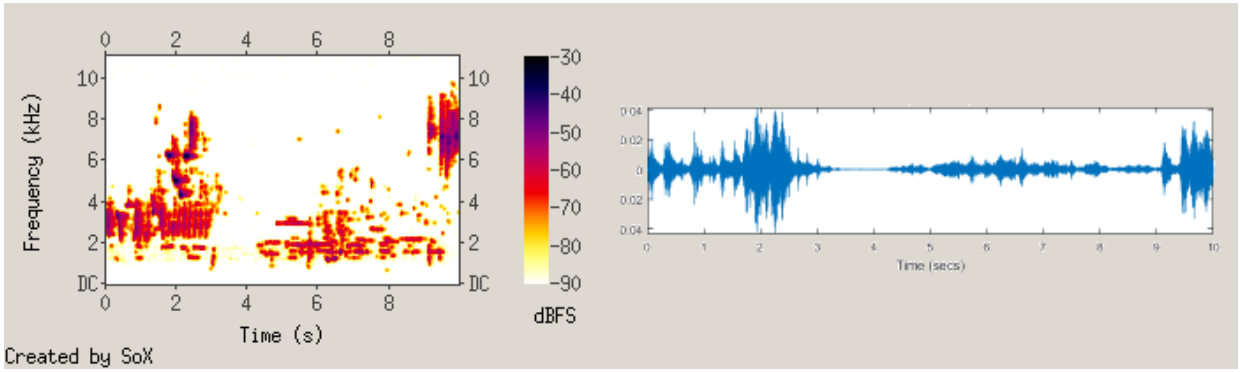
A. PERFORMANCE MEASURES

We measured the performance of our proposed method in **two** ways:

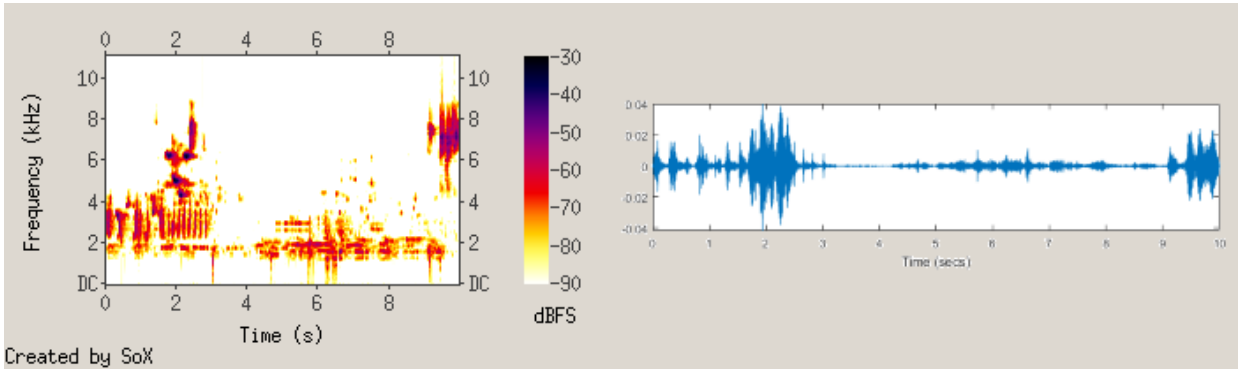
1) **Composite Evaluation Measures (SIG, BAK, and OVL)**: Composite measures [26] based on a linear combination of the Segmental SNR (SegSNR), Weighted-Slope Spectral Distance (WSS) [27], Perceptual Evaluation of Speech Quality (PESQ) [28], [29],



(a)



(b)



(c)

FIGURE 4. **Effect of high Shannon entropy on the Wavelet Transform algorithm in a Category 3 (High SNR) recording in terms of spectrograms (left) and waveforms (right). (a) Raw. (b) Clean. (c) Wavelet Transform (Entropy = 4.5).**

Log Likelihood Ratio (LLR), and Itakura-Saito (IS) distance [30] are evaluated for all filters and filter configurations. These are based on correlating these established evaluation metrics with a subjective evaluation of Signal Quality (SIG), Background Intrusiveness (BAK), and Overall Quality (OVL). The equations for these three metrics are:

$$C_{sig} = 3.093 - 1.029 \cdot LLR + 0.603 \cdot PESQ - 0.009 \cdot WSS \quad (11)$$

$$C_{bak} = 1.634 + 0.478 \cdot PESQ - 0.007 \cdot WSS + 0.063 \cdot SegSNR \quad (12)$$

$$C_{ovl} = 1.594 + 0.805 \cdot PESQ - 0.512 \cdot LLR - 0.007 \cdot WSS \quad (13)$$

