Template-based automatic recognition of birdsong syllables from continuous recordings

Sven E. Anderson, Amish S. Dave, and Daniel Margoliash
Department of Organismal Biology and Anatomy, 1027 East 57th Street, The University of Chicago, Chicago, Illinois 60637

(Received 31 August 1995; accepted for publication 26 February 1996)

The application of dynamic time warping (DTW) to the automated analysis of continuous recordings of animal vocalizations is evaluated. The DTW algorithm compares an input signal with a set of predefined templates representative of categories chosen by the investigator. It directly compares signal spectrograms, and identifies constituents and constituent boundaries, thus permitting the identification of a broad range of signals and signal components. When applied to vocalizations of an indigo bunting (*Passerina cyanea*) and a zebra finch (*Taeniopygia guttata*) collected from a low-clutter, low-noise environment, the recognizer identifies syllables in stereotyped songs and calls with greater than 97% accuracy. Syllables of the more variable and lower amplitude indigo bunting plastic song are identified with approximately 84% accuracy. Under restricted recording conditions, this technique apparently has general applicability to analysis of a variety of animal vocalizations and can dramatically decrease the amount of time spent on manual identification of vocalizations. © *1996 Acoustical Society of America*.

PACS numbers: 43.80.Ka, 43.72.Ne, 43.60.Lq [FD]

INTRODUCTION

The sound spectrograph is the principal tool used in analysis of bioacoustic signals. Based on visual inspection of sound spectrograms, most animal vocalizations are described as comprising discrete subunits in a hierarchical organization. This is reflected, for example, in a popular terminology for birdsong vocalizations: notes, syllables (or figures), phrases, and songs. Many bioacoustic studies require a description of the similarity of different units of vocalization within an animal (either as related to some experimental manipulation or during ontogeny), or the similarity of units of vocalization across individuals.

In most longitudinal studies, the analysis of sound spectrograms is largely based on manual inspection (e.g., Marler and Peters, 1982; Margoliash *et al*., 1994). This has the obvious problems of repeatability, subjectivity, and for large data sets can be extremely labor intensive. The last point is significant, in that it has inhibited research in cases such as vocal development in songbirds, which can span many months during which a vast number of songs are produced. In other cases, units of vocalizations are algorithmically compared by measuring heuristically chosen parameters (in some cases subjected to *post-hoc* analysis such as factor analysis), but these parameters may not reflect the overall ''shape'' or morphology of the vocalization. It is this overall shape that the human observer apparently uses in classifying vocal units.

Several studies have used computational techniques to analyze segmented vocalizations. Song syllable similarity has been measured using cross correlation of spectrograms (Clark *et al*., 1987) and dynamic programming (Williams, 1993). Buck and Tyack (1993) have demonstrated that single dolphin signature whistles can be recognized using dynamic time warping (DTW). These studies are representative in that they focus on the analysis or recognition of single vocalizations in isolation. In this study we focus instead on the automatic recognition of vocalization subcomponents from continuous recordings. We applied an existing DTW algorithm developed for speech recognition to the recognition of syllables and calls. The algorithm is particularly well suited to the recognition of vocalizations comprising highly regular units that may occur in unknown order. The vocalizations of many species are naturally segmented, and may meet these criteria far better than does speech; hence, DTW can be expected to perform better for such animal vocalizations than for human speech. It should be noted that the DTW algorithm was applied to automatically analyze unknown signals only after categories were first manually defined by human experts choosing representative tokens (''templates'') from the vocalizations. Thus a level of subjectivity remains (see Sec. IV).

We have tested the system with two species of song birds: indigo buntings (*Passerina cyanea*), which have relatively spectrally simple vocalizations, and zebra finches (*Taeniopygia guttata*), which have broadband vocalizations. Both species produce vocalizations comprising elements delimited by brief silent intervals. In addition, both species produce predominantly stereotyped vocalizations, but also exhibit vocalizations that are much more variable. All recordings were made in a low-noise, low-clutter environment, and the present results are only appropriate in such contexts. In what follows, we first describe the DTW algorithm. We then report syllable accuracy results for the two species tested, and assess the dependence of recognizer performance on template set size. Finally, the sensitivity of the recognizer to variation in the spectral and temporal resolution of the algorithm is explored.

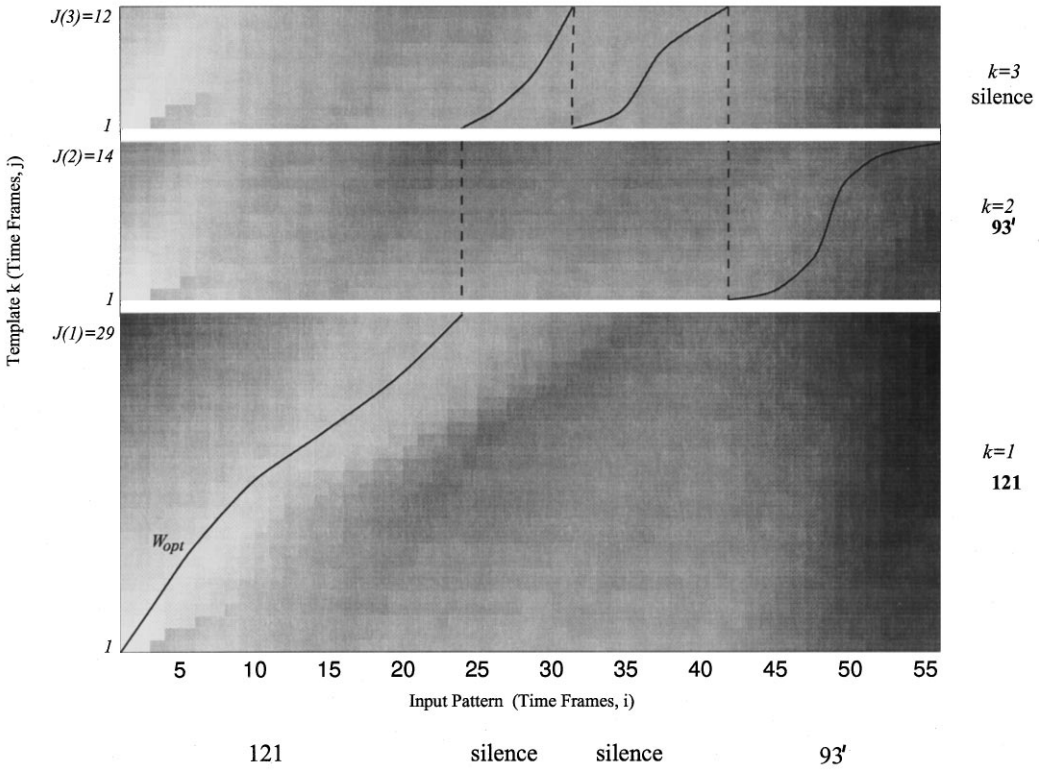

FIG. 1. The optimal path determined by the one-stage algorithm. The optimal warping path (W_{opt}) for the recognition of the sequence 121-93′ using the set of templates {121, 93′ and, (silence)} is overlaid on a representation of the cumulative distance for each grid point. The darkness of grid points increases with increasing values of the cumulative distance function, $D(i,j,k)$. The optimal path determines the sequence of templates as well as segment onset and offset times, and appears as the lightest path through the grid. Note the repeated match of the silence template between the two syllables.

