A versatile pitch tracking algorithm: From human speech to killer whale vocalizations

Ari Daniel Shapiro[a)]
Department of Biology, Woods Hole Oceanographic Institution, MS 50, Woods Hole, Massachusetts 02543

Chao Wang[b)]
Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar Street, Cambridge, Massachusetts 02139

(Received 1 November 2008; revised 6 April 2009; accepted 17 April 2009)

In this article, a pitch tracking algorithm [named discrete logarithmic Fourier transformation-pitch detection algorithm (DLFT-PDA)], originally designed for human telephone speech, was modified for killer whale vocalizations. The multiple frequency components of some of these vocalizations demand a spectral (rather than temporal) approach to pitch tracking. The DLFT-PDA algorithm derives reliable estimations of pitch and the temporal change of pitch from the harmonic structure of the vocal signal. Scores from both estimations are combined in a dynamic programming search to find a smooth pitch track. The algorithm is capable of tracking killer whale calls that contain simultaneous low and high frequency components and compares favorably across most signal to noise ratio ranges to the peak-picking and sidewinder algorithms that have been used for tracking killer whale vocalizations previously.
© *2009 Acoustical Society of America.* [DOI: 10.1121/1.3132525]

PACS number(s): 43.80.Ka, 43.72.Ar [WA] Pages: 451–459

I. INTRODUCTION

Robust pitch detection is a crucial first step in the analysis and modeling of human speech. The fundamental frequency (f_0) plays an important role in modeling linguistic attributes including lexical stress, tone, and intonation, as well as paralinguistic attributes such as emotion. However, it is difficult to build reliable statistical models involving f_0 because of pitch estimation errors and the discontinuity of the f_0 contour. Specifically, inaccurate voiced pitch hypotheses and erroneous voiced/unvoiced (V/UV) decisions can lead to noisy and undependable feature measurements. This is especially true for telephone speech due to inferior pitch detection performance caused by the noisy and band-limited telephone channel.

Previously, a pitch detection algorithm (PDA) was developed utilizing the discrete logarithmic Fourier transformation (DLFT) of the speech signal (Wang and Seneff, 2000b). This algorithm, which will be referred to as DLFT-PDA, is based on a robust pitch estimation method known as harmonic matching (Hess, 1983). Reliable estimates of both pitch and the temporal change of pitch are derived based on harmonic matching principles, which are then combined in a dynamic programming (DP) search to find a globally optimal solution. The DP search tracks pitch continuously, avoiding the propagation of V/UV decision errors to voiced pitch hypotheses. Evaluation results have demonstrated that the algorithm is particularly suitable for telephone speech and prosodic modeling applications (Wang and Seneff, 2000a, 2000b; Wang, 2001; Wang and Seneff, 2001a, 2001b).

Pitch tracking is also important for analyzing and quantifying features of the vocalizations of marine mammals. Several different manual and automatic approaches have been implemented previously. A labor-intensive but fairly reliable method is to trace the f_0 on the spectrogram by hand using a digital interface (Watwood *et al.*, 2004, 2005; Shapiro, 2006). Because the number and placement of (f_0, *time*) points will vary between contours, a subsequent interpolation is used to hold this number constant and represent all contours with a uniform number of evenly spaced points. Another commonly used automated method is to select the peak frequency value from a sliding power spectrum of the signal (i.e., peak-picking), followed by subsequent manual correction to remove pitch doubling and halving errors (Buck and Tyack, 1993; Janik *et al.*, 1994; McCowan, 1995). Often the signal is band-pass filtered over an appropriate frequency range before the frequency associated with the peak spectral energy is selected. Another automated technique, the sidewinder algorithm, is similar to the spectral autocorrelation method for tracking human speech (Lahat *et al.*, 1987). The algorithm computes an autocovariance sequence for each spectral slice [in contrast with human speech, no spectral flattening is necessary for killer whale (*Orcinus orca*) vocalizations] whose peaks occur at multiples of the spacing of the frequency bands (Deecke *et al.*, 1999). Two pitch extraction methods have been implemented for this technique: one searches for the second highest peak directly in the autocovariance sequence itself, while the other computes the real cepstrum of the autocovariance sequence to locate the highest peak. The fast Fourier transform FFT based comb-filter method described in Brown *et al.* (2006) applies the

[a)]Author to whom correspondence should be addressed. Electronic mail: ashapiro@whoi.edu
[b)]Present address: Vlingo Corporation, 17 Dunster Street, Cambridge, MA 02138.

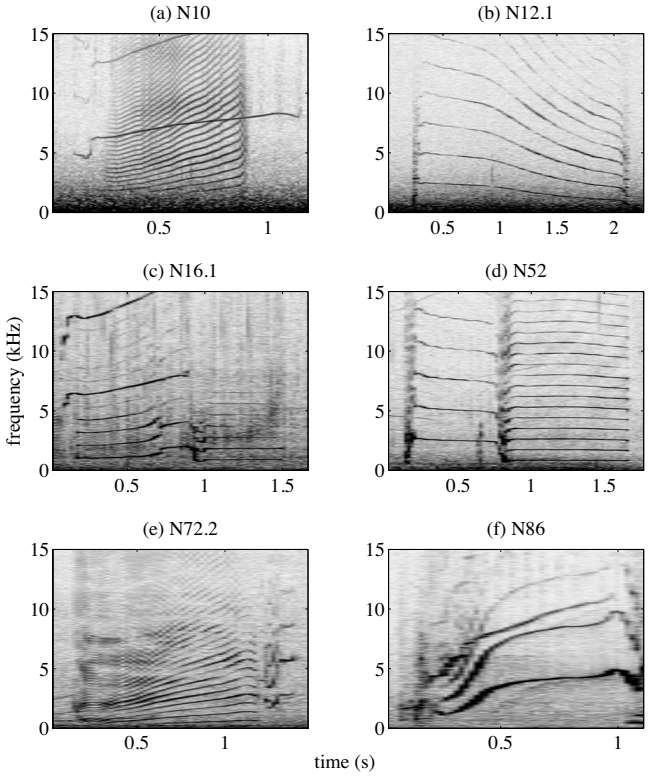

FIG. 1. Spectrograms of various killer whale calls, labeled by call type, and generated using a 2048-point FFT with 50% overlap and a Hamming window.

spectral comb method to tracking killer whale calls. It is the most similar in spirit to the DLFT-PDA algorithm.

