An adaptive filter-based method for robust, automatic detection and frequency estimation of whistles

A.Torbjörn Johansson[a),b)] and Paul R. White
Institute of Sound and Vibration Research, University of Southampton, Southampton SO17 1BJ, United Kingdom

(Received 29 April 2010; revised 5 May 2011; accepted 16 June 2011)

This paper proposes an adaptive filter-based method for detection and frequency estimation of whistle calls, such as the calls of birds and marine mammals, which are typically analyzed in the time-frequency domain using a spectrogram. The approach taken here is based on adaptive notch filtering, which is an established technique for frequency tracking. For application to automatic whistle processing, methods for detection and improved frequency tracking through frequency crossings as well as interfering transients are developed and coupled to the frequency tracker. Background noise estimation and compensation is accomplished using order statistics and pre-whitening. Using simulated signals as well as recorded calls of marine mammals and a human whistled speech utterance, it is shown that the proposed method can detect more simultaneous whistles than two competing spectrogram-based methods while not reporting any false alarms on the example datasets. In one example, it extracts complete 1.4 and 1.8 s bottlenose dolphin whistles successfully through frequency crossings. The method performs detection and estimates frequency tracks even at high sweep rates. The algorithm is also shown to be effective on human whistled utterances. © *2011 Acoustical Society of America*. [DOI: 10.1121/1.3609117]

PACS number(s): 43.66.Gf, 43.60.Bf, 43.30.Sf, 43.60.Mn [WMC] Pages: 893–903

I. INTRODUCTION

Passive acoustic monitoring of wildlife is a growing research field, which requires detection and identification of the calls of the species of interest. Manual call extraction is tedious and time-consuming, so an automatic method is desirable. We propose an adaptive filter-based method for analyzing tonal calls known as whistles, e.g., the vocalizations of some species of birds and marine mammals. This study deals with detection and frequency estimation, and the proposed method can serve as a front end to a pattern recognition stage aimed at, e.g., species identification.[1] Whistles are suitable for communication in difficult environments and are also used in human whistled languages such as that of La Gomera, Spain.[2] We focus on the analysis of marine mammal whistles, which are often observed in strong background noise and other difficult environmental conditions and are highly variable. Multiple whistles can occur simultaneously and can be mixed with interfering transients such as echolocation clicks. Whistles can also have higher harmonics in addition to the fundamental. The method presented here estimates the strongest tones at each time, which are likely to be fundamentals, but as will be shown the method can also track higher harmonics.

Most previous efforts at automatic analysis of whistled sounds have been based on a spectrogram time-frequency representation of the recording.[3–7] Whistle analysis typically commences by background noise compensation. The background noise spectrum can be estimated by averaging the amplitude of each spectrogram bin over a few seconds. For improved estimation of background noise in the presence of signal events such as whistles and clicks, we replace the average by the median or another so called order statistics estimator.[8] Denoising is commonly performed by subtracting the estimated noise spectrum from the spectrogram.[3,5] The proposed method instead divides by the noise spectrum to obtain a pre-whitened spectrogram in which the background noise is approximately white and of unit power. The proposed whistle detector relies on this property. A different approach to denoising is taken by Mallawaarachchi *et al.*[7] They combine the outputs of four different two-dimensional filters applied to a spectrogram representation to reduce the noise level and show that this method is adept at attenuating impulsive sounds, such as those due to snapping shrimp.

Spectrogram-based whistle detection then proceeds by finding all sufficiently strong peaks in the noise compensated spectrogram. Whistles, which appear as connected ridges of peaks, are then detected by connecting peaks across time and frequency. Detected whistle ridges can also be joined across short gaps if their characteristics match.[5–7] Datta and Sturtivant[3] run whistle detection from spectrogram peaks twice, first using a high threshold to decrease the rate of false detections and then using a lower threshold in an attempt to extract each detected whistle in its entirety. For the same

[a)]Also at Swedish Defense Research Agency, SE-164 90 Stockholm, Sweden
[b)]Author to whom correspondence should be addressed. Electronic mail: torbjorn.johansson@foi.se

reasons, Mallawaarachchi *et al.*[7] use morphological operations followed by region growing on detected whistle peaks. Their method can only extract a single whistle at each time and so is not appropriate for multiple whistles.

The rate of false whistle alarms from spurious peaks can be decreased by using prior knowledge of whistle characteristics during the contour extraction. Datta and Sturtivant use an inertial ridge following technique that, in selecting the next point of a peak ridge, is more likely to select peaks of a similar amplitude to the current peak and of a frequency that agrees with a sweep rate based estimate.[3] Whistles that fall below a given duration threshold are discarded. Mallawaarachchi *et al.*[7] employ a Kalman filter for whistle extraction. The Kalman model selects the next point of a whistle ridge by combining observed peak characteristics with a prediction based on averages of previous characteristics. The prediction is based on the first and second derivatives of the frequency evolution as a function of time. This can be seen as a different way of applying inertia to the whistle extraction process.

If clicks and other short-duration impulsive sounds are present in the recording, they can disturb the whistle detector. Datta and Sturtivant[3] attenuate clicks, which appear as vertical features in the spectrogram, by applying a "masked equalization" technique, which entails dividing the amplitude of each spectrogram bin by the average amplitude of its neighboring frequency bins. They then normalize the amplitude spectrogram by subtracting the mean of each partition and then dividing by the partition standard deviation. Gillespie *et al.*[5] and Halkias and Ellis[6] discard click and spurious noise peaks by requiring that at each time, a peak should not have too few (noise) or too many (click) neighbors that are peaks.

For frequency tracking, many non-spectrogram based techniques for instantaneous frequency estimation of non-stationary tonals are also applicable to whistles.[9] A recent example is the work of Ioana *et al.*,[10] who split the data into overlapping sections and use the polynomial phase transform[11] to estimate smooth frequency evolutions on each section. These are then merged into a single frequency track for each whistle. This study shows that whistles can be detected and tracked using an adaptive notch filter (ANF)[12–15] applied to the recorded waveform. However, this requires the development of methods for whistle detection, tracking through strong transients, and handling of multiple whistles the frequencies of which cross. Such methods are described here.

After pre-whitening, the proposed method applies an adaptive notch filter to estimate the dominant frequencies at each time. This is done on the whole recording. The method then uses internal filter variables to achieve detection, i.e., to determine when the filter was actually tracking a whistle. This provides independent detection of several simultaneous whistles. Detections that are too short to correspond to whistles are discarded, and short gaps between detections are bridged.

To cope with the difficult task of tracking simultaneous whistles through frequency crossings, previous authors have employed different techniques that all are based on connecting those segments the characteristics of which before and after the crossing provide the best match.[3,5,6] We proceed along similar lines, developing a method for sweep rate estimation from the estimated frequencies, and biasing the filter toward the current sweep rate close to frequency crossings.

Finally, we develop a heuristic waveform-based click detector that finds short bursts that are much stronger than their surroundings and estimate whistle tracks during clicks from the filter's state just before the click. This allows the proposed method to track whistles through clicks.

The method is evaluated by application to several recorded marine mammal whistles and a human whistled speech utterance. The results are compared to those from Datta and Sturtivant's[3] and Mallawaarachchi *et al.*'s[7] spectrogram-based methods. The comparison shows that the proposed method produces longer valid detections that are not disturbed at clicks while displaying fewer false alarms than the competing methods. The spectrogram-based methods detect some low SNR whistles that the proposed method does not detect but also miss some higher SNR whistles that are detected by the proposed method. The proposed method is applicable to recordings of many simultaneous whistles with strong clicks and consistently tracks whistles through frequency crossings. Because it employs a significantly shorter temporal analysis window, the proposed method also outperforms the competing methods at tracking rapidly sweeping whistles.