I. THE ONE-STAGE DTW ALGORITHM

Our song recognition system is based on the one-stage (parallel) DTW algorithm introduced by Vintsyuk (1971) and developed by Bridle *et al*. (1982). The algorithm segments an input pattern using a prespecified set of template patterns. In the one-stage algorithm, recognition of continuous input is recast into the problem of specifying a sequence of template patterns and nonuniform compression/dilation of the templates' time axes to achieve the best match to the input pattern. Other DTW methods applicable to the recognition of sequences of nonoverlapping templates exist (e.g., level building: see Myers and Rabiner, 1981; phrase-level matching: see Kato, 1980). The one-stage algorithm was selected because unlike other methods it performs a one-stage, parallel search over all templates and is, consequently, the most computationally efficient DTW algorithm (Ney, 1984; Silverman and Morgan, 1990).

The one-stage algorithm is applicable to the recognition of acoustic signals represented as sequences of multidimensional ''feature vectors'' (e.g., digital spectrograms, cepstral coefficients). Consider an input pattern consisting of N frames of feature vectors. In this context the goal of connected song recognition is to determine the sequence and timing of templates $\{T_1,T_2,\ldots,T_n\}$ that best match an input pattern. We assume that the input pattern is a nonoverlapping sequence of distorted template patterns; that is, the input pattern represents the nonoverlapping vocalizations of an individual animal. Signal ''distortion'' may arise from natural within-class variation as well as from ambient noise and uncontrolled variation in recordings.

We now summarize the one-stage algorithm as presented in Ney (1984). The time frames of the input pattern and time frames of the template patterns are used to define a set of grid points (i,j,k) as in Fig. 1. Here i indexes the time frames of the input pattern, j indexes the time frames of a single template, and k indexes the templates. Continuous paths among points of the grid that begin at the beginning of the input pattern and finish at the end of the input pattern determine a potential alignment between the input pattern and templates. We write an arbitrary path of length L indexed by r as $W^r=(w^r(1),w^r(2),\ldots,w^r(L))$ where $w^r(l) = (w_i^r(l),w_j^r(l),w_k^r(l))$ and l indexes the ordered steps along the path through the grid points (i,j,k). With each point (i,j,k) we can associate a local distance measure, $d(i,j,k)$, that measures the difference between time frame i of the input pattern, and time frame j of template k. The optimal path is found by making a series of locally optimal decisions [see Silverman and Morgan (1990) for discussion and examples]. Finding the best match between input pattern and templates is equivalent to finding the path that minimizes $\Sigma_l d(w(l))$ over the set of all allowed paths.

Let the number of templates be K, and $J(k)$ be the length of template k. For generality, in this section the frame rate (number of frames per second) for the input pattern and template patterns is unspecified. During testing, the frame rate was varied to determine its effect on accuracy (see Sec.

III). At each point (i,j,k), the minimum cumulative distance along any path from the beginning of the input pattern (and a template that is unknown beforehand) to (i,j,k) is $D(i,j,k)$. A set of constraints linking each $w(l-1)$ to $w(l)$ is adopted to ensure reasonable paths. The particular constraints employed allow a template and potential match within the input pattern to differ in duration by no more than a factor of 2 (Itakura, 1975). These are embodied by rules [Eq. (1)] that describe how the minimum cumulative path length $D(i,j,k)$ at $w(l)$ is iteratively computed using the local distance measure. Note that the constraints disallow paths that omit consecutive frames of the input or current template, the global effect being to limit paths within a parallelogram having sides with slopes one-half and two:

$$D(i,j,k)=d(i,j,k)+\min\begin{cases} D(i-1,j,k), & \text{iff } w_j(l-1)\neq w_j(l-2),\\ D(i-1,j-1,k),\\ D(i-1,j-2,k).\end{cases} \tag{1}$$

The local constraint does not warp two patterns in a symmetric manner; thus, for two arbitrary patterns A and B, the cumulative distance measure is not symmetric $(D(A,B)\neq D(B,A))$.

At template onsets, paths are constrained to arrive from the ends of other templates [Eq. (2)].

$$D(i,1,k)=d(i,1,k)+\min\begin{cases} D(i-1,1,k), & \text{iff } w_j(l-1)\neq w_j(l-2)\\ D(i-1,J(k^*),k^*): & k^*=1,2,\ldots,K.\end{cases} \tag{2}$$

All results presented here used the Euclidean metric to determine the distance between feature vectors. For sound spectrograms, the distance between two spectra was computed as the square root of the sum of the squared differences of log magnitudes at each frequency bin.

Once the minimum distance path from the beginning to the end of the input pattern $(N,J(\tilde{k}),\tilde{k})$ has been determined, it can be retraced from end to beginning, thereby determining the optimal sequence of matches between templates and input pattern. The complete constituent sequence can then be determined during the backtracking phase. During the calculation of $D(N,J(\tilde{k}),\tilde{k})$, for each frame of the input pattern, a record is kept of the template ending having minimum cumulative distance at that frame $(T(l))$. An array $B(i,j,k)$ also stores the frame at which that optimal partial path entered the template. The backtracking procedure then proceeds recursively from the template having minimal distance, to its predecessor, and so on. The Appendix presents a concise outline of how the algorithm is used to segment an input pattern into a sequence of template matches.

II. METHODS

A. Song collection

Acoustic data were obtained from animals housed individually in 45×35×25 cm wire cages placed in sound attenuation booths (IAC No. AC-1). The vocalization acquisition system has been previously described (Margoliash *et al.*, 1991, 1994). Prior to digitization, vocalizations were filtered with four-pole low-pass and high-pass filters, achieving a frequency response of the electronics of ±0.5 dB from 1 to 9.5 kHz, and −24 dB below 500 Hz and above 20 kHz. The incomplete high-frequency filtering resulted in some aliasing of the signal that is apparent in the spectral representation of digitized vocalizations. Vocalizations were collected either by manual triggering or automatically detected by a computer system that continuously analyzes the signal zero-crossings and amplitudes. ''Qualifying'' vocalizations were sampled at 20 kHz with 15-bit resolution. Each resulting waveform resides in a file that represents from several seconds to about 100 s of singing. These files may include numerous vocalizations, as well as noises caused by movement, beak wiping, etc., separated by periods of silence.