Without the availability of truly clean audio recordings to compare the filters' results against, these recordings are compared to samples processed using an aggressive spectral noise gating approach, with noise profiles specifically selected for each recording. These recordings represent a good approximation of the true signal. Files are also down-sampled to 16 kHz for this evaluation, which gives a Nyquist frequency of 8 kHz, which is lower than some bird sounds [17], but still captures

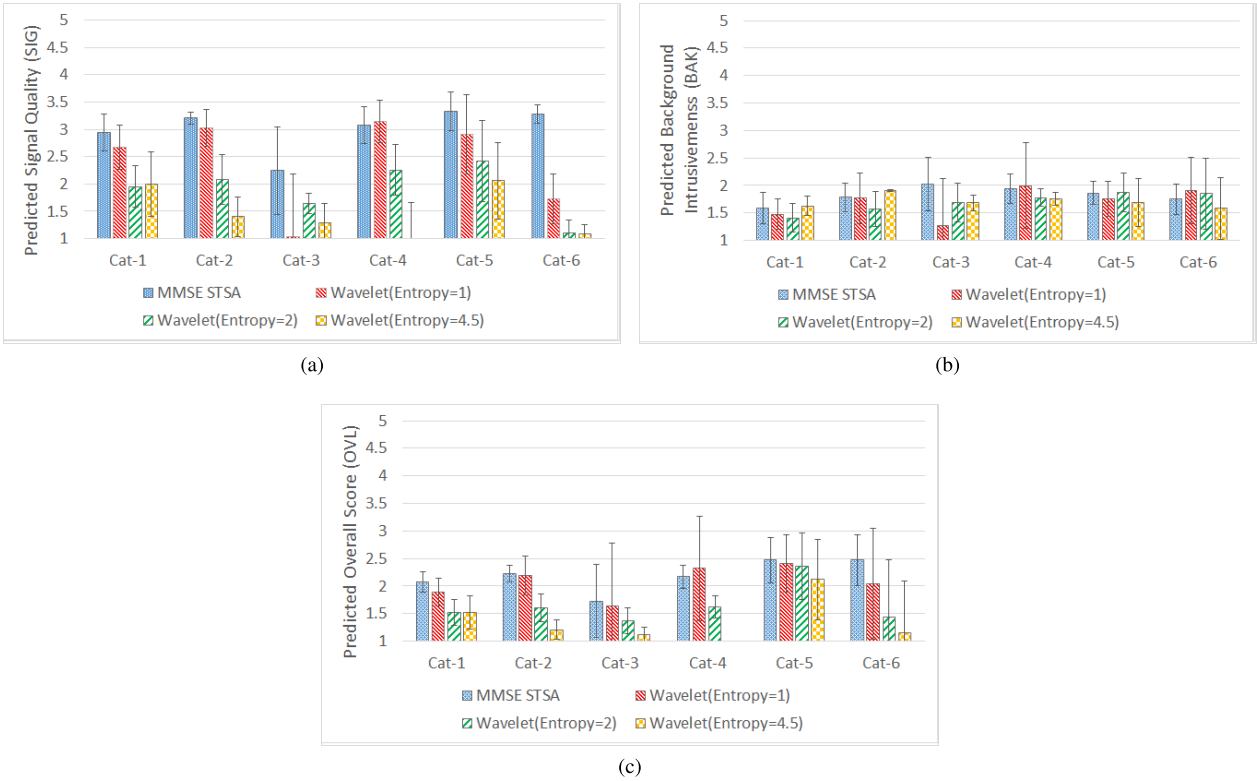


FIGURE 5. Comparison between the MMSE STSA and Wavelet Transform approaches at different Shannon entropies for each of the six categories. (a) SIG. (b) BAK. (c) OVL.

most of the soundscape. This is done to evaluate PESQ, which is needed to evaluate SIG, BAK, and OVL.

2) **Execution time:** The execution time is the time taken for denoising a bird acoustic recording. In general, the denoising step is just a pre-processing step in the whole automatic analysis of a recording. It is expected that it should take less time than the original recording to enable the overall analysis process to be efficient. Therefore, this metric is important to evaluate the practical usage of any denoising method.

V. EXPERIMENTAL EVALUATION

A. MMSE STSA PARAMETER ESTIMATION

We examined the MMSE STSA algorithm's sensitivity to two parameters that exhibit the largest impact on the audio output: window size and noise threshold.

