Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study

Joseph A. Kogan and Daniel Margoliash
Department of Organismal Biology and Anatomy, 1027 East 57th Street, The University of Chicago, Chicago, Illinois 60637

(Received 4 August 1997; revised 5 November 1997; accepted 2 January 1998)

The performance of two techniques is compared for automated recognition of bird song units from continuous recordings. The advantages and limitations of dynamic time warping (DTW) and hidden Markov models (HMMs) are evaluated on a large database of male songs of zebra finches (*Taeniopygia guttata*) and indigo buntings (*Passerina cyanea*), which have different types of vocalizations and have been recorded under different laboratory conditions. Depending on the quality of recordings and complexity of song, the DTW-based technique gives excellent to satisfactory performance. Under challenging conditions such as noisy recordings or presence of confusing short-duration calls, good performance of the DTW-based technique requires careful selection of templates that may demand expert knowledge. Because HMMs are trained, equivalent or even better performance of HMMs can be achieved based only on segmentation and labeling of constituent vocalizations, albeit with many more training examples than DTW templates. One weakness in HMM performance is the misclassification of short-duration vocalizations or song units with more variable structure (e.g., some calls, and syllables of plastic songs). To address these and other limitations, new approaches for analyzing bird vocalizations are discussed. [S0001-4966(98)02004-9]

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

INTRODUCTION

Many biological studies require identification of constituents of animal vocalizations from continuous recordings obtained in the field or in the laboratory (e.g., Payne *et al.*, 1981; Marler and Peters, 1982a). Most of these studies are based on manual inspection and labeling of sound spectrographs, which relies on agreement between human experts, often not explicitly described, for reproducible results, rather than a quantitative approach. In addition, many biological studies involve a large corpus of vocalizations and therefore are extremely labor extensive, yet substantial databases may be biologically important, for example, in identifying rarely observed if highly significant behaviors (e.g., Margoliash *et al.*, 1991, 1994), or in assessing variability of brain activity in relation to behavior (e.g., Yu and Margoliash, 1996).

Manual inspection of multiple vocalizations is prone to errors, which can be minimized by cross checking (duplicate scoring) but only at the cost of additional effort. On the other hand, the ability of a human expert in visual analysis of sound spectrographs supported by auditory playback cannot be outperformed on a small task. Since automated analysis techniques tend to reliably generate correct and erroneous constituent identification (labels), this suggests that a practical strategy for maximizing performance while minimizing manual effort should be based on training on a small data set manually prepared by experts, then automated recognition followed by limited inspection by human experts to correct obvious errors.

Recent progress in automated speech recognition (Makhoul and Schwartz, 1995) encourages expectations for achieving the goal of reliable automated recognition of animal vocalizations. Although the differences between human speech and animal vocalizations, and the different conditions under which they are recorded, are significant and have to be taken into account, the relative simplicity of animal vocalizations compared to speech can facilitate recognition of animal vocalizations. For example, most animal vocalizations consist of discrete subunits organized in stereotyped hierarchies. This is true for bird songs, which can be viewed as comprising notes, syllables (or figures), phrases (or motifs), and songs (Catchpole and Slater, 1995). On the other hand, many biological studies require recording animal vocalizations under adverse conditions (e.g., field recordings with contaminant vocalizations and nonhomogeneous noise backgrounds), whereas speech recognition is almost always focused on low-clutter high signal-to-noise (S/N) ratio conditions. Here we evaluate automated recognition techniques as applied to bird songs recorded under a variety of noise conditions in the laboratory. This may also be a first step towards developing suitable analysis techniques for field recordings (Larkin *et al.*, 1996).

A dynamic time-warping (DTW) template-based approach is potentially attractive for analyzing stereotyped vocalizations (Silverman and Morgan, 1990) such as the songs of adult birds. Thus we developed the DTW-based Long Continuous Song Recognition (LCSR) system (Anderson *et al.*, 1996). Similar dynamic programming techniques have also proven useful in analyzing and comparing manually extracted single animal vocalizations, for example, dolphin signature whistles (Buck and Tyack, 1993) and bird calls (Ito *et al.*, 1996).

 0001-4966/98/103(4)/2185/12/$10.00

In this study, we continue investigation of LCSR and compare its performance with hidden Markov models (HMMs) on a large database of songs. As would be expected with a template-based system, good LCSR performance depends on the degree of variability within and separability between classes represented by the templates. Thus LCSR performance can depend on the often quite subtle tradeoff between the effectiveness of templates in representing constituent vocalization variability while maintaining separability between vocalization classes (Anderson *et al.*, 1996).

In contrast to template matching, HMMs statistically represent and estimate the constituent vocalizations to be recognized. As a result of such estimation (training), these models can accumulate more information and generalize better than can template-based techniques [for recent reviews, see Rabiner and Juang (1993); Makhoul and Schwartz (1995)]. To explore the application of HMMs to bird song, we employed the Hidden Markov Model Toolkit (HTK) (Entropic Research Laboratory) developed by Young *et al.* (1995). HTK exhibits excellent performance for continuous speech recognition applications compared to other systems in this area (Kubala, 1995).

