A Method for Tracing Dolphin Whistles

Asitha Mallawaarachchi*, S.H Ong*, Mandar Chitre † and Elizabeth Taylor ‡

*Department of Electrical and Computer Engineering,
National University of Singapore,
E4-05-48, 4 Engineering Drive 3, Singapore 117576
{elema, eleongsh}@nus.edu.sg

†Acoustic Research Laboratory, Tropical Marine Science Institute,
National University of Singapore,12a Kent Ridge Road, Singapore 119223
mandar@arl.nus.edu.sg

‡Marine Mammal Research Laboratory, Tropical Marine Science Institute,
National University of Singapore, 12a Kent Ridge Road, Singapore 119223
mdcohe@leonis.nus.edu.sg

***Abstract*— Underwater acoustic recordings containing dolphin vocalizations are often analyzed in time-frequency domain using spectrograms. Spectrogram feature extraction techniques are widely adopted in whistle classification studies because they provide a visual representation of the whistle's frequency variation over time. However due to the low SNR of recordings, harmonics and frequency spread, most researchers use time consuming manual methods to trace whistles. The work presented in this paper attempts to automate this process.**

I. INTRODUCTION

Dolphins have impressive vocalization capabilities that can be categorized into three classes: (i) broadband short-duration clicks (ii) broadband pulsed sounds and (iii) narrowband frequency modulated whistles [1]. The third kind of signals (i.e whistles) are used in animal communication, and are chosen for this study. Whistles are non-stationary waves, and are of particular interest to researchers studying animal behavior and communication. Underwater acoustic recordings containing dolphin vocalizations are often transformed into time-frequency space by the short-time Fourier transform (STFT) for analysis as it allows visual inspection of whistles' frequency variation over time.

The STFT of a function $f(t)$ with respect to the window function $\phi(t)$ evaluated at the location (b, ξ) in time-frequency plane is defined as [2]:

$$G_\phi\ f(b,\xi) \equiv \int_{-\infty}^{\infty} f(t)\overline{\phi_{b,\xi}(t)}dt \tag{1}$$

where

$$\phi_{b,\xi}(t) \equiv \phi(t-b)e^{j\xi t} \tag{2}$$

the window function $\phi(t)$ used in our experiments is a periodic Hanning window.

The output of this operation is often interpreted as a gray-level intensity 2-D image called the spectrogram. Spectrogram feature extraction techniques are widely adopted in the studies of whistle classification because they provide a visual representation of the whistle's frequency variation over time. However due to the low SNR of recordings, harmonics and frequency spread, most researchers use time consuming manual methods to trace whistles ([1], [3]). The work presented in this paper attempts to automate this process.

We adopt a combination of signal and image processing techniques to remove acoustic noise, enhance the spectrogram images and spectrogram segmentation to perform the tracing of the fundamental frequency variation of a whistle.

II. METHOD

A. Acoustic Noise Filtering

Snapping shrimp are common in warm shallow waters and produce highly transient broadband noise. The statistics of this particular type of noise can be accurately modelled by the symmetric alpha-stable (SaS) family of distributions [4], which enables us to derive a statistical de-noising filter.

Let $x(t)$ be the time-series of the signal and $n(t)$ be the time series of the additive noise, then $y(t) = x(t) + n(t)$ is the received signal. The aim of the statistical filter is to produce an estimate of the signal $x(t)$ denoted by $\hat{x}(t)$ such that $E[\hat{x}(t)] = E[x(t)]$, where E is the statistical expectation.