Killer whale pods produce a set of stereotyped, harmonically-structured calls that often consist of multiple temporal and spectral components (Ford, 1987, 1989, 1991; Fig. 1). Individual killer whales tend to match the call types produced by other group members (Miller *et al.*, 2004). This kind of communication may facilitate group cohesion (Miller, 2002) and/or allow individuals to discriminate between one another (Miller *et al.*, 2007). Killer whales produce these vocalizations by varying the pulse repetition rate, which corresponds to the relative spacing between different frequency bands spectrographically (Watkins, 1967).

Each killer whale call type generally consists of a modulated low frequency component (LFC) with an f_0 typically ranging between 80 and 2400 Hz (Ford, 1987). Some call types also contain a high frequency component (HFC) [often with an f_0 between 2 and 12 kHz (Hoelzel and Osborne, 1986)] that is synchronously produced with the LFC but separately modulated to produce two sets of unique harmonics. The challenge of tracking the pitch of both of these components simultaneously is similar to that encountered by multi-pitch estimation of mixtures of music or speech signals (see Klapuri, 2008; Klapuri and Virtanen, 2008). In these scenarios, it is necessary to isolate and then track each constituent component of the additive signal. Automatic pitch tracking of killer whale calls would greatly facilitate the study and characterization of their stereotypy, cultural transmission (see Deecke *et al.*, 1999), and individual variability (see Miller and Bain, 2000; Nousek *et al.*, 2006).

Because multiple spectral components can be embedded within a single killer whale vocalization, a time domain representation of the signal would lead to a loss of the harmonic structure. For accurate pitch tracking, a frequency solution is required. This article examines the application of the DLFT-PDA to determine the fundamental frequencies of both the LFC and HFC of Norwegian killer whale stereotyped calls (see Nousek *et al.*, 2006, Miller *et al.*, 2007, for an application of this method). The DLFT-PDA has been designed especially for telephone speech, where the f_0 is often weak or missing and the signal to noise ratio (SNR) is usually low compared to microphone-recorded speech. Coincidentally, the recordings of marine mammal vocalizations are often characterized by the same features because of boat noise and substantial distances between the vocalizing animals and recording equipment. Several characteristics of our algorithm render it especially well suited for tracking the pitch of killer whale calls. First, the algorithm relies on the harmonic structure (i.e., spectral peaks at multiples of the f_0) to estimate pitch and deliberately ignores low-frequency spectrum, which makes it robust to interference from low-frequency boat noise. Second, the DLFT can be tuned to a sub-band in the spectrum, allowing the algorithm to track calls with simultaneous LFC and HFC that are somewhat separated in the frequency space.

In Sec. II, an overview and then the specifics of our algorithm are given (see Wang and Seneff, 2000b; Wang, 2001 for the full details), highlighting features that make it suitable for telephone speech and killer whale recordings. Adaptations of the algorithm are then described for killer whale calls. Finally, evaluation results of our algorithm are presented.

II. METHODS

A. Overview of algorithm

The DLFT-PDA is based on the observation that harmonic peaks will be spaced by a constant distance on a logarithmic frequency scale regardless of f_0. More formally, if a signal has harmonic peaks spaced by f_0, then on a logarithmic scale the peaks will occur at $\log f_0, \log f_0 + \log 2$, $\log f_0 + \log 3, \ldots$, etc. The fundamental frequency determines the position of the first peak and the subsequent harmonic peaks are at fixed distances from the first peak. Thus, harmonic spectra with different fundamental frequencies can be aligned by simple linear shifting. By correlating a spectrum sampled on the logarithmic frequency scale with a harmonic template (a logarithmic spectrum of an impulse train), a robust estimation of the $\log f_0$ of the signal can be obtained. The correlation of two logarithmic spectra from adjacent frames of a vocal signal leads to a very reliable estimation of the change in $\log f_0$ ($\Delta \log f_0$).

Instead of determining an f_0 value for each frame by picking the correlation maximum, a DP search is used to combine the $\log f_0$ and $\Delta \log f_0$ estimations to find an optimal solution overall. All values (quantized in the search space) are considered as possible f_0 candidates with different qualities. The quality of a pitch candidate P is indicated by the correlation between the spectrum and the template (Fig.

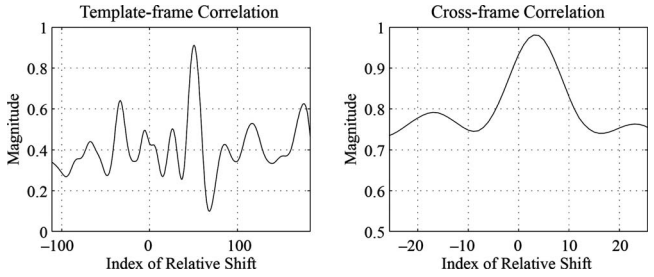

FIG. 2. Examples of "template-frame" and "cross-frame" correlations for DLFT spectrum.

2). The "consistency" of two consecutive pitch candidates is indicated by the correlation of the spectra of the adjacent frames at the position corresponding to the difference between the pitch candidates. These two constraints are used to define a score function for the DP search. The DP search algorithm solves the optimization problem iteratively (i.e., finding the optimal score at time t is achieved by finding the optimal score at time $t-1$). The path in the quantized (*frequency*, *time*) space with the highest score yields the optimum pitch track.

The algorithm requires defining a small set of parameters: the window size, the frequency range $[f_s,f_e]$ for the DLFT, and the f_0 search range and resolution. The pseudo-code of our pitch tracking algorithm is shown in Fig. 3 and the implementation details are discussed in Wang and Seneff, 2000b and Wang, 2001.