- **Window size** is the number of samples in each frame that is processed. A sample is equal to a part of audio representing 1/(sample rate) of audio. For a sample rate of 22.05 kHz, this is equal to 1/22050 seconds of audio per sample. Lower window sizes give the highest time resolutions, at the expense of having the lowest frequency resolutions. They also produce aliasing artifacts, as discontinuities between different windows can occur when processing each window separately and then recombining. This is the motivation for using overlapping Hamming windows, although this does not completely solve the problem. Audio clips processed with lower window sizes sound more crisp, but also suffer more distortion compared to higher window sizes, which tend to sound cleaner, but also more 'washed out'. With extremely high window sizes, a reverberation effect is heard.
- **Noise threshold** affects how much noise is removed from the audio. It is defined as the minimum difference in dB between the mean noise level and the current level to be detected as a signal. Smaller values remove less noise, but are less prone to unintentionally removing good signals compared to larger values.

The experiment is conducted using 10-second excerpts from a day-long bioacoustics recording. Excerpts are selected and placed into one of six categories which have different properties to each other. The categories are summarised in Table 1. These are processed using the MMSE STSA approach testing for different window sizes and noise thresholds. Three window sizes (128, 256, and 512 samples), and three noise thresholds (2 dB, 6 dB, 10 dB) are tested in combination with each other, for a total of 9 configurations.

Composite measures are evaluated for each of the six categories, for each of the MMSE STSA configurations evaluated in the subjective listening tests. The results of these are shown in Figure 2.

The results show that, in terms of average performance, there is little difference between configurations: differences