II. NOISE COMPENSATION

Ideally one would seek to measure the background noise in periods when whistles were absent from the recording. This can be achieved by performing spectral estimation on data sections when no whistle is detected. This assumes stationarity of the noise, an assumption that may not be valid over the period for which whistles are present, especially in the presence of a large group of vocal animals. To circumvent these issues, one can obtain robust estimates of the background spectrum even when whistles are present based on the use of order statistics. In particular, using the median of the spectrogram provides such a robust estimator.[8]

Order statistic estimators are based on sorting the data in order of increasing magnitude. The median estimate is the 50th percentile order statistic, i.e., the value which is larger than 50% of the data. For even better robustness to interfering calls, we use the 30th percentile value. Order statistic estimates of the average noise power are biased, but the bias can be estimated by making an assumption on the statistical distribution of the background noise. We assume a Gaussian distribution, so that the noise power spectrum follows a χ^2-distribution. Bias correction for the N^{th} order statistic can then be accomplished by multiplying each power spectral estimate by the data-independent factor ξ,[8]

$$\xi = -\frac{1}{\log(1 - N/100)}. \tag{1}$$

As the variance of the order statistic estimates increases with decreasing order,[8] the 30th percentile value was chosen as an empirical trade-off between accuracy and precision.

The presence of signals also introduces a bias to the proposed noise estimator if a significant part of the power spectral values in any given frequency bin are inflated. This bias is difficult to estimate without knowledge of signal characteristics and is typically significantly smaller than the noise-only bias of Eq. (1).

Figure 1 is a spectrogram of a short example recording with several strong whistles, calculated using Hann windowed 256-point partitions with 50% overlap. The sampling frequency is 44.1 kHz. Figure 2 shows background noise spectra of this recording, estimated using the mean, median, and 30th percentile methods. The bias compensated order statistic estimators provide results that are in agreement with those of the mean below 2 kHz, where whistles are absent. Above 5 kHz, where whistles dominate, the mean estimate is significantly affected by the whistles. However, the order statistic spectra exhibit no apparent distortion because of the whistles. There are no appreciable differences between the two bias compensated order statistic spectra.

Figure 3(a) shows the result of de-noising the recording of Fig. 1 by subtracting the noise spectrum. The method improves the local signal-to-noise ratio (SNR) of most whistles but fails to attenuate the low-frequency noise. Figure 3(b) shows the result of pre-whitening by dividing the spectrogram by the noise spectrum. The method of pre-whitening does not change the local SNR, but the low frequency noise is attenuated and the noise is now of equal power at all frequencies.

For comparison, the result of Datta and Sturtivant's masked equalization followed by denoising and normalization is shown in Fig. 3(c). Here a mask width of 30 frequency bins is used. The noise power spectrum is then estimated using the order statistic method described here. Near most whistle peaks, the noise is strongly attenuated. Clicks are also attenuated and only some remnants of them can be seen.

Mallawaarachchi *et al.*'s[7] transient suppression denoising consists of converting the spectrogram to a logarithmic scale and then filtering it with four different Gaussian two-dimensional filters of size 9×9. The filters respond to features that are horizontal, vertical, and oriented along the two diagonals, respectively. Denoising consists of applying a correction term, which is given by the difference between the largest output from the horizontal and diagonal filters and that of the vertical filter, to the log-scale spectrogram.

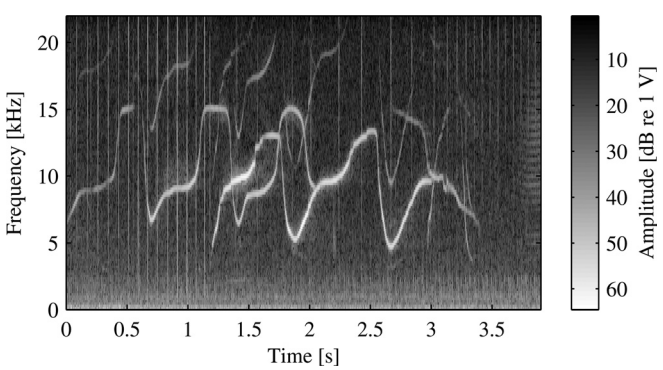

FIG. 1. Spectrogram of a bottlenose dolphin recording.

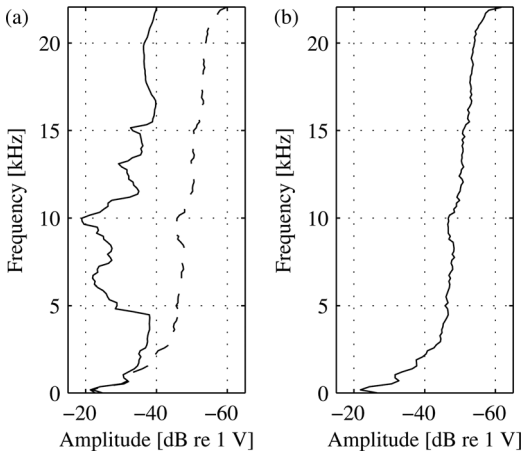

FIG. 2. Background noise spectrum estimates of the recording of Fig. 1. (a) Mean (solid) and median (dashed). (b) 30th order statistic.

The balance between the correction term and the original spectrogram is controlled by the parameters α and β. Using simulated data and varying the values of α and β, we found that using $\alpha = 0.3$ and $\beta = 0.7$ gave the best subjective performance. The results of applying this algorithm to the example recording are shown by Fig. 3(d). Clicks are strongly attenuated and the noise appears as a nearly constant background level.

The core of the proposed methodology is based in the time domain. The preprocessing (pre-whitening) is presently formulated in the frequency domain. At present, our implementation retains that structure, but we are actively seeking implementations that are solely based in the time domain.

III. FREQUENCY TRACKING

An adaptive notch filter method tracks whistle frequencies by applying a notch filter to the recording and adaptively minimizing its output. In the z domain, the input-output relationship for the notch filter is

$$E(z^{-1}) = H(z^{-1}, t)Y(z^{-1}), \tag{2}$$

where $E(z^{-1})$ is the z-transform of the filter output $e(t)$, $H(z^{-1}, t)$ is the filter's time-varying transfer function, with t representing discrete time, and $Y(z^{-1})$ is the z-transform of the pre-whitened recorded signal $y(t)$. The transfer function of the notch filter has one or several deep notches, each blocking a narrow band of frequencies, see Fig. 4.

The power of the output of the filter is minimized when the notches are placed at the frequencies of the strongest narrowband components in the signal. In a denoised whistle recording, these components correspond to whistles unless strong short duration transients are present. Section VI discusses how to deal with such disturbances. The center frequencies of the notches at each time should then correspond to the frequencies of the whistles. The form of time-varying notch filter employed here is[12,16]