B. Details of algorithm

1. Signal representation

To obtain a logarithmically spaced spectrum for the frequency region $[f_s,f_e]$, the discrete-time Fourier transform is directly sampled at linear intervals on the logarithmic frequency scale. This representation is defined as a DLFT. Assuming $x_t(n)$ is a Hamming-windowed audio signal centered at time t ($n=0,1,\ldots,N-1$, where N is the window size), the DLFT is computed as follows:

N: the total number of frames in an input waveform
M: the total number of quantized pitch candidates
P_i: the quantized pitch candidates ($i = 0, \ldots, M-1$)
T: the harmonic template
X_t: the logarithmic-frequency spectrum at the tth frame of the input
$S(t, i)$: the path score for the i^{th} pitch candidate at the t^{th} frame
begin
  compute T
  compute X_0
  compute the correlation of X_0 and T
  initialize $S(0, i)$ for all P_i ($i = 0, \ldots, M-1$)
  for $t = 1, \ldots, N-1$
    compute X_t
    compute the correlation between X_t and $X_t - 1$
    compute the correlation between X_t and T
    update the partial path score $S(t, i)$ and
    save the back trace pointer for all P_i ($i = 0, \ldots, M-1$)
  end
  back trace to find the best pitch contour $P(t)$ ($t = 0, \ldots, N-1$)
end

FIG. 3. Pseudo-code of the DLFT-PDA.

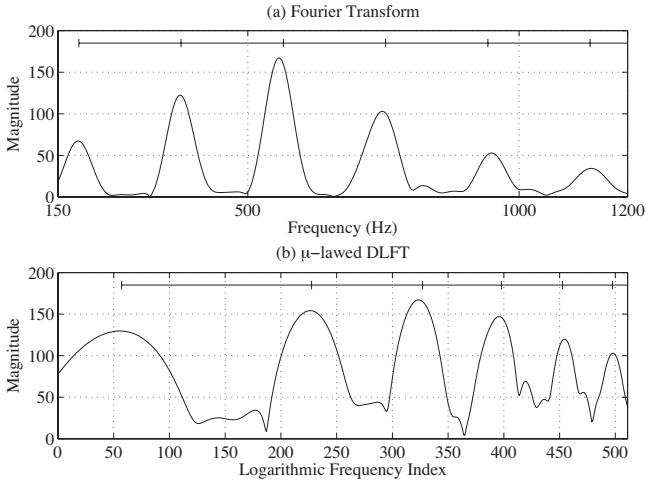

FIG. 4. (a) Fourier transform and (b) μ-law-compressed DLFT for a speech signal. Positions of harmonic peaks are indicated by ruler tick marks in the plots.

$$X_t(i) = \frac{1}{N}\sum_{n=0}^{N-1} x_t(n)e^{-j\omega_i n} \quad (i = 0,1,\ldots,N-1), \tag{1}$$

$$\omega_i = 2\pi e^{(\log f_s + i\cdot d\,\log f)} \cdot T_s, \tag{2}$$

$$d\log f = (\log f_e - \log f_s)/(N-1), \tag{3}$$

where T_s is the sampling period of the waveform. The term $d\log f$ can be viewed as the frequency resolution in the logarithmic domain.

The spectrum is normalized by a μ-law conversion to reduce the dynamic range of harmonic peak height due to formant influences:

$$X_t(i) = M_t \cdot \log(1+\mu\cdot X_t(i)/M_t)/\log(1+\mu)$$

$$(i = 0,1,\ldots,N-1), \tag{4}$$

where M_t is the maximum energy of the DLFT spectrum at the tth frame:

$$M_t = \max_i X_t(i). \tag{5}$$

The value $\mu=50$ was chosen in our implementation. The conversion holds the maximum value unchanged while promoting smaller values. Figure 4 shows the discrete Fourier transform (DFT) and μ-law-compressed DLFT for a Hamming-windowed voiced speech signal. Notice the dynamic range of the harmonic peaks in the DLFT spectrum is compressed due to the μ-law conversion.

2. Harmonic template

The harmonic template is constructed from an ideal periodic signal. The pulse train is first Hamming windowed, after which the DLFT spectrum is computed for the frequency range $[f'_s,f'_e]$. The parameters f'_s and f'_e can differ from those used for computing the DLFT of the signal. However, the equality $f'_e/f'_s=f_e/f_s$ must be ensured, so that the frequency resolution $d\log f$ [see Eq. (3)] of the signal and template's DLFT spectra match for the correlation operation.

TABLE I. Parameter settings for tracking killer whale calls. M is the number of pitch hypotheses in the search space. See text for details.

	Window	$[f_s, f_e]$	$[P_{\text{low}}, P_{\text{high}}]$	M
I	25 ms	[250,1750] Hz	[100,600] Hz	200
II	15 ms	[800,5600] Hz	[400,4000] Hz	500
III	2.5 ms	[6k,24k] kHz	[4,12] kHz	500

In our implementation, the harmonic template includes five complete harmonic lobes. The parameters of the template had been tuned using development data to achieve good performance for killer whale vocalizations (see Table I).

The harmonic lobes are wider for the low-frequency region on the logarithmic scale, which leads to a bias in reaching a correlation maximum when the first lobe in the template is included in the calculation. This is problematic when the harmonic component at f_0 is absent (very weak or out of range) in the signal's DLFT spectrum. Ideally, the template is also expected to match the signal maximally without the first component (i.e., $2P_T$ matches $2f_0$, $3P_T$ matches $3f_0$, and so on). However, due to the strong first lobe in the template, the correlation is likely to reach a maximum by matching the P_T, $2P_T$, and $3P_T$ components of the template with the $2f_0$, $4f_0$, and $6f_0$ components of the signal spectrum. This will result in pitch doubling errors.

To suppress this tendency, the energy of each harmonic lobe in the template is normalized, similar to the measure taken in Hermes, 1988. This is done by integrating over each lobe to find its area, followed by a scaling by the reciprocal of the area, subject to an exponential decay to tune the effect. The decay factor was determined empirically from development data to be 0.85.