At a given time t_1 we have have a measurement $y = y(t_1)$ and want to derive $\hat{x} = \hat{x}(t_1)$. We can use the maximum likelihood (ML) estimate for this derivation, i.e choose the x that maximizes $P(x|y)$ as $\hat{x}$. We use the Bayes Rule for this computation

$$P(x \mid y) = \frac{P(y \mid x)\, P(x)}{P(y)}. \tag{3}$$

Let $f_x(x)$ be the PDF of the signal and $f_n(n)$ be the PDF of the noise. Then $f_y(y) = f_x(x) \otimes f_n(n)$ is the PDF of the received signal, where $\otimes$ is the convolution operator. We can estimate the unknown probabilities as $P(x) \simeq f_x(x)$ and $P(y|x) \simeq f_n(n)$, as x is deterministic in expression $P(y|x)$

1-4244-0138-0/06/$20.00 ©2006 IEEE

which makes it only dependent on the stochastic behavior of n. Since $n = y - x$, (3) can be rewritten as

$$\begin{aligned} P(x \mid y) &= \frac{f_n(y-x)\, f_x(x)}{f_y(y)} \\ &= \frac{f_n(y-x) f_x(x)}{[f_x(x) \otimes f_n(y-x)]} \end{aligned} \tag{4}$$

The ML estimate of $x(t)$ is $\arg\max\{f(x|y)\}$ over all x. $P(y|x)$ is evaluated by substituting a pre-computed SaS distribution for f_n and assuming Gaussian distribution for f_x.

B. Spectrogram De-noising

1) Adjusting for Non-uniform Energy Distribution: The spectrogram is a special image that has its pixel intensities correlated to the instantaneous energy of the generating time series. The average energy contained in a given time window of the time series varies in a stochastic manner, an creates an effect similar to that of non uniform illumination in conventional images. Therefore we employ a simple adjustment operator to correct for non-uniform energy distribution as a pre-processing step. The local background energy is estimated for each 32×32 block by taking the minimum, and the whole block is normalized to this value to complete the adjustment.

2) Noise Model Normalization: The ambient noise characteristics of the recorded time series are dynamic. However for a small time interval, they can be assumed constant. Therefore a local time-average of the noise distribution over the discrete frequency bins can be computed, and used for local normalization of the spectrogram image.

In this implementation the small time interval is taken as 100 columns in the spectrogram. The columns within this time interval are normalized to a 'noise column', which is computed by randomly selecting 10 columns from the current time interval and computing a time-average.

3) Bilateral filtering: As a second step in spectrogram denoising, various image noise filtering algorithms (including mean, weighted-mean, median, adaptive Wiener and bilateral) were tested. Among the filters tested, bilateral filtering [5] produced the best results and hence adopted in the design.

The bilateral filter is a spatially adaptive filter and its kernel coefficients for a local neighborhood are computed based on both the geometric closeness and the gray level similarity between the neighborhood center and the other neighborhood pixels. Therefore the bilateral filter is better at preserving edges while smoothing effectively.

C. Iterative Harmonic Suppression

Whistles are not pure tones and therefore contain harmonics that are similar in shape to the fundamental frequency variation with only a shift in frequency and can potentially hinder its accurate tracing. The instantaneous frequency of a harmonic is an integer multiple of the fundamental frequency, and this can be exploited to remove them from the spectrogram image.

Each pixel row in a spectrogram represents a single frequency variation over time (technically it is the time variation of a discrete frequency bin), and from bottom to top, the rows represent linear frequency increase. Let us define a pixel intensity vector I_i that contains the pixels in the i^{th} row of the spectrogram, and let f_i be the frequency of the i^{th} row. Then the harmonic suppression update equation for the i^{th} row can be written as

$$I_i = I_i - k\, I_j, \tag{5}$$

where k is a user defined scalar constant and the vector I_j contains the pixel values of j^{th} row for which $f_j = f_i/N$. The integer N represents the set of the largest common divisors of the integer multiples of the fundamental frequency that has produced the harmonic pattern. In practise $N \in \{2, 3, 5\}$ is used, and (5) is applied to every row from top to bottom, and iterated for each value of N.