$$H(z^{-1}, t) = \frac{A(z^{-1}, t)}{A(\rho z^{-1}, t)}, \tag{3}$$

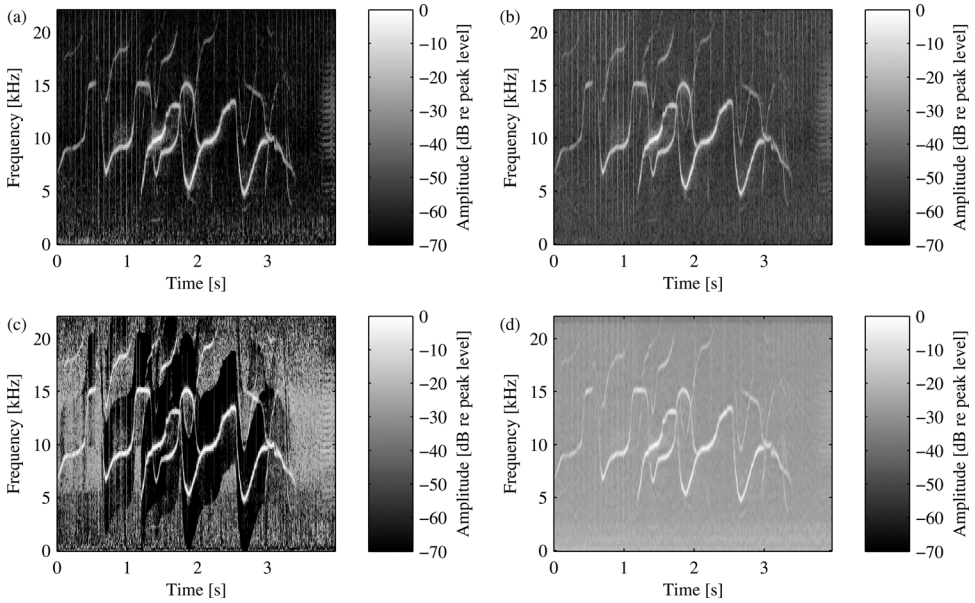

FIG. 3. Spectrogram of the recording of Fig. 1. (a) De-noised by subtracting the noise spectrum. (b) Pre-whitened by dividing by the noise spectrum. (c) De-noised using the method of Datta and Sturtivant.[3] (d) De-noised using the method of Mallawaarachchi *et al.*[7]

where the polynomial $A(z^{-1},t)$ is

$$A(z^{-1},t) = 1 + \sum_{i=1}^{n-1} a_i(t)\{z^{-i} + z^{-2n+i}\} + a_n(t)z^{-n} + z^{-2n}, \tag{4}$$

and the parameter ρ $(0 < \rho < 1)$ controls the notch width and is called the pole contraction factor. The order of $A(z^{-1},t)$ is $2n$ and n is the number of notches.

The coefficients of $A(z^{-1},t)$ have a symmetric form and are real-valued. This is a consequence of the fact that the roots of $A(z^{-1},t)$ are on the unit circle and occur in complex-conjugate pairs.[12] We can also express $A(z^{-1},t)$ as

$$A(z^{-1},t) = \prod_{i=1}^{n} \left(1 - z^{-1}e^{j\omega_i(t)}\right)\left(1 - z^{-1}e^{-j\omega_i(t)}\right) \tag{5}$$

where $\omega(t) = [\omega_1(t), \ldots, \omega_n(t)]^T$ are the notch center frequencies.

Nehorai presented an algorithm for estimating $\boldsymbol{a}(t) = [a_1(t), \ldots, a_n(t)]^T$ adaptively.[12] The properties of this algorithm are well known,[13,17] and it is applicable to whistle tracking. However, the whistle detection strategy of Sec. V requires that the filter is parametrized in terms of the notch center frequencies. Chen *et al.* presented such a direct frequency estimation ANF using a recursive prediction error (RPE) method.[16] The method updates $\omega(t)$ using

$$\omega(t) = \omega(t-1) + \boldsymbol{P}(t)\psi_\omega(t)e(t). \tag{6}$$

Here, $\psi_\omega(t)$ is the gradient of the prediction error with respect to $\boldsymbol{\omega}(t)$ and $\boldsymbol{P}(t)$ is the inverse of the so-called pseudo-Hessian matrix.[18] The pseudo-Hessian is a matrix of approximate second derivatives of $e(t)$:

$$\boldsymbol{P}(t) = \sum_{m=1}^{t} \lambda^{t-m}\psi_\omega(m)\psi_\omega^T(m) \tag{7}$$

where λ is the forgetting factor, which controls the duration of the analysis window. To use Eq. (6), we need to express $e(t)$ and $\psi_\omega(t)$ in terms of $y(t)$ and $\omega(t)$. From Eqs. (2) to (4), we have

$$e(t) = y(t) - y(t-2n) + \rho^{2n}e(t-2n) + \boldsymbol{\phi}^T(t)\boldsymbol{a}(t-1) \tag{8}$$

where $\boldsymbol{\phi}(t) = [\phi_1(t), \ldots, \phi_n(t)]^T$ is given by

$$\begin{aligned}\phi_i(t) &= -y(t-i) - y(t-2n+i) \\ &\quad + \rho^i(t)e(t-i) + \rho^{2n-i}(t)e(t-2n+i), \quad i < n \\ \phi_n(t) &= -y(t-n) + \rho^n(t)e(t-n).\end{aligned} \tag{9}$$

We estimate $\boldsymbol{a}(t)$ from $\boldsymbol{\omega}(t)$ in an iterative fashion.[19] These iterations are applied through increasing model order. So that $a_i^{(m)}(t)$ denotes the coefficients of order m, with $a_i(t) = a_i^{(n)}(t)$, and for $m = 1, 2, \ldots, n$ one calculates

$$a_i^{(m)}(t) = a_i^{(m-1)}(t) - 2a_{i-1}^{(m-1)}(t)\cos\omega_m(t) + a_{i-2}^{(m-1)}(t) \tag{10}$$

for $1 \le i \le n$, given that $a_0^{(m)}(t) = 1$ for all m and $a_i^{(m)}(t) = 0$ for all other i and m.

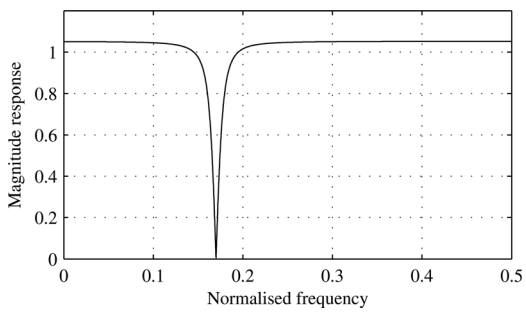

FIG. 4. Transfer function magnitude of a notch filter.

The gradient $\boldsymbol{\psi}_\omega(t)$ is expressed using the chain rule as

$$\boldsymbol{\psi}_\omega(t) = -\frac{\partial e(t)}{\partial \boldsymbol{\omega}^T} = -\frac{\partial e(t)}{\partial \boldsymbol{a}^T}\frac{\partial \boldsymbol{a}(t)}{\partial \boldsymbol{\omega}^T} = \psi_a(t)\frac{\partial \boldsymbol{a}(t)}{\partial \boldsymbol{\omega}^T}. \quad (11)$$

The gradient with respect to $\boldsymbol{a}(t)$, $\boldsymbol{\psi}_a(t)$, is given by[12]

$$\begin{aligned}\boldsymbol{\psi}_a(t) = &-\sum_{i=1}^{n-1} a_i(t)\{\boldsymbol{\psi}_a(t-i) + \boldsymbol{\psi}_a(t-2n+i)\} \\ &- a_n(t)\boldsymbol{\psi}_a(t-n) - \boldsymbol{\psi}_a(t-2n) + \boldsymbol{\phi}(t). \quad (12)\end{aligned}$$

The Jacobian $\partial \boldsymbol{a}(t)/\partial \boldsymbol{\omega}^T$ can be estimated using[19]

$$\begin{aligned}\frac{\partial a_0(t)}{\partial \omega_p} &= 0 \\ \frac{\partial a_1(t)}{\partial \omega_p} &= 2\sin\omega_p(t) \\ \frac{\partial a_i(t)}{\partial \omega_p} &= 2\cos\omega_p(t)\frac{\partial a_{i-1}(t)}{\partial \omega_p} - \frac{\partial a_{i-2}(t)}{\partial \omega_p} \\ &\quad + 2a_{i-1}(t)\sin\omega_p(t), \quad 2 \le i \le n,\end{aligned} \quad (13)$$

where $1 \le p \le n$.

