Automating identification of avian vocalizations using time–frequency information extracted from the Gabor transform

Edward F. Connor[a)]
Department of Biology, San Francisco State University, 1600 Holloway Avenue, San Francisco, California 94132

Shidong Li and Steven Li
Department of Mathematics, San Francisco State University, 1600 Holloway Avenue, San Francisco, California 94132

(Received 18 June 2011; revised 3 April 2012; accepted 2 May 2012)

Based on the Gabor transform, a metric is developed and applied to automatically identify bird species from a sample of 568 digital recordings of songs/calls from 67 species of birds. The Gabor frequency–amplitude spectrum and the Gabor time-amplitude profile are proposed as a means to characterize the frequency and time patterns of a bird song. An approach based on template matching where unknown song clips are compared to a library of known song clips is used. After adding noise to simulate the background environment and using an adaptive high-pass filter to de-noise the recordings, the successful identification rate exceeded 93% even at signal-to-noise ratios as low as 5 dB. Bird species whose songs/calls were dominated by low frequencies were more difficult to identify than species whose songs were dominated by higher frequencies. The results suggest that automated identification may be practical if comprehensive libraries of recordings that encompass the vocal variation within species can be assembled.
 [http://dx.doi.org/10.1121/1.4726006]

PACS number(s): 43.80.Ka, 43.80.Jz, 43.72.Ne, 43.72.Lc [JJF] Pages: 507–517

I. INTRODUCTION

In an effort to improve the efficiency, accuracy, and repeatability of point counts designed to monitor and assess breeding bird communities, we develop and test an approach to automatically identify birds from digital recordings of their vocalizations. Breeding birds are commonly monitored using their vocalizations to obtain their identities and count their numbers. Currently, human observers stand at a location for between 3 and 10 min and listen to the dawn chorus of singing birds while writing down which, and how many of, each species are heard or seen producing a so-called point count (Ralph *et al*., 1995). In forested habitats most detections are based on vocalizations, while in more open habitats visual detections are more common. The weak link in point count methodology is the human observer. The human observer is expected to be capable of recognizing numerous bird species from their vocalizations, and to estimate the direction and distance to each singing bird. This may be an impossible task for humans to do well even though the number of species singing at a site during a 10 min period is usually no more than a dozen and the number of individuals is usually no more than double that. Recent experiments by Alldredge *et al*. (2007a,b) and Simons *et al*. (2007) using song recordings played back on loudspeakers to simulate breeding bird communities show that expert human observers have much difficulty detecting avian vocalizations under ideal conditions. Even when observers knew which species to expect, the best of the seven human observers in their study only detected 65% of the birds for the most detectable species singing at a high rate and the worst only 19% of the birds for the least detectable species singing at a low rate.

Variation in abilities of human observers to accurately recognize bird species from their calls is substantial. This may arise either because of variation in the amount of training and experience or because of differences in auditory acuity among observers (Mayfield, 1966; Emlen and DeJong, 1992; Alldredge *et al*., 2007a). Not all observers are equally good at recognizing all species, which will lead to substantial inter-observer variability in point count estimates of species richness and abundance (Bart and Schoultz, 1984; Bart 1985).

Several studies have explored using digital field recordings of birds combined with transcription of the recordings by human observers in a lab setting (Haselmayer and Quinn, 2000; Hobson *et al*., 2002; Cunningham *et al*., 2004; Lindenmayer *et al*., 2004; Rempel *et al*., 2005; Acevedo and Villanueva-Rivera, 2006; Cellis-Murillo *et al*., 2009). Besides providing an archival copy of the bird songs/calls encountered, recordings provide the opportunity to replay sections to verify species detections. Replay could be particularly useful when multiple species are singing simultaneously or in rapid succession and might be missed by an observer trying to write down observations in the field. Recordings also provide the opportunity for review of the identifications by other experts to ensure accuracy of the transcription.

a)Author to whom correspondence should be addressed. Electronic mail: efc@sfsu.edu

Few studies have attempted to automatically identify bird species or count their numbers from recordings (Brandes, 2008a). An even smaller number of studies have attempted to use automatic identification based on recordings of the sounds produced by other animal groups (Riede, 1998; Horne, 2000; Chesmore and Ohya, 2004; Brandes *et al.*, 2006; Brandes, 2008b; Jennings *et al.*, 2008). Fully automated processing of field recordings involves two main steps: (1) Segmenting the recording into regions with bird vocalizations and those without, and (2) identifying the species recorded in each segment with bird vocalizations. Some studies perform manual segmentation and automatic identification (McIlraith and Card, 1997; Terry and McGregor, 2002). Others attempt to automatically segment and identify vocalizations (Anderson *et al.*, 1996; Kogan and Margoliash, 1998; Härmä and Somervuo, 2004; Kwan *et al.*, 2004; Fagerlund and Härmä, 2005; Lee *et al.*, 2006; Chen and Maher, 2006; Selin *et al.*, 2006; Fagerlund, 2007; Ranjard and Ross, 2008; Cai *et al.*, 2007; Swiston and Mennill, 2009). Most of these approaches focus on algorithms for identifying bird song or syllable clips (arbitrarily defined vocal segments), rather than the problem of segmenting field recordings.

A wide variety of identification algorithms have been proposed in the literature, although they may be grouped into three main approaches. A time–frequency decomposition of the recorded signal using the short-time Fourier transform (STFT) is the first step in the analysis. From this time–frequency decomposition, "features" are extracted from songs or syllables from known species to be compared to features from unidentified song clips. These features might include the highest and lowest frequency, song length, peak frequency, spectral coefficients, wavelet coefficients, mel frequency cepstral coefficients, or others (McIlraith and Card, 1997; Kwan *et al.*, 2004; Fagerlund and Härmä, 2005; Lee *et al.*, 2006; Chen and Maher, 2006; Selin *et al.*, 2006, 2007; Fagerlund, 2007; Brandes, 2008a,b). Features are extracted from a library of song or syllable clips and this library of features is used to develop classification functions using discriminant analysis, Bayesian classifiers, support vector machines, or neural networks (McIlraith and Card, 1997; Terry and McGregor, 2002; Lee *et al.*, 2006; Fagerlund, 2007; Ranjard and Ross, 2008). Features extracted from unknown clips are then classified using the previously developed classifiers. The second approach is template matching in which example sounds from a library of known species are compared to sounds from unknown clips and classified as the species it matches most closely. Both direct nearest neighbor matches and matches achieved using dynamic time warping (e.g., stretching or compressing an unknown song to determine how much modification is needed to achieve a match) are used in template matching (Anderson *et al.*, 1996; Härmä, 2003; Fagerlund and Härmä, 2005). Finally, stochastic sequence modeling has been used to examine short-time sound features and how they change through time using hidden Markov models (Kogan and Margoliash, 1998; Trifa *et al.*, 2008).

Tests of these identification algorithms have used a small number of species (no more than 14 species), although they use many syllable clips. No study has attempted to test an automated identification algorithm on a community of species as large as one might expect to encounter under field conditions. Studies with fewer species report higher rates of successful identifications with accuracies averaging above 90% per species (eight or fewer species: Anderson *et al.*, 1996; McIlraith and Card, 1997; Kogan and Margoliash, 1998; Tanttu *et al.*, 2003; Kwan *et al.*, 2004; Lee *et al.*, 2006; Chen and Maher, 2006; Selin *et al.*, 2006; Fagerlund, 2007; Selin *et al.*, 2007; Trifa *et al.*, 2008), while studies with more species report success rates with less than 50% of clips correctly identified to species (>12 species: Härmä, 2003; Somervuo *et al.*, 2006). Only a few studies have used whole songs for automatic identification (Kwan *et al.*, 2004; Somervuo *et al.*, 2006).