D. Spectrogram Segmentation

There is a vast amount of literature on Image segmentation; however the choice of an algorithm should depend on the characteristics of the input images. Since we are trying to segment spectrogram intensity images, we have the advantage of certain *a priori* knowledge but on the other hand the images often have low SNR. Therefore we propose a 2-stage algorithm which uses global thresholding followed by region growing.

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

$$J(x,y) = \begin{cases} 1 & \text{if } I(x,y) \geq T \\ \\ 0 & \text{else} \end{cases} \tag{6}$$

In view of the low SNR, we want to compute a global threshold T which slightly under-segments the image. The use of global thresholding instead of local thresholding and the computation of a higher than optimal threshold ensures that noisy pixels are less likely to be segmented as foreground.

Spectrograms vary in their gray level distributions, and noise statistics depending on the recording environment and the type of dolphin vocalizations. Therefore a robust algorithm must be able to compute threshold value T adaptively. Therefore we adopt a simplified variant of the method proposed in [6]. A global threshold is computed adaptively based on the fist and second order statistics of the image gray levels using the Niblack Method.

$$T = \mu + k * \sigma \tag{7}$$

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

For the purpose of calculating k we assume a normal distribution of gray values. Even though the actual distribution can differ this assumption is valid for computing k [6]:

$$\int_{-\infty}^{\mu+k*\sigma} N_{\mu,\sigma}(x)dx = \rho \tag{8}$$

where ρ is the percentage of background pixels. Using *a priori* knowledge that over 90% of a spectrogram contains background we compute a high threshold by setting $\rho = 0.96$. The calculation of k using (8) is done by the use of a pre-calculated Z-table per image.

TABLE I: Morphological Filtering and Enhancement

Step	Structuring Element	Morphological Operation
1	$s_1 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$	$J \bullet s_1$ (closing)
2	$s_2 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$	$J \circ s_2$ (opening)

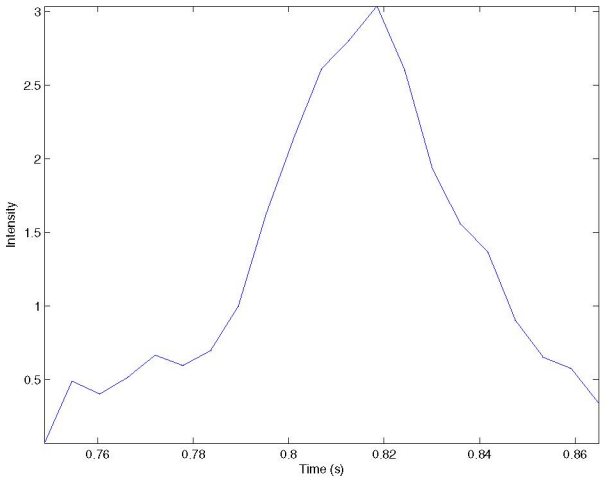


Fig. 1: The change of intensity of whistle pixels over time

As an additional precaution we use mathematical morphology to improve the segmentation and remove noise from J. This is done by fist applying morphological closing with a 2×2 square structuring element (SE) s_1 followed by opening with a 2×3 square SE s_2. The closing operator helps to fill gaps and holes smaller than the SE s_1 while opening removes noisy outliers smaller than the SE s_2. This intermediate operation is illustrated in Table I.

The whistles we are aiming to segment start off with low intensity and gradually increase with time before similarly decreasing and 'trailing-off'. This trend is illustrated in Fig. 1 and implies that with a high threshold the starting and end points of the whistle is likely to be left out as background. Therefore after removing outliers with the morphological operators, the pixels of the segmented image are input into a 2D region growing algorithm as seeds. which uses $0.9 \times T$ as the new threshold to include pixels connected to the segmented image. This helps to include the 'fading' ends of the whistle.