It is common to set $\lambda = \rho$. Previously, notch filters have been applied to tonals that start simultaneously at known times. The forgetting and pole contraction factors were then set to low values at the onset of a new tonal.[12,13,16,19] However, in automatic whistle analysis, we do not know *a priori* when the whistles start, so this strategy cannot be used. Instead we use constant values close to 1 for both λ and ρ. As our results show, the ANF can still pick up new tonals. This is essential to the operation of the proposed whistle analysis method. It permits us to use the ANF to estimate the dominant frequencies at each time and then detect whistles *after* frequency estimation by determining whether or not the ANF was tracking a whistle at each time.

Throughout this study, we use $\lambda = \rho = 0.94$. This value was arrived at by experimentation, applying the method to many whistle recordings. A lower value would make it easier to pick up new tonals but decrease the noise robustness.[12]

The ANF is initialized at time $t = 0$ by distributing the notch center frequencies evenly in the available frequency interval, with the pseudo-Hessian, $\boldsymbol{P}(t)$, being initialized as a diagonal matrix with elements equal to the reciprocal of the estimated signal variance.

A simulated signal of two linearly chirping whistles in additive noise, sampled at 50 kHz and with a duration of 1 s, is used to demonstrate the frequency tracking capabilities of the adaptive notch filter method. The first chirp has an amplitude of 1. It starts at 15 kHz at 0.2 s and ends at 5 kHz at 0.6 s. The second chirp starts at 5 kHz at 0.4 s and ends at 20 kHz at 0.8 s. Its amplitude decreases linearly from 10 to 1. To obtain a realistic onset, the chirps' amplitudes are ramped up linearly for 0.01 s.

Throughout this study, notch center frequencies are smoothed using a 10 tap averaging filter prior to display. This reduces sample-to-sample variations and is motivated by the fact that the effective temporal resolution of the ANF is not given by the sampling period, but is governed by the analysis window. The window is exponential with an effective width of approximately $1/(1-\lambda) = 16.7$ samples for $\lambda = 0.94$. Smoothing by a 10 tap rectangular window will decrease the temporal resolution to 25.7 samples; this is still far better than for the spectrogram-based methods. The averaging length was chosen heuristically.

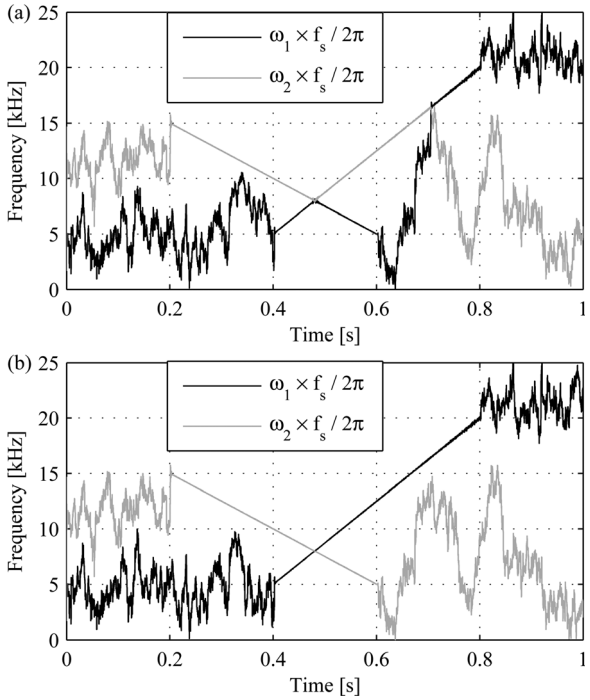

FIG. 5. Frequency tracks of the simulated signal, estimated using (a) an ANF and (b) an ANF with improved tracking through frequency crossings.

Figure 5(a) shows frequencies estimated from simulated signal using a fourth order ANF. The frequencies oscillate rapidly when not tracking, but when a whistle starts, it is rapidly detected by the frequency tracker and is successfully tracked until it ends. When a whistle ends, its frequency track returns to an oscillatory behavior.

Note that just before 0.5 s, the estimated frequencies for the two tracks cross. Before the crossing, notch 2 follows the upper frequency whistle, but at the crossing, it switches to tracking the other whistle. Then at 0.7 s, notch 1 takes over the whistle tracking from notch 2. The order of the tracks does not influence the prediction error, so the algorithm was modified to enhance its behavior through crossing points.

IV. TRACKING THROUGH NOTCH FREQUENCY CROSSINGS

To track whistles through notch frequency crossings, an improvement to the ANF-based frequency tracker that estimates and employs the sweep rate of each notch frequency evolution is developed. The frequency update rule Eq. (6) is replaced by

$$\begin{aligned}\boldsymbol{\omega}(t) = &(\mathbf{I}_n - \boldsymbol{\beta})[\boldsymbol{\omega}(t-1) + \boldsymbol{P}(t)\boldsymbol{\psi}_w(t)e(t)] \\ &+ \boldsymbol{\beta}\hat{\boldsymbol{\omega}}(t|t-1), \quad (14)\end{aligned}$$

where $\mathbf{I}_n$ $\hat{\boldsymbol{\omega}}(t|t-1)$ is a prediction of the current notch frequencies, derived from previously estimated sweep rates and $\boldsymbol{\beta}$ is a diagonal matrix that controls the balance between the

signal-driven and sweep rate-driven update terms. The elements of $\boldsymbol{\beta}$ are selected according to whether a frequency estimate is in the vicinity of other frequency estimates, i.e., whether trajectories are about to cross. When there are no frequency crossings, $\boldsymbol{\beta}=\mathbf{0}$, and Eq. (14) is equivalent to the standard frequency update of Eq. (6).

At each time step, the algorithm searches for notch frequencies that are closer than a prescribed threshold, here 0.02 times the sampling frequency. If notches i and j are closer than the threshold, we set $\beta_{ii}=\beta_{jj}=0.5$. When the frequency separation exceeds another threshold value (0.03 times the sampling frequency is employed in this implementation), the algorithm restores $\beta_{ii}=\beta_{jj}=0$ unless either of the notches are close to another notch. To avoid hysteresis, the frequency separation thresholds should be different.

Sweep rate estimation is conducted independently for each notch. In spectrogram-based processing, the sweep rate of a whistle track is typically estimated using first-order differences between adjacent partitions. In the case of ANF, the frequency estimates are available at a higher update rate, and a sweep rate estimate should be based on several samples.

A computationally simple calculation of the predicted frequency $\hat{\omega}_k(t|t-1)$ of the kth notch can be obtained by modeling the frequency evolution in a short analysis window as a linear function of time:

$$\hat{\omega}_k(m|t-1)=\tau_{k,1}(t-1)m+\tau_{k,2}(t-1), \tag{15}$$

where m is a temporal index. The parameters $\tau_k=\left[\tau_{k,1},\tau_{k,2}\right]^T$ of this linear model are updated adaptively at each time by minimizing the frequency prediction error $e_{\omega,k}(t)$,

$$e_{\omega,k}(t)=\omega_k(t)-\hat{\omega}_k(t|t-1), \tag{16}$$

using the Gauss–Newton RLS algorithm in a manner similar to Eq. (6). The result is[20,21]

$$\boldsymbol{R}(t)=\sum_{m=1}^{t}\lambda_T^{t-m}\begin{bmatrix}m^2 & m\\ m & 1\end{bmatrix} \tag{17a}$$

$$\boldsymbol{\tau}_k(t)=\boldsymbol{\tau}_k(t-1)+\boldsymbol{R}^{-1}(t)\begin{bmatrix}t & 1\end{bmatrix}^T e_{\omega,k}(t), \tag{17b}$$

where $\boldsymbol{R}$ is the pseudo-Hessian for estimation of $\boldsymbol{\tau}_k$, and λ_T is a forgetting factor for frequency trend estimation. Typically it is appropriate to use a longer estimation window for trend estimation than for ANF frequency estimation to reflect the inherently more noisy character of trend parameters. Here, $\lambda_T=0.99$ is used.