A second measure taken to discourage pitch doubling errors is to add negative lobes between the positive lobes in the template. If the P_T, $2P_T$, and $3P_T$ components of the template match with the $2f_0$, $4f_0$, and $6f_0$ components of the signal spectrum, then the negative lobes would match the $3f_0$ and $5f_0$ components and would reduce the magnitude of the correlation value. The negative lobes are obtained by computing the DLFT spectrum of the same pulse train with a frequency shift equivalent to half of its fundamental P_T:

$$\omega_i = 2\pi e^{(\log f_s' + i \cdot d \log f - \log P_T/2)} \cdot T_s \tag{6}$$

where $d \log f$ and T_s are the same as in Eq. (2). The shift $\log P_T/2$ causes the harmonic peaks in the new spectrum to fall precisely between those in the original one. The final harmonic template is constructed by combining the DLFT spectrum with a negatively weighted shifted DLFT spectrum. The weight for the negative lobes was determined empirically to be 0.35.

3. Two correlation functions

The normalized "template-frame" correlation function provides an estimation for $\log f_0$ by correlating the speech DLFT spectrum with the harmonic template, as shown in:

$$R_{TX_t}(n) = \frac{\Sigma_i T(i) X_t(i-n)}{\sqrt{\Sigma_i X_t(i)^2}} \quad (N_L < n < N_H). \tag{7}$$

The template, $T(i)$, is normalized to have unit energy in advance so the correlation is normalized by the signal energy only. The f_0 search range $[F_{\min}, F_{\max}]$ determines the bounds for the correlation, $[N_L, N_H]$.

The mapping between pitch candidate P and the corresponding index in the template-frame correlation function can be derived from Eq. (2). Assuming the index of the trial pitch P in the signal DLFT spectrum is i_P, according to Eq. (2):

$$\omega_{i_P} = 2\pi \cdot P \cdot T_s = 2\pi e^{(\log f_s + i_P \cdot d \log f)} \cdot T_s, \tag{8}$$

where f_s is the low-frequency bound for the signal DLFT spectrum and $d \log f$ is the logarithmic frequency resolution. The relationship of P and i_P can be further simplified as:

$$\log P = \log f_s + i_P \cdot d \log f, \tag{9}$$

$$i_P = (\log P - \log f_s)/d \log f. \tag{10}$$

Similarly, the index of the fundamental frequency (P_T) in the template, i_{P_T}, can be determined as:

$$i_{P_T} = (\log P_T - \log f_s')/d \log f \tag{11}$$

where f_s' is the low-frequency bound for the template.

The relative shift in the template-frame correlation to align the two harmonic structures is simply the difference of these two indices:

$$I_P = i_{P_T} - i_P = (\log P_T - \log f_s' - \log P + \log f_s)/d \log f. \tag{12}$$

Conversely, P can also be determined from the correlation lag I_P by:

$$P = \frac{P_T \cdot f_s}{f_s' \cdot e^{I_P \cdot d \log f}}. \tag{13}$$

By substituting P into Eq. (12) with the pitch range $[F_{\min}, F_{\max}]$, the bounds for template-frame correlation are obtained as:

$$N_L = (\log P_T - \log f_s' - \log F_{\max} + \log f_s)/d \log f, \tag{14}$$

$$N_H = (\log P_T - \log f_s' - \log F_{\min} + \log f_s)/d \log f. \tag{15}$$

By aligning two adjacent frames of the signal DLFT spectra, the normalized "cross-frame" correlation function provides constraints for $\Delta \log f_0$, as shown in:

$$R_{X_t X_{t-1}}(n) = \frac{\Sigma_i X_t(i) X_{t-1}(i-n)}{\sqrt{\Sigma_i X_t(i)^2}\sqrt{\Sigma_i X_{t-1}(i)^2}} \quad (|n| < N). \tag{16}$$

The correlation is normalized by the energy of both signal frames. Since f_0 should not change dramatically across two frames, the correlation bound N is set to be around 10% of the number of samples in the DLFT spectrum. A robust estimation of the $\log f_0$ difference across two voiced frames is given by the maximum of the correlation. See Fig. 2 for

examples of the template-frame and cross-frame correlation functions of a speech signal.

4. DP search

The advantage of using DP in pitch tracking is to incorporate continuity constraints across adjacent frames to reduce pitch doubling and halving errors (Secrest and Doddington, 1983; Talkin, 1995; Geoffrois, 1996; Droppo and Acero, 1998). This is typically achieved by incorporating a transition cost in the DP score function to penalize large changes in neighboring f_0 hypotheses. In our implementation, the transition cost is defined by the cross-frame correlation function [Eq. (16)]. It goes beyond enforcing continuity: it provides an estimation of the actual change in $\log f_0$. Given the score functions of $\log f_0$ and $\Delta \log f_0$, the target function S of our DP search is defined in an iterative manner as:

$$S(t,i) = \begin{cases} R_{TX_0}(i) & (t=0), \\ \max_j\{S(t-1,j)\cdot R_{X_tX_{t-1}}(i-j)\} + R_{TX_t}(i) & (t>0), \end{cases} \quad (17)$$

where i and j are the indices in the template-frame correlation function. The pitch value P_i can be converted from the index i by Eq. (13). We compute a score for each pitch candidate at $t=0$ according to Eq. (7). For each subsequent time point, we compute the scores iteratively using the scores from the previous frame. The pointer to the best past node is saved for backtracking upon reaching the last frame. Due to the logarithmic sampling of the DLFT, the search space for pitch values is naturally quantized logarithmically with constant $\Delta f_0/f_0$. Despite the first harmonic of the spectrum being fairly weak, the DP search is able to track f_0 whenever there is clear harmonic structure.

C. Adaptations of algorithm for killer whale vocalizations

This paper focuses on tracking the pulsed calls recorded from several Norwegian killer whale groups, each of which produces 3–16 call types (Strager, 1993, 1995; Van Opzeeland *et al.*, 2005). The LFCs of these calls can be characterized by a variety of f_0 patterns, including gradual or abrupt upsweeps and downsweeps, relatively constant frequencies, and abrupt transitions between these constant frequencies. The HFC is typically characterized by a gradual upsweep, though downsweeps and constant frequencies are also observed. In general, the frequency modulation of the LFC is more variable than that of the HFC. Figure 1 displays the spectrograms of some examples of these call type patterns, including simultaneous LFC and HFC [Figs. 1(a), 1(c), and 1(f)], abrupt frequency transitions [Figs. 1(a), 1(c), and 1(d)], and very dynamic f_0 range [as low as 230 Hz in Fig. 1(e) and as high as 11.25 kHz in Fig. 1(f)]. Notice that the low-frequency spectral energy of the killer whale calls is often masked by ambient boat noise.

