Semi-automatic classification of bird vocalizations using spectral peak tracks

Zhixin Chen[a] and Robert C. Maher[b]
Department of Electrical and Computer Engineering, Montana State University, Bozeman, Montana 59717-3780

(Received 22 June 2005; revised 23 May 2006; accepted 1 August 2006)

Automatic off-line classification and recognition of bird vocalizations has been a subject of interest to ornithologists and pattern detection researchers for many years. Several new applications, including bird vocalization classification for aircraft bird strike avoidance, will require real time classification in the presence of noise and other disturbances. The vocalizations of many common bird species can be represented using a sum-of-sinusoids model. An experiment using computer software to perform peak tracking of spectral analysis data demonstrates the usefulness of the sum-of-sinusoids model for rapid automatic recognition of isolated bird syllables. The technique derives a set of spectral features by time-variant analysis of the recorded bird vocalizations, then performs a calculation of the degree to which the derived parameters match a set of stored templates that were determined from a set of reference bird vocalizations. The results of this relatively simple technique are favorable for both clean and noisy recordings. © *2006 Acoustical Society of America.*
[DOI: 10.1121/1.2345831]

PACS number(s): 43.80.Ka, 43.60.Uv, 43.60.Lq [DOS] Pages: 2974–2984

I. INTRODUCTION

Several engineering applications require real time identification of birds while in flight, foraging, or roosting. These include systems to help avoid collisions between birds and aircraft (Pascarelle *et al.*, 2004), systems to track migratory birds in the vicinity of wind turbine generators (NWCC, 2004), and ornithological measurements systems to help understand avian behavior and migratory patterns, particularly at night and in unfavorable meteorological conditions. Among the possible means to identify the bird species are their vocalizations. Thus, there exists a need for research in on-line acoustical bird classification systems capable of running unattended and in real time.

In recent years many off-line techniques for classifying bird species based on recorded vocalizations have been proposed and developed. The most successful techniques are based on manual inspection and labeling of bird sound spectrographs by experts, but this process is tedious and dependent upon the subjective judgment of the observer (Kogan and Margoliash, 1998). The reliability of classification can be improved if a panel of experts is used, but this is expensive, time consuming, and unsuitable for real time classification. Nevertheless, the fact that the manual inspection of sound spectrographs tends to yield correct judgments has encouraged research into automatic classification using objective standards derived from expert opinions.

Several of the existing automatic off-line bird vocalization classification techniques are based on traditional speech recognition methods (Rabiner and Juang, 1993). Anderson, Dave, and Margoliash (1996) used dynamic time warping (DTW) for automatic recognition of birdsong syllables from continuous recordings. Their method directly compared the spectrograms of input bird sounds with those of a set of predefined templates representative of categories chosen by the investigator. They applied this method to vocalizations from two bird species recorded in a low noise environment and achieved 97% accuracy in the syllables of stereotyped songs and 84% accuracy in the syllables of plastic (variable or indistinct syllable) songs. The method did not use amplitude normalization, so the results may be sensitive to amplitude differences. Kogan and Margoliash (1998) used DTW and hidden Markov models (HMM) to classify bird sounds based on the syllables extracted from continuous recordings. Their method began by extracting linear predictive coding (LPC) coefficients or mel-frequency cepstral coefficients (MFCC) from a set of bird syllables and then used DTW or HMM for recognition. This method was found to perform well for two specific birds in a low noise environment. The method worked less well in noisy environments or with short duration bird vocalizations. Ito, Mori, and Iwasaki (1996) extracted two time-varying spectral features from syllables and used dynamic programming (DP) matching to classify budgerigar contact calls, and found that the method performed well. However, only the frequencies of the spectral features were used, not the spectral powers, so the method may not be appropriate for other bird sounds with different spectral structure.

Other researchers have developed classification methods specifically tailored to bird vocalizations. McIlraith and Card (1997) conducted research on the recognition of songs of six bird species. In their method the bird songs were represented with spectral and temporal parameters of the songs. They reduced the complexity of the search space by selecting features exhibiting the greatest discrimination, then used a neural network for classifying the bird songs. Their method achieved good performance but the neural classifier required

[a]Electronic mail: chen@montana.edu
[b]Corresponding author. Electronic mail: rob.maher@montana.edu

 © 2006 Acoustical Society of America

considerable computation. Härmä (2003) proposed an alternative method for bird sound classification. He observed that for many songbird vocalizations, a large class of syllables can be approximated as brief sinusoidal pulses with time varying amplitude and frequency. Although this model is too simple for certain bird sounds, the system provided good recognition results for species with tonal vocalizations. In a subsequent study, Härmä and Somervuo (2004) classified bird sounds into four classes based on their harmonic structure, where each harmonic component was modeled with one time-varying sinusoid. No classification statistics were reported, but they found that the signal models appropriately represented the spectral structure of 93% of the syllables in their database.

In summary, the existing methods are well suited to their specific application, but they also have some limitations. The DTW and HMM techniques did not perform well in noisy environments or for bird sounds with short duration and variable amplitude. The neural classifier required a very high computational complexity. The use of one or two spectral peak tracks is appealing for its simplicity and robustness to noise. Thus, in this paper we describe a spectral pattern detection method with relatively low computational complexity (i.e., suitable for a final implementation in real time) that can be used to classify in real time tonal bird vocalizations (harmonic or inharmonic) in the presence of realistic background noise levels.

The remaining sections of this paper are organized as follows. First, we review briefly the basic structure and terminology of bird vocalizations. We then describe the rationale and features of the proposed semi-automatic classification method, including our simulation results and interpretation, and compare the results to classification based on conventional DTW and HMM methods. Finally, we conclude with a summary and several suggestions for future research in this area.

II. BIRD SOUNDS

Birds are able to produce a wide variety of sounds. Air from the lungs is forced through the bronchi to the syrinx, which is the major source of vibratory modulation. Sound from the syrinx passes through the resonant structures of the trachea, larynx, mouth, and beak.