One can show that by assuming $t\gg 1$, i.e., ignoring the effects of initialization, the update Eq. (17b) can be simplified to[20,21]

$$\boldsymbol{\tau}_k(t)=\boldsymbol{\tau}_k(t-1)+\begin{bmatrix}(1-\lambda_T)^2\\ 1-\lambda_T^2-t(1-\lambda_T)^2\end{bmatrix}e_{\omega,k}(t). \tag{18}$$

Equations (14), (15), and (18) together with the framework for controlling the trade-off between signal-driven and trend-driven update comprise the proposed method for improved frequency tracking at crossings.

Figure 5(b) shows frequencies estimated by applying the ANF with improved tracking at frequency crossings to the simulated signal of Fig. 5(a). It is clear that the modified ANF manages to track the whistles through the frequency crossing.

V. WHISTLE DETECTION

The proposed method applies detection *after* frequency estimation. Whistle frequencies are estimated using the adaptive filter described in the preceding text, and a method for deciding when each notch is tracking a whistle will now be described.

The proposed detection method works by thresholding the diagonal elements of the inverse pseudo-Hessian matrix $\boldsymbol{P}(t)$ at each time. The detection statistic employed for each notch is the negative logarithm of the corresponding diagonal element of $\boldsymbol{P}(t)$:

$$\alpha_k(t)=-\log_{10}\boldsymbol{P}_{kk}(t),\quad 1\le k\le n. \tag{19}$$

The detection statistics $\alpha_k(t)$ provide a reliable quality measure of the tracking of each notch.[21] Some motivation for this is provided by the fact that in white Gaussian background noise, $\boldsymbol{P}(t)$, by construction, is similar to the Fisher information matrix for estimation of $\boldsymbol{\omega}(t)$. The diagonal elements of the Fisher information matrix give the Cramer–Rao lower bound, i.e., a lower bound on the variance of an unbiased estimator. Strictly, it applies only to estimation of deterministic parameters. It has been shown that the ANF algorithm of Chen *et al.* attains the Cramer–Rao lower bound,[16] wherefore it can be said that $\boldsymbol{P}_{kk}(t)$ are related to the variance of our frequency estimator. This establishes a link between the employed detection statistics and the variance of the frequency estimates, which in turn are directly related to the quality of the tracking.

Figure 6(a) shows the detection statistics for the simulation example of Fig. 5(b). A threshold of 2.2, indicated by a

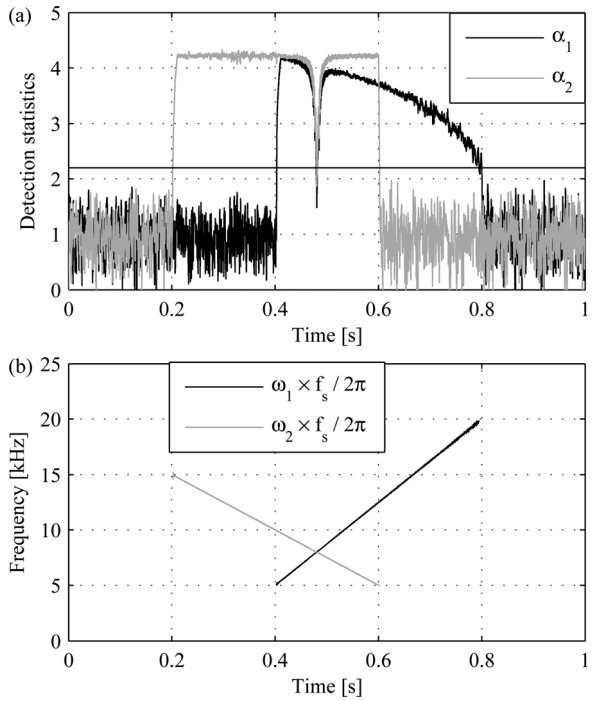

FIG. 6. Detection applied to the simulated signal. (a) Detection statistics. (b) Detected whistles.

horizontal line, is used for all data, simulated as well as recorded, presented here. This value was determined empirically using both simulated and real test signals. Experiments not reported here showed that on simulated linear chirps in additive noise, it corresponds to a whistle signal to noise ratio of approximately -3 dB. This can also be observed in Fig. 6, where the detection statistic crosses the threshold line just before 0.8 s. Here the amplitude of the tracked chirp is approximately 1. Note that the pre-whitening stage ensures that the threshold is independent of the signal gain, or spectral content of the noise.

Comparing Fig. 6(a) to Fig. 5(b), it is clear that the detection statistic for a notch is above the threshold when it is tracking a whistle and below it when not except near the frequency crossing. Tracking several notches through a frequency crossing is difficult, and the statistic frequently falls below the threshold.

Simple heuristic rules can assist in the construction of a suitable detection decision. In this case, a new detection is not declared until the threshold is exceeded for a fixed number of samples, and a track is not terminated until a set number of statistics fall below the threshold. In this work values of 500 and 50, respectively, were used for these parameters. In the vicinity of crossings, trajectories may not be terminated until the statistic is below the threshold for a greater period of time, in this case 1 000 samples.

The resulting detections for the test signal are shown in Fig. 6(b). When comparing it to Fig. 5(b), it is clear that at all times when a whistle is tracked, a detection is reported. The rules reported in the preceding text permit the detector to bridge the gap caused by a temporary fall below the threshold at the frequency crossing.

VI. CLICK COMPENSATION

Odontocete species (toothed whales) use short transients known as clicks mainly for the purposes of echolocation. If a transient is powerful enough that its magnitude spectrum in the notch filter analysis window is comparable to the magnitude of the tracked whistle, the notch filter can lose the whistle track. A similar artifact is observed when using Fourier based methods to track whistles. The effect is illustrated by Fig. 7(b), which shows frequencies obtained from a second order notch filter applied to the data of Fig. 7(a), which represents a section of the recording of Fig. 1.

A waveform-based click detector is developed as a first step to mitigating the influence of short-duration transients. The detector compares the square of the current sample to a short-time power estimate. If the ratio of these two values exceeds a threshold, in this case 10 is used, the current sample is deemed to be a click. It is also deemed a click if any of the 10 previous samples exceeded the threshold. This prevents the tracker from being dragged off course by click samples that do not quite meet the threshold.

The average power measure is calculated from a sliding window which is longer than the transients of interest. Here, a window of duration 1 ms, corresponding to 44 samples at a sample rate of 44.1 kHz, is used. If the current sample is a click, the average power measure is not updated.

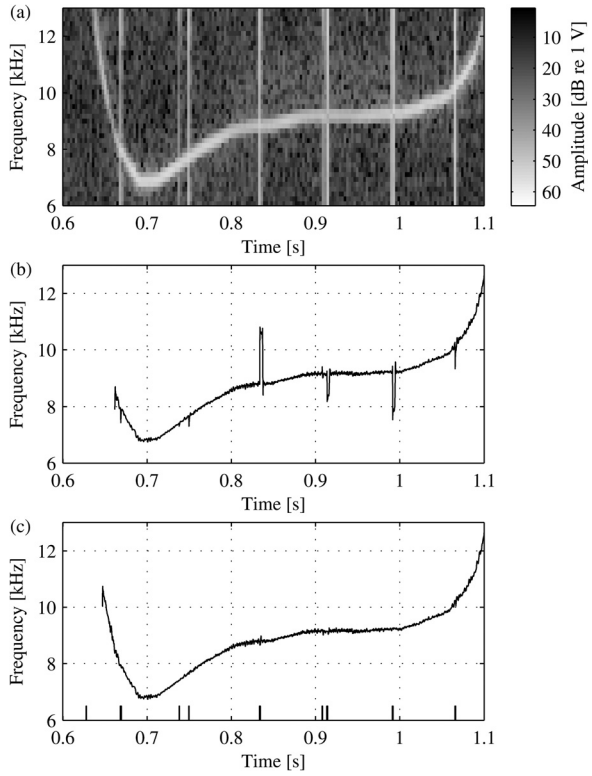

FIG. 7. Click detection. (a) Spectrogram of recording with clicks. Estimated frequency evolutions (b) without click detection and (c) with click detection. The bars on the temporal axis indicate click detections.

