Spectrogram denoising and automated extraction of the fundamental frequency variation of dolphin whistles

Asitha Mallawaarachchi and S. H. Ong[a)]
Department of Electrical and Computer Engineering, National University of Singapore, 9 Engineering Drive 1, Singapore 117576, Singapore

Mandar Chitre
Acoustic Research Laboratory, Tropical Marine Science Institute, National University of Singapore, 12a Kent Ridge Road, Singapore 119223, Singapore

Elizabeth Taylor
Marine Mammal Research Laboratory, Tropical Marine Science Institute, National University of Singapore, 14, Kent Ridge Road, Singapore 119223, Singapore

(Received 2 January 2007; revised 23 April 2008; accepted 27 May 2008)

Marine mammal vocalizations are often analyzed using time-frequency representations (TFRs) which highlight their nonstationarities. One commonly used TFR is the spectrogram. The characteristic spectrogram time-frequency (TF) contours of marine mammal vocalizations play a significant role in whistle classification and individual or group identification. A major hurdle in the robust automated extraction of TF contours from spectrograms is underwater noise. An image-based algorithm has been developed for denoising and extraction of TF contours from noisy underwater recordings. An objective procedure for measuring the accuracy of extracted spectrogram contours is also proposed. This method is shown to perform well when dealing with the challenging problem of denoising broadband transients commonly encountered in warm shallow waters inhabited by snapping shrimp. Furthermore, it would also be useful with other types of broadband transient noise. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2945711]

PACS number(s): 43.66.Gf [MCH] Pages: 1159–1170

I. INTRODUCTION

Marine mammal vocalizations are recorded and studied for a variety of purposes, including research on behavioral and contextual associations, animal detection and localization, and census surveys. Some of these natural signals of interest are nonstationary waves well suited for analysis using time-frequency representations (TFRs). In this paper, we focus on the analysis of dolphin vocalizations using spectrograms, which are TFRs based on the short-time Fourier transform (STFT).

The ability of dolphins to communicate acoustically has been instrumental in many field studies. For example, communication signals commonly referred to as "whistles" have been used to identify individuals or groups of animals,[1,2] and to determine the acoustic features salient to the animals.[3] These studies provide valuable insights into the dolphin's vocal repertoire and behavioral associations.

Dolphin whistles are commonly characterized by their time-frequency contours in spectrograms that highlight nonstationarities and provide effective visual means of differentiating whistles from other acoustic signals. Features extracted from spectrogram contours of dolphin whistles have been used in many classification studies.

However, dolphin whistle contours are often corrupted by other underwater acoustic sources (noise), making extraction difficult, and it is therefore done manually. However, when a large number of whistles are to be extracted, time spent on extraction could be significant. Therefore, an automated method is highly desirable.

The semiautomated method of Buck and Tyack[4] is able to extract whistle contours if the start and end points are known *a priori*. For each time bin between the beginning and end of a whistle, the algorithm selects the frequency bin with the highest energy as a pixel on the contour. A further check is performed to avoid choosing pixels on high-energy harmonics. However, this method only works well for recordings with a high signal-to-noise ratio (SNR), which can be a difficult requirement to satisfy in natural waters.

A noise removal method to facilitate contour extraction in natural waters has been proposed by Sturtivant and Datta.[5,6] In this method, dolphin echolocation clicks (broadband transient signals) are removed by the sequential application of a vertical edge suppression filter and an exponential smoothing filter. After denoising, potential whistles are detected by local segmentation, followed by connectivity thresholding to obtain tonal components that are longer than a predefined time duration. The contour is traced on the original spectrogram using an "inertial" whistle-following technique, starting from candidate points detected by the previous segmentation step. The start and end points of whistles are identified by drops in local SNR.

However, this algorithm is not widely known, and there is no indication in the original publication to how it would perform for warm shallow water recordings. Since we had

[a)]Also with Division of Bioengineering. Electronic mail: eleongsh@nus.edu.sg

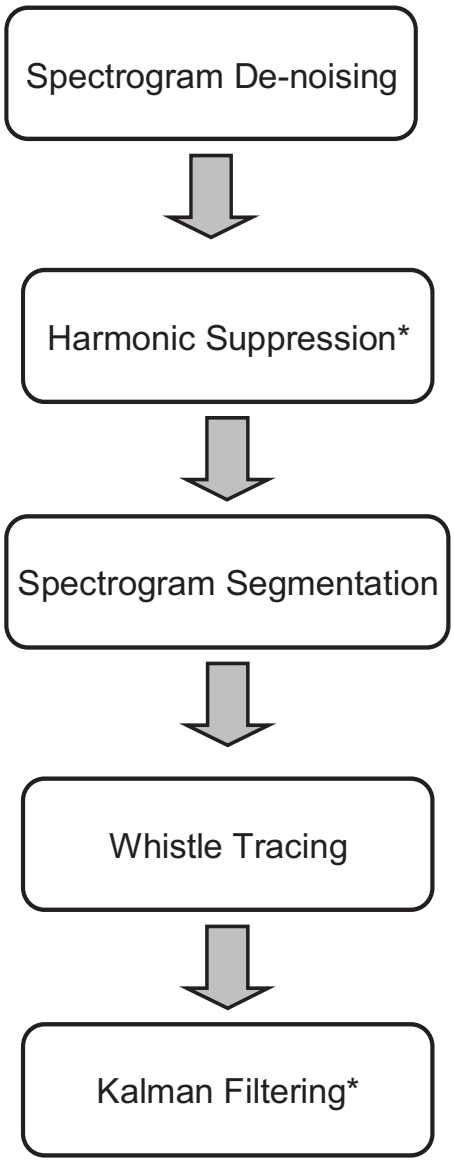

FIG. 1. Block diagram of the proposed algorithm. An asterisk denotes optional processing steps.

independently developed an algorithm consisting of (1) an image-based transient suppression filter to remove acoustic noise and (2) an adaptive image segmentation and tracing method to extract whistle contours from spectrograms, we conducted a comparative study to analyze the performance of the two methods. A flowchart which illustrates the method described in this work is given in Fig. 1.

Comparing the constituent elements of the two approaches, it may be noticed that while the denoising step of the Sturtivant and Datta method (SD) includes only a vertical edge suppression element, our proposed method also includes a horizontal smoothing element. Our experimental results (Sec. VII) indicate that this leads to better preservation of the signal while suppressing transient noise. The local normalizations used for denoising in SD are computed locally, whereas our method uses static directional filter kernels which are more efficiently implemented. Furthermore, the core algorithm components of denoising, segmentation, and tracing (excluding the optional steps—Kalman filtering and harmonic suppression) require fewer manual settings of parameters.

A denoised spectrogram may be converted back into a clean acoustic signal for play-back provided phase information is retained. Although the algorithm was developed for extracting dolphin vocalizations it has the potential to be used to extract other narrow-band nonstationary signals such as the humpback whale song.

Also introduced is an objective method of measuring the accuracy of tracing a vocalization spectrogram contour. This not only enables the comparison of tracing methods, but also helps tune the algorithm parameters such that optimal results can be obtained in a particular noise environment. Results obtained using recordings of vocalizations made by bottlenose dolphins (*Tursiops truncatus aduncus*) and Indo-Pacific humpback dolphins (*Sousa chinensis*) in warm shallow waters around Singapore confirm that our proposed method effectively denoises a wide variety of whistle contours and performs better than other methods in most instances tested.