E. Whistle Tracing

After segmentation, a one pixel thick trace is to be drawn on the segmented image to represent the shape of the whistle. This is a two-step process. As a first step we take the Euclidian distance transform of the segmented image. This creates an intensity ridge through the midpoints of the segmented whistle. And in the second step the candidate points for the trace are chosen by taking the maximum of each column of the distance transform (i.e., choosing the intensity ridge line).

TABLE II: Kalman Filter Model

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

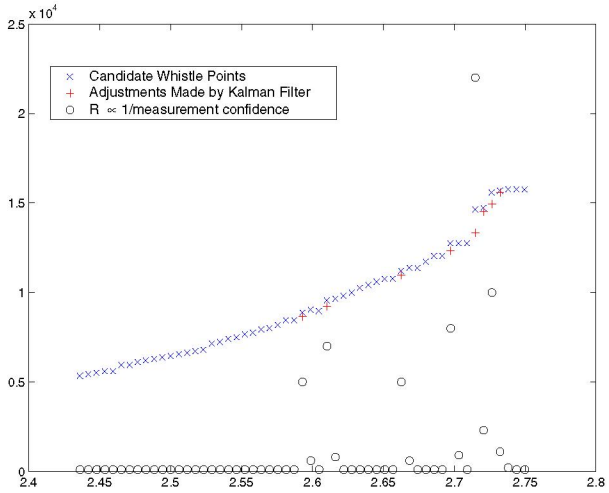
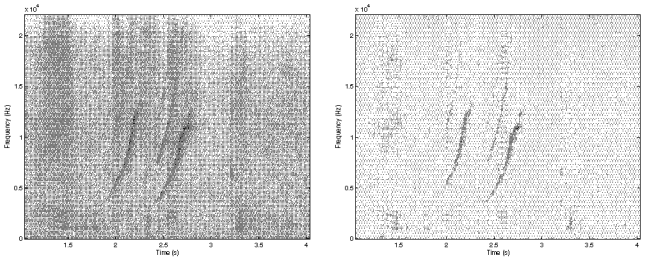


Fig. 2: whistle smoothing using the Kalman filter

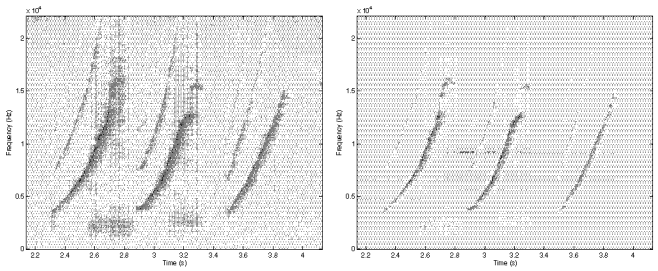
Since the points on the whistle should trace a smooth curve, the candidate points are then passed onto a 2^{nd} order Kalman filter for smoothing. The confidence values for each measurement are calculated and saved in the previous tracing step. Low confidence values are assigned to whistles points that exhibit sudden jumps in frequency. The assignment function has memory of 1, in the sense that if the previous measurement had low confidence the current measurement's confidence will be partially based on that previous value. Fig. 2 shows a selected set of candidate points, the Kalman filter corrections and the measure of confidence R for each measurement. High R values indicate low confidence in the measurement.

The second order Kalman filter is designed using the same process model that is used for trajectory tracking 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 the acceleration correspond the first and the second derivatives of frequency with respect to time. The



(a) (b)

Fig. 3: Effect of adjusting for non-uniform energy distribution. (a) shows the original image and (b) shows the adjusted image.



(a) (b)

Fig. 4: The effect of noise model normalization. (a) shows the original image and (b) shows the normalized image.

Kalman model is illustrated in Table II.

As one image may contain multiple whistles, the Kalman filter variables need to be reset at the start of a new whistle. This is done by taking note of their time discontinuities. The same technique is used in drawing the chosen whistle points as piece-wise continuous curves.

F. Spectrogram Reconstruction

The enhanced spectrogram can be converted back into audio by implementing a inverse spectrogram function, which attenuates the complex FFT coefficients that correspond to the background pixels. The points on the trace are amplified and provides audible improvement over the original recording.