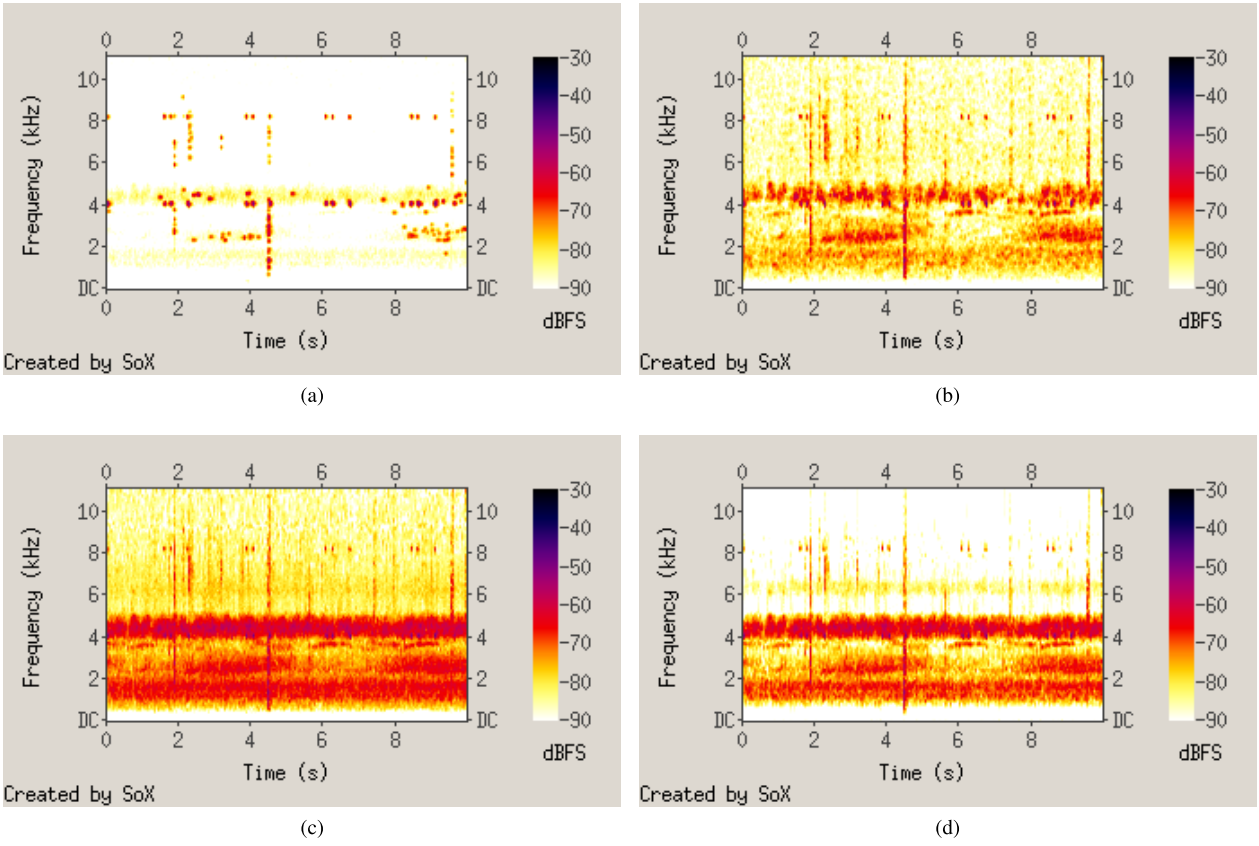


FIGURE 6. Spectrogram comparison of Filters for Category 1 recordings. (a) Clean. (b) MMSE STSA. (c) Wavelet Transform (Entropy = 1). (d) Wavelet Transform (Entropy =2).

TABLE 1. Categories of different experimental recordings.

Category	Recording characteristics
1	Low Noise, Medium Signal
2	Low Noise, Low Signal
3	Low Noise, High Signal
4	Light Rain, High Signal
5	Loud Cicadas (species 1, predominantly in 2 kHz region), Medium Signal
6	Loud Cicadas (species 2, predominantly in 1 kHz region), Medium Signal

between the best and worst configurations are within one standard deviation. Additionally, the 'clean' recordings are not truly clean, but are in fact denoised using a different approach, which introduces a confounding variable. Nonetheless, with a low standard deviation and equal highest average for the 'overall' metric, the configuration with a window size of 512 and noise threshold of 2 is identified as a strong configuration, and is selected for comparison between the MMSE STSA algorithm and the Wavelet Packet Decomposition approach with Shannon Entropy Threshold by Priyadarshani *et al.* [9]

The filter preserves signal more effectively than it removes background noise, as indicated by the much higher average values of SIG compared to those for BAK; surprisingly, BAK does not correlate with higher noise thresholds. Additionally, overall scores are low throughout. Average SIG is approximately 3, indicating *somewhat natural, somewhat degraded* sound, while the average BAK is approximately 2, indicating *fairly conspicuous, somewhat intrusive* background noise [26], although it is unclear whether this is because the MMSE STSA filter is poor, or the 'clean' comparison audio is problematic.

B. COMPARISON WITH WAVELET TRANSFORM

A MATLAB implementation of the Wavelet Transform technique by Priyadarshani *et al.* [9] is openly available to use, and is evaluated on the same target audio samples as the MMSE STSA algorithm. Using default settings, the noise filtering (indicated by Shannon entropy) removes too much information from the audio recordings, which may be observed in Figure 3. Better results are observed if the original SNR is sufficiently high (e.g., Figure 4), although a human listening test reveals that some signal information is lost or degraded in most cases.

Accordingly, for the following experiments, lower Shannon entropy thresholds are used to reduce the amount of signal degradation. The processing results with lower thresholds somewhat similar to MMSE STSA, although they tend to degrade more of the signal, contain more artifacts, and reduce less noise.