II. SPECTRAL PATTERNS OF COMMON ACOUSTIC SIGNALS

A spectrogram is produced by converting a time-domain signal to the joint time-frequency domain by the STFT. Formally, the STFT of a discrete-time function $x[n]$ with respect to the window function $w[n]$ evaluated at the location $[\omega,m]$ in the frequency-time plane is defined as

$$X[\omega,m] = \sum_{n=-\infty}^{\infty} x[n]w[n-m]\exp(-j\omega n). \quad (1)$$

The columns of the matrix $X[\omega,m]$ contain the time-localized frequency content of the discrete signal $x[n]$. The values of X are generally complex, and it is customary to log-compress the absolute value of this transform for visual inspection due to the large dynamic range. This log-compressed gray-level two-dimensional (2D) image is called the spectrogram of $x[n]$, and is denoted by $\hat{X}[\omega,m]$.

In underwater environments where dolphin vocalizations are recorded using hydrophones, there are various types of acoustic sources: mechanical, such as produced by ships; natural physical sources such as waves; and biological sources. It is important that the spectral characteristics of this noise are taken into consideration when attempting to isolate dolphin whistles.

The two most commonly encountered dolphin vocalizations are narrow-band, frequency-modulated whistles and short-duration, broadband echolocation clicks. Whistles give rise to smooth frequency-localized contours [Fig. 2(a)] in a spectrogram, while clicks create vertical line patterns [Fig. 2(b)]. Signals generated by mechanical processes usually have low, constant frequencies, resulting in spectral patterns consisting of horizontal lines in the lowest regions of a spectrogram [Fig. 2(c)].

Ambient noise in warm shallow waters worldwide is dominated by the short-duration broadband crackling or popping sounds made by snapping shrimp, which has been shown to have a non-Gaussian energy distribution.[7] This

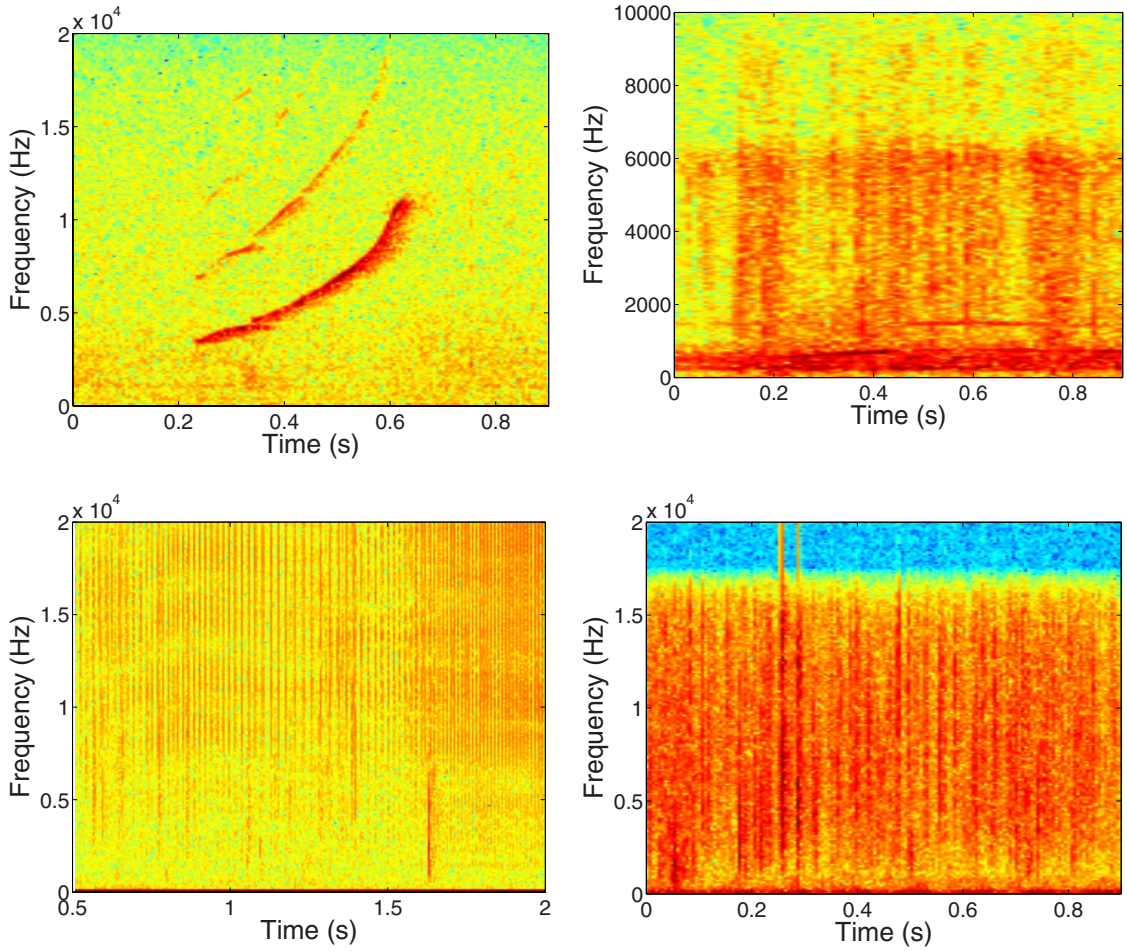

FIG. 2. (Color online) Common spectral patterns.

sound creates spectral patterns resembling narrow vertical lines. However, since many snapping shrimp produce these sounds simultaneously, individual "snaps" overlap and the resulting vertical structures are not as clearly defined as dolphin clicks [Fig. 2(d)].

III. SPECTROGRAM DENOISING

Spectrogram images are a special class of images that are not the product of conventional optical imaging. Noise in these images must be defined according to the higher-level detection task one is attempting. Unlike optical images, where the image noise usually arises from imperfections in acquisition or transmission, spectrogram noise is due mainly to the presence of "undesirable" acoustic sources. Some of these noise sources are introduced by human activities, while others are inherent in the recording environment and are described in Sec. II.

When detecting dolphin whistles all other acoustic sources are treated as noise sources; hence the spectral patterns to which they give rise in the resulting spectrograms are defined as noise. The aim of denoising a spectrogram is to facilitate the extraction of the desired type of spectral patterns by attenuating all other patterns. This paper introduces image processing methods to achieve this objective.

As a preprocessing step, the low-frequency tonal sounds created by mechanical devices such as motors and engines can easily be removed by high-pass filtering with a cut-off frequency set slightly below the lowest frequency at which dolphin whistles are expected ($\sim$1.5 kHz). However, if this method were to be used for denoising other types of signals, such as the low frequency calls of baleen whales, the band of interest would have to be modified accordingly. In the latter case, the use of a low-pass filter might be appropriate.

A. Denoising in nonimpulsive noise environments

The quality of spectrogram images of recordings made in pool environments and not excessively corrupted by transient noise can be improved significantly by an edge-preserving local-smoothing filter such as the bilateral filter.[8] This is essentially a neighborhood averaging filter with the kernel coefficients computed from the geometric closeness and the gray level similarity between the neighborhood center and the other neighborhood pixels.

If the center pixel of the local neighborhood being processed in spectrogram $\hat{X}$ is $\hat{X}[\omega,m]$ and any other pixel belonging to the same local neighborhood is denoted by $\hat{X}[\omega',m']$, the geometric closeness function, c, depends only on the relative positions of the two pixels $x=[\omega,m]$ and $\xi=[\omega',m']$. The gray level similarity function s, on the other hand, is a function of the relative pixel intensities $\hat{X}[x]$ and $\hat{X}[\xi]$.