Bird vocalizations can be divided into the general categories of elements, syllables, phrases, calls, and songs (Krebs and Kroodsma, 1980; Catchpole and Slater, 1995). *Elements* can be regarded as the elementary sonic units in bird vocalizations. A *syllable* includes one or more elements and is usually a few to a few hundred milliseconds in duration. *Phrases* are short groupings of syllables. *Calls* are generally compact sequences of phrases, while *songs* are long and complex vocalizations. An example of the hierarchy is shown in Fig. 1. The reader is cautioned to be aware that the details of each category often show individual and geographic variations and temporal plasticity, even within a single species (Kroodsma and Miller, 1996; Krebs and Kroodsma, 1980). Thus, a classification strategy consisting of a simple tone sequence detector tuned to match a specific

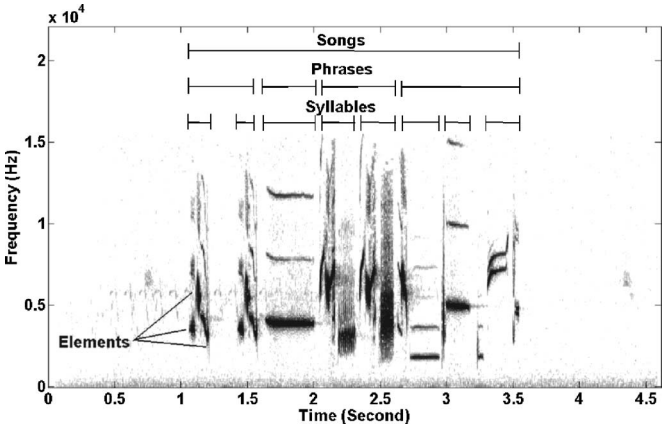

FIG. 1. Descriptive hierarchy of bird vocalization. The bird vocalization can be divided into four hierarchical levels: element, syllable, phrase, and song (or call). The classification experiment reported in this paper operates at the syllable level in the hierarchy.

example bird song is unlikely to be useful in engineering applications due to natural variation in the song details.

Syllables can range from being nearly tonal whistles, to harmonic sounds with a distinct fundamental frequency, to inharmonic bursts, or even to being noise like (Nowicki, 1997; Kahrs and Avanzini, 2001; Fagerlund, 2004). In the case of voiced harmonic sounds, the fundamental frequency range typically lies between 500 and 5000 Hz. The spectral content of a harmonic sound can also vary with time as the bird changes the shape and length of the vocal tract and beak. Figure 2 shows examples from songs and calls from different species, illustrating only a small variety of the complex sounds birds can produce.

III. CLASSIFICATION METHOD AND EXPERIMENTAL RATIONALE

The bird sound classification problem is similar to many existing pattern detection and classification problems. The classification procedure described in this paper follows the standard design cycle of Duda *et al.* (2001), as depicted in Fig. 3.

The *data collection* process included the acquisition of audio recordings using recording apparatus similar to what is anticipated in a real time classification system. The data were obtained in the field with a variety of ambient noise sources such as wind, motor traffic, and aircraft.

The *feature choice* process was based on our examination of the spectrotemporal characteristics of the bird sounds of interest. This examination was supplemented by a review of prior research into bird sound recognition and our prior experience with a variety of signal processing methods useful for data extraction in speech, environmental sounds, and music. For the purposes of classification in the presence of background noise, we have found that subsyllable elements are difficult to use in practice simply due to the difficulty in extracting reliable signal parameters for such short segments. Meanwhile, ornithologists report that higher level phrases, calls, and songs often contain regional and individual variations that will add additional degrees of freedom to the classification problem (Krebs and Kroodsma, 1980). This led us

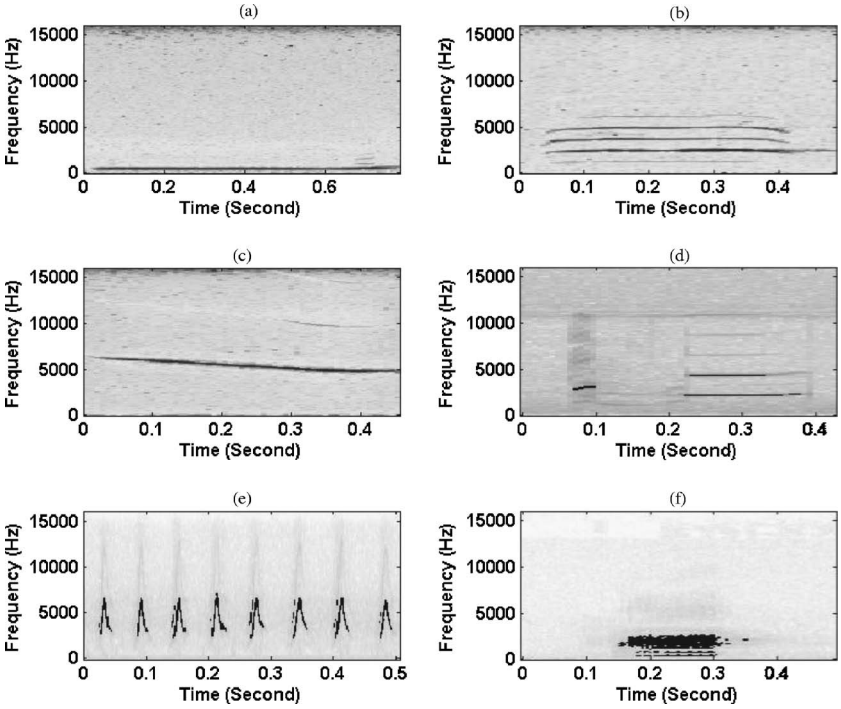

FIG. 2. Example time-variant spectra of bird sounds from several bird species. (a) Mourning dove (*Zenaida macroura*), (b) Red-tailed hawk (*Buteo jamaicensis*), (c) Red-winged blackbird (*Agelaius phoeniceus*), (d) Herring gull (*Larus argentatus*), (e) Brown-headed cowbird (*Molothrus ater*), (f) Mallard (*Anas platyrhynchos*).

to focus on syllable-level classification as the starting point for our proposed method. Restricting the process to single syllables allows the problem to be tractable, but a fully practical automatic classification system will need to accommodate the many different syllables associated with a particular bird species. Nevertheless, the characteristic strengths and weaknesses of our proposed classification method can be assessed using a single syllable study.

We observed the patterns and trends in our bird sound database and developed a set of discrimination parameters including spectral frequencies, frequency differences, track shape, spectral power, and track duration. The resulting framework uses measurements of the principal peak tracks in each sound syllable based on our observations and measurements of the recorded sound set.

The classifier *model choice* and *training* processes (described in the following section) consisted of selecting a representative syllable from each desired bird species in the database, deriving the features, or discrimination parameters, according to the basic peak track model, and then determining an allowable discrepancy between the derived features and the data for which a match would still be allowed.

The *evaluation* process, described in Sec. V, consisted of an iterative procedure comparing the error rate of the proposed classifier under a range of parameter range adjustments and additive noise conditions.

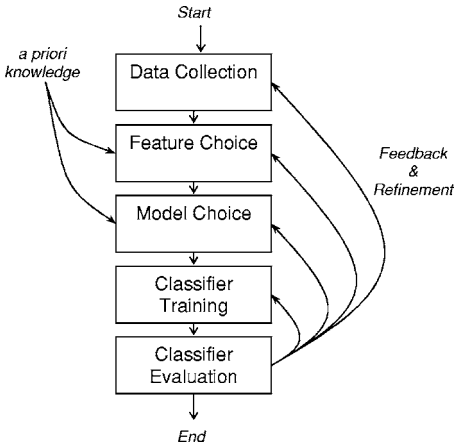

FIG. 3. Pattern classification design cycle (Duda *et al.*, 2001). Although numerous pattern matching procedures have been proposed and demonstrated, no universal system yet exists. Nonetheless, the design cycle depicted here is a useful common starting point for classification system design.