Given that killer whale calls have distinctive f_0 dynamics for different call types, it was unrealistic to expect that a single set of parameters would work well for all call types. More importantly, some call types contain both LFCs and HFCs, which clearly could not be tracked with one set of parameter settings. To solve this multi-pitch problem, three sets of parameters aimed at tracking f_0 in three frequency ranges for different types of killer whale calls were identified. The first setting is used to track LFCs that have an f_0 below 600 Hz [e.g., Fig. 1(e)]. The second setting aims to track LFCs between 400 and 4000 Hz. These include calls that have a rising or falling f_0 in that frequency range [e.g., Figs. 1(a) and 1(b)], as well as calls with a relatively flat f_0 contour that can contain abrupt changes [e.g., the LFC in Figs. 1(c) and 1(d)]. The third setting is used to track the HFCs, which typically range between 4 and 12 kHz [e.g., Figs. 1(a), 1(c), and 1(f)].

For killer whale recordings, the optimal set of parameters can depend on the dynamics of the f_0 contour (e.g., slow vs abrupt changes). Parameterization issues can be alleviated if the harmonic matching principle is incorporated and brute force approach to estimating the pitch is adopted. Recall that the frame-based pitch estimation is obtained by shifting the harmonic template linearly to find the correlation maximum with the signal's DLFT spectrum. Correlation of finite-length sequences tends to taper off as the relative shift between these sequences increases. With normalized cross-correlation (i.e., the correlation is normalized by the energy in the overlapped region), the harmonic template parameters are important in balancing the bias between shifting left and shifting right. The problem would be resolved if, instead of shifting the harmonic template, the f_0 of the pulse train is changed and its DLFT is recomputed to obtain a new harmonic template. In this way, the correlation is always computed on the same (full) length of the signal and pulse train's DLFT spectra. The drawback of this approach is that M DLFT spectra must be computed and stored as harmonic templates, where M is the number of pitch hypotheses in the search space and could be large to achieve a refined resolution. Given that pitch tracking for killer whale recordings is typically not done in real time, currently the added computation requirement is not likely to be a serious issue. Again, the three sets of parameters used here are summarized in Table I. The parameters were selected to optimize performance of the algorithm on a training set.

D. Data collection

Free-ranging killer whales were tagged with digital archival tags containing acoustic and movement sensors (Johnson and Tyack, 2003) in Tysfjord, Norway in November 2005. These tags sampled sound at 96 kHz and were recovered upon their scheduled release for data offload. Eight orcas were tagged in all, but only the vocalizations recorded by six of these tags were evaluated for performance here. These six tags recorded for 20.6 h in total. Manual auditing documented the times and durations of calls that were clearly audible. Clearly audible calls were excerpted from the recordings, down-sampled to 48 kHz to match the human data on which the algorithm had been trained, and saved as separate files. If the call type had been observed before, it was labeled according to its earlier designation

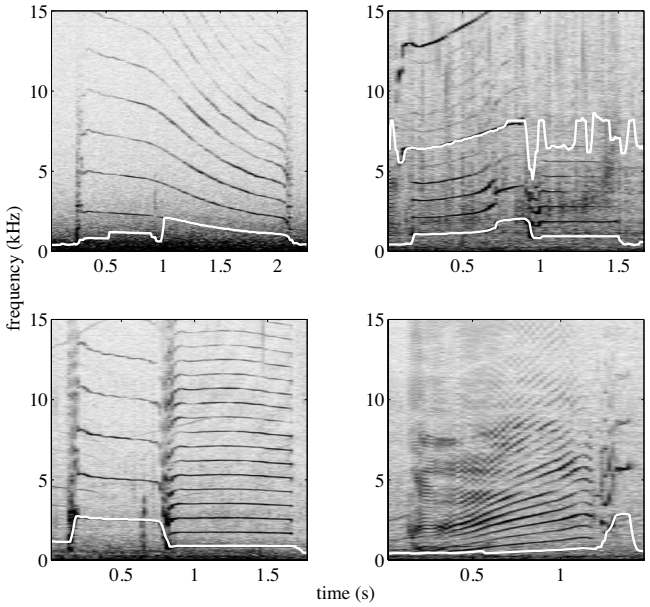

FIG. 5. Spectrograms of killer whale calls overlaid with pitch tracks computed by the DLFT-PDA. All calls contain a LFC but only the call in the upper right contains a HFC.

(Strager, 1993, 1995). Otherwise, new assignments were made (Shapiro, 2008). The DLFT-PDA was run on each of these files with the three parameter settings described previously, after which the traces were manually corrected to provide references for subsequent evaluation (see Fig. 5 for sample pre-corrected traces). During this post-processing stage, each call was also marked with beginning and ending times and labeled by its type. Each call took between 1 and 60 s to correct depending on the accuracy of the automatic trace.

E. Evaluation of performance

The algorithms for tracking killer whale calls (see below) were compared using gross error rate (GER) and fine error (FE) metrics. Specifically, GER is the percentage of f_0 hypotheses that deviate from the reference value (i.e., the manually-determined value) by more than 20%. The FE is characterized by the mean absolute value of the percentage deviation from the reference, excluding frames deviating from the reference by more than 20%. The f_0 of killer whale calls spans a large dynamic range, and so the pitch hypotheses were normalized by the reference values in assessing the FE. Because there are two outputs for the LFC (due to two parameter settings), the more accurate one (i.e., the contour with smaller GER) was chosen for evaluation, similar to the strategy a human labeler would use for post-editing.

To generate a profile of performance on different SNRs, the SNR for all non-overlapping calls was calculated. The SNR was calculated using a segment of background noise of the same duration (without vocalizing or surfacing sounds) occurring as close to the signal as possible within about 30 s. Before the SNR was computed, the recording was band-pass filtered with a two-pole Butterworth filter using the cutoff ranges of $[400,3000]$ Hz and $[4,12]$ kHz for the LFCs and HFCs, respectively. The SNR was measured in terms of energy flux density (Madsen, 2005).