In the shift-invariant Gaussian implementation of the bilateral filter, both the closeness and similarity functions are Gaussian functions of their respective arguments. Thus, we have

$$c(\xi,x) = \exp\left[-\frac{1}{2}\left(\frac{d(\xi,x)}{\sigma_d}\right)^2\right], \tag{2}$$

where the Euclidian distance $d(\xi,x)$ between ξ and x is

$$d(\xi,x) = d(\xi - x) = \|\xi - x\|$$

and

$$s(\xi,x) = \exp\left[-\frac{1}{2}\left(\frac{\delta(\hat{X}[\xi],\hat{X}[x])}{\sigma_r}\right)^2\right] \tag{3}$$

where

$$\delta(\hat{X}[x],\hat{X}[\xi]) = \|\hat{X}[\xi] - \hat{X}[x]\|.$$

Using the the aforementioned definitions, bilateral filtering can be described by

$$\hat{X}_{\mathrm{BF}}(x) = \frac{1}{k(x)}\sum_{\xi}[\hat{X}[\xi]c(\xi,x)s(\hat{X}[\xi],\hat{X}[x])], \tag{4}$$

where $\hat{X}_{\mathrm{BF}}[x]$ is the output of the filter operation on pixel $\hat{X}[x]$, and the summation is performed over all the neighborhood points ξ. The normalization factor $k(x)$ is obtained by

$$k(x) = \sum_{\xi}[c(\xi,x)s(\hat{X}[\xi],\hat{X}[x])]. \tag{5}$$

The parameter values of the bilateral filter depend on the size of the features one desires to preserve and the amount of smoothing preferred. The window size should be larger than the whistle-contour thickness, which depends on the time-frequency resolution set by the fast Fourier transform (FFT) window size. The Gaussian kernel width σ_d should be a fraction of the window size, while σ_r should be a fraction of the range of gray levels.

B. Short-duration transient suppression

Recordings made in open waters are typically more challenging to denoise. The vertical line patterns created by dolphin clicks and snapping shrimp can overlap whistles and complicate the tracing process. Pixels that are part of these spectral patterns should be detected and attenuated before whistle tracing.

The dominant direction of energy distribution in the local neighborhood of each pixel is detected using a set of four asymmetric kernels generated from the Gaussian functions (7),

$$G_1(p,q) = \exp\left[-\frac{1}{2}\left(\left(\frac{p}{\sigma_p}\right)^2 + \left(\frac{q}{\sigma_q}\right)^2\right)\right] \tag{6}$$

and

TABLE I. Asymmetric Gaussian kernels.

Kernel	Generating function	Value of σ_p	Value of σ_q	Orientation
$\nu_1(p,q)$	$G_1(p,q)$	a	$6a$	Horizontal
$\nu_2(p,q)$	$G_1(p,q)$	$6a$	a	Vertical
$\nu_3(p,q)$	$G_2(p,q)$	a	$6a$	Diagonal
$\nu_4(p,q)$	$G_2(p,q)$	$6a$	a	Diagonal

$$G_2(p,q) = \exp\left[-\left(\left(\frac{q-p}{\sigma_p}\right)^2 + \left(\frac{p+q}{\sigma_q}\right)^2\right)\right]. \tag{7}$$

The four Gaussian kernels ν_i $i \in \{1,\ldots,4\}$ given in Table I are oriented in the horizontal, vertical, and two diagonal directions. The values of σ_p and σ_q are chosen according to the kernel size, and in Table I, a is used to indicate their relative magnitudes. Generally, for a kernel of size $M \times M$, $a \approx M/10$ is recommended. In this work, a 9×9 window is used.

The spectrogram image $\hat{X}$ is filtered by ν_i to produce four intermediate images $\hat{X}_i$. Denoting the neighborhood center by $[p,q]$ and a pixel in the neighborhood by $[p',q']$, the intermediate images are given by

$$\hat{X}_i[p,q] = \frac{1}{t(p,q)}\sum_{p'}\sum_{q'}[\hat{X}[p',q']\nu_i(\|p'-p\|,\|q'-q\|)] \tag{8}$$

with the normalization factor $t(p,q)$ defined by

$$t(p,q) = \sum_{p'}\sum_{q'}\nu_i(\|p'-p\|,\|q'-q\|). \tag{9}$$

Since the output of the functions ν_i can be precomputed for a given size of the local neighborhood, the above operations can be efficiently implemented.

To remove the vertical spectral patterns, pixels belonging to local neighborhoods with vertical energy distributions are attenuated. Let $r(\xi)$ be the highest nonvertical energy average, and $v(\xi)$ the vertical energy average (Algorithm 1, see Table II). Therefore the expression $r(\xi) - v(\xi)$ evaluates to a positive value when the primary direction of energy distribution is nonvertical, and negative when it is vertical. Adding the expression $r(\xi) - v(\xi)$ to the original pixel value therefore has the effect of attenuating the pixels with a vertical energy distribution. The constants α and β are used to control the degree of attenuation.

Only the relative values of α and β are important. Higher relative α preserves more of the original detail while higher relative β increases the amount of attenuation of ver-

TABLE II. Algorithm 1: Transient suppression.

1:	**for all** pixel $\xi=(p,q)$ of $\hat{X}(\xi)$ **do**
2:	$r(\xi)=\arg\max \hat{X}_1(\xi),\hat{X}_3(\xi),\hat{X}_4(\xi)$
3:	$v(\xi)=\hat{X}_2(\xi)$
4:	$\hat{X}_{TS}(\xi)=[\alpha\times\hat{X}(\xi)+\beta\times(r(\xi)-v(\xi))]/[\alpha+\beta]$
5:	**end for**