IV. DESCRIPTION OF THE SPECTRAL PEAK TRACKS METHOD

As mentioned above, spectral features have been found to be useful for bird syllable classification. Specifically, the use of spectral peak tracks provides a compact, distinctive, and computationally tractable basis for classification (Härmä, 2003). A spectral peak track is formed by segmenting the input signal into a set of overlapping short-time frames, calculating the Fourier transform magnitude for each short-time frame, and then matching peaks from one spectral magnitude frame to the next. If the input signal contains underlying sinusoidal partials, the resulting spectral peaks will persist for several successive frames. The corresponding peaks (e.g., peaks nearest in frequency on successive frames) are identified and linked from one frame to the next to form a connected sequence, known as a *peak track*, indicating the amplitude and frequency trajectory of the underlying partial. The vocal sounds of many birds are found to be well modeled by the peak track model (Härmä and Somervuo, 2004).

We extend the basic peak track method by including a variable number of peak tracks, which allows tonal, harmonic, or inharmonic combinations. We also enhance the

robustness of the peak track method by encoding the power and shape of the spectral peak tracks to capture the additional characteristic behavior of each vocalization.

In the proposed method we extract a variable number of spectral peak tracks from one syllable of the desired bird sound. The number of peak tracks selected is determined by the relative power of the tracks from the entire syllable. A set of descriptive parameters is then derived from the selected spectral peak tracks. These parameters include the frequencies, frequency differences, track shape, relative powers, and the track duration. We have found these parameters to work well in the presence of moderate noise and competing background sounds. The method also has low computational complexity and ease of implementation.

In the following sections we describe the three major steps of the spectral peak tracks method: spectral peak tracks search, feature extraction, and target and recognition.

A. Spectral peak tracks search

Step A1: The first processing step is to determine the spectral peak tracks of the syllable. A digital recording of the syllable with 16-bit samples and a 16 kHz sample rate is first high pass filtered with a 100 Hz cutoff in order to remove the low frequency background noise attributable to wind or mechanical sounds. The filtered signal is then segmented into Hamming windowed 256-sample frames with 128-sample overlap (8 ms frame hop). This is a reasonable frame rate based on the expected spacing of the spectral peaks. The *raw spectral representation* (i.e., the short-time discrete Fourier transform) is obtained with a Fast Fourier Transform (FFT) algorithm.

15 frames (120 ms) of the sample recording from the time interval prior to the start of the vocalization are used to estimate the background noise level. The average noise level for the 15 frames is calculated and used to set the amplitude threshold for the subsequent peak track detection. The estimated background noise level is also used to clarify the onset and release boundary of the syllable, and to distinguish the temporal boundaries of syllables with two or more parts. A simple energy-based detection algorithm is used to identify syllables with more than one part. If more than one part is found, only the part with the largest energy is retained, but the multipart detection is noted for use during the matching and recognition process.

Step A2: Next, the spectral peak tracks are derived using the McAulay and Quatieri (MQ) procedure (McAulay and Quatieri, 1986; Smith and Serra, 1987; Ellis, 2003). We refer to this step as the *coarse search*. In each frame the magnitude of the FFT data is examined to locate peaks in the spectrum. A peak is identified by three adjacent spectral magnitude coefficients within a frame where the middle coefficient is larger than both its higher and lower frequency neighboring coefficients, and the magnitude also exceeds the noise threshold determined in Step A1. This is done to remove peaks with very weak magnitude that are likely to be caused by background noise or by sidelobes of the Hamming window spectrum. We use a quadratic fit for the three spectral coefficients defining the peak in order to refine the frequency and magnitude estimate for the peak (Smith and Serra, 1987).

The refined peaks in one frame are compared to the peaks in the subsequent frame, and those spectral peaks that match well from one frame to the next are connected to form candidate peak tracks. A good peak-to-peak match is determined by locating a peak in the subsequent frame that is closest in frequency to each peak in the current frame. Acceptable matches are also restricted by a maximum rate of change in frequency corresponding to the most rapid frequency sweep found in actual bird syllables. We have found that a 200 Hz range works well for spectral peaks with frequencies above 2 kHz, while a 100 Hz range is used as the allowable frequency difference for peaks below 2 kHz. In some cases there will not be a suitable matching peak in the subsequent frame, while in other cases there may be several possible matches. Any conflicts are resolved by finding the match that minimizes the frequency and magnitude difference between the tentatively matched peaks. If no good match is found, we assume that the current peak is not continuous with any of the existing peak tracks, and the peak track associated with that peak is marked as "dead." Similarly, if a peak in the subsequent frame is not found to be a match for any of the peaks in the current frame, the new peak is considered as the "birth" of a new peak track. Thus, the coarse search process continues for each frame of spectral data, creating sets of candidate peak tracks.

Step A3: The spectral peak tracks produced by the coarse search in Step A2 may be discontinuous, too short, or otherwise inconsistent and poorly constructed. Therefore, a *fine search* is conducted to seek a higher level of structure in the syllable peak track data. We first identify any brief gaps (three or fewer frames) in the candidate peak tracks and reconnect across the gap under the assumption that the gap is due to noise or thresholding. Next, we prune away tracks that are inconsistent with the expected fundamental frequency range of the bird vocalization, such as tracks below 500 Hz. We then eliminate any closely spaced tracks, keeping the track with the larger magnitude, again under the assumption that the weaker track is due to spurious noise. Finally, we calculate the mean or median frequency, track length, and track power for every peak track and store the results of the fine search in the track analysis database.

Step A4: If the track list derived for the syllable has only one spectral peak track, the peak track search is terminated. If the track list includes two or more tracks, a second fine search is conducted. First, any relatively weak peak tracks are eliminated from the track list, thereby retaining only the tracks with the greatest signal to noise ratio. These strong peak tracks are sorted according to power and duration. If any of the strong peak tracks are found to be much shorter or longer in duration than the strongest peak track, or if the onset or offset of one of the weaker tracks is considerably different from the strongest peak track, those marginal tracks are eliminated from the track list. The resulting list of primary peak tracks (usually eight or fewer tracks) is retained for use in the matching procedure.