B. Preprocessing and recognition

The most significant preprocessing step is the creation of templates for the segments to be identified from the input stream. Selection of templates depends on the sound environment in which recognition will take place. In the present study the sound environment included song syllables, calls, noises, and periods of silence. We chose a set of templates that included exemplars of all salient types of sounds encountered. In addition to templates for vocalizations, we created templates for a variety of ''silent'' intervals of background noise as well commonly occurring noises.

Template waveforms were manually edited from several songs using the WAVES+ (Entropic Research Laboratory, Inc.) signal-processing software. Templates were then created from the spectra of the template waveforms computed using the fast Fourier transform (FFT). Frequencies below 500 Hz (where vocalizations have little power) were omitted to remove fan and other low-frequency noise evident in much of the data. Feature vectors are thus the log magnitude FFT bins from 0.5 to 10 kHz, each of which is weighted equally by the local distance measure. We explored the sensitivity to choice of templates and template parameters including FFT size and frame rate; in some cases, the size of the spectral representation was reduced by linearly reinterpolating the FFT.

The initial preprocessing component of the recognition system isolated vocalizations in the input stream on the basis of signal average magnitude and the durations of preceding and following silence. Such preprocessing is convenient for eliminating long periods of silence but is not necessary. Vocalization onsets were detected when the average magnitude exceeded a threshold (two standard deviations above the mean of the background level) twice in a specified duration corresponding to the shortest expected vocalization [40 ms for the zebra finch (ZF001) and 20 ms for the indigo bunting (IB007)]. Following detection of an onset, vocalization off-

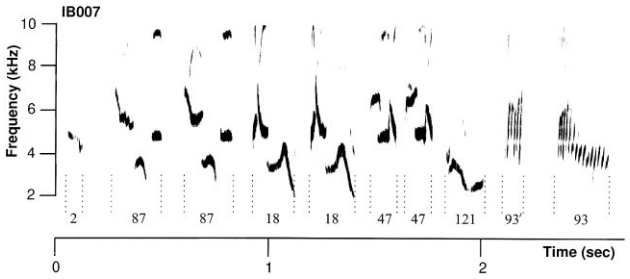

FIG. 2. One song of IB007 with syllable identifiers and boundaries as determined by the recognizer. Syllable onsets and offsets are indicated by dotted lines. For clarity, the intersyllable ''silence'' identifications have been removed. The onsets and offsets are close but not identical to those derived from manual scoring.

sets were assumed if the average magnitude was subthreshold for 400 ms. Spectral representations of the input patterns were then computed with FFTs using the same parameters as those used to create templates. The output of this module is the basis for connected song recognition reported here.

Song recognition occurred off-line on several different computer systems [Sun Microsystems Sparcstation 2, Silicon Graphics, Inc. (SGI) Challenge, and SGI 4*D*/380]. Recognizer output is a file of onset and offset times and associated segment identifiers represented graphically as in Fig. 2.

C. Scoring

The performance of the recognizer was compared against a baseline of manually labeled vocalizations. For manual labeling, constituent syllables were identified by experts simultaneously viewing digitally computed song spectrograms and waveforms displayed on a computer monitor. These procedures are well established in our laboratory (see Margoliash *et al*., 1991, 1994). For zebra finches, vocalization units of song (''notes'' and ''syllables,'' see Sossinka and Böhner, 1980) tend to be highly stereotyped and are unambiguously classified, whereas calls produced in isolation are often graded and can only be classified into one or a few broad categories. The vocalizations of zebra finch ZF001 automatically analyzed here had been previously subjected to exhaustive manual analysis in the context of a neurobiological experiment (Yu and Margoliash, 1995). Indigo buntings produce both ''stereotyped'' and ''plastic'' songs, the latter exhibiting more variable syllables, ''*S* notes,'' and indistinct figures (see Margoliash *et al*., 1991). The analysis of indigo bunting songs was aided by the extended catalog of syllable types (Thompson, 1970; Payne and Payne, unpublished data). The bunting vocalizations analyzed here resulted from a bird (IB007) whose vocalizations were extensively analyzed and subjected to cross verification by human experts (Margoliash *et al*., 1991, 1994).

Scoring relies on matching onset and offset times for each manually labeled (reference) segment with each segment determined by the recognizer. Because there were many segments and recognizer runs, we automated the scoring procedure. The scoring procedure first removes silence and noise labels from the recognizer output. For each reference segment, it then finds the temporally nearest recognizer segment. If either of the segment boundaries differs by more than δ milliseconds, then the match is rejected. Each reference segment is allowed to match at most one segment in the input vocalization. Once an attempt has been made to match all reference segments, any remaining reference segments are counted as deletions, and any unmatched test segments are counted as insertions. Substitution errors are those matches for which segment identity is incorrect. For this report the value of δ was 50 ms for the ZF001 data and 100 ms for the IB007 data. The larger value for scoring the IB007 data was chosen to permit greater flexibility in matching the segment boundaries of plastic syllables.

It should be noted that the scoring procedure will count errors that cause segment boundaries to mismatch more severely than errors that merely misclassify one segment as another. For example, if reference segment X is recognized as segments Y and Z, and the boundaries of segments Y or Z are not within δ of segment X, then the error will be scored as two insertions (segments Y and Z) as well as the single deletion of X. The likelihood that boundaries of test and reference segments will differ by more than δ will increase if two or more test segments are mapped onto one reference segment, and vice versa. This method of counting errors reflects their severity. Scores are reported as average segment accuracy which is defined to be 100% minus the percentage of substitution, insertion, and deletion errors for all song syllables and calls.

III. RESULTS

The recognizer was tested with continuous, unsegmented recordings of a zebra finch and an indigo bunting. The acoustic properties of the vocalizations of these two species are markedly different. During singing, zebra finches produce a repeated sequence of broadband vocalizations often with a clear harmonic structure, while indigo buntings produce doublets or triplets of relatively narrowband vocalizations, except for buzzy syllables (Thompson, 1970).