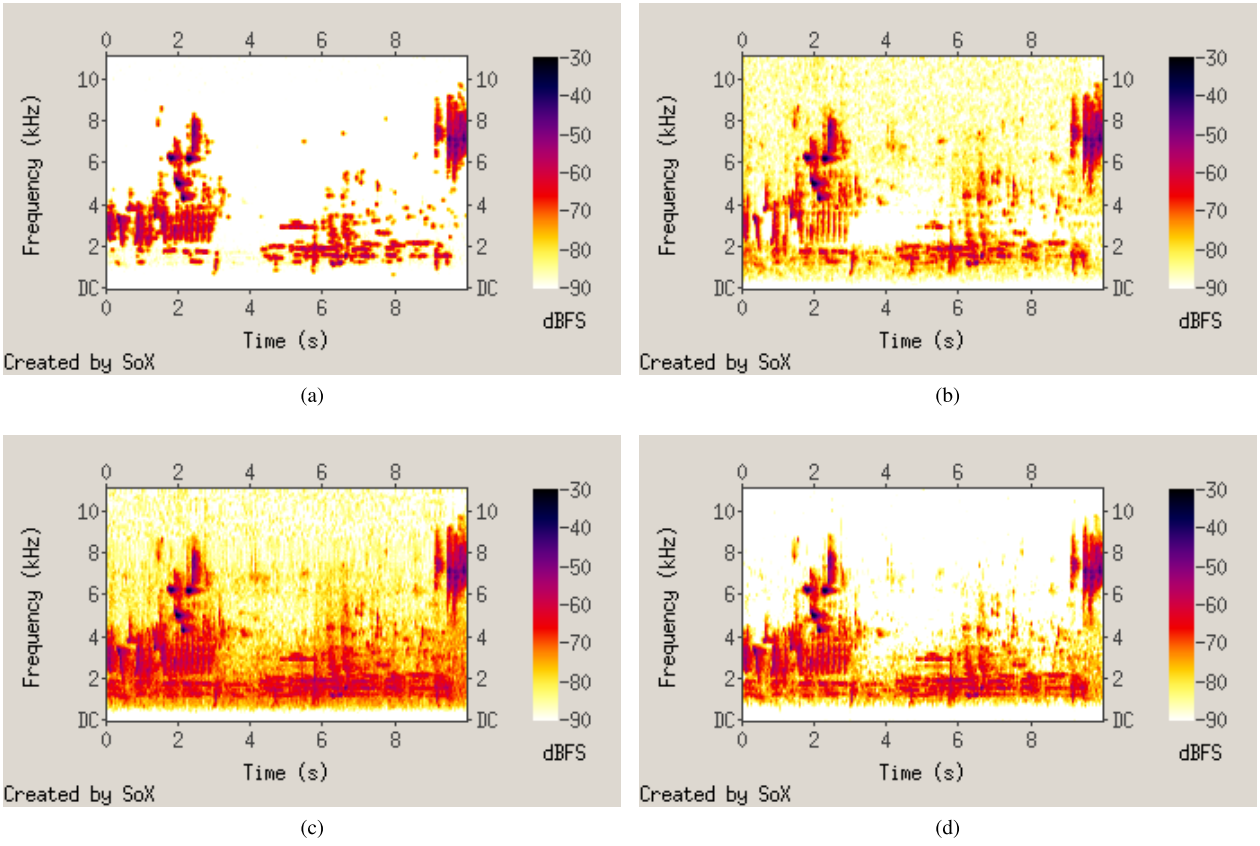


FIGURE 7. Spectrogram comparison of Filters for Category 3 recordings. (a) Clean. (b) MMSE STSA. (c) Wavelet Transform (Entropy = 1). (d) Wavelet Transform (Entropy = 2).

1) COMPOSITE EVALUATION METRICS

A test is conducted comparing the composite evaluation measures SIG, BAK, and OVL for MMSE STSA (Window Size = 512, Noise Threshold = 2) and the Wavelet Transform approach with different Shannon entropies, the results of which are shown in Figure 5. Each category is tested using 5 files each.

The results show that these composite indices vary significantly between files in the same category, as indicated by the large standard deviations (error bars in Figure 5). However, it appears likely that, for most categories, the MMSE STSA filter and wavelet transform-based filter with Shannon entropy equal to 1 outperform the wavelet transform approaches with higher Shannon entropies in terms of signal preservation (SIG). However, as shown in Figure 3 in some cases, the wavelet transform-based technique can significantly damage the signal at high Shannon Entropies. In some categories, most notably Category 6, MMSE STSA outperforms the wavelet transform technique for all Shannon entropy thresholds. For background noise intrusiveness (BAK), there is little difference between any filter for any category, which seems to contradict the spectrograms, which show large variations in the amount of background noise removed (see, for example, Figure 6).

Overall, these results indicate that it is unlikely that the wavelet transform technique is better than the MMSE STSA filter in improving the quality of a noisy bioacoustics recording. Additionally, Figures 6–8 suggest that the MMSE STSA filter is more effective in removing noise, while preserving signal, compared to the wavelet transform approach.

2) EXECUTION TIME

Table 2 shows the execution times of the proposed MMSE STSA and the Wavelet Transform algorithms. The experiment is conducted using a MATLAB implementation of both algorithms. Each algorithm is applied to one 10-second-long sample for each category and used a machine with an Intel Core i5-5200U @ 2.2 GHz (64-bit) processor and 8 GB RAM. The test is repeated five times for each file and an average is calculated. The MMSE STSA algorithm tested is a MATLAB implementation using its default settings. Note that this is different to the implementation used for evaluating the quality of the algorithm, which is Java-based and significantly faster. We use the MATLAB implementation here because the existing implementation of the Wavelet Transform algorithm provided by Priyadarshani *et al.* [9] is MATLAB-based. Default settings (with the Shannon Entropy set to 4.5) are used for testing, although in informal observations, changing the Shannon Entropy does not appear to have a significant effect on execution times. The algorithm is set to not perform band-pass filtering, as this is done to the raw audio prior to processing by the two algorithms.