An example of the primary peak tracks for a harmonic syllable is shown in Fig. 4. The syllable has five spectral

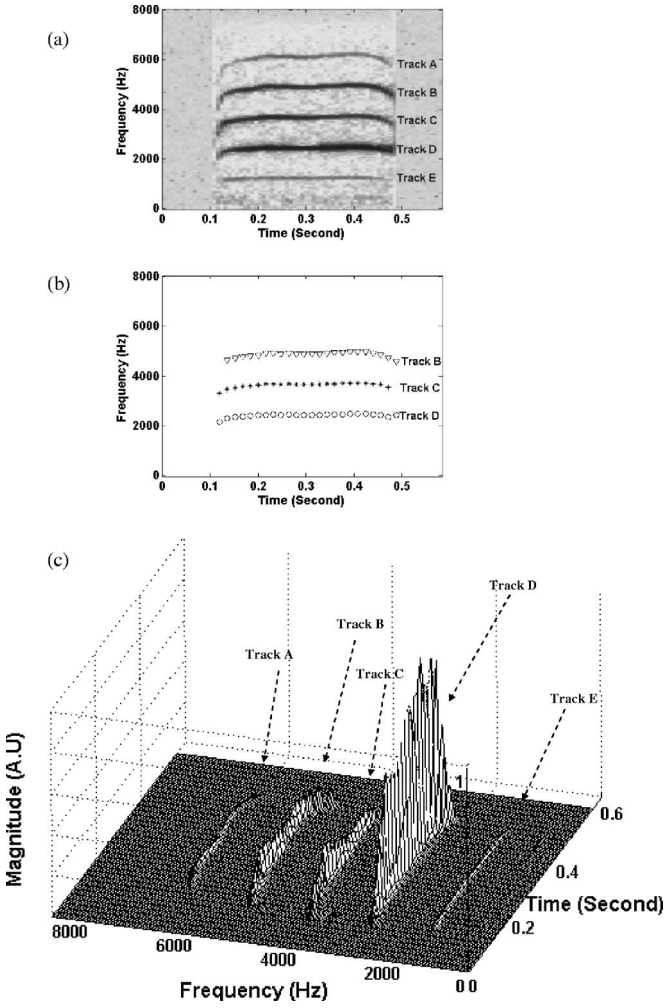

FIG. 4. Result of the spectral peak tracks search. (a) Frequency versus time representation (spectrogram) of one syllable of a red-tailed hawk (*Buteo jamaicensis*), (b) Spectrogram showing the search results from the spectral peak track method, (c) A three-dimensional representation of the spectrogram (amplitude versus frequency versus time).

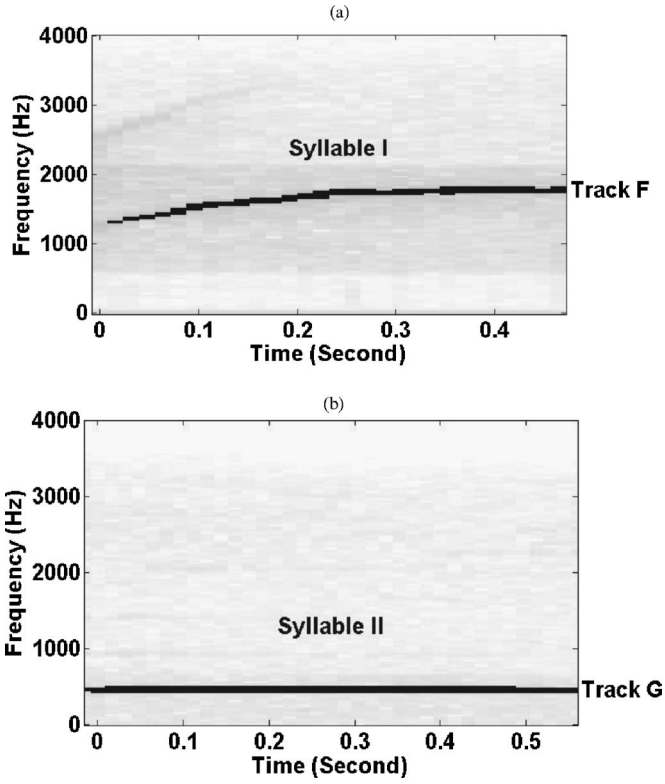

FIG. 5. Frequency range and shape example. (a) Peak track representation of a syllable from bird species I, (b) Peak track representation of a syllable from bird species II. The classification algorithm compares both the range (frequency extent) and the shape of the spectral peak tracks to improve the reliability of the syllable classification.

peak tracks. Among these tracks, Track D is the most significant, Track B is the next most significant, followed by Track C, Track A, and finally Track E. Since Tracks A and E are sufficiently weak to be immersed in the background noise, they are not retained in the search phase. Consequently, only Tracks D, B, and C are selected in the spectral peak tracks search phase.

B. Feature extraction

Once the primary spectral peak tracks are determined, we analyze the track contours to obtain a set of descriptive parameters. The parameters include the frequencies, the frequency differences, the relative power, the shape, and the duration of the spectral peak tracks. These parameters were selected based on several key insights. First, the peak track frequencies are observed to be distinctive for many of the bird species in our recorded database, and are therefore considered useful for pattern classification (see the example in Fig. 5). Second, we recognize that there are different bird species with syllables in overlapping spectral ranges, meaning that calculating only the mean frequency will be insufficient to distinguish between the species. In some cases the presence of several peak tracks with specific frequency relationships can be a key discriminator between species, while in other cases the shape and trend of the peak tracks may be the most important distinguishing feature. Finally, it may be that two different syllables are otherwise similar in frequency range, track frequency spacing, and track shape. In these cases we have found that comparisons between the relative power and the total duration of the tracks are necessary to obtain a reliable discrimination between the species.

For each primary track we first calculate the frequency at the start, end, midpoint, first quartile and third quartile point, which are F_1, F_2, F_3, F_4, and F_5 respectively. Next, we calculate the difference between the mean frequencies of the three primary tracks, which we denote as FD_1 and FD_2. Note that if there is only one primary track in the syllable, both FD_1 and FD_2 will be set to the mean frequency of the track, while if there are only two tracks in the syllable, both FD_1 and FD_2 will be set to the difference between the mean frequencies of these two tracks.

Next, a simple calculation is performed to describe the *shape* of each track. The shape is described by three parameters, S_1, S_2, and S_3. The first parameter, S_1, is expressed as

$$S_1 = \frac{F_3 - F_1}{\text{Frame\_Num\_1}}, \tag{1}$$

where F_1 is the starting frequency, F_3 is the midpoint frequency, and Frame_Num_1 is the number of frames be-

TABLE I. Output parameters for bird syllables. The first column represents the number of syllable parts and peak tracks. The second column represents the twelve parameters for each peak track.

Overall description	For each track
Number of parts in the syllable	Starting frequency (F_1)
Number of peak tracks	Ending frequency (F_2)
	Middle frequency (F_3)
	1st quartile frequency (F_4)
	3rd quartile frequency (F_5)
	Frequency difference (FD_1)
	Frequency difference (FD_2)
	Shape (S_1)
	Shape (S_2)
	Standard deviation (S_3)
	Relative power
	Duration

tween the start point and the middle point of the peak track. Similarly, the second parameter, S_2, is expressed as

$$S_2 = \frac{F_3 - F_2}{\text{Frame\_Num\_2}}, \tag{2}$$

where F_2 is the ending frequency and Frame_Num_2 is the number of frames between the endpoint and the middle point of the peak track. The third parameter, S_3, is the standard deviation of the frequency of the track calculated over the entire duration.