A. Zebra finch song

The songs of zebra finches consist of a number of introductory elements followed by one or more motifs (Sossinka and Böhner, 1980). Motifs tend to be highly stereotyped and comprise ordered sequences of two or more syllables separated by brief (5–50 ms) silent intervals of characteristic duration. Syllables are the subunits of motifs of at least 20 ms in duration that are separated by baseline energy lasting at least 5 ms. The most common song of ZF001 is shown in Fig. 3. This bird was observed to produce one to several introductory notes, two types of motifs, two stereotyped calls, and a large set of calls that could not be manually classified into distinct categories. Calls that were not part of motifs sometimes occurred between motifs. The inventory of syllable and call types is shown in the first two columns of Table I and displayed in Fig. 4.

A set of segment waveforms was extracted from five different 100-s recordings comprising a total of 275 syllables and calls. Templates were made by transforming these waveforms to log magnitude spectral representations (Hamming window, 256-point FFTs, 128-point step size). A sample of

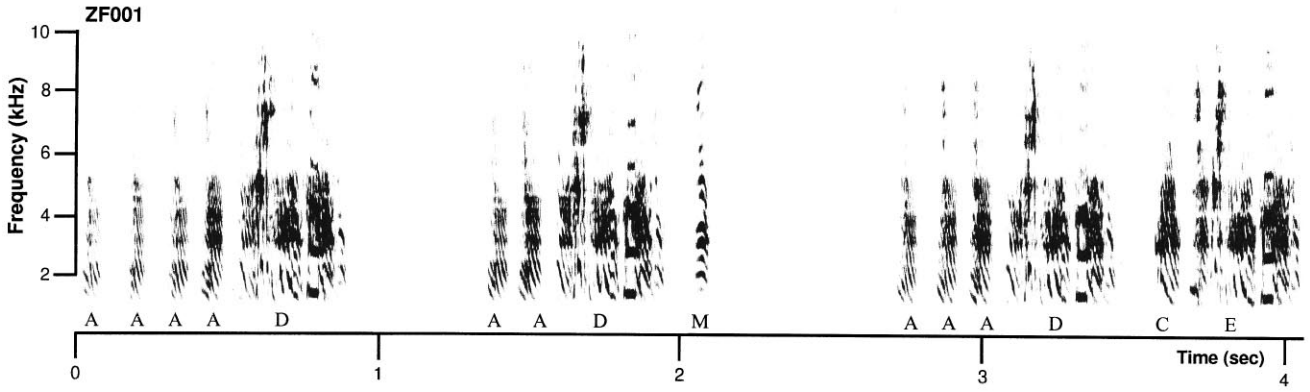

FIG. 3. The song of ZF001 judged to occur most frequently. The song comprises four groupings of syllables (motifs) each of which contains the syllable D or the variant E. The first three motifs are preceded by a variable number of introductory notes (A). The final motif comprises two syllables, C and E.

cage noises and background-silence intervals was also obtained from which 9 silent intervals and 27 noise segments were retained. All four syllables and two of the calls from ZF001 were highly stereotyped and therefore could be reliably labeled. From an initial set of seven templates for each stereotyped call and song syllable, subsets of size one, three, five, and seven were selected. For subsets of size three and five an effort was made to select templates that were most dissimilar from one another. The remaining calls (42% of all calls) varied greatly and a further breakdown into subclasses was deemed impossible (see Sec. II). By analogy with our treatment of noise, therefore, this call class was labeled using a large number (13) of exemplar templates chosen to span the range of variation. The complete database for ZF001 comprised 112 min of recordings stored in 96 files which resulted in 5180 labeled segments and 302 labeled songs. The testing data included the five recordings from which templates were extracted. Thus there was a maximum overlap of 55/5180=1.1% between the largest set of templates and test database.

Using a set of five templates per syllable type, the highest segment accuracy score for ZF001 was 98.1%, which resulted from a feature vector size of 128 and step duration of 3.2 ms (128/3.2 ms). The corresponding segment confusion matrix is shown in Table I. Segment identity is indicated in the first column, followed by the number of that segment type in the data. Entries along the row labeled INS are insertions; those below the DEL column are deletions of the type indicated by the row label. The final three segment types are calls, the last of which does not form a type that can be accurately categorized. This confusion matrix is typical for this bird in that all segments except O are recognized with high accuracy. The 95.5% correct recognition of the O call is not directly comparable to the other scores because these sounds are highly variable and are therefore represented by an unusually large set of templates. Note that calls or syllables misidentified as noises will appear as deletions (DEL), and noises misidentified as calls or syllables will appear as insertions (INS).

TABLE I. Identification confusion matrix for 5180 syllables and calls of ZF001. Null entries contain 0.

Syllable	N	DEL	A	C	D	E	M	N	O
INS	26	⋯	7	3	0	0	1	1	14
A	1752	2	1733				12		5
C	340	0		340					
D	589	1			586		2		
E	336	2				334			
Call									
M	434	0					428		6
N	825	1						824	
O	904	14	7				20		863

The accuracy of recognizer-assigned onsets and offsets relative to manual marking was calculated for all correctly recognized segments. The average difference for onsets was 8.5±1.9 ms; the average difference for offsets was 4.4±3.8

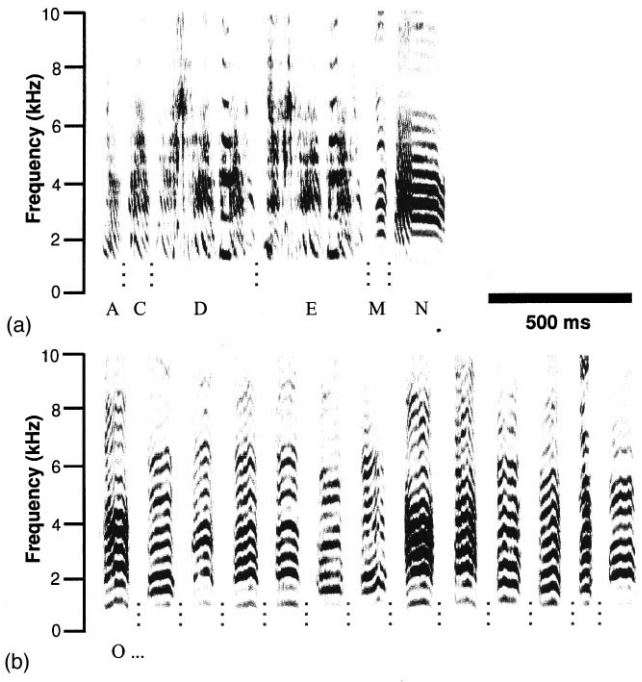

FIG. 4. (a) Repertoire of the four ZF001 song syllables (A,C,D,E) and two stereotyped calls (M,N). Dotted lines indicate syllable or call boundaries. Note the similarity between syllables D and E except at their onsets. (b) The 13 O calls used as templates. The calls occur in isolation, have a fairly consistent duration of 50–70 ms, and have a chevron-shaped frequency modulation; they differ in the details of their temporal structure, frequency modulation, and timbre.