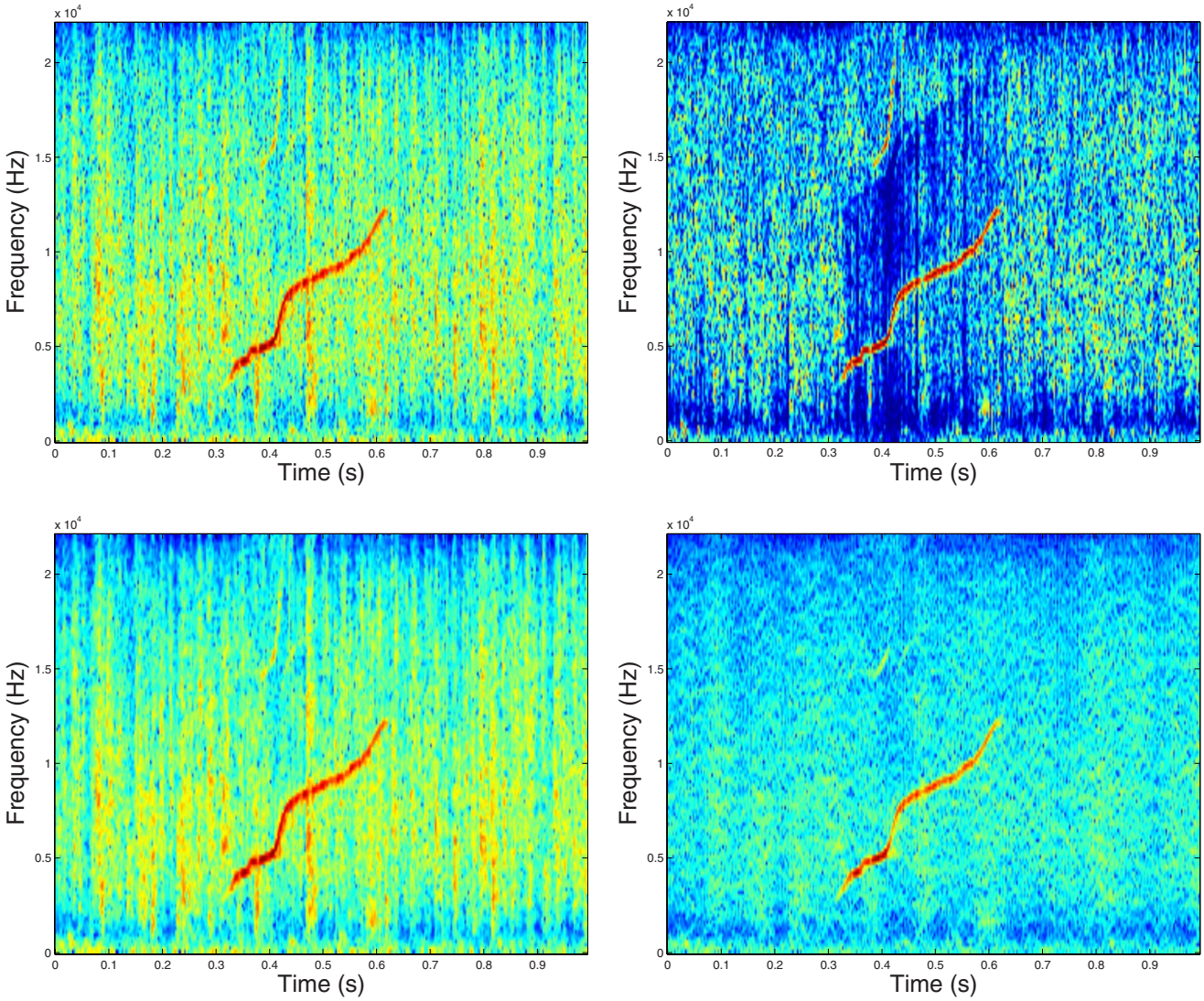

FIG. 3. (Color online) Denoising in the presence of transient noise.

tical spectral patterns. Although the procedure is described here in its noniterative mode of operation, there is no restriction on using the filter iteratively.

An example of applying this filter to a whistle corrupted by snapping shrimp noise is shown in Fig. 3 together with the output of bilateral filtering and an implementation of the method of SD.[6] Bilateral filtering [Fig. 3(b)] enhances connected high-energy regions irrespective of orientation and therefore produces poor denoising performance. The method of SD [Fig. 3(c)] reduces most of the transients and creates a "noise trough" around the whistle, but leaves behind a significant amount of noise pixels. In comparison, the proposed transient suppression filter [Fig. 3(d)] removes most of the undesired vertically oriented spectral patterns while preserving the whistle contour.

IV. HARMONIC SUPPRESSION

Whistles often contain harmonics similar in shape to the fundamental frequency variation with only a shift in frequency, and these can potentially hinder accurate tracing of the fundamental. The instantaneous frequency of a harmonic is an integer multiple of the fundamental, a property that can be exploited to automatically remove it from the spectrogram image.

A row of pixels in a spectrogram represents the time variation of a discrete frequency bin f_i, and, from bottom to top, the rows represent a linear increase in frequency. Let us define a pixel intensity vector $\mathbf{I}_i$ that contains the pixels in the ith row of the spectrogram. The harmonic suppression update equation for the ith row is expressed as

$$\mathbf{I}_i - \mathbf{I}_i - k_h \mathbf{I}_j, \tag{10}$$

where k_h is a user-defined scalar constant and the vector $\mathbf{I}_j$ contains the pixel values of the jth row for which $f_j = f_i/N$, where N is an integer. For example, if $N=2$ is used, harmonics which are $\{2, 4, 6,\ldots\}$ multiples of the base frequency will be attenuated. Therefore by repeating the procedure for different values of N, most harmonic patterns may be attenuated. A good choice for N is the set of the largest common divisors of the integer multiples of the fundamental that produced the harmonic pattern. In practice, $N \in \{2,3,5\}$ is used, and Eq. (10) is applied to every row from top to bottom, and iterated for each value of N.

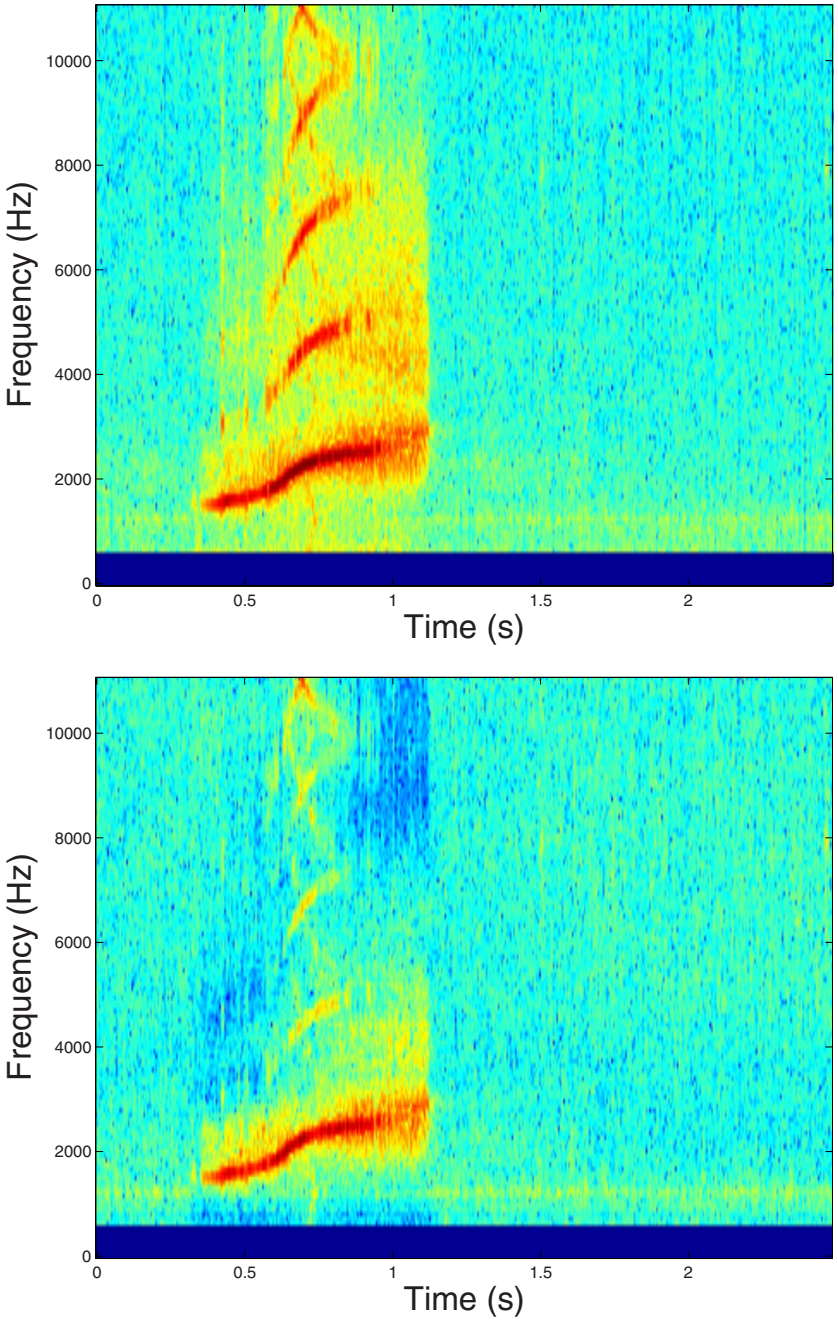

FIG. 4. (Color online) Iterative harmonic suppression.