The performance of DLFT-PDA was compared with the peak-picking and sidewinder algorithms (Deecke *et al.*, 1999) for tracking killer whale calls. The peak-picking method employed here selects the maximum frequency value of each power spectrum slice of the spectrogram followed by a smoothing step to remove any outliers. The sidewinder algorithm was implemented in both manners mentioned previously (i.e., the peak of the real cepstrum of the autocovariance sequence, and the second highest peak of the autocovariance sequence itself), and the better output of the two methods was selected for each call in the evaluation. In both the peak-picking and sidewinder algorithms, a restricted frequency range was considered for the LFC ($[400,3000]$ Hz) and the HFC ($[4,12]$ kHz). Before being implemented, these two alternative methods were checked against several call type exemplars to ensure their accuracy. In addition, the code used to deploy the sidewinder algorithm by Deecke *et al.* (1999) was kindly furnished by Deecke for this analysis.

TABLE II. Performance of DLFT-PDA, peak-picking, and sidewinder algorithms on the LFCs and HFCs of killer whale calls with different SNRs. The average SNR and the number of contours evaluated in each condition are also provided in the table. GER is the 20% gross error rate. FE is the mean of the normalized absolute fine error. See text for details.

			DLFT-PDA		Peak-picking		Sidewinder	
SNR range (dB)	N	Mean SNR (dB)	GER (%)	FE (%)	GER (%)	FE (%)	GER (%)	FE (%)
			LFC					
0–10	1180	4.7	29.0	1.5	64.3	4.9	49.8	6.7
10–20	647	14.4	16.8	0.9	51.7	3.6	37.8	4.3
>20	268	25.2	12.8	0.8	34.8	3.3	21.1	3.0
All	2095	10.3	22.5	1.2	55.9	4.2	41.7	5.3
			HFC					
0–10	182	6.6	24.3	5.6	27.1	2.0	25.1	9.5
10–20	346	15.2	13.3	2.7	28.2	1.9	22.9	8.4
>20	318	26.3	7.6	2.4	38.2	2.5	28.9	9.3
All	846	17.5	13.6	3.2	31.7	2.2	25.5	9.0

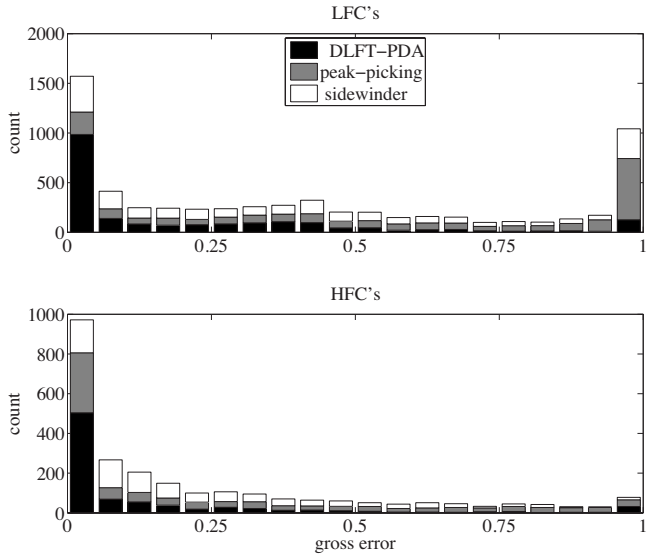

FIG. 6. Histogram of GERs for LFCs (top) and HFCs (bottom).

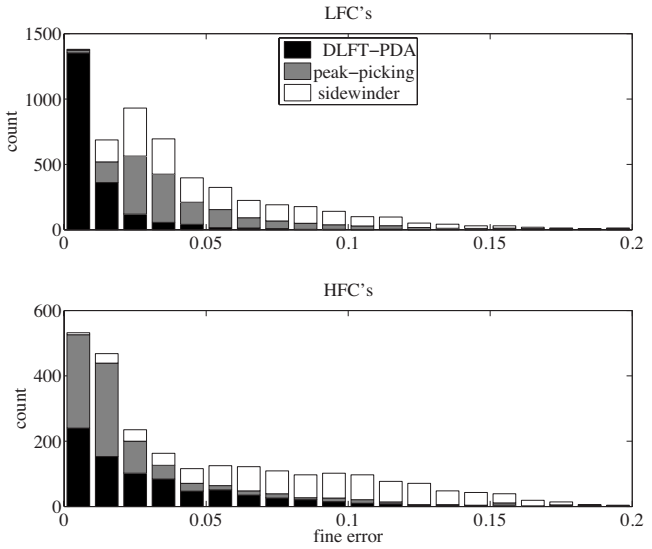

FIG. 7. Histogram of FE for LFCs (top) and HFCs (bottom).

III. RESULTS

The results from this comparison are reported in Table II for the LFCs and HFCs for three different SNR ranges: below 10 dB, between 10 and 20 dB, and above 20 dB. The DLFT-PDA performed better than the peak-picking and sidewinder algorithms on the LFCs in terms of both gross and FEs for all SNR ranges. The time required for manual post-editing will therefore be substantially reduced using the DLFT-PDA for initial contour tracing. The DLFT-PDA also outperformed the other two algorithms on the HFCs for all SNR ranges, except that the peak-picking algorithm had better FE results. The peak-picking algorithm performed much worse on the LFCs than on the HFCs. This is likely because the spectrogram of the HFCs normally has only one harmonic peak in the pitch search range for the HFC, in contrast with the spectrogram of LFCs that is characterized by multiple harmonic peaks in the search range (see Fig. 1). Consequently, it was easier for the peak-picking algorithm to locate f_0 in the HFC, except for the case when the LFCs had stronger harmonic peaks than the HFC in the search range. A potential refinement to our method is to use peak-picking to fine-tune the contour traced by DLFT-PDA prior to manual correction.