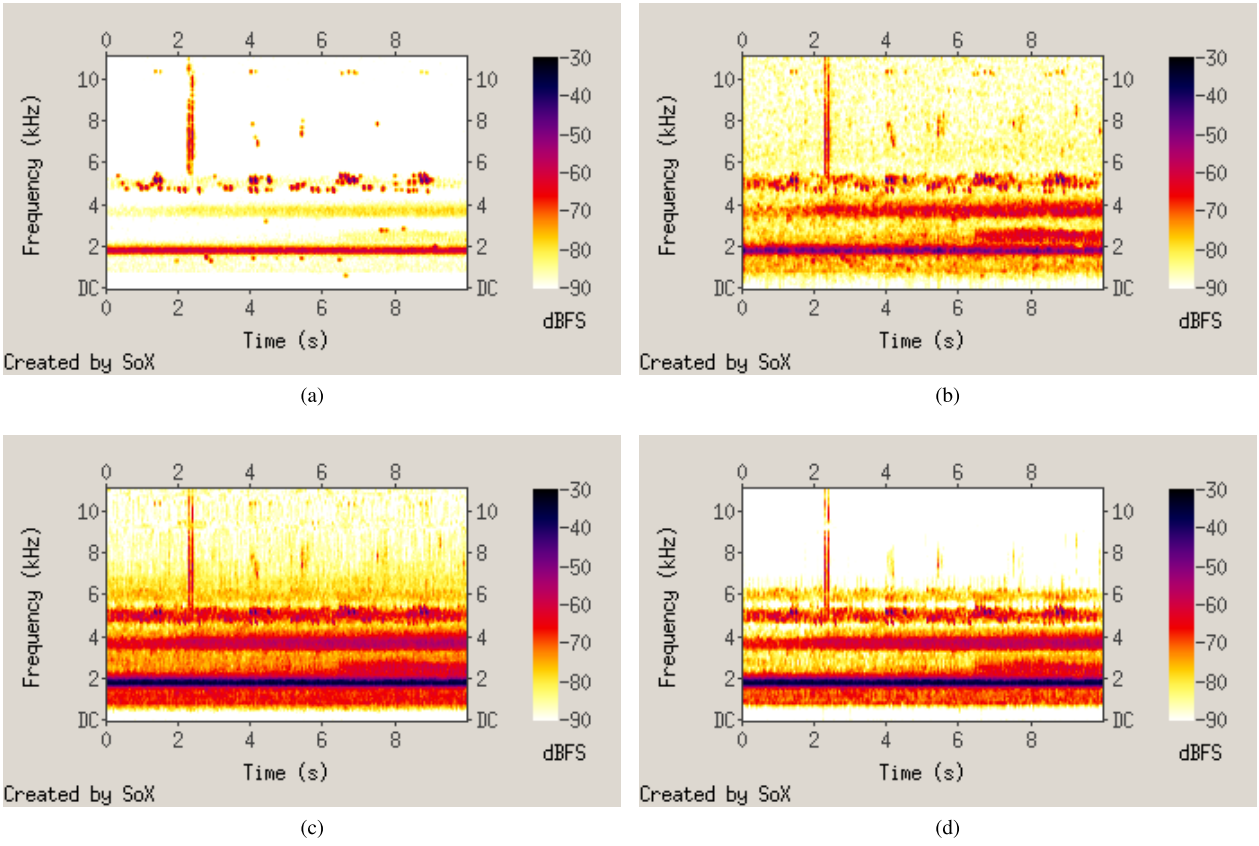


FIGURE 8. Spectrogram comparison of Filters for Category 5 recordings. (a) Clean. (b) MMSE STSA. (c) Wavelet Transform (Entropy = 1). (d) Wavelet Transform (Entropy = 2).

TABLE 2. Comparison of execution time in seconds.

Category	Average Execution time ± Standard Deviation	
	MMSE STSA	Wavelet Transform
1	2.20 ± 0.04	17.71 ± 3.10
2	2.11 ± 0.07	9.69 ± 1.37
3	1.98 ± 0.05	16.97 ± 3.03
4	1.97 ± 0.04	9.67 ± 0.81
5	1.98 ± 0.02	625.18 ± 17.64
6	1.98 ± 0.02	190.39 ± 14.77

The results indicate that the execution time of MMSE STSA is stable and takes about 2.0 seconds to denoise a sample of 10 seconds. In comparison, the wavelet transform approach's runtime is highly variable (from 9.67 to 625.18 seconds) depending on the file tested. Across the range of samples, the wavelet algorithm's average runtime to process 10 seconds of audio is 89 seconds, which is unacceptably high given that in practical scenarios recordings are of at least 24 hours. Hence, MMSE STSA appears better suited for denoising audio recordings in practical automated systems.