Figure 4 shows how harmonics are suppressed from the spectrogram by this iterative procedure; the harmonics now have lower energy, while the fundamental signal retains intensity values similar to the original. This step should be used selectively since not all whistle recordings capture strong harmonic patterns. If multiple whistles overlap in time, and the fundamental (frequency variation) of one whistle intersects the harmonic pattern of another whistle, the harmonic suppression algorithm will attenuate the pixels belonging to the fundamental of the intersecting whistle. This algorithm should therefore not be used in such situations.

V. SPECTROGRAM SEGMENTATION

After the spectrogram has been denoised to remove unwanted spectral patterns, whistles can be extracted using image segmentation techniques. Although many segmentation techniques have been proposed, the objective of the current work is to choose a method applicable to spectrogram images and simple enough to be efficiently implemented. The latter objective is important because an on-line tool for whistle extraction would be very helpful to field scientists studying dolphins and other marine mammals.

Thus, a three-stage image segmentation technique is proposed. Thresholding is chosen as the first step as it can be efficiently implemented and requires only one parameter (the threshold) to be determined. Furthermore, there are several known methods of adaptively computing a threshold. Having the ability to adaptively select a threshold is crucial in spectrogram images because the ambient intensity values can greatly vary from one image to the other, even in spectrograms taken from different sections of the same recording.

Factors contributing to these variations include the dynamic nature of underwater acoustic sources, and the settings of the recording equipment used. Further details on thresholding are given in Sec. V A.

As an intermediate step, a morphological clean-up operation is performed on the output of the thresholding operation (Sec. V B).

Dolphin whistles are generally continuous contours with fading starts and ends. Therefore, the start and end segments are likely to be misclassified as background by thresholding. However, these segments are connected to other high intensity parts of the whistle that are readily identified by thresholding. We, therefore, propose the use of an intensity-based region-growing algorithm for the final stage as described in Sec. V C.

A. Adaptive thresholding

A thresholding function f operates on an *intensity* image I and produces a *binary* image J, using a global threshold T,

$$J(x,y) = \begin{cases} 1 & \text{if } I(x,y) \geqslant T \\ 0 & \text{else.} \end{cases} \tag{11}$$

In the first segmentation stage, we use a higher global threshold T value to ensure noisy pixels are unlikely to be segmented as foreground. This creates the possibility of coclassifying some segments of the whistle as background noise, but the subsequent region growing step compensates for this.

Since spectrograms vary significantly in average energy level, depending on the recording environment and the type of dolphin vocalizations, the threshold T has to be calculated adaptively. The *Niblack* method[9] is a well-known and simple method for computing an adaptive threshold,

$$T = \mu + k\sigma, \tag{12}$$

where μ is the mean gray value, σ the standard deviation, and k a user-defined constant. We determine k from

$$\int_{-\infty}^{\mu+k\sigma} N_{\mu,\sigma}(x)dx = \rho, \tag{13}$$

where ρ is the percentage of background pixels. A normal distribution of gray values may be assumed even though the actual distribution can differ.[10] Using the *a priori* knowledge that over 90% of a spectrogram is background, our proposed method computes a high threshold by setting $\rho=0.96$. The calculation of k using Eq. (13) is implemented by using precalculated values for a Gaussian cumulative distribution function, stored in a table format. This is also referred to as a Z table.

Note that a simple and fast algorithm is preferred because, in the first stage of segmentation, we are mostly interested in obtaining a set of seed points that we can feed into the region-growing algorithm.

B. Morphological clean-up operations

As an intermediate step, mathematical morphology is used to improve the segmentation and remove any noisy outlying pixels from J. In morphology, the structure of a group of pixels is considered rather the the pixel intensity values. All morphological operators are therefore defined with respect to a "structuring element." It is possible to use morphology to remove noisy outlying pixels based on structural characteristics such as size, by using an appropriate structuring element.

Two basic morphological operators are opening and closing. Opening with a structuring element S will erase image structures (groups of connected pixels) smaller than S, while closing will fill gaps (holes) smaller than S. Morphological closing was first performed on J with a 2×2 square structuring element (SE) s_1, followed by opening with a 2×3 SE s_2. This dual operation preserves original detail and removes any remnant noisy outlying pixels.

C. Region growing

After removing outlying pixels with the morphological operators, the pixels of the segmented image are input as seed points in a 2D region-growing algorithm. Region-growing algorithms take one or more seed points together with a threshold value T' as inputs, and initially the output image contains only the seed(s). As the algorithm progresses, the neighborhood pixels of the seed(s) above the threshold T' are located and added to the output image, and then the neighbors of the newly added points are searched. This recursive algorithm continues until all the pixels meeting the criteria are added to the output image.

VI. WHISTLE TRACING

The final goal of the algorithm is to extract the time-frequency contour of the fundamental frequency variation of a whistle. A trace of this extracted contour should also be drawn on the spectrogram for visual inspection.

The candidate points for the trace consist of the strongest peaks in each time bin of the segmented spectrogram.[4] This simple process is used for whistle tracing as the denoising and segmentation steps contain most of the algorithmic intelligence. However, strong harmonics or remnant background noise can cause incorrect segmentation and, hence, outlying points in the automatically obtained trace. In such situations, an additional step which can correct outlying points is beneficial.

Based on the assumption that dolphin whistles are smooth curves without sudden jumps in frequency, we propose the use of a Kalman filter,[11] which is a model-based estimator. A Kalman filter takes imprecise measurements of a process that can be mathematically modeled, and produces, as output, a weighted average of the model prediction and the measurement. The weights depend on the relative confidence in the measurements and the model prediction.

The confidence values for each measurement (a point on the automated trace) are calculated and saved during tracing. Low confidence values are assigned to whistle points that exhibit sudden jumps in frequency. The confidence assignment function has a memory of 1 in the sense that if the

TABLE III. Second-order Kalman filter model used for whistle smoothing.

State vector	$x=[f\ v\ a]$
	f—frequency (position)
	v—rate of change of frequency (velocity)
	a—second derivative of frequency (acceleration)
State equations	$f(k)=ut(k)+\frac{1}{2}at^2(k)$
	$v(k)=u+at(k)$
	where u is the initial velocity
Discrete update equations	$f(k+1)=f(k)+v(k)\delta_T+\frac{1}{2}a\delta_T^2$
	$v(k+1)=v(k)+a\delta_T$
	$a(k+1)=a(k)$ where $\delta_T=t(k+1)-t(k)$
	$\begin{bmatrix} f_{k+1} \\ v_{k+1} \\ \alpha_{k+1} \end{bmatrix} = \begin{bmatrix} 1 & \delta_T & \frac{1}{2}\delta_T^2 \\ 0 & 1 & \delta_T \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f_k \\ v_k \\ a_k \end{bmatrix}$

previous measurement had low confidence, the confidence attributed to the next measurement will be partially based on the previous value.

The same process model was used to "track" a smooth dolphin whistle as when tracking the trajectory of a particle moving in a straight line under constant acceleration. Frequency (f), which is the filtered variable with respect to time, is analogous to the position of the particle. In this context, velocity and acceleration correspond, respectively, to the first and second derivatives of frequency with respect to time. The Kalman model is summarized in Table III. The filter output is continuous and will be quantized to the nearest discrete Fourier transform (DFT) bin value. Figure 5 shows the plot of a traced whistle overlaid with the Kalman filter output. The whistle points adjusted by the filter are marked by ellipses.

One shortcoming of the Kalman filter is the slight shift of the trace to the right (on the time axis) caused by the filter. This indicates a "filter inertia," which resists changes in direction contrary to the model prediction. Effective use of Kalman filtering requires an appropriate trade-off between smoothness and flexibility (to change direction), and this can be done by setting the process and measurement error covariances. Another problem is that in situations where a large number of continuous outlying points are present, the filter will gradually adapt to those incorrect values and will fail to make the expected filter corrections. Despite these imperfections, Kalman filter corrections are beneficial in some situations.