We propose a method for identifying bird songs from digital recordings based on template matching using frequency–amplitude spectra and time–amplitude profiling from a time–frequency decomposition using the Gabor transform (Gabor, 1946; Emlen and De Jong, 1992). We examine the performance of our approach for a large set of species and a range of levels of environmental noise.

II. METHODS

A. Overview of our approach

Our overall goal is to develop methods for automating identification of birds from field recordings of their songs/calls. To this end, we develop and test a metric based on the Gabor time–frequency decomposition of such signals. We outline the development and rationale for the proposed metric, and test the accuracy of this metric for a large set of bird songs/calls from 67 different species of birds, and a total of 568 song/calls. The performance of the proposed metric for song/calls that are largely free of noise, after adding environmental noise at specified signal-to-noise ratios [(SNRs) expressed in dB], and after applying de-noising techniques is examined.

B. The Gabor transform

The Gabor transform provides a time–frequency representation of a signal similar to a windowed Fourier transform or the STFT. If we let $g(t) \in H$, a general Hilbert space, be a Gaussian-type window function, a sequence of functions,

$$\left\{ g_{n,k}(t) \equiv g(t - kT)e^{-2\pi int/N} \right\}_{k=0,\,n=0}^{(L/T)-1,\,N-1}$$

is called a Gabor sequence with t indexing time. Here, for practical applications, we have assumed that t is a finite discrete time index, say $0 \leq t \leq L-1$ with L being the length of the signal.

The Gabor transform for all $f \in H$ is defined as

$$G_f(n,\,k) = (Gf)(n,\,k) = \{\langle f,\, g_{n,k}\rangle\}_{n,k},$$

provided $\{g_{n,k}\}$ forms a mathematical *frame* in H.

A mathematical frame in a Hilbert space H is a basis-like sequence $\{x_n\} \subseteq H$. We say $\{x_n\}$ is a frame of H if there are constants $0 < A \leq B < \infty$ such that

$$\forall f \in H, \quad A \| f \|^2 \leq \sum_n |\langle f, x_n \rangle|^2 \leq B \| f \|^2.$$

Frames differ from bases in that they are overcomplete (complete but generally redundant). By complete, we mean that the Gabor time–frequency expansion preserves all time–frequency information of a signal. In addition, the degree of redundancy of a Gabor expansion can be controlled by the choice of parameters T and N. Generally speaking, as long as $N > T$, a Gabor sequence $\{g_{n,k}\}$ would form a complete mathematical frame for most window functions with the property that

$$0 < a \leq \sum_k |g(t - kT)|^2 \leq b < \infty.$$

The concept of a frame is relevant because the STFT is also a frame expansion. The corresponding STFT frame sequence is formed by translations and complex modulations of a window function, very much like Gabor sequences. Yet, the difference between the Gabor frame expansion and the STFT is that while both are complete, the Gabor frame is not as redundant as the STFT. Indeed, if the Gabor parameters are $T = 1$ and $N = L$, a Gabor transform becomes a STFT. However, since the Gabor parameters can be $T \gg 1$ and $N \ll L$, the Gabor frame expansion can be much less redundant. More specifically, the amount of data in a discrete Gabor transform is given by $(L/T) \times N$, where typically $T \gg 1$ and $N \ll L$. In the case of the STFT, this amounts to an $L \times L$ matrix.

While most studies on automatic identification of bird song use the traditional spectrogram or other information extracted from the STFT, this may be cumbersome. The time–frequency information in the spectrogram is overly redundant, computationally burdensome, and demands large amounts of data storage. For a typical bird song, if window g is selected to have the support of about 1000 non-zero indices with parameters set at $T = 400$ and $N = 1000$, and a clip length of $L = 60\,000$, the spectrogram would produce a matrix of 3.6×10^9 entries. However, for the Gabor transform the matrix would have $60\,000/400 = 1.5 \times 10^5$ entries, a considerable saving in data storage needs.

Given the large size of the STFT for a single song, data storage for a reasonably sized library of bird songs/calls would be challenging. As such, automatic identification algorithms based on such a large time–frequency feature may not be efficient or even practical. However, as a mathematical frame expansion, the Gabor transform can result in a substantial reduction in redundancy over that of the conventional STFT. For applications to automatic identification of avian vocalizations where one would need a sizable library comprising many bird species and the range of vocal variation within each species, reduced redundancy would be a great advantage since it would require less calculation and data storage. The saving in data storage is not without cost. While there is no loss of information using the Gabor transform with parameter choices such as those suggested earlier, the Gabor transform will inevitably have reduced time–frequency resolution compared to that of a spectrogram. The pertinent question is whether for reasonable values of T and N, the information in the Gabor time–frequency decomposition is sufficient for identification of bird songs/calls.

1. Gabor frequency–amplitude spectrum (GFAS)

In addition to differences in the time–frequency patterns between species, time–frequency intensities also differ dramatically among species. Such differences in time–frequency intensities are important because they suggest that we may further simplify bird song/call identification based on the Gabor transform. We define a GFAS as a time-cumulative result of the Gabor transform. This is motivated by the observation of striking differences in time–frequency intensities that result in distinct patterns among species in a time-cumulative GFAS.

Given a bird song/call f, let $G_f(n, k)$ be the Gabor transform of f, where the indices n and k are the frequency and time variables, respectively. Our formulation presupposes that the signal f has been either manually or automatically segmented from some longer continuous recording. The GFAS of f is calculated by summing the magnitude of the Gabor transform G_f over the time variable k for every frequency variable n. Namely, let $\mathcal{G}_f$ be the GFAS of f, then

$$\mathcal{G}_f(\cdot) = \sum_k |G_f(\cdot, k)|.$$

$\mathcal{G}_f$ is a much simpler one-dimensional feature than the complete two-dimensional Gabor transform for a bird song/call. $\mathcal{G}_f$ is a profile of the frequency intensities within an entire song/call clip, collapsing out the time-varying information. The GFAS is the Gabor transform-based analog of what Emlen and deJong (1992) calculated from spectrograms and termed the *frequency–amplitude spectrum*.

Having only the frequency variable, the GFAS $\mathcal{G}_f$ is substantially different from the direct Fourier transform. Because of its global integration/summation nature, both time-varying frequency information and frequency-intensity variation are typically averaged together in a direct Fourier transform. Furthermore, since time-varying frequencies have no unique frequency to correspond to when mapped to the frequency domain, the direct Fourier transform of the signal must identify a range of frequencies often with highly variable intensities. As a result, the Fourier transform tends to demonstrate significant oscillatory behavior.

On the other hand, the GFAS first involves a complete time–frequency decomposition of the signal by the Gabor transform, which contains all the time-varying frequencies and time-varying frequency-intensity information. When averaged over the time index, the GFAS would still contain the variable information that was initially detected from the localized Gabor transform. Consequently, the GFAS is capable of reflecting both the frequency range and the frequency intensity within this range. The ability to portray both the frequency range and variation in intensities within this range is key to distinguishing bird songs/calls because different

songs/calls are likely to have both different frequency ranges and also different frequency-intensity patterns. It is highly unlikely that two species have songs that are permutations of both the frequency patterns and frequency intensities of each other.