To provide a detailed view of how well the DLFT-PDA performs on individual calls, a histrogram of the GER for calls is plotted in Fig. 6. The algorithm is able to trace many of the contours accurately: 46.9% and 59.6% of all extracted LFCs and HFCs, respectively, were characterized by a GER of 5% or less. The percentage of LFCs that the DLFT-PDA seriously mistraced was small. For example, 9.1% of LFCs had a GER of over 70%, in contrast to 45.2% for the peak-picking algorithm and 23.9% for the sidewinder algorithm. A histogram of the FE for calls is plotted in Fig. 7.

IV. DISCUSSION AND CONCLUSIONS

In this article, a pitch tracking algorithm originally designed for human telephone speech was modified for killer whale vocalizations. The algorithm derives reliable estimations of pitch and the temporal change of pitch from the entire harmonic structure. The correlation of the DLFT spectrum with a carefully constructed harmonic template provides a robust estimation of the f_0, especially for low frequencies. The correlation of two DLFT spectra from adjacent frames gives a very reliable estimation of the f_0 change. The estimations of both f_0 and the temporal change in f_0 are then combined in a DP search to find a smooth pitch track. Evaluation results have demonstrated that the DLFT-PDA is capable of tracking killer whale calls that contain simultaneous LFC and HFC, and compares favorably to several other algorithms available for tracking killer whale vocalizations.

The challenges to DLFT-PDA plague all current pitch tracking algorithms. For example, it performs more poorly at low SNRs. In addition, it is unable to track multiple calls produced by multiple animals simultaneously. Currently, manual extraction and tracing of each call is necessary. To perform this task automatically, triangulation of the position of the callers using multiple recorders would be required at a minimum.

For the peak-picking and sidewinder algorithms, some GER and FE measurements that were associated with the HFC actually increased with a higher SNR (Table II). This might be explained by the relative energies of the LFC and HFC. Ambient boat noise occupied the lower frequency range. For calls with higher SNRs, more LFC energy would be visible to the tracker, making the HFC less obvious. In this situation, the peak-picking and sidewinder algorithms might have locked onto an upper harmonic of the LFC instead of selecting the fundamental frequency of the HFC, which would account for the poorer performance associated with increasing SNRs. This highlights one of the difficulties of tracking vocalizations with multiple simultaneous frequency components.

In addition to serving human telephone speech, the DLFT-PDA provides a reliable and unbiased approach toward determining the fundamental frequency of the pulsed calls of killer whales (see Nousek *et al.*, 2006; Miller *et al.*, 2007 for its successful application). A strongly performing

pitch tracker can lead to more objective results, which can then be used to characterize and classify vocalizations dependably. Future work should apply such a tracker to the calls of other species to automate, accelerate, and standardize the process. More generally, this manuscript highlights the benefit of introducing techniques developed for analyzing human speech into the realm of marine mammal vocalizations. The field of speech recognition has negotiated numerous challenges in signal processing. Rather than replicating these efforts, research on marine mammal acoustics will be well served by incorporating such advances in human speech processing.

ACKNOWLEDGMENTS

The authors would like to thank Stephanie Seneff and Peter L. Tyack for their support and advice, Ghinwa F. Choueiter for helpful comments on multiple drafts of the manuscript, Volker B. Deecke for providing implementations of the peak-picking and sidewinder algorithms, and two anonymous reviewers who improved the flow and transparency of this document. Special thanks to Sara Kim, Gary Matthias, Rebecca McGowan, Maitagorri Schade, and Ivan Dimitrov for their assistance with the post-processing corrections of the contours. We are grateful to Tiu Similä, Mads Christoffersen, Geoff Magee, Patrick Miller, Petter Helgevold-Kvadsheim, Sanna Kuningas, Sari Oksanen and Filipa Samarra for assistance in the field collecting the killer whale data. Financial support for collection of the Norwegian killer whale vocal data was provided by the Ocean Life Institute of the Woods Hole Oceanographic Institution and the National Geographic Society. The MIT Undergraduate Research Opportunities Program (UROP) provided support for some of the post-editing work. C.W. was supported by DARPA under Contract No. N66001-96-C-8526, monitored through Naval Command, Control, and Ocean Surveillance Center and by the National Science Foundation under Grant No. IRI-9618731. A.D.S. was supported by a National Defense Science and Engineering Graduate Fellowship.

Brown, J. C., Hodgins-Davis, A., and Miller, P. J. O. (**2006**). "Classification of vocalizations of killer whales using dynamic time warping," J. Acoust. Soc. Am. **119**, EL34–EL40.

Buck, J. R., and Tyack, P. L. (**1993**). "A quantitative measure of similarity for *Tursiops truncatus* signature whistles," J. Acoust. Soc. Am. **94**, 2497–2506.

Deecke, V. B., Ford, J. K. B., and Spong, P. (**1999**). "Quantifying complex patterns of bioacoustic variation: Use of a neural network to compare killer whale (*Orcinus orca*) dialects," J. Acoust. Soc. Am. **105**, 2499–2507.

Droppo, J., and Acero, A. (**1998**). "Maximum *a posteriori* pitch tracking," in Proceedings of ICSLP, Sydney, Australia, pp. 943–946.

Ford, J. K. B. (**1987**). "A catalogue of underwater calls produced by killer whales (*Orcinus orca*) in British Columbia," Canadian Data Report of Fisheries and Aquatic Sciences **633**, 1–165.

Ford, J. K. B. (**1989**). "Acoustic behavior of resident killer whales (*Orcinus orca*) off Vancouver Island, British Columbia," Can. J. Zool. **67**, 727–745.

Ford, J. K. B. (**1991**). "Vocal traditions among resident killer whales (*Orcinus orca*) in coastal waters of British Columbia," Can. J. Zool. **69**, 1454–1483.

Geoffrois, E. (**1996**). "The multi-lag-window method for robust extended-range f_0 determination," in Proceedings of ICSLP, Philadelphia, PA, pp. 2399–2402.

Hermes, D. J. (**1988**). "Measurement of pitch by subharmonic summation," J. Acoust. Soc. Am. **83**, 257–264.

Hess, W. (**1983**). *Pitch Determination of Speech Signals* (Springer-Verlag, Berlin, Germany).