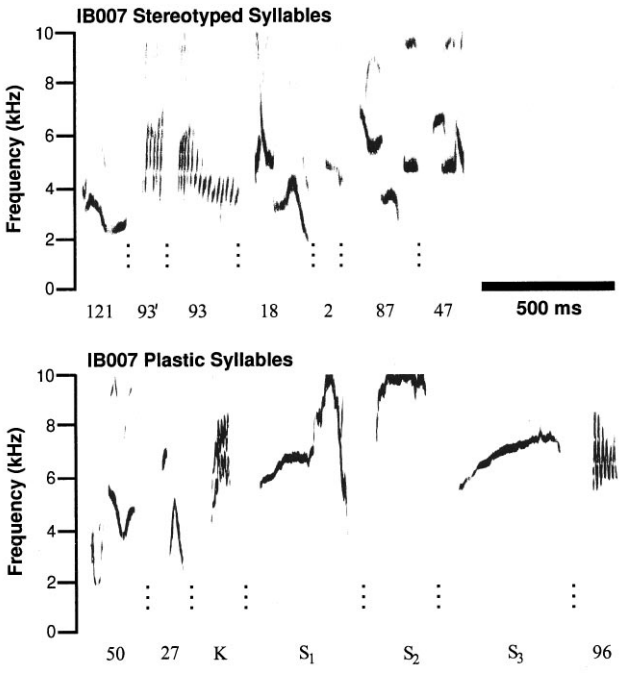

FIG. 5. Syllable repertoire of IB007. Syllables are separated by dotted lines. Syllables of stereotyped song are shown on the left; additional syllables from plastic song are shown on the right. The three common types of *S* syllables have been shown for comparison.

ms. Assuming actual segment boundaries are uniformly distributed between 3.2-ms frames, we anticipate average discretization error introduced by window step size to introduce 1.6 ms of error. The recognizer's boundary difference values are much larger than predicted by these assumptions. The boundary locations are acceptable for segment identification purposes, but cannot be used when highly accurate temporal alignment is necessary (e.g., when aligning acoustic and neurophysiological data). We have verified that accurate timing can be obtained using a procedure that slightly adjusts the algorithm's segment boundaries using simple criteria for the average magnitude of the waveform.

B. Indigo bunting song

The indigo bunting (IB007) was a yearling male who sang distinct bouts of two types of song. The first type, stereotyped song, comprised an initial note followed by several different pairs of syllables (phrases), and then one or several buzzy syllables (Fig. 2). The second type of song, termed plastic song, can be distinguished from stereotyped song by its inclusion of indistinct sounds and syllable types not found in stereotyped song, more variable ordering of syllables, shorter duration song, and decreased amplitude (Margoliash *et al*., 1991). The syllable repertoire of IB007 is shown in Fig. 5. A number of syllables did not appear to belong in any class and were labeled indistinct. In 154 stereotyped songs only 5 indistinct syllables were found, whereas in 61 plastic songs 62 indistinct syllables were found. The present system has no means to report indistinct syllables; for scoring purposes indistinct syllables are counted as insertion errors and attributed to the segment type with which they are labeled.

Templates for commonly occurring syllables in stereotyped song were chosen from the first five stereotyped songs by selecting a maximally dissimilar set based on visual inspection of their spectrograms. Templates for other syllables were selected by choosing the first three robust occurrences of a syllable type in the bout of plastic songs. Three templates were made for each of 15 syllable types, including three templates for silence. The 215 songs (495 s) of this bird had been previously excised from much longer recordings and each song had been placed in a separate file, so no templates were necessary for nontarget sounds. In these recordings there were 1061 syllables. There was an overlap of 45/1061=4.2% between the set of templates and testing data.

Unlike the song of zebra finches, indigo bunting song is spectrally simple. After poor performance was initially obtained we opted to enhance input representations for indigo bunting song. For each vocalization, the mean of all feature vector components was determined. Then feature vector components of magnitude less than one standard deviation above the mean value were set to zero. This enhancement procedure removes variance from the local distance measure that results from mismatch at low-power frequencies. A comparison of recognizer performance with and without this enhancement is shown in Table II (64/6.4 ms). Enhancement dramatically improved performance on stereotyped songs. The concomitant decrease in performance for plastic songs probably reflects the fact that much of the signal is at very low amplitude, nearer to noise levels than stereotyped song, and thus a significant proportion of the signal may be lost during enhancement (see Margoliash *et al*., 1991). Spectral enhancement may therefore be attractive for some research questions but not others. Also, in some cases (e.g., indigo bunting) it may be easy to objectively assign signals to the optimal procedure. The remaining results reported for IB007 employed enhancement of templates and test vocalizations.

The highest segment accuracy for the 154 stereotyped songs (1061 syllables and 5 indistinct syllables) was 97.8% at the highest spectral resolution (128/1.6 ms). The confusion matrix associated with this parameter choice is shown in Table III. From Table III we see that most errors (ten insertions and two substitutions) were related to identification of syllable type 93′. A review of recognition output revealed

TABLE II. Comparison of recognition accuracy for IB007 with and without signal enhancement (64/6.4 ms).

	Stereotyped				Plastic			
	Accuracy	Sub	Ins	Del	Accuracy	Sub	Ins	Del
Not enhanced	54.6	16.8	28.0	0.7	87.6	1.5	8.9	1.9
Enhanced	97.3	0.9	1.8	0.1	82.6	3.4	11.5	2.4

TABLE III. Identification confusion matrix for 1061 stereotyped song syllables produced by IB007 (128/1.6 ms).

Syll	N	DEL	2	87	18	47	121	93′	93	96	K	50	27	S
INS	15	⋯	2	0	0	1	0	10	0	1	0	0	1	0
2	152	0	152											
87	355	0		351									4	
18	295	0			295									
47	154	0				152				1				1
121	43	0					43							
93′	16	0						14		2				
93	33	0							33					
96	10	0								10				
K	0	0									0			
50	3	0										3		
27		0											0	
S	0	0												0

that the majority of the insertion errors (eight) occurred when syllable type 87 was misrecognized as the sequence 93′–27–93′. This misrecognition also explains the origin of the four substitution errors for syllable 87.