If the current sample is a click, it is not useful for whistle frequency tracking. We therefore replace it with its prediction obtained by setting $e(t) = 0$ in Eq. (8). No other changes are made to the algorithms for frequency tracking, sweep rate estimation, and detection presented in Secs. III to V.

The frequency tracking example of Fig. 7(b) is re-run with click compensation, and the results are shown in Fig. 7(c). The figure shows that click disturbances are greatly reduced. Click detections are indicated by bars on the temporal axis of Fig. 7(c).

VII. RESULTS

We will now evaluate the performance of the proposed method for whistle detection and tracking by application to several recorded whistles. Results from application of the competing Fourier based methods of Datta and Sturtivant[3] and Mallawaarachchi *et al.*,[7] both described in Sec. I, are also presented.

Processing parameters are the same for all results presented here. Most of the parameters and settings applied to the three methods have been specified in previous sections, but some have not yet been given. Parameter values were taken from the reports on the competing methods.[3,7,22] Where the value of a parameter was not given, we optimized it for best performance using simulated as well as recorded data. For the proposed method, we use a notch filter with $n = 3$ notches, so we can extract at most three simultaneous whistles. Spectrograms for pre-whitening and for the analysis by the spectrogram-based methods are calculated using 256 point Hann windows with 50% overlap.

For Datta and Sturtivant's method, track inertia with a parameter of 0.95 is applied to improve whistle tracking through frequency crossings during extraction. We allow a maximum frequency jump of three bins between adjacent partitions and discard detections shorter than 16 partitions. This corresponds to a minimum detection length of 2 048 samples. The exact peak detection thresholds were not given in Refs. 3 and 22, but it is stated that the thresholds should be set using the mean and standard deviation of each partition of the denoised spectrogram. We have set the peak detection threshold equal to the mean plus three times the standard deviation. For ridge extraction from the detected peaks, a lower threshold of two standard deviations above the mean was used.

For Mallawaarachchi *et al.*'s method, the initial peak detection threshold is calculated from the mean and standard deviation of the denoised spectrogram. Following Mallawaarachchi *et al.*, we use a threshold of 1.75 times the standard deviation above the mean. Region growing uses a threshold of 0.95 times the detection threshold. Whistle tracing and Kalman filter post-processing are then applied according to Ref. 7. Additionally enforcing a maximum frequency jump of 10 bins between neighboring partitions in the same manner as for Datta and Sturtivant's method reduced the number of false alarms while leaving the true detections unaffected. Requiring a minimum detection length of 16 partitions had the same effect, so these modifications were introduced.

The dolphin whistle recordings are sampled at 44.1 kHz. The spectrogram of the first recording, which contains three single Atlantic spotted dolphin (*Stenella frontalis*) whistles in strong background noise, is illustrated in Fig. 8(a). Frequency tracks of whistle detections from the proposed method are given by Fig. 8(b). The algorithm has detected all three whistles in their entirety and has not produced any false alarms.

Figure 8(c) shows that Datta and Sturtivant's method has detected only half of the first whistle and missed the rapidly sweeping part of the second whistle. However, it has picked up the weak potential whistle around 2.8 s. A very faint additional whistle at 0.8 s is detected by the method, but the extracted frequency track appears to extend outside the edges of the whistle. It is difficult to see the detection at 0.9 s in the spectrogram. Consequently, Datta and Sturtivant's method has failed to extract all stronger whistles in their entirety, but managed to pick up some more faint whistles. In our experience, this behavior is typical of Datta and Sturtivant's method.

The results of Mallawaarachchi *et al.*'s method are presented in Fig. 8(d). It has fully extracted the three whistles and also the faint additional whistle at 0.8 s. However, there are some spurious edge effects on the frequency tracks. Note that the reason that Mallawaarachchi *et al.*'s method can track the rapidly swept part of the second whistle is that it uses a higher maximum frequency jump between partitions than Datta and Sturtivant's method.

The proposed method gives a new set of frequency estimates every sample, while the spectrogram-based methods give a new set for every spectrogram partition. Small

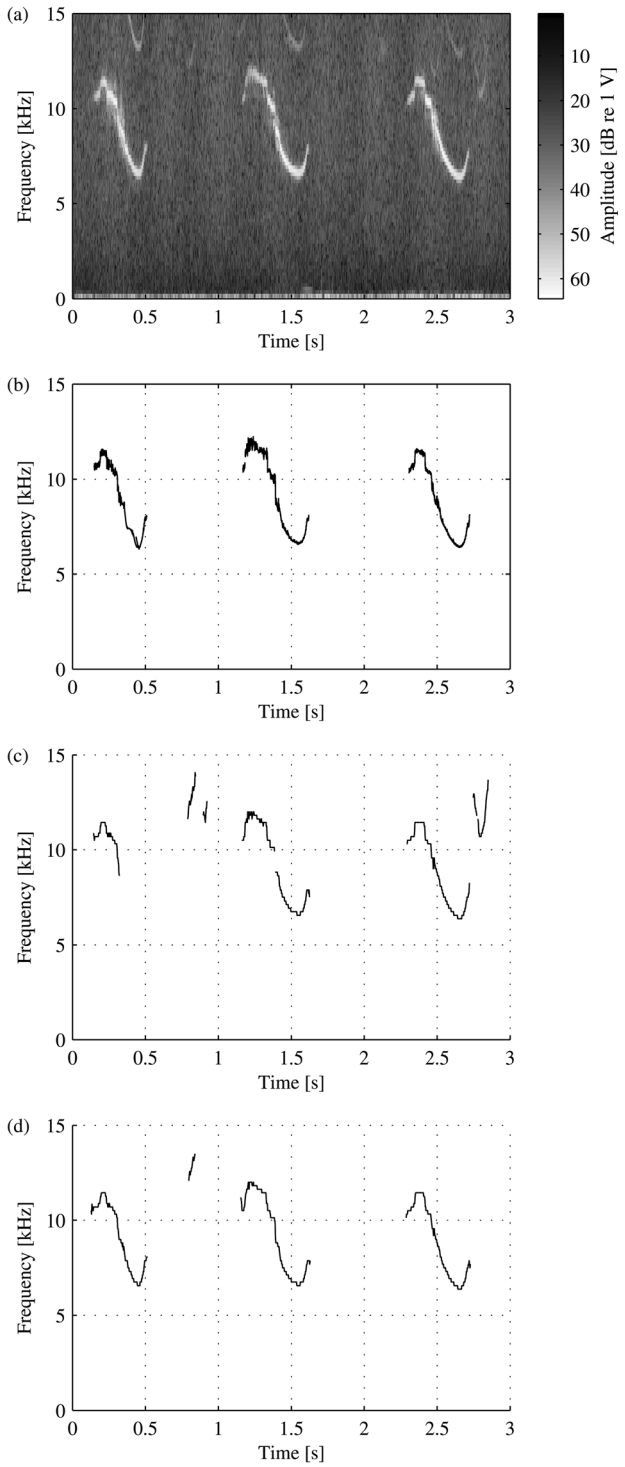

FIG. 8. Analysis of the spotted dolphin recording. (a) Spectrogram. (b) Detections from the proposed method. (c) Detections from Datta and Sturtivant's method.[3] (d) Detections from Mallawaarachchi *et al.*'s method.[7]

sample-to-sample frequency variations give the ANF frequency estimate curves their "thick" appearance when plotted on a dense temporal axis as in Figs. 8 and 9.