Finally, we also calculate the relative power of each track by determining the fraction of the total power in each frame due to the current track's peak, then averaging these power fractions over the length of the track.

The 12 parameters derived for each peak track have different units and magnitudes. Therefore, the parameters are normalized so that the parameter vector can be used in a distance calculation for classification. Of course, any number of mathematical functions could be invented for the distance calculation. We have developed a set of empirical coefficients that balance the variation of each dimension: scaling the frequency and frequency difference parameters by 62.5, scaling the shape parameters by 1.6, scaling the power by 2, and scaling the duration by 125 frames provides this balance.

The final step for feature extraction is to collect the derived parameters into a data structure ready for the subsequent matching process. Table I shows the output parameter summary.

C. Target and recognition

The target phase is conducted with representative vocalizations for each bird species to be tested. The representative vocalizations are intended to represent the desired syllable in the best possible manner, so the selection is conducted manually. The spectral peak tracks and features are extracted for each representative vocalization, creating a template for each bird syllable in the test set. These syllable templates then become the target database for the matching process, as described next.

The recognition phase begins with a recorded segment of an unknown bird sound. In our study the sound is assumed to be edited manually to contain a suitable segment for the matching process.

Step C1: Determine the number of syllable parts and peak tracks and store them in the output parameter list for the unknown sound.

Step C2: Calculate the *frequency distance* and the *number of syllable parts* match between the unknown sound and each template in the target database. The first five parameters (frequencies F_1 through F_5) are used in this step because in our experience the frequency information has been found to be the most important parameter of the matching process. The strongest peak track from the template is compared to each peak track in the test sound since the strongest peak track may differ from one example to the next even within the same species. If the number of syllable parts in the test sound differs from the number of parts in a template, the frequency distance is increased as a "penalty" to account for the lower likelihood of a match.

Step C3: (a) If the minimum distance between the unknown sound and at least one of the target templates is less than an empirically determined threshold (e.g., 8 Hz, or approximately 1% of the typical fundamental frequency component), we assert that a match has been found and the template with the minimum distance from the unknown sound is deemed to be the recognition result. Note that this small difference is unlikely, considering the noise and level differences inherent in real signals. Otherwise, the potential target matches are sorted and the two templates with the smallest distances from the unknown sound are identified for further comparison.

(b) Conversely, if the frequency distance between the unknown sound and one of the target templates is large, e.g., the mean frequency difference is greater than 800 Hz, the target template is classified as a very poor match to the unknown sound, and the recognition process proceeds with the next potential match.

(c) If the number of peak tracks in the unknown sound differs substantially (e.g., a difference greater than four) from the target template, the target is considered a poor match and the recognition process moves onto the next target template. For example, if the unknown sound is nearly a pure tone with a single peak track while a target template contains seven peak tracks, we assume the unknown sound cannot be the same bird species as the template.

Step C4: If the process has identified several candidate target templates based on the frequency distance calculations, now all 12 parameters are used to calculate the total weighted distance between the unknown sound and the candidate templates. We have found that better matching results are obtained if a penalty is included for mismatched tracks: 10% multiplied by the absolute track count difference between the unknown sound and the target template is added to the distance calculation.

The target template with the minimum distance is deemed the best match to the unknown sound.

V. SIMULATION AND DISCUSSION

The original raw simulation database of bird vocalizations contained sounds with varying formats and signal integrity. Some of the sounds came from recordings in the vicinity of an airport and therefore contain substantial background noise, while other sounds came from archival records with extremely low noise. The original sample rates ranged from 8 to 24 kHz, and both mono and stereo recordings were part of the raw database.

To facilitate the simulation all recordings were converted to monophonic, 16 kHz sample rate, and 16 bits per sample. For the purposes of verifying the performance of the peak track algorithm, recordings containing multiple syllables were manually edited to isolate a single syllable with a gap at the head and tail of the syllable. The rationale for this editing step was that if the algorithm failed under these optimal conditions it would not make sense to treat the continuous syllable case, nor the real time parameter extraction process.

A. Test database

Twelve bird species were part of the test database. Each bird species had 20 sound files, including similar syllables from separate raw recordings. The species included Mallard (*Anas platyrhynchos*), American Crow (*Corvus brachyrhynchos*), Canada Goose (*Branta canadensis*), Baltimore Oriole (*Icterus galbula*), Common Nighthawk (*Chordeiles minor*), Killdeer (*Charadrius vociferous*), Osprey (*Pandion haliaetus*), Northern Cardinal (*Cardinalis cardinalis*), Blue Jay (*Cyanocitta cristata*), Great Horned Owl (*Bubo virginianus*), Trumpeter Swan (*Cygnus buccinator*), and Herring Gull (*Larus argentatus*). These bird species are common in North America and relevant to the aviation and wind turbine bird strike avoidance problem. With the exception of the Mallard syllables, all the sounds contained strong tonal components that are well matched to the spectral peak track model. Despite the lack of predominantly tonal components in the Mallard examples, the method still achieved satisfactory results on the database, as described below.

As an additional assessment measure, we augmented the database with 16 synthesized syllables: 5 single tones (or chirps), 5 harmonic sounds, 5 inharmonic sounds, and 1 two-part syllable. The augmented database elements, or *distracters*, are deterministically generated and therefore provide a means to determine the sensitivity of the matching procedure. Thus, the total database contained 28 different syllable sets, designated Category 0 through Category 27, with 20 syllable recordings in each set. The bird species category designations are shown in Table II.

In this study, we first evaluated the database with two conventional classification methods used in speech recognition (Rabiner and Juang, 1993), then with the peak track method described in this paper.

B. Classification by DTW and HMM

The first conventional classification system was based on linear prediction cepstral coefficients (LPCC) and dynamic time warping (DTW) methods (Rabiner *et al.*, 1978).

TABLE II. Bird species category designations. Category 0 through 11 are natural bird species. Category 12 through 27 are synthetic test signals with deterministic parameters.