The highest segment accuracy for the 61 plastic songs (582 syllables and 62 indistinct syllables) was 84.4% (64/3.2 ms). Relative to stereotyped song, plastic song was consistently recognized with a greater overall proportion of insertion errors. Table IV shows that syllable *K* is the single greatest source of substitution errors. The highly variable *S* syllables are the greatest source of deletions (10) and insertions (22), a result similar to that obtained with the highly variable O call of ZF001. Note, however, that for ZF001 call O was also the primary source of substitution errors. Unlike call O, syllable S is acoustically distinct from all other syllables in the bird's repertoire, and its variants are constituents of its full realization (Fig. 5).

For stereotyped song (128/1.6 ms), the mean of the absolute value of the segment onset error (mean absolute error) was 5.0±5.4 ms, and the mean absolute offset error was 5.2±6.4 ms. For plastic song (128/3.2 ms) the mean absolute onset error was 8.3±10.6 ms, and the mean absolute offset error was 11.8±16.2 ms. The variances of mean boundary error for stereotyped and plastic song appear to be normally distributed about zero and differ significantly [$F(1103,2105) = 5.4, p < 0.01$]. Since segmental boundary error was only calculated for correctly recognized syllables, the difference in variability is presumably due to variation in plastic song that is unlike that found in stereotyped song.

C. Sensitivity analysis

1. Template set size

The performance of the song recognizer was assessed as a function of the number of templates using set sizes of one, three, five, and seven templates per syllable and stereotyped call of ZF001. (The number of templates for the type O call, and for noises, was held constant.) Table V summarizes recognizer performance and indicates that segment accuracy increases with the number of templates per syllable. The error rate using seven templates per syllable is approximately one-half of its value when using one template per syllable. The dependence of performance on template set size rapidly decreases with increasing set size. For template sets having at least three templates per syllable, the increased performance is entirely due to a decrease in the number of substitutions. The largest decrease in substitution errors occurs between three and five templates per syllable. Whereas performance is only modestly affected by template set size, the number of computations increases as the square of the number of input patterns and template time frames (see the Appendix). In any given application, these factors must be weighed against each other to determine how large a set of templates is necessary.

TABLE IV. Identification confusion matrix for 582 plastic song syllables produced by IB007 (128/3.2 ms).

Syll	N	DEL	2	87	18	47	121	93′	93	96	K	50	27	S
INS	65	⋯	14	1	2	2	4	4	0	8	3	4	1	22
2	4	0	4											
87	68	1		67										
18	44	1			41	1				1				
47	32	0	1			31								
121	14	0					14							
93′	17	0						17						
93	28	1						2	25					
96	57	0								57				
K	43	0								7	36			
50	70	0	1									69		
27	109	0											109	
S	96	10												86

TABLE V. Accuracy as a function of template set size for ZF001.

Templates/ Syllable	No. Templates	Accuracy	Substitution	Insertion	Deletion
1	55	96.2	2.2	0.9	0.7
3	67	97.0	2.1	0.5	0.4
5	79	97.8	1.3	0.5	0.4
7	91	97.9	1.2	0.5	0.4

2. Template resolution

Several parameter choices affect recognizer performance and computational cost of the algorithm. The most important of these, feature vector size and temporal step size, determine template resolution.

a. Zebra finch song and calls. Step durations of 3.2, 6.4, 9.6, and 12.8 ms were assessed with feature vector sizes of 32, 64, and 128 [Fig. 6(A)]. Computational cost prevented investigation of shorter step durations with the data from ZF001. All graphs in Fig. 6 only display recognition as a function of step duration for a feature vector size of 64; for ZF001 performance depends very weakly on feature vector size. Each bar is subdivided to reveal how average error (the sum of bar sizes) depends on substitution, insertion, and deletion error types. Performance depends most strongly on step duration. In particular, average error increases substantially with step durations between 6.4 ms and 9.6 ms. Performance using step durations of 3.2 and 6.4 ms is nearly the same. From comparison of the three error types we conclude that the dependence of accuracy on step duration is the result of increased substitutions, deletions, and insertions, with the greater number of errors attributed to substitutions.

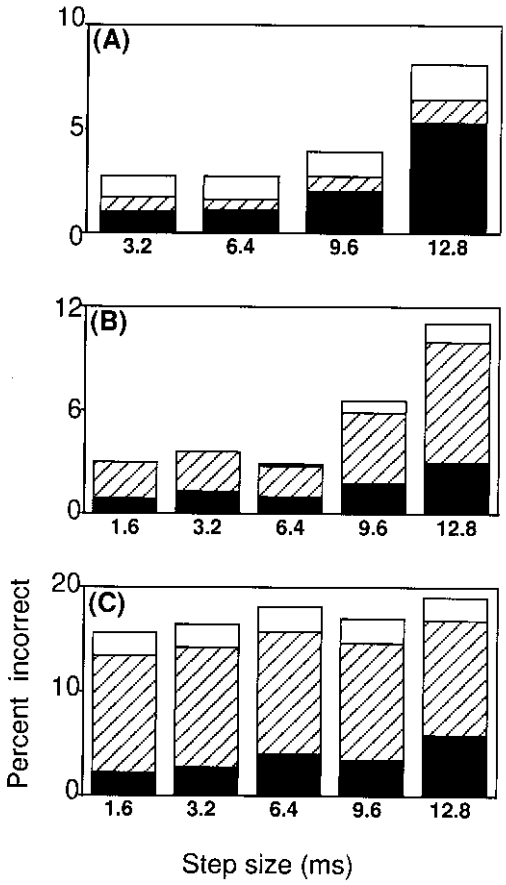

FIG. 6. Error as a function of template resolution for (A) ZF001 and (B) the stereotyped and (C) the plastic song of IB007. Average error (100%-accuracy) is the total height of each bar. Errors: substitutions (▬); insertions (▨); deletions (▭).

b. Indigo bunting stereotyped song. Recognition results were obtained using the IB007 stereotyped song for step durations of 1.6, 3.2, 6.4, 9.6, and 12.8 ms assessed with feature vector sizes of 32, 64, and 128 [Fig. 6(B)]. As for the ZF001 data, performance decreases markedly at a step duration of 9.6 ms. For each vector size the average error rate for step durations of 1.6, 3.2, and 6.4 ms is similar. For all step durations average error decreases 2%-3% as feature vector size is varied from 32 to 128 (not shown). For example, using a step duration of 1.6 ms, average error falls from 5.1% to 2.1% as feature vector size increases from 32 to 128. All types of errors increase with decreasing template resolution. In contrast with the ZF001 results [compare Fig. 6(A) and 6(B)], substitution errors are not the dominant error type, nor do they increase as dramatically with increasing step size. The greater proportion of insertion errors results from the morphology of bunting syllables, parts of which are sometimes misclassified. We have not observed this type of error in the recognition of finch vocalizations.

c. Indigo bunting plastic song. Response to plastic song, shown in Fig. 6(C), exhibits a very different pattern of sensitivity. Overall accuracy is very weakly dependent on vector size, and it does not depend strongly on step duration. Deletions and insertions show little variation with step duration, but increase slightly with feature length. The preponderance of insertion errors is presumably due to matching to more variable syllable types than for stereotyped song. These scores indicate that the accuracy of DTW does not immediately extend to more variable song types.