To illustrate the difference between the GFAS and the direct Fourier transform of a chirping signal, we calculated both the GFAS and direct Fourier transform from a song clip of the Hammond's flycatcher [Figs. 1(a) and 1(b)]. The GFAS clearly shows the variation in frequency intensity of the song and appears to behave as an envelope on the frequency-intensity pattern. However, the Fourier transform shows rapid oscillation over the 4–6 kHz range (e.g., high variation in frequency intensity) and is unable to resolve the two major peaks in frequency just above 5 kHz and just below 7 kHz [Fig. 1(a)]. When using a template matching approach, the pattern shown by the GFAS would be easier to match, while the violent oscillations of the Fourier transform would be very difficult to align and match.

We also note that if two bird songs have different chirping behaviors in the time domain, but are similar in frequency range and in the pattern of frequency intensity then the GFAS would not be capable of distinguishing them. Without variation in frequency intensity, the GFAS tends to produce a smooth curve within the frequency range, which would be indistinguishable among songs/calls.

2. Gabor time–amplitude profile (GTAP)

Somervuo *et al.* (2006) report that temporal information was not useful in automatic identification of bird songs/calls. A lack of temporal alignment between the features extracted from known and unknown songs/calls or differences in song length may be responsible for their result. However, we included information contained in the temporal structure of a signal since it is conceivable that when aligned, temporal information may be useful in identifying bird songs/calls. We defined the GTAP as a frequency-cumulative Gabor transform, similar to the time-cumulative GFAS. If we call the GTAP of $f\,\mathcal{T}_f(k)$, then

$$\mathcal{T}_f(\cdot) = \sum_n |G_f(n,\cdot)|.$$

That is, the GTAP cumulates the frequency-varying information from a signal's Gabor transform by summing all the frequency indices. The GTAP reflects the temporal variation in sound intensity of a signal regardless of the frequency or time-varying frequency that the signal inherits. To improve the alignment of the GTAP of known and unknown song/call clips, we calculated the magnitude of the Fourier transform of the GTAP of each song/call and then took the inverse Fourier transform. If, in some instances, bird songs/calls have similar Gabor frequency–amplitude spectra, then

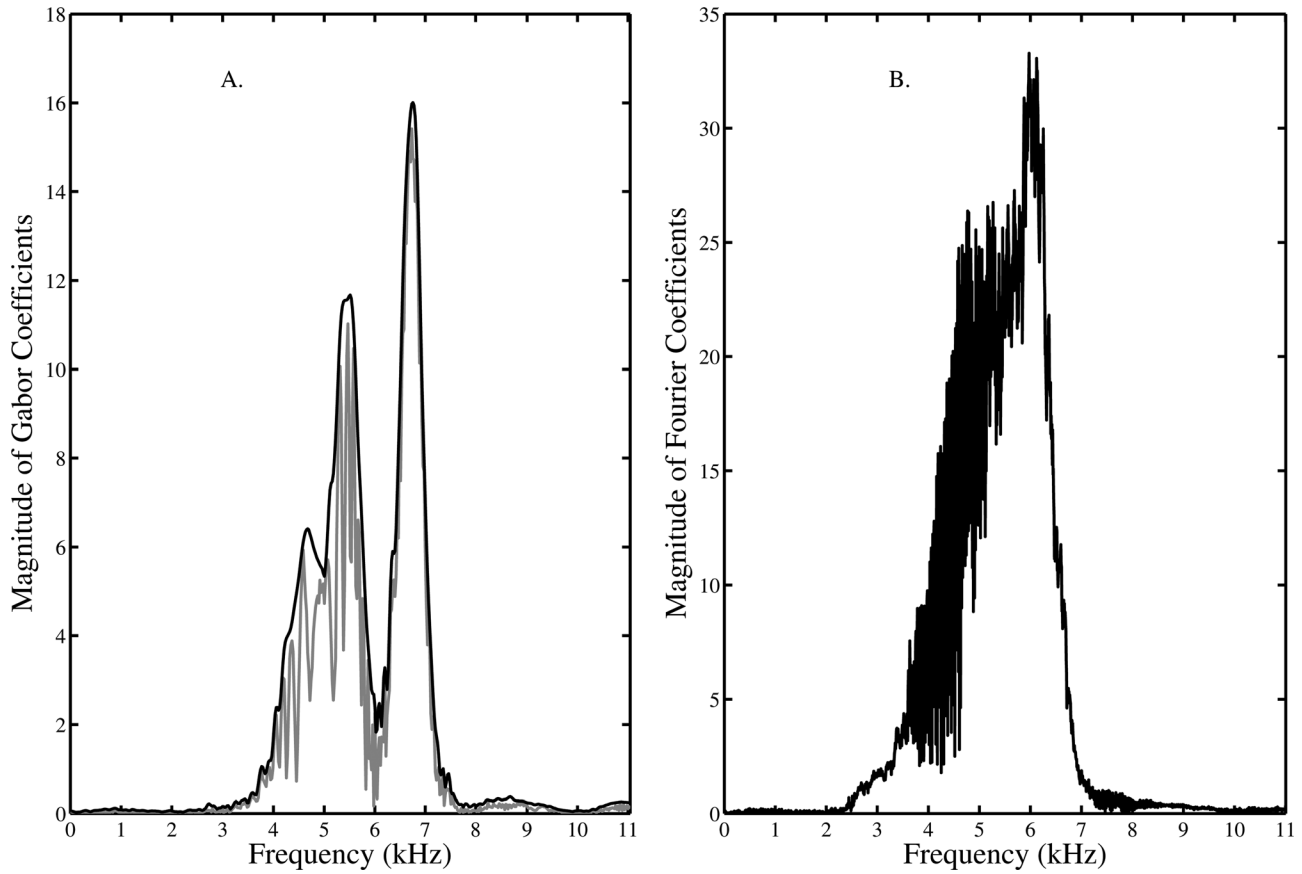

FIG. 1. (a) GFAS of Hammond's flycatcher (solid black line). The black line depicting the GFAS is the sum of the magnitudes of the Gabor coefficients. The GFAS is an envelope for the inner gray line, which is the magnitude of the sum of the Gabor coefficients: $|\sum_k G_f(\cdot, k)|$. (b) Frequency spectrum of the same bird call in (a).

GTAPs can serve secondarily to discriminate between species.

C. Bird song/call identification using GFAS and GTAP

We use a distance metric to compare each bird song/call to a library of known bird songs/calls. The metric is based on the correlation between unknown and known songs/calls of the GFAS and the GTAP. Specifically, if d indicates the distance between a known (a) and unknown (b) song/call, then

$$d(\mathrm{a}, \mathrm{b}) = [(1 - r(\mathcal{G}_\mathrm{a}, \mathcal{G}_\mathrm{b})) + (1 - r(\mathcal{T}_\mathrm{a}, \mathcal{T}_\mathrm{b}))],$$

where $r(f, g)$ is the correlation between two functions f and g, and in calculating $r(\mathcal{T}_\mathrm{a}, \mathcal{T}_\mathrm{b})$ the length of the two GTAPs is set to be the smaller of the two signal lengths. Hence, the distance metric involves the sum of the correlations between the intensities of the known and unknown song/call in the frequency and the time domains. Values of d range between 0 and 2, with 0 indicating a perfect match of frequency and time-varying components (negative correlations are set equal to zero). In practice, we have not encountered values of $d > 0.6$. Furthermore, the signals are initially downsampled by 2, so that the Nyquist frequency is 11.025 kHz. Most of the frequency content of bird songs/calls is well below this upper bound.

D. Noise addition

To determine the accuracy of our identification algorithm for bird song recordings with various SNR levels, we added noise to each song clip in our test set. Rather than using the sampled background noise, we felt it was more prudent to stochastically simulate background noise with the same spectral properties as natural background noise. If we had used the sampled backgound noise, we might have developed a de-noising algorithm optimized only for this particular noise sample.