VII. EXPERIMENTAL RESULTS

A. Experiment design

To evaluate the effectiveness of the automated tracing methods, 26 dolphin whistles were chosen from underwater recordings of vocalizations made by bottlenose dolphins (*Tursiops truncatus aduncus*) and Indo-Pacific humpback dolphins (*Sousa chinensis*) in waters surrounding Singapore. The sampling rate was 44.1 kHz, and spectrograms were created with an FFT window size of 256, with successive windows overlapping by 50%. Each whistle was approximately 0.5–1 s in duration and contained natural background noise dominated by snapping shrimp

The Marine Mammal Research Laboratory (MMRL) at the National University of Singapore provided the underwater recordings and the reference traces for the selected whistles. To obtain reference traces of the whistle contours, the spectrograms of the selected whistles were manually traced by the MMRL, and quantized to fit the spectrogram bins using a nearest neighbor algorithm. The whistle traces obtained by automated methods were then compared with the reference traces for quantitative evaluation of their performance using the metrics defined in the following.

B. Performance metrics

A quantitative method to evaluate the accuracy of an extracted whistle contour trace has not previously been proposed. This paper defines three metrics that can be used to gauge the completeness and accuracy of tracing a known whistle contour. They are defined with respect to a reference contour $\omega(m)$, with associated times $t(m)$ obtained by

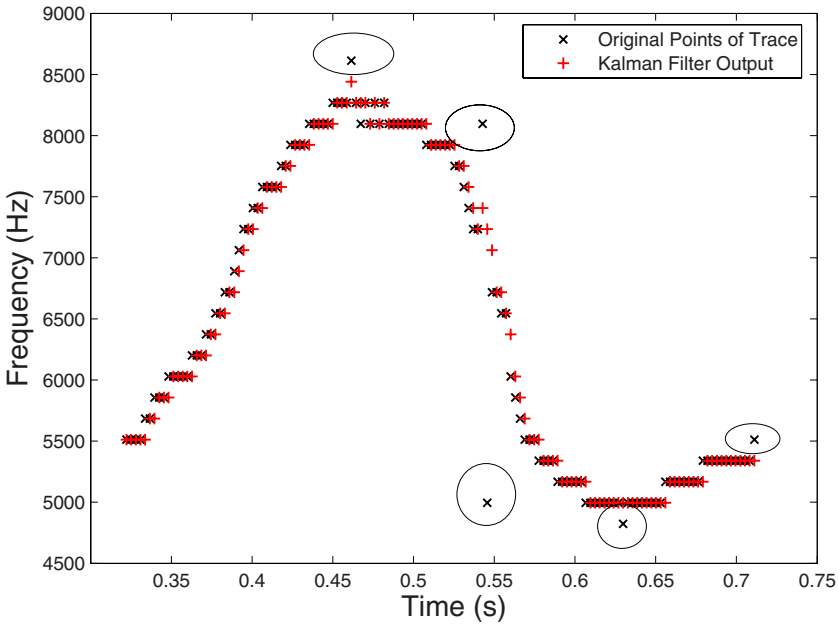

FIG. 5. (Color online) Plot of a traced whistle overlaid with the Kalman filter output. The whistle points adjusted by the filter are marked with ellipses.

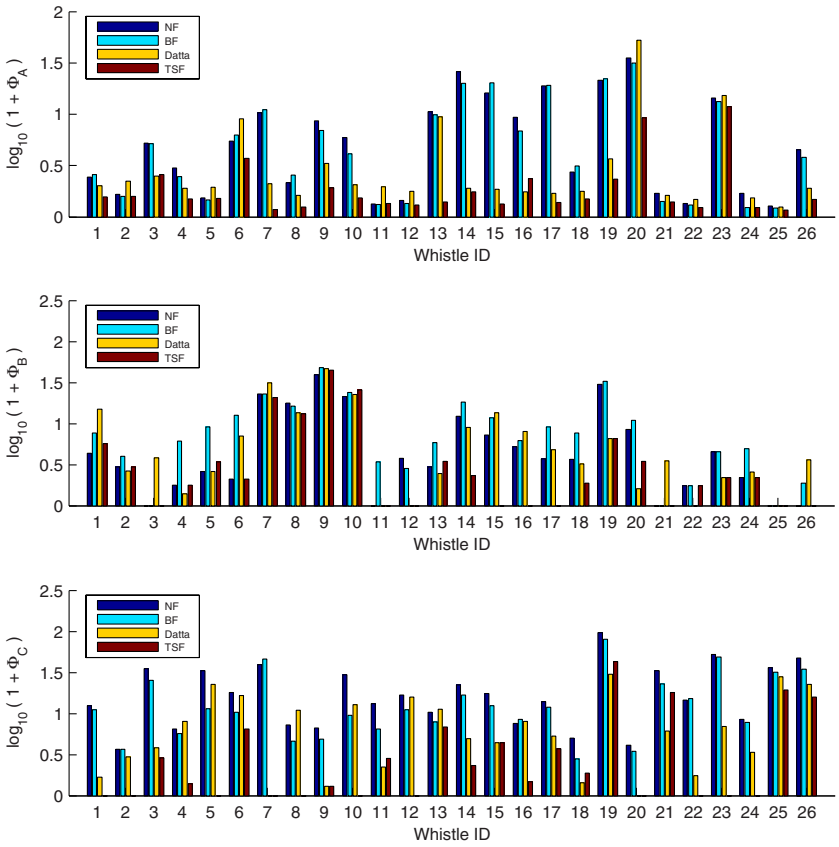

FIG. 6. (Color online) Performance metrics of tracing 26 recorded whistles.

manual tracing (Sec. VII A). Similarly, a whistle contour obtained using an automated method can be denoted by $\hat{\omega}(m)$, with associated times $\hat{t}(m)$.

Metric 1: Mean percentage error of tracing,

$$\Phi_A(\omega,\hat{\omega}) = \frac{1}{N}\sum_m \frac{\|\omega(m)-\hat{\omega}(m)\|}{\omega(m)} \times 100\,\%\,, \tag{14}$$

where N is the number of matching points in ω and $\hat{\omega}$. This is the relative percentage displacement of the contour $\omega(m)$ with respect to $\hat{\omega}(m)$, averaged over all the whistle points. The displacements are calculated only over the matching points of the trace, i.e., where $t(m)=\hat{t}(m)$. Here the absolute relative displacement is used instead of squared error since it provides a more direct measurement of the actual deviation from the reference trace.

Metric 2: Percentage of missing points,

$$\Phi_B(\omega,\hat{\omega}) = \frac{\beta}{|t(m)|} \times 100\,\%\,. \tag{15}$$

This is computed by using the number of missing time bins in $\hat{t}(m)$ compared to $t(m)$, denoted by β.

Metric 3: Percentage of extra points,

$$\Phi_C(\omega,\hat{\omega}) = \frac{\eta}{|t(m)|} \times 100\,\%\,. \tag{16}$$

This is computed by using the number of extra time bins in $\hat{t}(m)$ compared to $t(m)$, denoted by η.

The average of all three metrics provides a measure of the total tracing error in terms of the extent and accuracy of the traced contour. This metric is defined as follows.