The second recording, discussed in Sec. II and shown in Fig. 1, has several simultaneous strong bottlenose dolphin (*Tursiops truncatus*) whistles and many strong clicks. Figure 9(a) shows that our proposed method extracts many of the whistles that can be visually identified from the spectrogram.

It appears undisturbed by clicks and extracts the two long whistles that last from 0.6 to 2.0 and 1.2 to 3.0 s, respectively, in their entirety and through frequency crossings. Moreover, all its detections can be matched to whistle tracks that can be visually identified in the spectrogram. The methods of Datta and Sturtivant, see Fig. 9(b), and Mallawaarachchi *et al.*, see Fig. 9(c), both give false alarms at low frequencies. They also fail to extract many of the whistles, and only extract parts of the two long whistles mentioned in the preceding text. Mallawaarachchi *et al.*'s method can only extract one simultaneous whistle, but this result is more surprising for Datta and Sturtivant's method, which can extract multiple simultaneous whistles. One cause of this failure is the masked equalization operation, which attenuates multiple whistles if their frequency separation is less than the width of the frequency mask. The masked equalization is however necessary for click attenuation. None of the methods appear disturbed by the clicks.

Figure 1 shows that around 1.3, 1.7, 2.0, and 2.5 s, a whistle exhibits a rapid frequency sweep. The proposed method can track the whistles through these rapid frequency sweeps, but both spectrogram-based methods fail. This is because these sweeps are so rapid that the whistle's energy is split between several frequency bins. This could be alleviated by selecting a shorter spectrogram partition length but that would lead to a decreased frequency resolution, which would make peaks from slowly sweeping whistles less apparent. It would also make the analysis less noise robust; the whistle peaks would not be as high above the background noise. This is an example of the fundamental trade-off

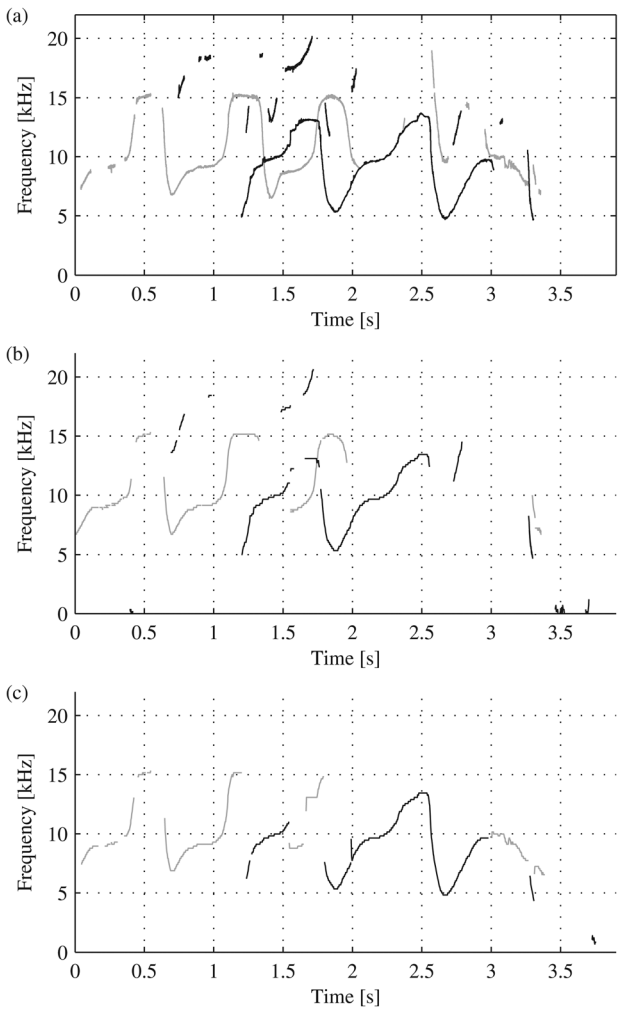

FIG. 9. Analysis of the bottlenose dolphin recording of Fig. 1. (a) Detections from the proposed method. (b) Detections from Datta and Sturtivant's method.[3] (c) Detections from Mallawaarachchi *et al.*'s method.[7]

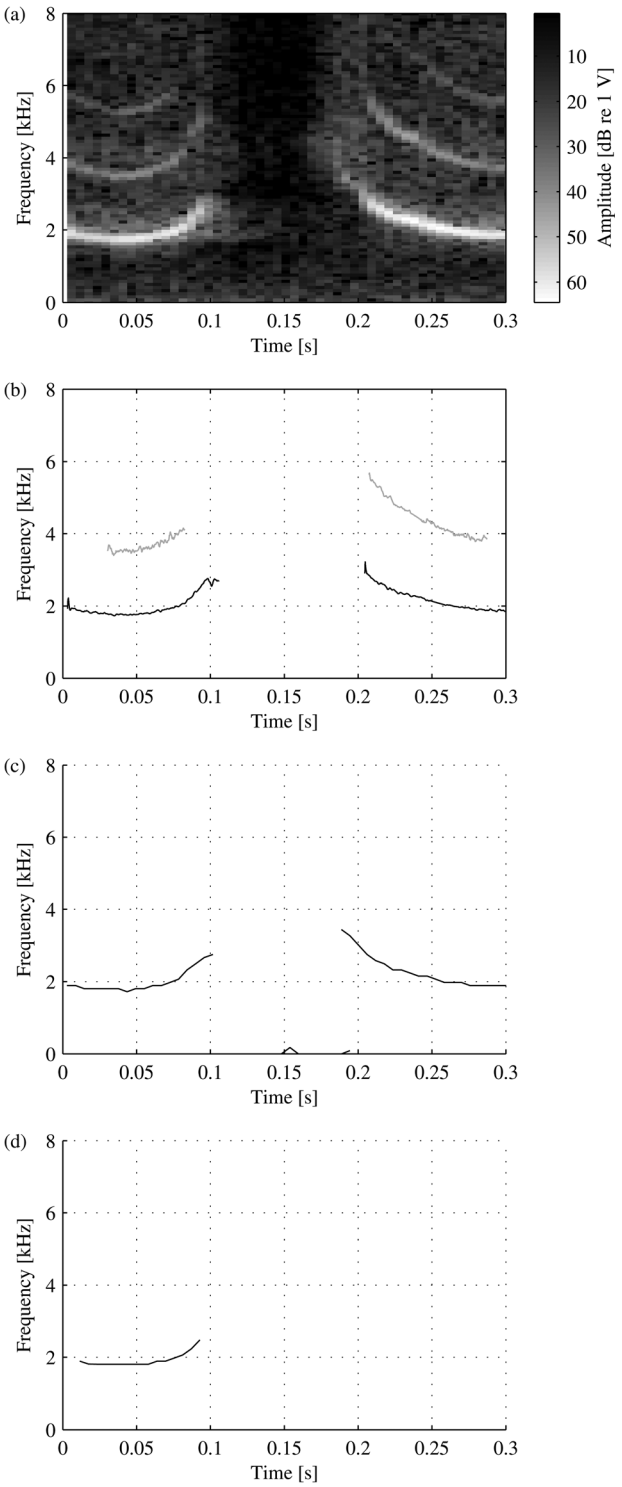

FIG. 10. Analysis of a human whistled speech utterance: "ata" in whistled Turkish. (a) Spectrogram. (b) Detections from the proposed method. (c) Detections from Datta and Sturtivant's method. (d) Detections from Mallawaarachchi *et al.*'s method.

involved in selecting spectrogram parameters and an effect of the limited time-frequency resolution of the spectrogram. Most authors use 256- or 512-point spectrograms for whistle analysis of recordings sampled at rates of 44.1 or 48 kHz;[3,5–7] our selection of 256 points is typical. Compared to the 256-point spectrograms employed here, the ANF method uses a much shorter temporal window, resulting in a higher temporal resolution. Therefore it is more effective at extracting rapidly swept whistles.