Category	Description
0	Mallard (*Anas platyrhynchos*)
1	American Crow (*Corvus brachyrhynchos*)
2	Canada Goose (*Branta canadensis*)
3	Baltimore Oriole (*Icterus galbula*)
4	Common Nighthawk (*Chordeiles minor*)
5	Killdeer (*Charadrius vociferous*)
6	Osprey (*Pandion haliaetus*)
7	Northern Cardinal (*Cardinalis cardinalis*)
8	Blue Jay (*Cyanocitta cristata*)
9	Great Horned Owl (*Bubo virginianus*)
10	Trumpeter Swan (*Cygnus buccinator*)
11	Herring Gull (*Larus argentatus*)
12	Single chirp. Frequency linearly increases.
13	Single chirp. Frequency linearly decreases.
14	Single chirp. Frequency linearly increases and then decreases.
15	Single chirp. Frequency linearly decreases and then increases
16	Single tone.
17	Harmonic chirp. Frequency linearly increases.
18	Harmonic chirp. Frequency linearly decreases.
19	Harmonic chirp. Frequency linearly increases and then decreases.
20	Harmonic chirp. Frequency linearly decreases and then increases
21	Harmonic tone.
22	Inharmonic chirp. Frequency linearly increases.
23	Inharmonic chirp. Frequency linearly decreases.
24	Inharmonic chirp. Frequency linearly increases and then decreases.
25	Inharmonic chirp. Frequency linearly decreases and then increases
26	Inharmonic tone.
27	One syllable has two parts: inharmonic tone plus inharmonic chirp.

The second conventional system was based on mel-frequency cepstral coefficients (MFCC) and hidden Markov models (HMM) (Rabiner, 1989).

The conventional classifiers have been successful for speech recognition because both the linear prediction model and the cepstral model are well suited to the human speech production mechanism. Specifically, the speech spectrum is characterized by a harmonic sequence attributable to the glottis excitation (vocal fold vibrations) and a set of relatively broad resonances (formants) due to the vocal tract. As mentioned in Sec. II, bird vocalizations are similarly produced by an excitation source that is spectrally shaped by resonances of the trachea, larynx, mouth, and beak. However, the differences in the excitation signal and the generally smaller dimensions of the bird vocal structure compared to the human speech production system result in more widely spaced spectral partials and more narrowly spaced resonances in typical bird sounds compared to human speech patterns. Thus, a conventional speech recognition technique will not necessarily work well with bird sounds, and this concern is borne out in the following test results.

1. Description of the classification system based on LPCC and DTW

Bird vocalization is a time-dependent process. Two similar bird syllables may have a different duration because these syllables may be pronounced at different rates. Consequently, a straightforward method that compares the value of the first syllable at time t to that of the second bird syllable at the same relative time may not correctly classify a given syllable. Instead, an algorithm must be used to search the space of mappings from the time sequence of the first bird syllable to that of the second bird syllable. Dynamic time warping is a standard technique used to perform time alignment of two syllables with different duration (Vintsyuk, 1971).

In order for the DTW technique to be tested for bird syllable classification, a reference template must first be chosen and stored for every bird species. The classification process entails matching the incoming bird syllable with the stored templates. The templates and the incoming bird syllable are represented as a sequence of parameter vectors and the best matching template is the one that exhibits the minimum path aligning the input bird syllable to the template. The search space for the DTW method is constrained in such a way that the mapping function between the time axis of the input signal and the time axis of the template must be monotonically increasing with time so that the ordering of events in both the input and the template are preserved. The global distance score for a mapping path is simply the sum of local distances that make up the path.

The first preprocessing step in using DTW is the creation of bird syllable templates to be identified from the input bird syllable. In the current study, templates were selected manually to achieve the best recognition results. The second preprocessing step is to extract feature vectors for the templates and the input bird syllable. In this step, a digital recording of the syllable with 16-bit samples and a 16 kHz sample rate is first high pass filtered with a 100 Hz cutoff in order to remove the low frequency background noise attributable to wind or mechanical sounds. The filtered signal is then segmented into Hamming windowed 320-sample (20 ms) frames with 160-sample overlap (10 ms frame hop). The energy-based detection method mentioned in Sec. IV is used to detect the onset and offset of the sound signal. The frames between the onset and offset are transformed into the feature vectors on a frame-by-frame basis. Thirteen LPC coefficients $[a_1\ a_2 \ldots a_P]$ ($P=13$) are calculated using Durbin's recursive algorithm for every frame. To improve the recognition accuracy, the LPC coefficients are transformed into 18 LPCC coefficients $[c_1\ c_2 \ldots c_P \ldots c_L]$ ($L=18$) using the following equations:

$$c_1 = a_1, \tag{3}$$

$$c_n = a_n + \sum_{k=1}^{n-1} \left[\frac{n-k}{n} \cdot a_k \cdot c_{(n-k)} \right], \quad 2 \le n \le P, \tag{4}$$

$$c_n = \sum_{k=1}^{P} \left[\frac{n-k}{n} \cdot a_k \cdot c_{(n-k)} \right], \quad P+1 \le n \le L. \tag{5}$$

Note that the first order LPCC coefficient represents the spectral energy, which is generally not normalized between input signals, so c_1 is not used for the distance computation described next.

After generating the feature vectors for the templates and input bird syllables, the comparison is made by nonuniformly adjusting the time axis of the input syllable to achieve the best match to the template. In the matching process for an input bird syllable with M frames and a template with N frames, the time frames of the input bird syllable and the time frames of the templates are organized in a lattice (i,j), where i and j are the indexes of the time frames of the input syllable and the template, respectively. The quality of the match is measured recursively by the formula

$$D(i,j) = \min[D(i-1,j-1), D(i-1,j), D(i,j-1)] + d(i,j), \tag{6}$$

where $d(i,j)$ is the Euclidean distance between the two multidimensional vectors of the input signal at time frame i and the template at time frame j, and therefore $D(i,j)$ is the global distance up to (i,j). Given the initial condition $D(1,1)=d(1,1)$, we have the basis for an efficient recursive algorithm for computing $D(i,j)$. The final global distance $D(M,N)$ gives us the overall matching score of the template with the input. The input syllable is then classified as the species corresponding to the template with the lowest matching score.

2. Description of the classification system based on MFCC and HMM

The Hidden Markov Model is a doubly stochastic process with an underlying "hidden" stochastic process that is not observable, but can only be observed through another set of stochastic processes that produce the sequence of observed symbols (Rabiner, 1989). The HMM comprises a finite set of states, each of which is associated with a specific probability distribution. Transitions among the states are governed by a set of transition probabilities. In a particular state an observation can be generated according to the associated probability distribution.

The HMM can be used to model the bird sound generation statistically. It has been used for bird sound recognition due to its ability to characterize bird sounds in a mathematically tractable manner (Kogan and Margoliash, 1998). In contrast to the deterministic template matching of DTW, HMM uses a statistical representation. Therefore, this model can accumulate more information and possibly generalize better than techniques based on fixed templates.

The HMM procedure implemented for this study includes the same preprocessing as for the DTW procedure: 16-bit samples, 16 kHz sample rate, and 100 Hz high pass filtering, and the energy-based onset/offset detection described in Sec. IV. The reference templates are selected manually and include three sound files from each Category as the training cadre. Each preprocessed signal is segmented

TABLE III. Performance of a classification system based on linear prediction cepstral coefficients (LPCC) and dynamic time warping (DTW) methods (23/240 means that the error count was 23 out of 240).