VI. CONCLUSION AND FUTURE DIRECTIONS

With the rapid growth in the number of audio recorders installed to continuously monitor different natural locations, automating the process of identifying bird species from bioacoustics recordings is a pressing need. However, these recorders are often unattended and the noise level is quite high, which makes reliable identification of bird volcalisations a difficult and time consuming task. In this paper, we proposed using the MMSE STSA filter in combination with a high-pass filter to efficiently and accurately denoise such recordings.

The MMSE STSA filter depends on two input parameters, window size and noise threshold. We first estimated the most appropriate settings using real bioacoustics recording samples with varying noise and bird call characteristics by evaluating composite measures for processing with different settings. We found that a window size of 512 and a noise threshold of 2 gave the highest average with the lowest standard deviation, though standard deviations for MMSE STSA are high, meaning this is not a definitive result.

We then compared the performance of our proposed method with a Wavelet Transform-based approach, one of the most recently proposed denoising method for bird acoustic recordings. Composite index testing showed that there is little difference between the wavelet transform with a Shannon Entropy of 1 and the MMSE STSA filter, and higher Shannon Entropy thresholds failed to preserve the signal as effectively. This can be observed in Figure 4. However, the execution time for MMSE STSA is considerably shorter than that of Wavelet Transform, by one to two orders of magnitude, and this increased execution time is not justified by any corresponding increase in filtering quality. In particular, MMSE STSA's

execution time is much lower than the length of the original audio being processed, which is essential if continuous recordings are to be processed in a reasonable time.

Even though MMSE STSA gives better results, there is still space to improve given there will be further processing of audio files which may be more complex and time consuming. In future, we will try to develop a parallel and scalable implementation of MMSE STSA utilizing GPUs to further reduce the processing time.

ACKNOWLEDGMENT

The authors would like to thank Scott Whitemore, and Adel Toosi for their valuable feedback. They also thank SERF for sharing with them their recordings which help them to successfully test their proposed method for denoising.