III. RESULTS AND DISCUSSION

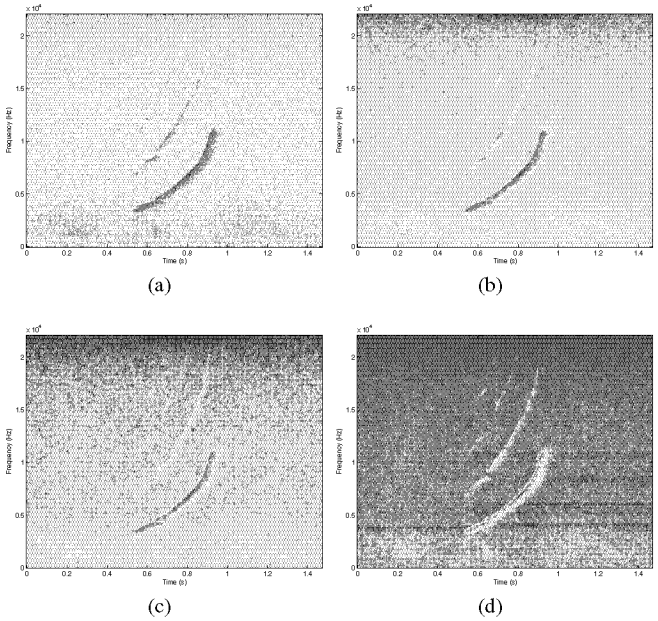
In spectrogram denoising, the time domain statistical filtering is useful in reducing non-Gaussian noise characteristics introduced by snapping shrimp noise, and thus improves the subsequent image processing tasks. However finding the correct statistical parameters are a computationally intensive process.

Adjusting for non-uniform energy distribution is a useful pre-processing step. Fig. 3 illustrates the effectiveness of this method.

Local noise model normalization is useful particularly in spectrograms with low SNR or transient noise patterns (Fig. 4), while for most images bilateral filtering alone gives good results (Fig. 5). Iterative harmonic suppression is very effective, and helps with robust tracing of the fundamental frequency variation as show in (Fig. 6).

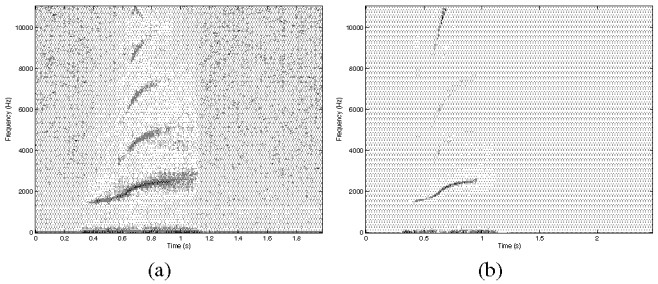
Global thresholding is the more robust than the local thresholding as it is less sensitive to local maxima. Local thresholding usually causes some background pixels to be misclassified as foreground points, and this complicates the tracing process. Local Thresholding results are also heavily dependant on the choice of the local window size. These effects are illustrated in Fig. 7.

The whistle tracing depends on the segmentation quality, and becomes more complicated when multiple whistles are present. To make the process more robust, a discontinuity-detection mechanism and Kalman filtering is used. This enables us to trace whistles as smooth piecewise continuous



(a) (b) (c) (d)

Fig. 5: Effect of bilateral filtering. (a) shows the original image and (b) shows the bilateral filtered image. (c) shows illumination adjustment followed by bilateral filtering and (d) noise model normalization followed by bilateral filtering



(a) (b)

Fig. 6: Iterative harmonic suppression. (a) shows the original image with harmonics and (b) shows the resulting image with harmonics erased.

curves. However some whistles that have complicated shapes such as the whistles shown in Fig. 8 cause problems for