An examination of the spectral properties of background noise present in avian habitats, based on an early morning recording of birds singing in a forest in the Sierra Nevada mountains of California, suggests that the signal content between 2 and 7 kHz is predominantly bird song, while the background noise is mostly low frequency noise [Fig. 2(a)].

The low frequency background noise consists of two components: A pink noise component, $p(\gamma)$, plus a component whose *power spectral density* has the form of a Gaussian function, $n(\gamma)$. We approximate the background noise to have a *power spectral density function* of the form $p(\gamma) + n(\gamma)$, where

$$p(\gamma) = \frac{1/100}{(\gamma - 1/100)^{3/4}},$$

$$n(\gamma) = \frac{100}{192\sqrt{2\pi}} e^{-(\gamma-500)^2/2(192^2)}.$$

The units of γ are hertz and the units of $p(\gamma)$ and $n(\gamma)$ are energy. Using these spectral density approximations, we generate random noise in MATLAB [Fig. 2(b)]. Because the simulated background noise was stochastically generated, we added the noise, applied an adapative high-pass filter to remove noise, and identified each bird call 50 times for each SNR level. The match rate for an individual bird song clip at each SNR level is then taken to be the proportion of these 50 trials that were correctly identified.

We added this simulated background noise to each of the 568 bird call recordings in the data set at four SNR levels (5, 10, 15, and 20 dB). We measured the target SNR in dB across all frequencies in our downsampled clip (0–11.25 kHz) because the set of species we were using included species with very low and also fairly high center

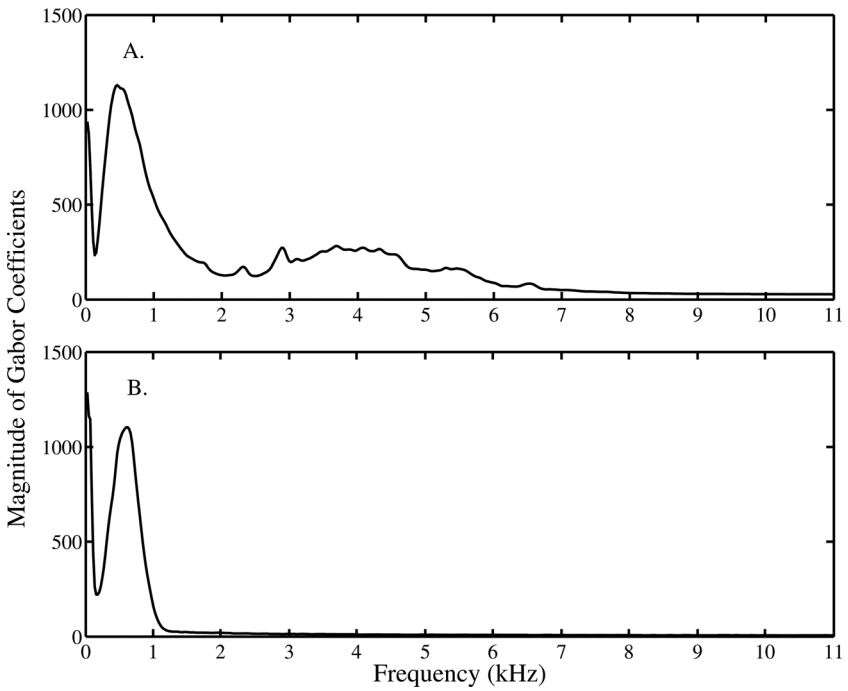

FIG. 2. (a) GFAS of a Sierra recording. (b) Example of the GFAS of simulated background environmental noise.

frequencies. We define our practical observation as the incident signal received at the microphone in the field, which encompasses mostly low frequency environmental noise, higher frequency bird sounds in the background, as well as the bird songs/calls we are attempting to identify. If we used the Great Blue Heron, the species in our set with the lowest mean frequency to define the lower limit of the signal bandwidth, then our measured SNR would only increase from 10 to 10.44 dB. We measured the SNR ratio as follows: We calculated the L2-norm of our CD- (compact disc recording) extracted bird vocalization clip and the L2-norm of the simulated background noise separately. Upon adding the noise to the bird vocalization clip, we defined the SNR of the resulting clip as $20\log_{10}(\|s\|_2/\|n\|_2)$ where $\|s\|_2$ is the L2-norm of the bird vocalization clip s and $\|n\|_2$ is the L2-norm of the noise n.

E. Noise removal via an adaptive high-pass filter

We investigated a simple method of de-noising bird call recordings through an adaptive high-pass filter. Using the spectral density distribution of the background noise from our Sierra recording [Fig. 2(a)] and the bird calls from our data set, we developed a Butterworth high-pass filter whose stop band frequency, passband frequency, and stop band attenuation are functions of the mean and standard deviation of frequency of the input signal. We calculated the mean and standard deviation of frequency based on the GFAS of the input signal. This method exploits the fact that most bird calls have frequency ranges much higher than the frequency range of the background noise. The parameters of the adaptive high-pass filter were designed to remove as much low frequency background noise without removing more than 3% of the signal energy. For bird calls with center frequencies less than 1 kHz, no filter was applied since there was too much overlap in frequency between the signal and noise. Our adaptive high-pass filter employed four filter ranges (Table I). The choice of filter to apply was determined by the mean and standard deviation of the input signal and the decision thresholds are given in Table II.

F. Bird song data set

We selected 568 bird vocalizations from two commercially available CD collections of birds from western North America and California (Cornell Laboratory of Ornithology, 1992; Keller, 2003). All recordings were CD quality with bit depth of 16 bits and a sampling rate of 44 100 samples/s. All recordings had very low background noise levels. We selected 67 species that occur in the northern Sierra Nevada mountains, including several orders of both passerine (song birds) and non-passerine birds. We included at least two replicate vocalizations for each of several song or call types for each species. We defined song and call types by ear from the available vocal clips. We made no effort to be comprehensive in representing variation in the vocalizations of each species. However, we included the territorial vocalizations of each species and variants on it. The average number of vocal clips and call types for each species was 8.48 ± 0.62 and 2.07 ± 0.14, respectively ($\pm$standard error). The maximum number of vocal clips and song or call types for a species was 29 and 5, respectively. The average length of each vocal clip was 1.55 ± 0.05 s. A complete list of the species used and the number of vocal clips and song or call types of each is included in the supplementary material.[1]

TABLE I. Parameters for the four filter ranges for the adaptive high-pass filter.

Filter type	Stop band (Hz)	Passband (Hz)	Stop band attenuation (dB)
No filter			
Low filter	300	700	4
Mid filter	700	1250	6
High filter	1250	1700	6

TABLE II. Decision thresholds for application of the adaptive high-pass filter.

Mean frequency (kHz)	Standard deviation of frequency (kHz)	Filter choice
<1	All	No filter
1–2	All	Low filter
2–3	<1	Mid filter
2–3	≥1	No filter
3–4	<1.5	Mid filter
3–4	≥1.5	No filter
4–5	All	Mid filter
5–6	<2	High filter
5–6	≥2	Mid filter
6–7	<1.3	High filter
6–7	≥1.3	Mid filter
7–8	All	High filter

G. Testing the effectiveness of bird song identification using the Gabor transform

To determine how successful our approach to bird song/call identification might be, we used our library of 568 song/call clips to produce a confusion matrix for the songs/calls without adding noise and for each of the 50 replicates of the song/calls with noise added at SNR 5, 10, 15, and 20 dB. Self matches were not allowed, so each song/call clip was compared to the remaining 567 songs/calls and d was estimated. We classified each song/call clip to be from the species for which it had the lowest d value. For the analysis without noise addition, we report whether or not we correctly identified the song/call. For the simulations involving noise addition, we report the success rate for each song/call clip as the percentage of correct identifications among the 50 replicates for each SNR.