IV. DISCUSSION

We caution that the results have been obtained for a small number of individuals and species. (The recognizer has thus far been used to segment the song of four additional zebra finches with results similar to those we report in this paper.) Zebra finch, however, are particularly popular in behavioral and especially neurobiological research. Successful application of our technique to zebra finch song may be welcomed by these communities. The harmonic stacks and other broadband vocalizations of zebra finch are representative of relatively spectrally complex bird vocalizations. The spectrally compact vocalizations of indigo buntings are representative of the large class of birds with narrow-band songs (Greenewalt, 1968). The degree of stereotypy exhibited in these two species songs relative to interclass variability is common in birds songs generally. Thus it is likely the excellent performance reported here will extend to the advertisement songs of many species of birds.

A. Automated analysis of animal vocalizations

The DTW algorithm studied here is unique with respect to other animal vocalization recognition methods because it

was developed to accurately identify vocalizations from continuous records. Other approaches to computer recognition and analysis of animal vocalizations have emphasized their viability as metrics for the quantification of acoustic similarity (Symmes *et al*., 1979; Bradley and Bradley, 1983; Clark *et al*., 1987; Williams, 1993). These approaches generally assume a corpus of vocalizations that have been manually segmented to isolate the target vocalization type (e.g., whistles, calls). Except for Clark *et al*. (1987) all of these studies employed a small set of acoustic parameters that only specify certain aspects of the signal. The distance score calculated by DTW methods warrants quantitative comparison with the measure of similarity based on cross correlation (Clark *et al*., 1987), but is beyond the scope of this paper.

Dynamic programming has been used by several researchers to analyze birdsong. Bradley and Bradley (1983) and Williams (1993) applied dynamic programming based on simplified symbolic representations of sounds and human assessment of them. In the former study internote distances were generated by dendograms of binary human judgments of note similarity which were then combined with three spectral parameters. Williams (1993) employed a digitization tablet to trace the contour of spectrograms of song syllables. In both instances, subjective judgments are employed to determine which aspects of a vocalization are relevant to classification. Although this may be supported and even desirable in contexts where the vocalizations and their behavioral significance are understood, in other cases an approach that does not require such assumptions is desirable. The present system requires human judgment in the preselection of templates, but the one-stage algorithm does not require the researcher to make prior judgments about the importance of particular spectral features.

Buck and Tyack (1993) obviated the subjectivity inherent in human scoring of dolphin signature whistles by using an algorithm to isolate spectral contours. Like the algorithm presented here, the dynamic time warping algorithm was based on the local constraint of Itakura (1975). The enhancement of indigo bunting song we employed also creates spectral contours because it enhances the high-amplitude frequencies, those associated with the contour. Such simplifications of the signal are of limited use, since they do not adequately represent broadband sounds. Alternately, one may wish to consider other less intuitive encodings of acoustic data (e.g., linear predictive coefficients), for example, when the behavioral significance of some aspect of the acoustic signal becomes known.

B. Limitations of dynamic time warping

Dynamic time warping is a form of template-based recognition that has been used in speech recognition applications for several decades. Because of its failure to adequately model speech in all but the most ideal situations (i.e., isolated utterances spoken by a single speaker), it has been largely superseded in research and commercial systems by hidden Markov models and, to a lesser extent, by neural ''connectionist'' networks. Despite its success recognizing stereotyped syllables, more variable calls, plastic song, and adverse acoustic environments pose problems that may be best solved using more sophisticated modeling.

The relatively lower recognition accuracy for nonstereotyped vocalizations reflects some of the several shortcomings of dynamic time warping and template-based algorithms generally. Template-based algorithms are not easily applied to variable vocalizations, unless the variability can be incorporated into the set of templates. Amplitude variation reduces recognizer accuracy, but it can be reduced using normalization of the input signal (Clark *et al*., 1987), a technique we have not explored. A more difficult problem arises with vocalizations that exhibit a high degree of context dependence on preceding and/or following vocalizations. Context dependence is not modeled by a system that weights all temporal variation equally. Rather, one must select a set of templates that captures important context dependence explicitly, a requisite that may be impractical if vocalizations are constituted from context dependent segments that combine in a large number of ways.

Application to real-world acoustic environments entails further difficulties. Recordings obtained under suboptimal conditions may contain nonstationary distortions, noises, and overlapping vocalizations from other animals that degrade performance on target vocalizations. Preliminary recognition results of simultaneously recorded male and female zebra finches indicate that the vocalizations of individual birds can be accurately discriminated and recognized, but only if they do not overlap.

A related problem arises when one tries to identify a class of sounds that cannot be accurately represented by a finite number of templates. For example, the indistinct figures of plastic bunting song are integral parts of song. However, our results show that indistinct syllables are not accurately modeled by a small number of templates. Whereas during manual marking one can label indistinct types, the present approach must label the indistinct with some template label. A potential solution to this problem might determine when the distances between an input pattern and each of two templates are too near, thereby marking poorly recognized patterns for manual interpretation. A more robust mechanism for internal assessment, and thus a prediction of confidence, would be desirable. Statistical validation may be possible by examining the distributions of the template-internal warp paths. Such a procedure would be complex and memory intensive (see the Appendix), and has not been attempted.

Finally, the local distance metric we employed treats the temporal dimension as one that can be nonlinearly compressed and dilated without affecting the similarity calculation. The algorithm does not show a high degree of accuracy at pinpointing segment boundaries, since single frames at the boundaries may be omitted altogether by the warping algorithm. In cases where precise segment boundaries are required, in a postprocessing stage the boundaries output by the DTW stage are easily adjusted based on the average magnitude of the waveform.

V. SUMMARY

Dynamic time warping is attractive for the recognition and analysis of stereotyped vocalizations for several reasons: (1) it returns an objective quantitative measure of similarity; (2) it measures temporal and spectral differences; (3) it permits the selection of a feature representation most useful for analysis; (4) it automates analysis, increasing researcher efficiency. Massive databases are common in bioacoustic studies. Automated analysis may be helpful, for example, facilitating studies that track vocalization of individual birds over many months. While our results are indicative of the power of the algorithm for song recognition, and suggestive of its use in song analysis, determination of its success will require application to larger numbers of individual animals in response to specific research objectives.