REFERENCES

[1] J. B. Alonso *et al.*, "Automatic anuran identification using noise removal and audio activity detection," *Expert Syst. Appl.*, vol. 72, pp. 83–92, 2017.
[2] I. Potamitis, S. Ntalampiras, O. Jahn, and K. Riede, "Automatic bird sound detection in long real-field recordings: Applications and tools," *Appl. Acoust.*, vol. 80, pp. 1–9, Jun. 2014.
[3] J. Cheng, Y. Sun, and L. Ji, "A call-independent and automatic acoustic system for the individual recognition of animals: A novel model using four passerines," *Pattern Recognit.*, vol. 43, no. 11, pp. 3846–3852, 2010.
[4] M. Towsey, J. Wimmer, I. Williamson, and P. Roe, "The use of acoustic indices to determine avian species richness in audio-recordings of the environment," *Ecol. Informat.*, vol. 21, pp. 110–119, May 2014.
[5] J. Xie, M. Towsey, A. Truskinger, P. Eichinski, J. Zhang, and P. Roe, "Acoustic classification of australian anurans using syllable features," in *Proc. IEEE 10th Int. Conf. Intell. Sensors, Sensor Netw. Inf. Process. (ISSNIP)*, Apr. 2015, pp. 1–6.
[6] P. M. Stepanian, K. G. Horton, D. C. Hille, C. E. Wainwright, P. B. Chilson, and J. F. Kelly, "Extending bioacoustic monitoring of birds aloft through flight call localization with a three-dimensional microphone array," *Ecol. Evol.*, vol. 6, no. 19, pp. 7039–7046, 2016.
[7] J. Salamon *et al.*, "Towards the automatic classification of avian flight calls for bioacoustic monitoring," *PloS ONE*, vol. 11, no. 11, p. e0166866, 2016.
[8] R. J. Willacy, M. Mahony, and D. A. Newell, "If a frog calls in the forest: Bioacoustic monitoring reveals the breeding phenology of the endangered Richmond Range mountain frog (Philoria richmondensis)," *Austral Ecol.*, vol. 40, no. 6, pp. 625–633, 2015.
[9] N. Priyadarshani, S. Marsland, I. Castro, and A. Punchihewa, "Birdsong denoising using wavelets," *PloS ONE*, vol. 11, no. 1, p. e0146790, 2016.
[10] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.
[11] B. M. Mahmmod, A. R. bin Ramli, S. H. Abdulhussain, S. A. R. Al-Haddad, and W. A. Jassim, "Signal compression and enhancement using a new orthogonal-polynomial-based discrete transform," *IET Signal Process.*, Aug. 2017. [Online]. Available: http://digital-library.theiet.org/content/journals/10.1049/iet-spr.2016.0449
[12] L. Neal, F. Briggs, R. Raich, and X. Z. Fern, "Time-frequency segmentation of bird song in noisy acoustic environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 2012–2015.
[13] M. C. Baker and D. M. Logue, "Population differentiation in a complex bird sound: A comparison of three bioacoustical analysis procedures," *Ethology*, vol. 109, no. 3, pp. 223–242, 2003.
[14] R. Bardeli, D. Wolff, F. Kurth, M. Koch, K.-H. Tauchert, and K.-H. Frommolt, "Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring," *Pattern Recognit. Lett.*, vol. 31, no. 12, pp. 1524–1534, 2010.
[15] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*, vol. 2. Berlin, Germany: Springer, 2009.
[16] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.
[17] B. C. Pijanowski *et al.*, "Soundscape ecology: The science of sound in the landscape," *BioScience*, vol. 61, no. 3, pp. 203–216, 2011.
[18] Y. Ren, M. T. Johnson, and J. Tao, "Perceptually motivated wavelet packet transform for bioacoustic signal enhancement," *J. Acoust. Soc. Amer.*, vol. 124, no. 1, pp. 316–327, 2008.
[19] A. Patti and G. A. Williamson, "Methods for classification of nocturnal migratory bird vocalizations using Pseudo Wigner-Ville transform," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2013, pp. 758–762.
[20] C. Bedoya, C. Isaza, J. M. Daza, and J. D. López, "Automatic recognition of anuran species based on syllable identification," *Ecol. Informat.*, vol. 24, pp. 200–209, Nov. 2014.
[21] C. Bagwell and U. Klauer. *Sox-Sound Exchange*. Accessed: Dec. 22, 2017. [Online]. Available: http://sox.sourceforge.net/
[22] A. D. Team. (2008). *Audacity (Version 1.2. 6) [Computer Software]*. Available: https://audacity.sourceforge.net/download
[23] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.
[24] B. M. Mahmmod, A. R. Ramli, S. H. Abdulhussain, S. Al-Haddad, and W. A. Jassim, "Low-distortion MMSE speech enhancement estimator based on laplacian prior," *IEEE Access*, vol. 5, pp. 9866–9881, 2017.
[25] A. Truskinger, M. Cottman-Fields, P. Eichinski, M. Towsey, and P. Roe, "Practical analysis of big acoustic sensor data for environmental monitoring," in *Proc. IEEE 4th Int. Conf. Big Data Cloud Comput. (BdCloud)*, Dec. 2014, pp. 91–98.
[26] Y. Hu and P. C. Loizou, "Evaluation of objective measures for speech enhancement," in *Proc. Interspeech*, 2006, pp. 1447–1450.
[27] D. Klatt, "Prediction of perceived phonetic distance from critical-band spectra: A first step," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 7. May 1982, pp. 1278–1281.
[28] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 2. May 2001, pp. 749–752.
[29] *Subjective Test Methodology for Evaluating Speech Communication Systems That Include Noise Suppression Algorithm*, document ITU-T Rec P.835, P. ITU-T, 2003.
[30] S. R. Quackenbush, T. P. Barnwell, III, and M. A. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1988.

ALEXANDER BROWN received the B.Sc. degree in applied mathematics and the B.Comp. degree in 2016, and the B.ICT. (Hons.) degree with a focus on the effects of noise interference on bioacoustics recordings from the University of Tasmania in 2017, where he is currently pursuing the Ph.D. degree.

SAURABH GARG received the Ph.D. degree from The University of Melbourne. He is currently a Lecturer with the University of Tasmania, Australia. He has authored over 40 papers in highly cited journals and conferences. His research interests include resource managements, scheduling, utility and grid computing, cloud computing, green computing, wireless networks, and ad hoc networks. He received various special scholarships for his Ph.D. candidature.

JAMES MONTGOMERY received the BInfTech degree (Hons.) and the Ph.D. degree in computer science from Bond University, Gold Coast, Australia, in 2000 and 2005, respectively. He held post-doctoral positions with the Swinburne University of Technology and The Australian National University. He is currently a Lecturer with ICT, University of Tasmania, Hobart, Australia. His research interests span evolutionary computation, machine learning, and web services.