III. RESULTS

A. Successful identification rate without added noise

While the GFAS of different examples of particular bird songs are not identical, they appear to be similar enough

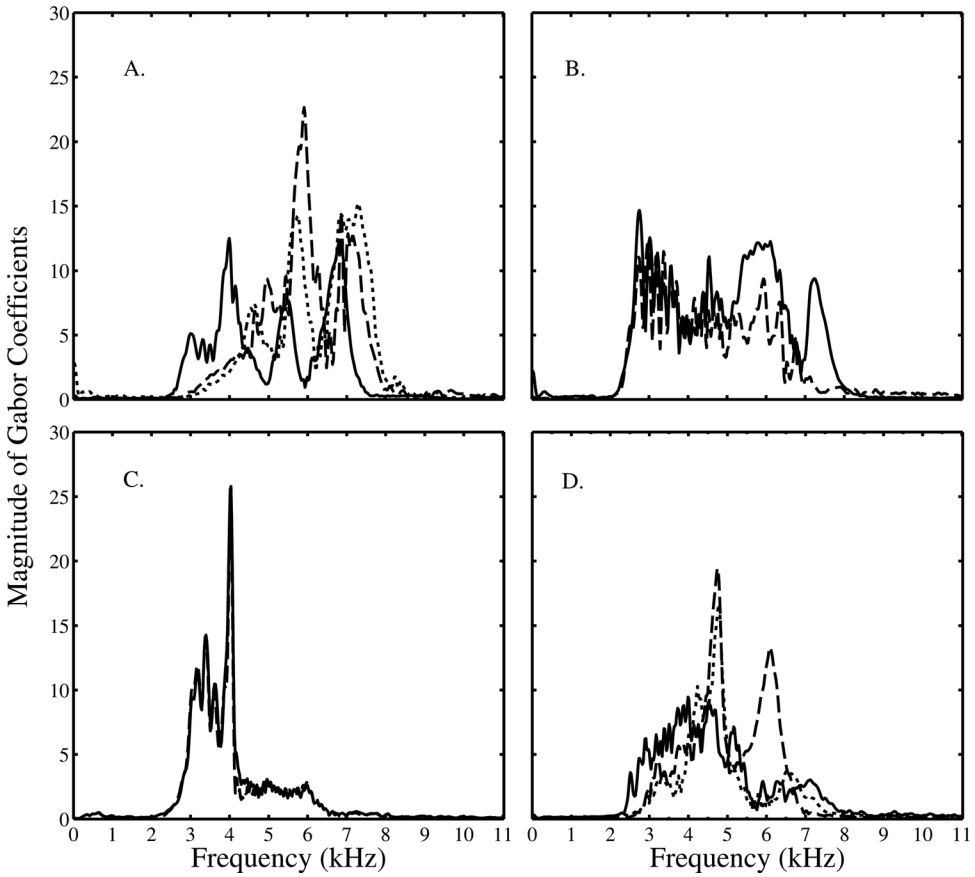

FIG. 3. GFAS of replicate songs of the Hammond's flycatcher. (a)–(c) Illustrative examples of three different call types all correctly identified and (d) illustrates three songs clips that we misidentified and appear not to fall into any of our identified song types.

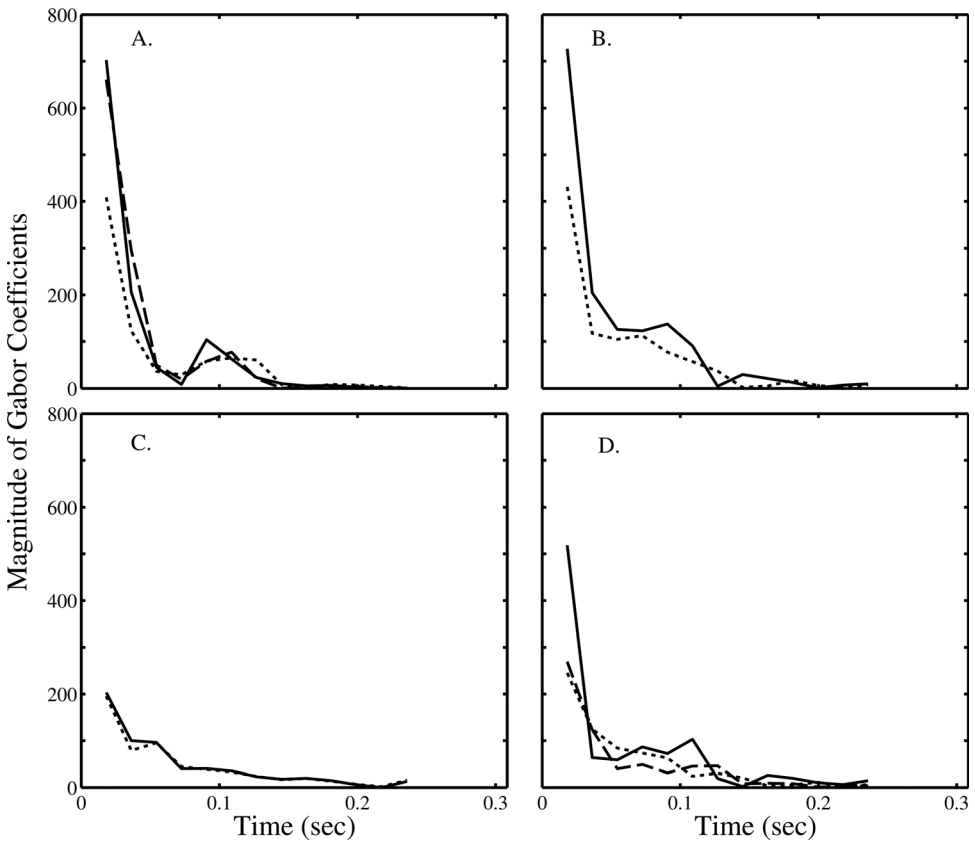

FIG. 4. GTAP of several replicate songs of the Hammond's flycatcher. (a)–(c) Illustrative examples of three different call types all correctly identified and (d) illustrates three songs clips that we misidentified and appear not to fall into any of our identified song types.

to permit accurate identification. For example, in Figs. 3(a)–3(c), we illustrate multiple examples of the GFAS for three different call types for the Hammond's flycatcher. For each call type there is similarity in the patterns of frequency intensities with call type 3 [Fig. 3(a)] showing the greatest similarity among replicate songs. The three calls depicted in Fig. 3(d) are examples of song clips for the Hammond's flycatcher that were not correctly identified. These three song clips are clearly distinct from the other three song types. For the Hammond's flycatcher the GTAP also showed strong similarity among replicate calls in each call type [Figs. 4(a)–4(c)], and greater differences between the misidentified calls [Fig. 4(d)].

Overall, we correctly identified 544 of the 568 songs/calls when no noise was added (95.77%). Examination of the particular songs that we misidentified suggests no clear pattern for our failures. The missed songs were not concentrated among any particular order of birds, or concentrated within a single species. There was a tendency for the missed songs to have a wider bandwidth than correctly identified songs, but this pattern was not statistically significant (Mann-Whitney $U = 5225.5$, $p = 0.098$). When we played back the audio versions of these song clips and reexamined plots of their GFAS, in many instances we concluded that these missed calls had no close match in our library and should have been classified as distinct call types.