Our last sound example demonstrates that the proposed method also performs well on non-marine mammal whistles. It is a high quality recording of a human whistled speech utterance, namely, the whistled emulation of "ata" from the word "Yatan" of the whistled Turkish language. The recording is sampled at 22.05 kHz. All parameters are the same as for the above dolphin whistles sampled at 44.1 kHz.

Figure 10(a) presents a spectrogram of the whistled utterance. Compared to the dolphin whistles, the human whistles have stronger harmonics and a larger bandwidth relative to the sampling frequency. When whistling, some flow noise is also generated. The flow noise has a broadband character and is responsible for the apparent increase in background noise level when a whistle is present.

As shown by Fig. 10(b), the proposed method extracts nearly all of the fundamental and most of the second harmonic. Again, it displays no false alarms. Figure 10(c) shows that Datta and Sturtivant's method fails to extract the second harmonic and reports low frequency false alarms. Figure 10(d) shows that Mallawaarachchi *et al.*'s method[7] detects only the first part of the fundamental.

VIII. CONCLUSIONS

An automatic whistle analysis method based on adaptive notch filters has been described. The whistle detection and frequency estimation method presented here has been shown to be applicable to real-world whistle recordings of different characteristics recorded in different settings. Consequently, it is suitable for use in an automatic whistle processing system. The performance of the method was compared to two spectrogram-based methods, and it was found that the proposed method is both capable of extracting several simultaneous whistles and of accurately extracting all whistles of sufficient signal-to-noise ratio provided that there are enough notches in the filter. The results also showed that the method can track simultaneous whistles through frequency crossings and produce an unbroken detection from whistles as long as 1.8 s. The detection methodology reported here was found satisfactory, producing detections that correspond to the known start and end times of simulated whistles and not reporting any false alarms on real data. An interesting direction for future research is to develop an improved detector based on the proposed detection statistics. Moreover, a drawback in common to the proposed method and the two competing spectrogram-based methods applied here is that they all report whistle detections on periodic non-whistle calls, which spectrally can be characterized as fundamentals with strong harmonics. Perhaps the most important direction for future research is to develop a method to avoid such non-whistle detections or alternatively to separate whistles and periodic call detections in a post-processing stage.

ACKNOWLEDGMENTS

The authors would like to acknowledge the financial support to this project by the Engineering and Physical Sciences Research Council, the Institute of Sound and Vibration Research, and QinetiQ, Ltd. We thank QinetiQ for providing the marine mammal data, Julien Meyer for providing the human whistled speech recording, and M. Chitre and coauthors of Ref. 7 for sharing the software implementation of their analysis method. Finally, we thank Leif Persson and Julien Meyer for helpful comments on early versions of this manuscript.

[1] J. N. Oswald, S. Rankin, J. Barlow, and M. O. Lammers, "A tool for real-time acoustic species identification of delphinid whistles," J. Acoust. Soc. Am. **122**(1), 587–595 (2007).
[2] J. Meyer, "Typology and acoustic strategy of whistled languages: phonetic comparison and perceptual cues of whistled vowels," J. Int. Phonetic Assoc. **38**(1), 69–94 (2008).
[3] S. Datta and C. Sturtivant, "Dolphin whistle classification," Signal Process. **82**(2), 251–258 (2002).
[4] J. R. Buck and P. L. Tyack, "A quantitative measure of similarity for Tursiops Truncatus signature whistles," J. Acoust. Soc. Am. **94**(5), 2497–2506 (1993).
[5] D. Gillespie, J. Gordon, R. McHugh, D. Mclaren, D. Mellinger, P. Redmond, A. Thode, P. Trinder, and X. Y. Deng, "PamGuard: Semiautomated, open source software for real-time acoustic detection and localization of cetaceans," in *Proceedings of the Institute of Acoustics* (Institute of Acoustics, St Albans, UK, 2008), Vol. 30, Pt. 5.
[6] X. C. Halkias and D. Ellis, "Call detection and extraction using bayesian inference," Appl. Acoust. **67**(11-12), 1164–1174 (2006).
[7] A. Mallawaarachchi, S. H. Ong, M. Chitre, and E. Taylor, "Spectrogram denoising and automated extraction of the fundamental frequency variation of dolphin whistles," J. Acoust. Soc. Am. **124**(2), 1159–1170 (2008).
[8] T. S.-T. Leung and P. R. White, "Robust estimation of oceanic background noise spectrum," in *Mathematics in Signal Processing IV*, edited by J. G. McWhirter and I. K. Proudler (Clarendon, Oxford, UK, 1998), pp. 369–382.
[9] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal. II. Algorithms and applications," Proc. IEEE **80**(4), 540–568 (1992).
[10] C. Ioana, C. Gervaise, Y. Stéphan, and J. I. Mars, "Analysis of underwater mammal vocalisations using time-frequency-phase tracker," Appl. Acoust. **71**(11), 1070–1080 (2010).
[11] S. Peleg and B. Porat, "Estimation and classification of polynomial phase signals," IEEE Trans. Inf. Theory **37**(2), 422–429 (1991).
[12] A. Nehorai, "A minimal parameter adaptive notch filter with constrained poles and zeros," IEEE Trans. Acoust., Speech, Signal Process. **33**(4), 983–996 (1985).
[13] Y. Xiao, Y. Takeshita, and K. Shida, "Tracking properties of a gradient-based second-order adaptive IIR notch filter with constrained poles and zeros," IEEE Trans. Signal Proces. **50**(4), 878–88 (2002).
[14] P. A. Regalia, *Adaptive IIR Filtering in Signal Processing and Control* (Dekker, New York, 1995), pp. 704.
[15] M. Niedzwiecki and A. Sobocinski, "Generalized adaptive notch smoothers for real-valued signals and systems," IEEE Trans. Signal Process. **56**(1), 125–133 (2008).
[16] B. S. Chen, T. Y. Yang, and B. H. Lin, "Adaptive notch filter by direct frequency estimation," Signal Process. **27**(2), 161–176 (1992).
[17] P. Stoica and A. Nehorai, "Performance analysis of an adaptive notch filter with constrained poles and zeros," IEEE Trans. Acoust., Speech, Signal, Process. **36**(6), 911–919 (1988).
[18] L. Ljung, *System Identification: Theory for the User*, 2nd ed. (Prentice Hall PTR, Upper Saddle River, NJ, 1999), pp. 672.

[19]A. Nehorai and D. Starer, "Adaptive pole estimation," IEEE Trans. Acoust., Speech, Signal Process. **38**(5), 825–838 (1990).
[20]A. T. Johansson and P. R. White, "A comparison of two methods for cetacean tonal detection and characterisation, with emphasis on performance at crossing frequencies," in *Proceedings of the Institute of Acoustics* (Institute of Acoustics, St Albans, UK, 2004), Vol. 26, Pt. 6.
[21]A. T. Johansson, "Parametric modelling of cetacean calls," Ph.D. thesis, Faculty of Engineering, Mathematics, and Applied Sciences, (University of Southampton, UK, 2004).
[22]C. Sturtivant, "Extraction and recognition of tonal sounds produced by small cetaceans and identification of individuals and behaviour," Ph.D. thesis, Department of Electronic and Electrical Engineering, (Loughborough University, UK, 1997).