ACKNOWLEDGMENTS

The authors thank Albert Yu and Cynthia Staicer for collecting and labeling the zebra finch and indigo bunting vocalizations, respectively. S.E.A. was supported by NIH NRSA Grant No. 1-F32-MH10525.

APPENDIX

In algorithmic form the procedure to determine the sequence of syllable matches $\{\sigma_1,\sigma_2,...\}$ is:

(1) Initialize.

$$D(1,j,k)=\sum_{n=1}^{j} d(1,n,k)$$

(2) Time Warp.

(a) For $i=2:N$ do steps (b)–(f).

(b) For $k=1:K$ do steps (c)–(e).

(c) Warp across boundaries using Eq. (2).
If a template transition is made [i.e., $w_j(l)=1$ and $w_j(l-1)=J(w_k(l-1))$], then $B(i,j,k)=i-1$.
Otherwise, $B(i,1,k)=B(i-1,1,k)$.

(d) For $j=1:J(k)$ do step (e).

(e) Warp within template using Eq. (1).
$B(i,j,k)=B(w_i(l-1),w_j(l-1),k)$.

(f) $T(i)=\text{argmin}_k\ D(i,J(k),k)$
$F(i)=B(i,J(T(i)),T(i))$

(3) Backtrack. Set $n=N$. Do steps (g) and (h) until $n=1$.

(g) $\sigma_n=T(n)$.

(h) $n=F(n)$.

The memory requirements of the algorithm can be minimized in two ways. First, the calculation of $D(i,j,k)$ and update of backpointers at each time frame only require access to transitions at two previous time frames. Therefore, only small portions of the arrays $D(i,j,k)$ and $B(i,j,k)$ are needed at any step. Second, the simple recognizer does not rely on details of the template-internal path, so the backtracking step need only maintain information about template-match boundaries. In some applications it may be desirable to retain the optimal warping path in order to gather statistics regarding how the input pattern and template are temporally aligned. This is possible if at each time frame of the input pattern, for the template ending recorded as having the minimum cumulative distance, the partial optimal path leading to that point is recorded. The backtracking step can then recover the complete optimal path.

The one-stage algorithm is computationally intensive; its computational speed depends on parameter choices and the number of templates. Recognition requires one warp of the test vocalization for each template in the set. If J_k is the duration of template k, N is the duration (in frames) of the input vocalization, and there are K templates, the total number of warps required is $KN\Sigma_k^K J_k$. Thus the computation required depends on the size of feature vectors as well as the duration in frames of templates and the test vocalization. Computation grows linearly as a function of feature vector size, but depends quadratically on the temporal resolution, since this increases the number of time frames in the templates and test vocalization. Nonetheless, the one-stage algorithm is the most computationally efficient continuous dynamic time-warping algorithm (Ney, 1984; Silverman and Morgan, 1990). On a SparcStation 2 our unoptimized implementation requires about 1 CPU second per second of templates per second of song (64/6.4 ms). Thus the technique is practical on modern personal computers and workstations, but in our laboratory it is not usually used interactively. With software optimization, we anticipate that real-time performance on modern personal computers should be possible. After code enhancement, we expect to make the programs available electronically.

Bradley, D. W., and Bradley, R. A. (**1983**). ''Application of sequence comparison to the study of bird songs,'' in *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, edited by D. Sankoff and J. B. Kruskal (Addison-Wesley, Reading, MA), pp. 55–91.

Bridle, J. S., Chamberlain, R. M., and Brown, M. D. (**1982**). ''An algorithm for connected word recognition,'' in *Proceedings of IEEE Conference on Acoustics, Speech, Signal Processing*, Paris, France (IEEE, New York), pp. 899–902.

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

Clark, C. W., Marler, P., and Beeman, K. (**1987**). ''Quantitative analysis of animal vocal phonology: an application to swamp sparrow song,'' Ethology **76**, 101–115.

Greenewalt, C. H. (**1968**). *Bird Song: Acoustics and Physiology* (Smithsonian Institute, Washington, DC).

Itakura, F. (**1975**). ''Minimum prediction residual principle applied to speech recognition,'' IEEE ASSP Mag. **23**, 67–72.

Kato, Y. (**1980**). ''Words into action III: A commercial system,'' IEEE Spectrum **17**, 29.

Margoliash, D., Staicer, C. A., and Inoue, S. A. (**1991**). ''Stereotyped and plastic song in adult indigo buntings, *Passerina cyanea*,'' Animal Behav. **42**, 367–388.

Margoliash, D., Staicer, C. A., and Inoue, S. A. (**1994**). ''The process of syllable acquisition in adult indigo buntings (*Passerina Cyanea*),'' Behaviour **131**, 39–64.

Marler, P., and Peters, S. (**1982**). ''Developmental overproduction and selective attrition: New processes in the epigenesis of birdsong,'' Dev. Psychobiol. **15**, 369–378.

Myers, C. S., and Rabiner, L. R. (**1981**). ''A level building dynamic time warping algorithm for connected word recognition,'' IEEE Trans. Acoust. Speech. Signal Process. **29**, 284–297.

Ney, H. (**1984**). ''The use of a one-stage dynamic programming algorithm for connected word recognition,'' IEEE Trans. Acoust. Speech Signal Process. **32**, 263–271.

Silverman, H. F., and Morgan, D. P. (**1990**). ''The application of dynamic programming to connected speech recognition,'' IEEE ASSP Mag. **7**, 7–24 (July).

Sossinka, R., and Böhner, J. (**1980**). ''Song types in the zebra finch *Poephila guttata castanotis*,'' Z. Tierpsychol. **53**, 123–132.
Symmes, D., Newman, J. D., Talmage-Riggs, G., and Lieblich, A. K. (**1979**). ''Individuality and stability of isolation peeps in squirrel monkeys,'' Animal Behav. **27**, 1142–1152.
Thompson, W. L. (**1970**). ''Song variation in a population of indigo buntings,'' Auk **87**, 58–71.
Vintsyuk, T. K. (**1971**). ''Element-wise recognition of continuous speech composed of words from a specified dictionary,'' Kibernetika **7**, 133–143.
Williams, J. M. (**1993**). ''Objective comparisons of song syllables: A dynamic programming approach,'' J. Theor. Biol. **161**, 317–328.
Yu, A. C., and Margoliash, D. (**1995**). ''Functional hierarchy defined by single units in singing birds: HVc represents syllables and RA represents notes,'' Soc. Neurosci. Abstr. **21**, 958.