Hoelzel, A. R., and Osborne, R. W. (**1986**). "Killer whale call characteristics: Implications for cooperative foraging strategies," in *Behavioral Biology of Killer Whales*, edited by B. C. Kirkevold, and J. S. Lockard (Alan R. Liss, Inc., New York), pp. 373–403.

Janik, V. M., Dehnhardt, G., and Todt, D. (**1994**). "Signature whistle variations in a bottlenosed dolphin, *Tursiops truncatus*," Behav. Ecol. Sociobiol. **35**, 243–248.

Johnson, M. P., and Tyack, P. L. (**2003**). "A digital acoustic recording tag for measuring the response of wild marine mammals to sound," Inf. Sci. (N.Y.) **28**, 3–12.

Klapuri, A. (**2008**). "Multipitch analysis of polyphonic music and speech signals using an auditory model," IEEE Trans. Audio, Speech, Lang. Process. **16**, 255–266.

Klapuri, A., and Virtanen, T. (**2008**). "Progress towards automatic music transcription," in *Handbook of Signal Processing in Acoustics*, edited by D. Havelock, S. Kuwano, and M. Vorlander (Springer-Verlag, Berlin).

Lahat, M., Niederjohn, R. J., and Krubsack, D. A. (**1987**). "A spectral autocorrelation method for measurement of the fundamental frequency of noise-corrupted speech," IEEE Trans. Acoust., Speech, Signal Process. **35**, 741–750.

Madsen, P. T. (**2005**). "Marine mammals and noise: Problems with root mean square sound pressure levels for transients," J. Acoust. Soc. Am. **117**, 3952–3957.

McCowan, B. (**1995**). "A new quantitative technique for categorizing whistles using simulated signals and whistles from captive bottlenose dolphins (Delphinidae, *Tursiops truncatus*)," Ethology **100**, 177–193.

Miller, P. J. O. (**2002**). "Mixed-directionality of killer whale stereotyped calls: A direction of movement cue?," Behav. Ecol. Sociobiol. **52**, 262–270.

Miller, P. J. O., and Bain, D. E. (**2000**). "Within-pod variation in the sound production of a pod of killer whales, *Orcinus orca*," Anim. Behav. **60**, 617–628.

Miller, P. J. O., Samarra, F. I. P., and Perthuison, A. D. (**2007**). "Caller sex and orientation influence spectral characteristics of "two-voice" stereotyped calls produced by free-ranging killer whales," J. Acoust. Soc. Am. **121**, 3932–3937.

Miller, P. J. O., Shapiro, A. D., Tyack, P. L., and Solow, A. R. (**2004**). "Call-type matching in vocal exchanges of free-ranging resident killer whales, *Orcinus orca*," Anim. Behav. **67**, 1099–1107.

Nousek, A. E., Slater, P. J. B., Wang, C., and Miller, P. J. O. (**2006**). "The influence of social affiliation on individual vocal signatures of northern resident killer whales (*Orcinus orca*)," Biol. Lett. **2**, 481–484.

Secrest, B. G., and Doddington, G. R. (**1983**). "An integrated pitch tracking algorithm for speech synthesis," in Proceedings of ICASSP, Boston, MA, pp. 1352–1355.

Shapiro, A. D. (**2006**). "Preliminary evidence for signature vocalizations among free-ranging narwhals (*Monodon monoceros*)," J. Acoust. Soc. Am. **120**, 1695–1705.

Shapiro, A. D. (**2008**). "Orchestration: The movement and vocal behavior of free-ranging Norwegian killer whales (*Orcinus orca*)," in *Biological Oceanography*, Ph.D. thesis (MIT/WHOI).

Strager, H. (**1993**). "Catalogue of underwater calls from killer whales (*Orcinus orca*) in northern Norway," (University of Århus, Århus, Denmark).

Strager, H. (**1995**). "Pod-specific call repertoires and compound calls of killer whales, *Orcinus orca* Linnaeus, 1758, in the waters of northern Norway," Can. J. Zool. **73**, 1037–1047.

Talkin, D. (**1995**). "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, edited by W. B. Kleijn and K. K. Paliwal (Elsevier, New York), pp. 495–518.

van Opzeeland, I. C., Corkeron, P. J., Leyssen, T., Similä, T., and Van Parijs, S. M. (**2005**). "Acoustic behaviour of Norwegian killer whales, *Orcinus orca*, during carousel and seiner foraging on spring-spawning herring," Aquat. Mamm. **31**, 110–119.

Wang, C. (**2001**). "Prosodic modeling for improved speech recognition and understanding," Ph.D. thesis (MIT, Cambridge, MA).

Wang, C., and Seneff, S. (**2000a**). "Improved tone recognition by normalizing for coarticulation and intonation effects," in Proceedings of the International Conference on Spoken Language Processing, Beijing, China.

Wang, C., and Seneff, S. (**2000b**). "Robust pitch tracking for prosodic modeling in telephone speech," in Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, pp. 1143–1146.

Wang, C., and Seneff, S. (**2001a**). "Lexical stress modeling for improved speech recognition of spontaneous telephone speech in the JUPITER domain," in Proceedings of EUROSPEECH, Aalborg, Denmark.
Wang, C., and Seneff, S. (**2001b**). "Prosodic scoring of recognition outputs in the JUPITER domain," in Proceedings of International Speech Communication Association Workshop on Prosody in Speech Recognition and Understanding, Red Bank, NJ.
Watkins, W. A. (**1967**). "The harmonic interval: Fact or artifact in spectral analysis of pulse trains," in *Marine Bioacoustics*, edited by W. N. Tavolga (Pergamon, New York), pp. 15–43.
Watwood, S. L., Tyack, P. L., and Wells, R. S. (**2004**). "Whistle sharing in paired male bottlenose dolphins, Tursiops truncatus," Behav. Ecol. Sociobiol. **55**, 531–543.
Watwood, S. L., Owen, E. C. G., Tyack, P. L., and Wells, R. S. (**2005**). "Signature whistle use by temporarily restrained and free-swimming bottlenose dolphins, *Tursiops truncatus*," Anim. Behav. **69**, 1373–1386.