	12 natural sounds		*28 total sounds*	
(SNR) (dB)	*Noisy training set*	*Clean training set*	*Noisy training set*	*Clean training set*
Original	23/240	23/240	33/560	33/560
30	23/240	24/240	33/560	36/560
24	27/240	28/240	38/560	40/560
18	30/240	34/240	43/560	48/560
12	38/240	40/240	54/560	57/560
9	50/240	61/240	68/560	82/560
6	59/240	90/240	81/560	116/560
3	68/240	115/240	93/560	148/560

TABLE IV. Performance of a classification system based on the mel-frequency cepstral coefficients (MFCC) and hidden Markov models (HMM) (12/240 means that the error count was 12 out of 240).

	12 *natural sounds*		28 *total sounds*	
SNR (dB)	*Noisy training set*	*Clean training set*	*Noisy training set*	*Clean training set*
Original	12/240	12/240	20/560	20/560
30	12/240	12/240	20/560	21/560
24	15/240	16/240	23/560	25/560
18	17/240	19/240	27/560	30/560
12	23/240	28/240	33/560	39/560
9	34/240	51/240	45/560	65/560
6	45/240	83/240	62/560	99/560
3	58/240	126/240	87/560	158/560

into Hamming windowed 256-sample (16 ms) frames with 128-sample overlap (8 ms frame hop). The frames between the onset and offset are transformed into 13-dimensional mel-frequency cepstral coefficient feature vectors using 26 mel filterbanks, on a frame-by-frame basis. The MFCC coefficients are a compact representation, which is the result of a cosine transform of the real logarithm of the short-term energy spectrum expressed on a mel-frequency scale (Zheng *et al.*, 2001). As with the first order LPCC coefficient in the DTW method, the first order MFCC coefficient is not considered in the distance computation.

A continuous HMM representation was used in the traditional HMM-based classification system employed for this study. The HMM was a left-to-right model with six states. The densities of the observation probabilities in the emitting states were modeled as mixtures of two multidimensional Gaussian distributions with diagonal covariance matrices. The Baum-Welch algorithm was used in the training step employing the three manually selected bird syllables. One set of HMM parameters were generated for every bird species. The trained HMM parameters were then used for recognition using the Viterbi algorithm (Viterbi, 1967).

3. Simulation tests using the conventional classifiers

The sensitivity of the classification algorithms to noise was evaluated with an additional group of simulations. For each simulation the 560 example files (including the Template files) were deliberately contaminated with separate segments of uniformly distributed (white) noise to achieve a specified signal-to-noise ratio (SNR). The simulations included SNRs from 30 to 3 dB in steps of 3 or 6 dB.

A more likely practical situation would occur when using a set of "clean" Templates but testing with noisy Category recordings, i.e., field recordings that would ordinarily be obtained under less than ideal conditions. Thus, a separate simulation was used to examine this condition.

The DTW performance on the bird sound database is shown in Table III. The results indicate a likelihood of error for DTW with the original recorded signals of approximately 1/10 (i.e., 90% correct matches), with performance degrading to 71% correct classification under low signal-to-noise conditions.

The performance on the sound database for the conventional HMM system is shown in Table IV. The results show a likelihood of error for HMM of approximately 1/20 (95% correct matches) under high signal-to-noise conditions, degrading to 76% correct with noisy data.

Although the conventional procedures are moderately successful on the sound database used in this study and correspond well to prior results reported in the literature, the performance degradation with decreasing SNR is a serious shortcoming for a practical, robust system.

C. Classification by the spectral peak track method

The spectral peak track method requires a reference template for each bird species in the database. One representative syllable from each of the 28 Categories was selected manually, designated Template 0 through Template 27, and used as the training cadre.

The first test of the peak track procedures described in Sec. IV was to distinguish between single part and two-part syllables. The simple energy based detection algorithm performs well with the bird sounds in the database: all 560 example sounds were correctly classified. However, only 40 examples (Category 11 and Category 27) contained two-part syllables, so additional verification will be needed before assuming the simple algorithm is sufficient.

In the classification experiment, each of the 560 syllable recordings was processed using the peak track method described above. Here 238 out of 240 natural bird sounds were correctly classified (error rate 2/240), while the error rate was 5/560 (99% correct matches) for the entire natural and synthetic database at a high SNR. These results indicate that the peak track method was well suited to this particular database.

The five misclassified examples were examined. The first misclassification was a Category 3 example mistaken for Category 0. The particular example contained a frequency change from 3000 to 2500 Hz, while the Template 3 reference had a 3000 to 2000 Hz change. The resulting distance was closer to Template 0, causing the mismatch.

The second error was due to an example from Category 0 being mistaken for Category 1. In this case there are two peak tracks in Template 0, while the processed input example was found to have three significant peak tracks. The differing

TABLE V. Performance of the spectral peak tracks method (2/240 means that the error count was 2 out of 240).

	12 *natural sounds*		28 *total sounds*	
SNR (dB)	*Noisy training set*	*Clean training set*	*Noisy training set*	*Clean training set*
Original	2/240	2/240	5/560	5/560
30	3/240	3/240	6/560	6/560
24	5/240	5/240	8/560	8/560
18	4/240	4/240	7/560	7/560
12	3/240	3/240	6/560	6/560
9	6/240	6/240	9/560	9/560
6	8/240	14/240	9/560	14/560
3	11/240	18/240	11/560	18/560

number of peak tracks caused this example to have a greater distance from Template 0 than from Template 1 that has three peak tracks, therefore causing the misclassification.

The other errors were because of examples from Category 11 being mistaken for Category 27. The Category 11 examples are a natural bird sounds, while Category 27 contains synthesized examples with relatively strong peak track powers that were included specifically as distracters (successful in triggering a misclassification in this case). The Template 11 data contains two peak tracks, while the processed input examples were found to have four significant peak tracks. The differing number of peak tracks caused these examples to have a greater distance from Template 11 than from Template 27, thereby causing the misclassifications.

The classification performance degraded as the SNR decreased from 30 to 24 dB. When the SNR decreased from 24 to 12 dB, the classification accuracy unexpectedly improved, despite the degraded signal quality. An examination of the Category 11 signal revealed that the additive noise slightly altered the relative power of the peak tracks, causing the distance between the Category 11 signal and Template 11 to be less than the distance between Category 11 and Template 27. A similar subtle change occurred between a Category 3 signal and Template 0, causing a correct classification. Decreasing the SNR to 9, 6, and 3 dB degraded the classification accuracy, although the overall performance was still good (95% correct at 3 dB SNR), as shown in Table V.