When we used a metric based solely on the GFAS, we correctly identified 86% of the bird songs. The success of the GFAS alone suggests that the information content in the temporal structure of the bird songs, as we have represented it via the GTAP, enhances our ability to identify bird songs only very slightly.

B. Successful identification rate with added noise and noise removal

Among the 544 songs/calls that we could identify, the addition of simulated background noise reduced our algorithm's ability to correctly identify the bird species producing the songs (Fig. 5). However, our adaptive high-pass filter was sufficiently effective at removing the added noise, so that our rate of successful identification was only slightly degraded in the face of added noise. The overall mean rate of successful identification at SNR 5 dB was still

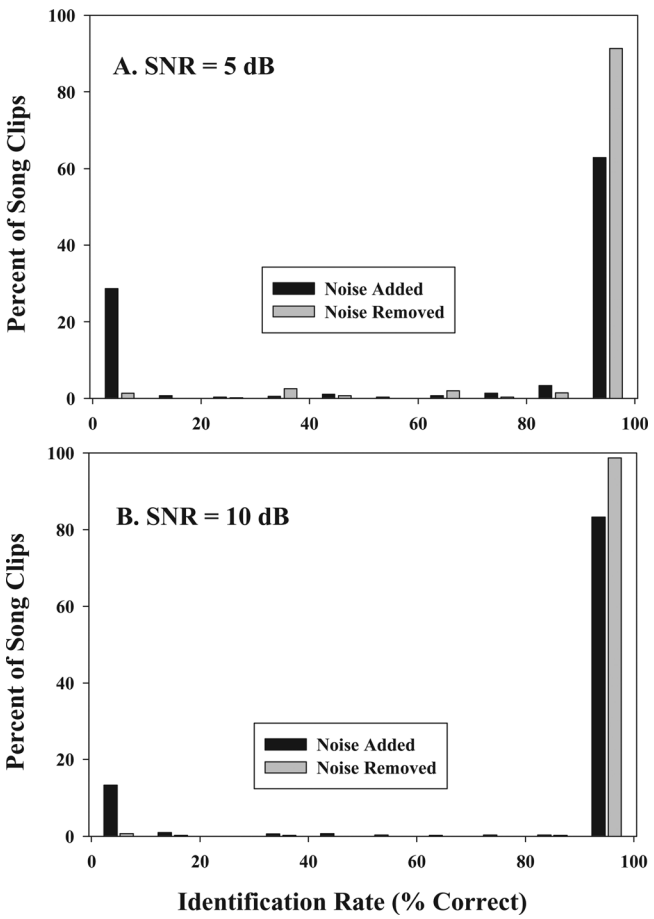

FIG. 5. Effect of noise addition and noise removal on the correct identification rate for song clips. The x axis represents the identification rate out of the 50 trials with randomly added noise, and the y axis is the percentage of our set of 544 song clips having particular identification rates. The effect of noise addition was minimal at SNR 15 and 20 dB, so we only show the results for (a) SNR 5 dB and (b) SNR 10 dB. Notice that when noise is added the distribution of identification rates is bimodal, with some song clips with high identification rates and some that are unidentifiable (black bars). However, after noise removal (gray bars) almost all song clips had high rates of correct identification.

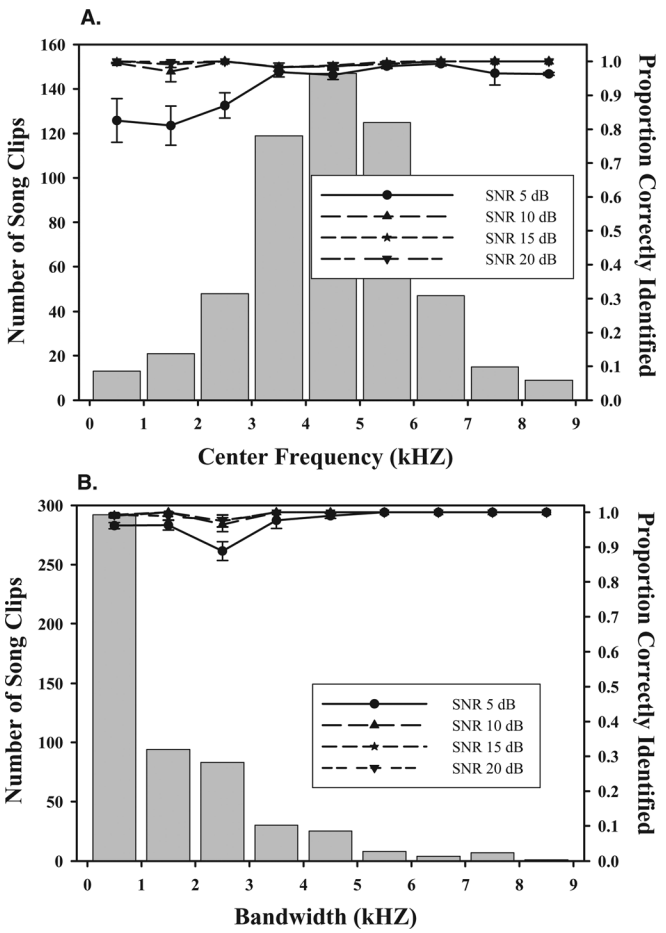

FIG. 6. Identification rates for songs with noise added and noise removed as a function of (a) center frequency and (b) bandwidth. The solid bars indicate the frequency of songs in each center-frequency or bandwidth class (left vertical axis), and the lines illustrate the successful identification rate (right vertical axis). Error bars are ± standard error.

TABLE III. Identification rates for de-noised bird songs at SNR 5 dB and characteristics of those songs for different avian orders.

Avian order	Number of species	Number of songs	ID rate	Center frequency	Bandwidth
Apodiformes (hummingbirds)	1	3	0.98 ± 0.02	6.7 ± 0.08	0.26 ± 0.12
Charadriiformes (killdeer)	1	4	0.80 ± 0.20	5.1 ± 0.12	0.41 ± 0.15
Ciconiiformes (herons)	1	5	0.74 ± 0.05	1.8 ± 0.06	1.8 ± 0.52
Columbiformes (doves)	2	13	0.92 ± 0.08	1.15 ± 0.09	0.37 ± 0.02
Coracciformes (woodpeckers)	3	17	0.94 ± 0.04	4.0 ± 0.31	2.3 ± 0.75
Falconiformes (hawks)	5	24	0.94 ± 0.04	3.9 ± 0.19	1.5 ± 0.32
Galliformes (quail)	2	22	0.87 ± 0.06	2.2 ± 0.07	0.43 ± 0.10
Gruiformes (cranes)	1	4	0.59 ± 0.12	1.9 ± 0.05	1.5 ± 0.39
Passeriformes (songbirds)	50	443	0.93 ± 0.01	4.8 ± 0.06	1.5 ± 0.08
Strigiformes (owls)	1	9	0.75 ± 0.08	1.5 ± 0.59	0.25 ± 0.01

$95.44 \pm 0.007\%$. We illustrate our results in Figs. 6(a) and 6(b), where we plot the successful identification rate as a function of the center frequency of the song/call and its bandwidth, respectively. Note that we only have a substantially reduced rate of correct identification at SNR 5 dB for songs with center frequencies below 3 kHz. Our rate of correct identification for songs/calls with higher center frequencies is consistently above 97%, even at SNR 5 dB. We were also less successful with songs with bandwidth in the 2–3 kHz range, but the degradation in identification rate was much lower than for songs with low center frequency [Fig. 6(a)]. Among song birds that are the most common and species rich of the avian orders in our study (Passeriformes), our success rate was $96.53 \pm 0.007\%$. The Passeriformes have high center frequency unlike the Gruiformes, Strigiformes, and Charadriiformes, which all have the lowest center frequencies and the lowest rates of successful identification (Table III).