Metric 4: Average percentage tracing error,

$$\Phi(\omega,\hat{\omega}) = \frac{\Phi_A+\Phi_B+\Phi_C}{3}. \tag{17}$$

C. Results and discussion

The set of 26 whistles was traced using four different approaches, three of which use the general procedure proposed in this work, with differing denoising methods. The fourth approach is an implementation of the method proposed by Sturtivant and Datta.[2,5] Specifically, the methods are:

(1) NF: No spectrogram denoising is performed in this method. The rest of the algorithm takes the original spectrogram images and performs segmentation and tracing.
(2) BF:Spectrogram denoising is performed using bilateral filtering.[8]
(3) TSF:Spectrogram denoising is performed using transient suppression filtering proposed in this work.
(4) SD: Whistle tracing is performed by implementing the method proposed by Sturtivant and Datta.[2,5]

Each algorithm must first be calibrated to determine the parameter values for optimal performance. This is done by a combination of visual verification of tracing results and computing the performance metrics. Note that the parameter settings used here are chosen to be generic to the recording environment and not optimized for each individual whistle. The first three metrics (Φ_A,Φ_B,Φ_C) are calculated on the

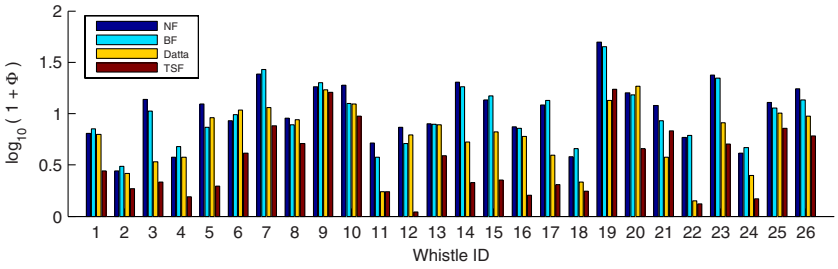

FIG. 7. (Color online) The average of the three metrics Φ_A, Φ_B, Φ_C for 26 whistle recordings.

tracing results of the whistle set and are graphically presented in Fig. 6. The average metric Φ is presented in Fig. 7.

NF is included as a "control sample" to quantify the effect of image denoising on tracing performance, particularly for BF and TSF. Compared to NF, BF tends to decrease the mean percentage error (Φ_A) and the number of extra trace points (Φ_C), indicating a reduction of outlying points in the traces [Figs. 6(a) and 6(c)]. However, the average percentage of missing points (Φ_B) rises in general, indicating that a significant number of signal pixels are also attenuated by the filter.

SD performs better compared to the BF method (on all metrics), and better than NF on Φ_A and Φ_C (but not Φ_B). From the tracing metrics and the visual inspection of the spectrogram plots, it can be inferred that the vertical edge suppression filter of SD is more effective in removing transients than the smoothing operation performed by BF.

TSF has an even greater impact on reducing the number of outlying points; at the same time, it manages to retain more of the whistle points of $\omega(m)$ (overall reduction of all metrics compared to NF, BF, and SD). This behavior is attributable to the vertical suppression and horizontal smoothing property of the directional filter kernels, which manage to effectively suppress high energy broadband transient spectral patterns while retaining frequency modulated narrowband whistles.

Statistical analysis of the results were performed on the results using STATISCA 6.0 (STATSOFT, Inc., 2001, Tulsa, OK). One-way multivariate analysis of variance (MANOVA) was performed simultaneously on the three scoring categories for the four methods. Log $(x+1)$ transformation successfully reduced homogeneity of variances for Φ_A (percentage error: Cochran's $C_{df3}=0.328$, $p>0.05$), Φ_B (percentage missed points: Cochran's $C_{df3}=0.264$, $p>0.05$), and Φ_C (percentage extra points: Cochran's $C_{df3}=0.348$, $p>0.05$). The results showed significant effects of the various methods on the scores (Wilk's $s_{9,239}=0.619$, $p<0.001$): univariate decomposition demonstrated that effects were significant for all three scoring categories [$\log(\Phi_A+1)$: $F_{3,100}=5.592$, $p<0.01$; $\log(\Phi_B+1)$: $F_{3,100}=5.418$, $p<0.05$; $\log(\Phi_C+1)$: $F_{3,100}=8.873$, $p<0.01$].

The SD method was slightly modified to improve its performance. First, a low-pass filter was introduced to remove high frequency tonals in addition to the high-pass filter stipulated in the original work. Second, when the inertial whistle-following algorithm stops upon encountering a sudden drop in SNR, it is automatically restarted to increase detection probability. The metrics for the SD method applied to filtered and unfiltered data were: square-root transformed "error" ($F_{1,50}=11.38$, $p<0.01$); and log (average+1) ($F_{1,50}=4.110$, $p<0.05$); and were significantly lower after imposing a low-pass filter than in unfiltered data.

Post-hoc Tukey's honest significant difference test demonstrated that TSF scores were consistently significantly lower (at $p=0.05$) than BF scores; the Φ_A and Φ_C (but not Φ_B) were significantly lower for TSF than NF; TSF also scored consistently lower than SD for all metrics but the difference was statistically significant only for Φ_B (see Fig. 8).

In summary, the statistical analysis of the results indicate that TSF consistently performed better than the other denoising methods. All metrics derived from the TSF method were consistently lower than the SD method, although this was only statistically significant for the Φ_B. SD metrics significantly outperformed the BF and NF methods only for Φ_C.

However, for some whistles, e.g., whistles 19 and 21 [Figs. 9(b) and 9(c), respectively], TSF does not perform well. In both cases a closely packed cluster of transients has caused an incomplete segmentation, resulting in a higher number of extra points on the trace. In these two situations, SD has not included the extra portion due to the SNR drop between the two segments and hence performs better.

Both SD and TSF methods perform poorly on whistle 9 [Fig. 9(a)], which contains rapid rising and falling segments. In such scenarios, methods employing vertical edge suppression filters, while useful for removing broadband patterns, also attenuate sections of the signal with similar characteristics. Additionally, due to its low SNR, all methods miss a large percentage of the whistle. Another reason for the poor performance for this whistle is due to the nature of the reference trace. Human vision is known to be remarkably adept at linking disconnected line segments to form a "complete"

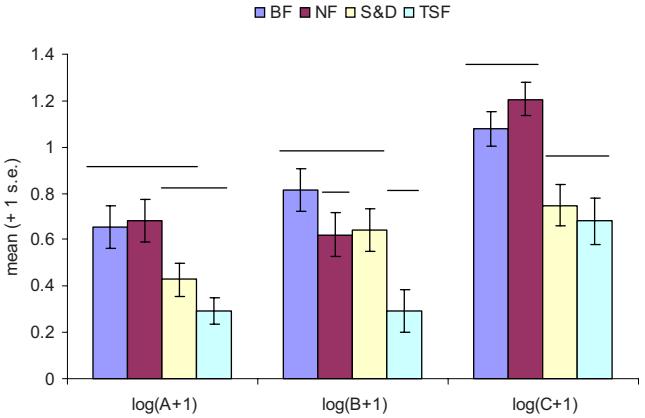

FIG. 8. (Color online) Statistical analysis of tracing results.