When using "clean" Templates but testing with noisy Category recordings, the classification accuracy was unaffected until the SNR decreased to 6 and 3 dB. The performance differed between the noisy and clean Template cases, but not in a significant manner. Thus, based on this simulated noise test the peak track method does not appear to be particularly sensitive to the SNR difference between the Template and the test samples.

D. Discussion

As indicated by the results in Tables III, IV, and V, the spectral peak track method provided better overall results than the conventional DTW and HMM methods, particularly in the low signal-to-noise ratio tests. There are two expected reasons for this result. First, the conventional methods based on a linear prediction model have difficulty with the sparse spectrum of the bird syllables in the test database: insufficient information is present in each syllable recording to create a unique and easily distinguishable LPC model. The frequency spacing of the spectral components is relatively wide compared to the underlying resonances of the bird vocal tract, indicating that the linear prediction methods are mismatched to the signals in the database. Second, the conventional model parameters are quite sensitive to existing background noise, reverberation, and competing sounds in the recordings. The results in Tables III and IV for the 12 natural sounds and for the 28 total sounds show that the conventional systems operated better on the artificial distracter signals than on the natural sounds with inherent background noise. For these reasons it is useful to consider alternative pattern classification techniques, such as the spectral peak track method presented in this paper.

VI. CONCLUSION

The spectral peak track method described in this paper appears to work as designed for isolated bird syllables, and the simulation results are better in comparison to conventional DTW and HMM methods used to classify the same database. The proposed method extends the prior peak track methods by using a variable number of tracks to represent each syllable by determining the relative power of each detected spectral track. This method accommodates tonal, harmonic, or inharmonic syllables, and bases the pattern classification on the strongest tracks present in each syllable. The set of parameters derived for each significant peak track was quite robust in the presence of simulated additive noise, which is an encouraging result for future applications of this technique in classifying bird vocalizations.

However, the proposed method is inappropriate for use with bird vocalizations containing aperiodic or noise-like components because the assumption of connected peak tracks is violated in these cases. The proposed method is also inappropriate if the underlying spectral components change too rapidly in frequency or fluctuate in amplitude such that the peak tracks cannot be determined reliably. Research is continuing on methods to classify such rapidly varying signals, and also to identify broadband and noisy sounds.

In the current system we manually extract one syllable from the recorded bird sounds and save the data in a separate sound file. However, in an on-line real time production application it will be necessary to demonstrate an automatic syllable extraction method. Thus, there remain a variety of engineering challenges to deploying this system for real-time classification.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the assistance of three anonymous reviewers in improving this manuscript. The work described in this paper was supported by Advanced Acoustic Concepts, Inc.

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

Catchpole, C. K., and Slater, P. J. B. (**1995**). *Bird Song: Biological Themes and Variation* (Cambridge University Press, Cambridge, UK).
Duda, R. O., Hart, P. E., and Stork, D. G. (**2001**). *Pattern Classification*, 2nd ed. (Wiley, New York).
Ellis, D. P. W. (**2003**). "Sinewave and sinusoid+noise analysis/synthesis in MATLAB," Electronic document, URL: ⟨http://www.ee.columbia.edu/~dpwe/resources/matlab/sinemodel⟩.
Fagerlund, S. (**2004**). "Automatic recognition of bird species by their sounds," Masters thesis, Laboratory of Acoustics and Audio Signal Processing, Helsinki Univ. of Technology, Laboratory of Acoustics and Audio Signal Processing.
Härmä, A. (**2003**). "Automatic identification of bird species based on sinusoidal modeling of a syllable," *IEEE Int. Conf. Acoust. Speech and Signal Processing (ICASSP 2003)*, **5**, 545–548.
Härmä, A., and Somervuo, P. (**2004**). "Classification of the harmonic structure in bird vocalization," *IEEE Int. Conf. Acoust. Speech, Signal Processing (ICASSP 2004)*, **5**, 701–704.
Ito, K., Mori, K., and Iwasaki, S. (**1996**). "Application of dynamic programming matching to classification of budgerigar contact calls," J. Acoust. Soc. Am. **100**, 3947–3956.
Kahrs, M., and Avanzini, F. (**2001**). "Computer synthesis of bird songs and calls," *Proc. Conf. Digital Audio Effects (DAFx-01)*, pp. 23–27.
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.
Krebs, J. R., and Kroodsma, D. E. (**1980**). "Repertoires and geographical variation in bird song," Journal of Advances in the Study of Behavior **11**, 143–177.
Kroodsma, D. E., and Miller, E. H. (**1996**). *Ecology and Evolution of Acoustic Communication in Birds* (Comstock, Ithaca, NY).
McAulay, R. J., and Quatieri, T. F. (**1986**). "Speech analysis/synthesis based on a sinusoidal representation," IEEE Trans. Acoust., Speech, Signal Process. **34**, 744–754.
McIlraith, A. L., and Card, H. C. (**1997**). "Birdsong recognition using backpropagation and multivariate statistics," IEEE Trans. Signal Process. **45**, 2740–2748.
National Wind Coordinating Committee (NWCC) (**2004**). "Wind-turbine interactions with birds and bats: a summary of research results and remaining questions," RESOLVE, Washington, DC.
Nowicki, S. (**1997**). "Bird acoustics," in *Encyclopedia of Acoustics*, edited by M. J. Crocker (Wiley, New York), Chap. 150, pp. 1813–1817.
Pascarelle, S. M., Pinezich, J., Merritt, R. L., Kelly, T. A., Roman, B., and Maher, R. C. (**2004**). "Automated acoustic monitoring of bird strike hazards," *6th Annual Meeting of the Bird Strike Committee USA/Canada*, Baltimore, MD, September, **2004**.
Rabiner, L. R., Rosenberg, A. E., and Levinson, S. E. (**1978**). "Considerations in dynamic time warping algorithms for discrete word recognition," IEEE Trans. Acoust., Speech, Signal Process. **26**, 575–582.
Rabiner, L. R. (**1989**). "A tutorial on hidden markov models and selected applications in speech recognition," Proc. IEEE **77**, 257–286.
Rabiner, L. R., and Juang, B. H. (**1993**). *Fundamentals of Speech Recognition* (Prentice–Hall, Englewood Cliffs, NJ).
Smith, J. O., and Serra, X. (**1987**). "PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation," *Proc. Int. Computer Music Conf.*, San Francisco, Computer Music Association.
Vintsyuk, T. K. (**1971**). "Element-wise recognition of continuous speech composed of words from a specified dictionary," Journal of Cybernetics and Systems Analysis, **7**(2), 361–372.
Viterbi, A. J. (**1967**). "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," IEEE Trans. Inf. Theory **13**, 260–269.
Zheng, F., Zhang, G. L., and Song, Z. J. (**2001**). "Comparison of different implementations of MFCC," Journal of Computer Science & Technology, **16**(6), 582–589.