We built logistic regression models with success defined as having a 70% or higher identification rate to determine if center frequency, bandwidth, dominant frequency, the ratio of center frequency and bandwidth, and the number of files of the same song type in the library could explain the observed variation in identification rate. No terms were added to any model, except for the model for SNR 5 dB, where center frequency was the only significant predictor variable. Songs with higher center frequency had higher rates of successful identification.

IV. DISCUSSION

Our results suggest that it is possible to automatically identify a large number of bird species from digital recordings of their vocalizations. The approach we propose is essentially a template matching method that uses information extracted from the Gabor transform of the unknown song clip to characterize both the frequency-intensity composition and temporal structure of the song. Comparison of this information to a comprehensive library of similarly treated song clips yielded high rates of correct identification. Our success at identification suggests that the Gabor transform using the parameters we suggest, $N = 1000$ and $T = 400$, provides sufficient resolution to identify bird songs. Our simulations of added noise and noise-removal using an adaptive high-pass filter indicate that even at a SNR as low as 5 dB, we can identify over 90% of our test sample of bird songs/calls.

Although our proposed metric d is *ad hoc*, we examined many other metrics, all of which were less successful. The rate of correct identification using matrix correlation or cross correlation was substantially lower than the rate we report for d and took substantially longer time to compute. We suspect that these alternative metrics suffer from the problem we previously mentioned of achieving adequate temporal alignment between the reference song and the unknown song, and that when using a matrix approach, misalignment in the temporal domain creates misalignment in the frequency domain as well.

Our results are in agreement with the conclusions of Somervuo *et al.* (2006) that, in general, the frequency-intensity composition of the bird songs were more useful for identification than information about the temporal structure of the songs. Again, we believe that the potential for misalignment renders the temporal information less useful.

The drawbacks of our approach are that as currently applied, we set no threshold for d below which we do not attempt an identification, we have no means to make any probabilistic inference concerning the likelihood of correct identification, and the technique requires a large library of song clips potentially including regional song dialects. However, further study of the behavior of d for a wide range of species and song clips might suggest appropriate thresholds for successful identification. Furthermore, we have yet to explore extensively other ways to use the information contained in the GFAS or GTAP to achieve accurate species identification. For example, we have not attempted to use these features in classification functions, neural networks, or hidden Markov models, as has been common in other approaches to the automatic identification of bird songs (Terry and McGregor, 2002; Ranjard and Ross, 2008; Trifa *et al.*, 2008). The fact that we could not identify 24 of 568 song clips, even with no added noise, is evidence of the critical importance of having a comprehensive library encompassing the variation in the songs/calls of each bird species that might be encountered at a sample site. Our examination of these misidentified song clips suggests that we did not

have a homolog of each in our library. To the extent that a species sometimes drops a note, alters the pace, or varies its song in any way, examples of these alternatives must be in the library to permit identification. For species with strong regional song dialects, libraries might need to be customized to contain the local dialect.

Many other improvements to reduce computing time and potentially improve identification accuracy could be applied to our approach. At present we use a brute force examination of the match between the unknown song and every library member to achieve identification. However, limiting the search to songs with similar center frequency, bandwidth, or other attributes would speed the identification process and potentially reduce spurious correlations arising from chance similarity in the GTAP. Applying wavelet-based de-noising techniques might also improve identification accuracy, particularly at low SNR (Frisch and Messer, 1992; Donoho, 1995).

While our results suggest that automatic identification of birds from digital recordings might be feasible, the work we report here represents only one of the obstacles that must be overcome to successfully automate avian point counts from digital recordings. Under field conditions, noise levels, the problem of segmenting out the sections of the recording with avian vocalizations, and the "cocktail party" problem might be more difficult hurdles to overcome.

For example, point counts are often conducted with a 100 m radius for the sample area. However, for a four microphone array configured to minimize the maximum distance between a microphone and a singing bird, the maximum distance to a microphone is approximately 70 m. For a small bird (8 g) its sound pressure level is approximately 78 dB (Calder, 1990). If the environmental noise has sound pressure level of 45 dB (i.e., we estimated noise levels of 35–45 dB from field recordings in the Sierra Nevada mountains and in Alameda County, CA), and we only account for reduction in sound pressure level due to spherical spread, then at 70 m from the nearest microphone the incident SNR of this bird's song would be only 7 dB. If we also take into account signal attenuation due to the vegetation and the potential for the bird to be facing away from the microphone, the incident SNR might be even less. A denser array of microphones would increase the SNR as would using a smaller sample radius. In any event, care must be taken to make sure that point counts based on microphone arrays are designed to maximize detection probabilities for resident birds.

The analyses we performed involved using song clips that had been manually segmented from continuous recordings. Human observers listened to the song clips and clipped out that part to be used for identification. The development of automatic segmentation methods is also a challenge that must be overcome before the approach we outline can have practical application. Previous attempts have met with only limited success (Lee *et al.*, 2006).

The cocktail party problem might be the most difficult to overcome. During the dawn chorus when bird communities are sampled it is not unusual to hear different birds or bird species singing simultaneously—like people talking at a cocktail party. If we can only use the sections of the recording with unique vocalizations then it might be difficult to get sufficient vocal clips to detect and identify each individual bird. However, if the vocal contributions of simultaneously signing birds could be resolved, then perhaps the endeavor of automating point counts is feasible. Further work on identification algorithms, recording segmentation, and microphone array design will ultimately be necessary before point counts could be successfully automated.

ACKNOWLEDGMENTS

We wish to thank Adam Paganini, Hugh-Krogh Freeman, and Amanda Jobbins for assistance in building our library of song clips. This research was supported by National Science Foundation Undergraduate Biology-Math Grant No. EF-046313 and by a grant from the Office of the Vice Provost for Research at San Francisco State University.

[1]See supplementary material at http://dx.doi.org/10.1121/1.4726006 for a complete list of the species used and the numbers of vocal clips and song or call types for each.

Acevedo, M. A., and Villanueva-Rivera, L. J. (**2006**). "Using automated digital recording systems as effective tools for the monitoring of birds and amphibians," Wildl. Soc. Bull. **34**, 211–214.

Alldredge, M. W., Simons, T. R., and Pollock, K. H. (**2007a**). "A field evaluation of distance measurement error in auditory avian point count surveys," J.Wildl. Manage. **71**, 2759–2766.

Alldredge, M. W., Simons, T. R., and Pollock, K. H. (**2007b**). "Factors affecting aural detections of songbirds," Ecol. Appl. **17**, 948–955.

Anderson, S. E., Dave, A. S., and Margoliash, D. (**1996**). "Template-based automatic recognition of bird song syllables from continuous recordings," J.Acoust. Soc. Am. **100**, 1209–1219.

Bart, J. (**1985**). "Causes of recording errors in singing bird surveys," Wilson Bull. **97**, 161–172.

Bart, J., and Schoultz, J. D. (**1984**). "Reliability of singing bird surveys: Changes in observer efficiency with avian density," Auk **101**, 307–318.

Brandes, T. S. (**2008a**). "Automated sound recording and analysis techniques for bird surveys and conservation," Bird Conserv. Int. **18**, S163–S173.

Brandes, T. S. (**2008b**). "Feature vector selection and use with hidden Markov models to identify frequency-modulated bioacoustic signals amidst noise," IEEE Trans. Audio, Speech, Lang. Process. **16**, 1173–1180.