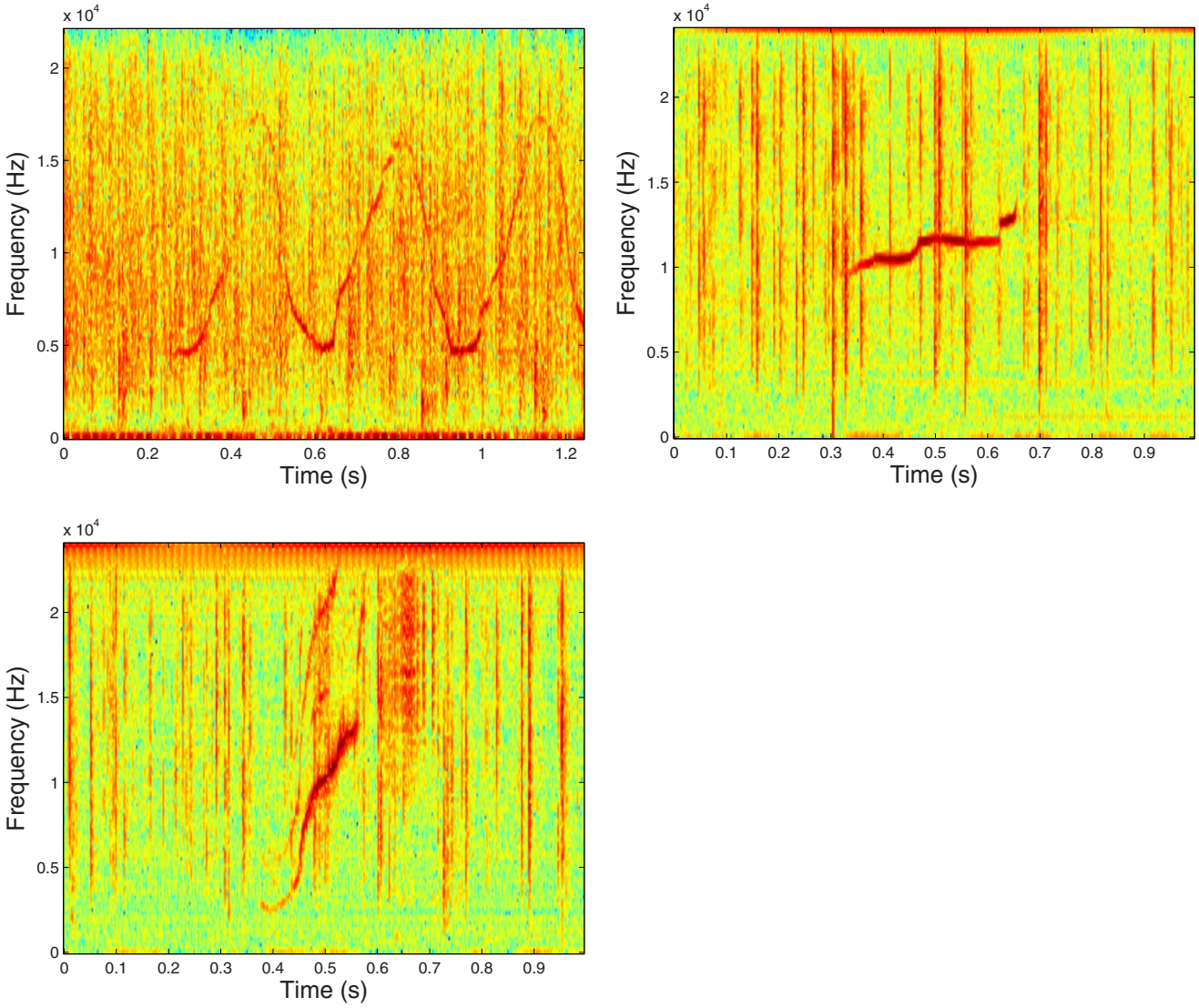

FIG. 9. (Color online) Spectrograms of problematic whistles for tracing.

picture. Therefore the manual trace contains pixels that are actually "breaks" in the whistle but are "artificially" linked together. However, the automated methods fail to identify these pixels as being part of the whistle and this leads to a higher number of missing points for such whistles.

Apart from measuring tracing performance, metrics Φ_B and Φ_C shed light on the values of operational parameters used in the tracing algorithm. For example, a high percentage of missed points Φ_B coupled with a low percentage of extra points Φ_C indicate that the segmentation (or detection) thresholds are probably set too high, and vice versa. Appropriate action can then be taken to determine the optimal point at which both metrics are sufficiently low. This method was followed in order to obtain the best performance for all the algorithms.

This study demonstrates that several different image processing-based approaches can be considered for extracting whistle contours from recordings. It also confirms that TSF performs better for recordings made in warm shallow waters.

VIII. CONCLUSION

This paper introduces an algorithm based on image processing techniques to denoise and extract contours of dolphin whistles from spectrograms. The algorithm presented in Sec. III is well suited for denoising recordings made in warm shallow waters, where the ambient noise is dominated by snapping shrimp. It exceeds the performance of existing algorithms[5,6] by incorporating the pixel values of adjacent time bins for greater preservation of the signal while effectively attenuating transient spectral patterns. The use of static directional smoothing kernels yields more efficient implementations and reduces the processing time of denoising. The objective method introduced for testing the tracing performance using known time-frequency contours not only enables the selection of a particular algorithm, but can also be used to tune its parameters.

The modularity of the algorithm enables easy integration of other techniques at any stage of the process, and further studies may be carried out to examine combinations that produce the best results. Although an effort has been made to

make the algorithm self-adaptive as far as possible, some parameters still have to be manually set depending on any particular acoustic recording. However, suggestions are made to guide the choice of the values of those parameters. Features extracted from whistle contours are available for classification tasks, paving the way for faster analysis of dolphin vocalizations. Although this implementation has not yet been optimized for speed, and currently works on off-line data, an on-line system coupled with a whistle recognition module would be an invaluable tool for field biologists. We believe the techniques presented here can be applied to extract other animal vocalizations such as whale calls and bird songs from acoustic recordings. Further studies are recommended to determine the feasibility of such extensions.

ACKNOWLEDGMENTS

The authors wish to express their gratitude to Gao Rui and Suranga Chandima Nanayakkara of the Department of Electrical and Computer Engineering, NUS, Dr. Sin Tsai Min of TMSI, and Yeo Kian Peen of MMRL, for their assistance and encouragement at various stages, including recording of dolphin whistles, manual whistle tracing, and statistical analysis.

[1]V. M. Janik, "Pitfalls in the categorization of behaviour: A comparison of dolphin whistle classification methods," Anim. Behav. **57**, 133–143 (1999).
[2]S. Datta and C. Sturtivant, "Dolphin whistle classification for determining group identities," Signal Process. **82**, 251–258 (2002).
[3]J. R. Buck, H. B. Morgenbesser, and P. L. Tyack, "Synthesis and modification of the whistles of the bottlenose dolphin, tursiops truncatus," J. Acoust. Soc. Am. **108**, 407–416 (2000).
[4]J. R. Buck and P. L. Tyack, "A quantitative measure of similarity for tursiops truncatus signature whistles," J. Acoust. Soc. Am. **94**, 2497–2506 (1993).
[5]C. Sturtivant and S. Datta, "Techniques to isolate dolphin whistles and other tonal sounds from background noise," Acoust. Lett. **18**, 189–193 (1995).
[6]C. Sturtivant and S. Datta, "The isolation from background noise and characterisation of bottlenose dolphin (tursiops truncatus) whistles," J. Acoust. Soc. India **23**, 199–205 (1995).
[7]M. Chitre, J. Potter, and S. H. Ong, "Optimal and near-optimal signal detection in snapping shrimp dominated ambient noise," IEEE J. Ocean. Eng. **31**, 497–503 (2006).
[8]C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proceedings of the IEEE International Conference on Computer Vision, Bombay, India*, Jan. 4–7, 1998.
[9]W. Niblack, *An Introduction to Digital Image Processing* (Prentice Hall, Englewood Cliffs, NJ, 1986).
[10]F. Yan, H. Zhang, and C. R. Kube, "A multistage adaptive thresholding method," Pattern Recogn. Lett. **26**, 1183–1191 (2005).
[11]R. E. Kalman, "A new approach to linear filtering and prediction problems," ASME J. Basic Eng. **82**, 35–45 (1960).