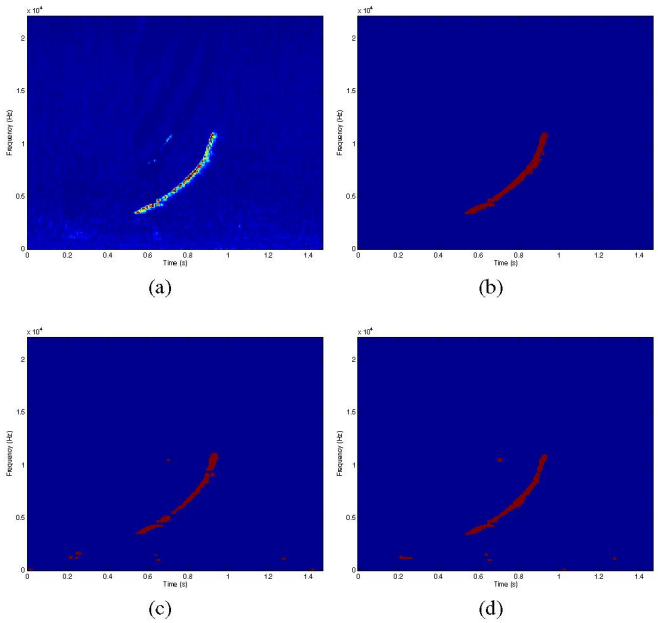


Fig. 7: Segmentation. (a) shows the enhanced image and (b) shows the result of global thresholding. (c) and (d) show the results of local thresholding using (c) window size of 35 and (d) window size of 53

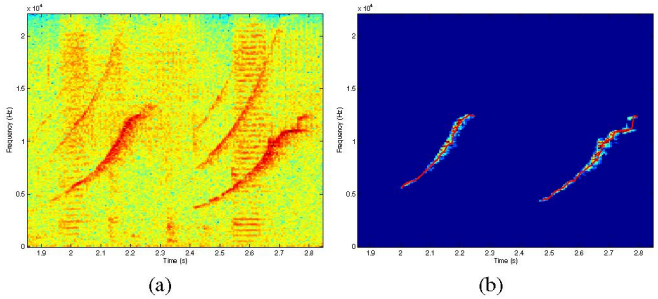


Fig. 8: The 'fork' shape at the end of the second whistle in (a) proves to be difficult to trace. (b) shows the tracing result accurate tracing as they have more than one frequency value corresponding to one discrete time bin.

IV. CONCLUSION AND FUTURE WORK

We have presented a multi stage automated method for tracing dolphin whistles in spectrograms. In the next phase of our work, this software will be used to automatically extract pertinent features from the spectrogram, which will be used as inputs to a whistle classification system. It is hoped that this work will aid to speed-up the work of researchers in the area of dolphin communication and similar acoustic processing tasks.

REFERENCES

[1] M. Greco, F. Gini, and L. Verrazzani, "Analysis and modeling of acoustic signals emitted by mediterranean bottlenose dolphins," in *Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology*, Dip. di Ingegneria dell'Informazione, Pisa Univ., Italy, December 2003, pp. 122 – 125.

[2] J. C. Goswami and A. K. Chan, *Fundamentals of Wavelets: Theory, Algorithms, and Applications*. Wiley, 1999.

[3] V. M. Janik, "Pitfalls in the categorization of behaviour: a comparison of dolphin whistle classification methods," *Animal Behaviour*, 1999.

[4] M. Chitre, J. Potter, and S. H. Ong, "Optimal and near-optimal signal detection in snapping shrimp dominated ambient noise."

[5] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proceedings of the IEEE International Conference on Computer Vision, Bombay, India*, 1998.

[6] F. Yan, H. Zhang, and C. R. Kube, "A multistage adaptive thresholding method," *Pattern Recognition Letters*, vol. 26, pp. 1183–1191, June 2005.