Brandes, T. S., Naskrecki, P., and Figueroa, H. K. (**2006**). "Using image processing to detect and classify narrow-band cricket and frog calls," J. Acoust. Soc. Am. **120**, 2950–2957.

Cai, J., Ee, D., Pham, B., Roe, P., and Zhang, J. (**2007**). "Sensor network for the monitoring of ecosystem: Bird species recognition," in *IEEE International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*, pp. 293–298.

Calder, W. A. (**1990**). "The scaling of sound power output and territory size: Are they matched?" Ecology **71**, 1810–1816.

Cellis-Murillo, A., Deppe, J. L., and Allen, M. F. (**2009**). "Using soundscape recordings to estimate bird species abundance, richness, and composition," J.Field Ornithol. **80**, 64–78.

Chen, Z., and Maher, R. C. (**2006**). "Semi-automatic classification of bird vocalizations using spectral peak tracks," J. Acoust. Soc. Am. **120**, 2974–2984.

Chesmore, E. D., and Ohya, E. (**2004**). "Automated identification of field-recorded songs of four British grasshoppers using bioacoustic signal recognition," Bull. Entomol. Res. **94**, 319–330.

Cornell University of Ornithology. (**1992**). *A Field Guide to Western Bird Songs*, 2nd ed. (Houghton Mifflin, New York).

Cunningham, R. B., Lindenmayer, D. B., and Lindenmayer, B. D. (**2004**). "Sound recording of bird vocalisations and point interval counts of bird numbers—A case study in statistical modeling," Wildl. Res. **31**, 195–207.

Donoho, D. L. (**1995**). "De-noising by soft-thresholding," IEEE Trans. Inf. Theory **41**, 613–627.
Emlen, J. T., and DeJong, M. J. (**1992**). "Counting birds: The problem of variable hearing abilities," J. Field Ornithol. **63**, 26–31.
Fagerlund, S. (**2007**). "Bird species recognition using support vector machines," EURASIP J. Adv. Signal Process. **2007**, 1–7.
Fagerlund, S., and Härmä, A. (**2005**). "Parametrization of inharmonic bird sounds for automatic recognition," in *13th European Signal Processing Conference (EUSIPCO 2005)*, September 4–8, 2005, Antalya, Turkey.
Frisch, M., and Messer, H. (**1992**). "The use of the wavelet transform in the detection of an unknown signal," IEEE Trans. Inf. Theory **38**, 892–897.
Gabor, D. (**1946**). "Theory of communication," J. IEE (London) **93**, 429–457.
Härmä, A. (**2003**). "Automatic identification of bird species based on sinusoidal modeling of syllables," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003)*, Vol. V, pp. 545–548.
Härmä, A., and Somervuo, P. (**2004**). "Classification of the harmonic structure in bird vocalization," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004)*, Vol. V, pp. 701–704.
Haselmayer, J., and Quinn, J. S. (**2000**). "A comparison of point counts and sound recording as bird survey methods in Amazonian southeast Peru," Condor **102**, 887–893.
Hobson, K. A., Rempel, R. S., Greenwood, H., Turnbull, B., and Van Wilgenburg, S. L. (**2002**). "Acoustic surveys of birds using electronic recordings: New potential from an omnidirectional microphone system," Wildl. Soc. Bull. **30**, 709–720.
Horne, J. K. (**2000**). "Acoustic approaches to remote species identification: A review," Fish. Oceanogr. **9**, 356–371.
Jennings, N., Parsons, S., and Pocock, M. J. O. (**2008**). "Human vs. machine: Identification of bat species from their echolocation calls by humans and by artificial neural networks," Can. J. Zool. **86**, 371–377.
Keller, G. A. (**2003**). *Bird Songs of California* (Cornell Laboratory of Ornithology, Ithaca, NY).
Kogan, J. A., and Margoliash, D. (**1998**). "Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study," J. Acoust. Soc. Am. **103**, 2185–2196.
Kwan, C., Mei, G., Zhao, X., Ren, Z., Xu, R., Stanford, V., Robert, C., Aube, J., and Ho, K. C. (**2004**). "Bird classification algorithms: Theory and experimental results," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. V, pp. 289–292.
Lee, C.-H., Chou, C.-H. Han, C.-C., and Huang, R.-Z. (**2006**). "Automatic recognition of animal vocalizations using averaged MFCC and linear discriminant analysis," Pattern Recogn. Lett. **27**, 93–101.
Lindenmayer, D. B., Cunningham, R. B., and Lindenmayer, B. D. (**2004**). "Sound recording of bird vocalisations in forests. II. Longitudinal profiles of vocal activity," Wildl. Res. **31**, 209–217.
Mayfield, H. (**1966**). "Hearing loss and bird song," Living Bird **5**, 167–175.
McIlraith, A. L., and Card, H. C. (**1997**). "Birdsong recognition using backpropagation and multivariate statistics," IEEE Trans. Signal Process. **45**, 2740–2748.
Ralph, C. J., Droege, S., and Sauer, J. R. (**1995**). "Managing and monitoring birds using point counts: Standards and applications," USDA Forest Service General Technical Report No. PSW-GTR-149 (Pacific Southwest Research Station, Forest Service, U.S. Department of Agriculture, Albany, CA), pp. 161–169.
Ranjard, L., and Ross, H. A. (**2008**). "Unsupervised bird song syllable classification using evolving neural networks," J. Acoust. Soc. Am. **123**, 4358–4368.
Rempel, R. A., Hobson, K. A., Holburn, G., Van Wilgenburg, S. L., and Elliot, J. (**2005**). "Bioacoustic monitoring of forest songbirds: Interpreter variability and effects of configuration and digital processing methods in the laboratory," J. Field Ornithol. **76**, 1–11.
Riede, K. (**1998**). "Acoustic monitoring of Orthoptera and its potential for conservation," J. Insect Conserv. **2**, 217–223.
Selin, A., Turunen, J., and Tanttu, J. T. (**2006**). "Bird song classification and recognition using wavelets," Razprave IV. Razreda Sazu **47**, 185–204.
Selin, A., Turunen, J., and Tanttu, J. T. (**2007**). "Wavelets in recognition of bird sound," EURASIP J. Adv. Signal Process. **2007**, 1–9.
Simons, T. R., Alldredge, M. W., Pollock, K. H., and Wettroth, J. M. (**2007**). "Experimental analysis of the auditory detection process on avian point counts," Auk **124**, 986–999.
Somervuo, P., Härmä, A., and Fagerlund, S. (**2006**). "Parametric representations of bird sounds for automatic species recognition," IEEE Trans. Audio, Speech, Lang. Process. **14**, 2252–2263.
Swiston, K. A., and Mennill, D. J. (**2009**). "Comparison of manual and automated methods for identifying target sounds in audio recordings of Pileated, Pale-billed, and putative Ivory-billed woodpeckers," J. Field Ornithol. **80**, 42–50.
Tanttu, J. T., Turunen, J., and Ojanen, M. (**2003**). "Automatic classification of flight calls of the common crossbills," *Proceedings of the International Conference on Acoustic Communication in Animals*, July 27–30, College Park, MD, pp. 239–240.
Terry, A. M. R., and McGregor, P. K. (**2002**). "Census and monitoring of individually based vocalizations: The role of neural networks," Anim. Conserv. **5**, 103–111.
Trifa, V. M., Kirshel, A. N. G., and Taylor, C. (**2008**). "Automated recognition of antbirds in a Mexican rainforest using hidden Markov models," J. Acoust. Soc. Am. **123**, 2424–2431.