We compared the performance of HTK and LCSR on a large database of songs of two species: stereotyped songs of zebra finches (*Taeniopygia guttata*), which have broadband vocalizations, and stereotyped and plastic songs of indigo buntings (*Passerina cyanea*), which have relatively spectrally compact vocalizations. Both techniques were tested over a broad range of regimes, and we attempted to make balanced comparisons. Our results show that HTK usually outperforms LCSR when tested with relatively noisy recordings or variable bird vocalizations with confusing calls and notes. For optimal performance we find HTK in general requires considerably more training examples than LCSR requires templates, while LCSR requires greater expert knowledge and repeated trials in selecting an optimal set of templates, especially for noisy recordings. For practical laboratory usage, an efficient approach may be to use LCSR to prepare a training set, correct that training set manually, then train HTK and use it for subsequent recognition.

I. METHODS AND TECHNIQUES

Here we emphasize specific differences in the two techniques employed for the recognition task. Technical details can be found in the cited references.

A. Long continuous song recognition

The LCSR system segments and labels an acoustic input (bird songs) using a prespecified set of template patterns of bird song constituents (Anderson *et al.*, 1996). In our application we have used digital spectrographs to represent the signals.

LCSR is based on the one-stage synchronized search which is known as one of the most computationally efficient DTW algorithms. This algorithm was first introduced by Vintsyuk (1971) for automated word recognition and subsequently extended by Bridle *et al.* (1982) and Ney (1984) for connected word recognition. For this search, the time frames of the input acoustic signal and the time frames of the templates are organized in a lattice (i,j,k), where i and j are the indexes of the time frames of the input song and each individual template, respectively, and k is the template counter. The quality of the match is measured through the sum of local metrics $d(i,j,k)$ which represent the distances between the two multidimensional vectors of the input signal at time frame i and template k at time frame j. In Anderson *et al.* (1996), $d(i,j,k)$ was computed as the Euclidean distance between the log magnitudes of the fast Fourier transform (FFT) bins of those vectors. Based on Bellman's principle of optimality, the one-stage algorithm ''recognizes'' the continuous input as the optimal sequence of selected template patterns. This minimizes the total distance between the sequence of warped templates and the continuous input under the allowable nonlinear compression–dilation transformations (warping) of the templates. In the present realization of LCSR, the range of such warping is constrained by a factor of 2. Note that in LCSR compression–dilation of the templates is the only mechanism to compensate for the variability of sounds (see Anderson *et al.*, 1996).

B. Hidden Markov toolkit

In HTK, the process of recognition of the input signal can be divided into several steps including preprocessing and representation, modeling the acoustic patterns present in the input signals, training these models, and ultimately, recognition. Here we focus on the HTK statistical models for acoustic patterns (HMMs) and their training, which distinguishes HTK from LCSR and other template based approaches in automated recognition. Though HTK is primary designed for speech recognition, its 18 major tools and programs can be adapted and used for analysis and recognition of any acoustic time series, including bird vocalizations. During this work, all major tools and most auxiliary programs of HTK were extensively tested under a variety of regimes.

1. Modeling

Using HTK, each sound category (calls, syllables, cage noises) was modeled by an HMM. The HMM is a finite state machine consisting of two levels. The hidden level operates in time, t $(1 \leqslant t \leqslant T)$ as a Markov chain with a finite number of states $X=\{1,\ldots,n\}$ and the transitions from state i to state j occur with a transition probability a_{ij}. The level of observations of the acoustical signals are represented by a sequence of vectors $\mathbf{O}=\{o_t\}$ which are assumed to be emitted from the hidden state j with some probability density $b_j(o_t)$. If the transition and the output probabilities of the HMM, **H**, are known, then the hidden state sequence $X=\{x(1),x(2),\ldots,x(T)\}$ of interest can be estimated as the most likely sequence that has generated the observed output **O**:

$$\hat{X}=\text{argmax}_X\left\{a_{x(0)x(1)}\prod_{t=1}^{T} b_{x(t)}(o_t)a_{x(t)x(t+1)}\right\}, \qquad (1)$$

where $x(0)$ is the model entry state and $x(T+1)$ is the model exit state. In our typical model, the HMM was a left-to-right model $(a_{ij}=0, j<i)$ with five states, with only the middle three states {2,3,4} emitting signals. The first and the

last states {1,5} served to glue these elementary HMMs into a compound model.

The densities of observation probabilities in the emitting states are usually modeled as mixtures of multidimensional Gaussian distributions:

$$b_j(o_t) = \sum_{m=1}^{M} c_{jm} N(o_t; \mu_{jm}, \Sigma_{jm}), \quad (2)$$

where M is the number of mixtures, c_{jm} is the weight of the mth component, $\Sigma_{m=1}^{M} c_{jm} = 1$, and $N(o_t; \mu, \Sigma)$ is some n-dimensional Gaussian density with mean vector μ and covariance matrix Σ where n is the dimensionality of observed vector o_t.

In practice, the parameters of HMMs are unknown and have to be estimated from data during the training phase. The choice of model parameters is important for reliable performance, and will reflect the data to be modeled. For example, the number of mixtures should reflect, roughly speaking, the number of different subcategories of acoustic patterns combined within the same signal class. This interacts with the amount of data available for reliable estimation (training) of the model parameters. In our experiments we used from 2 to 4 mixtures with a diagonal covariance matrix for modeling syllables and calls.

2. Training

Although direct computation by Eq. (1) is intractable, the efficient and well-known Baum–Welch algorithm (see Rabiner and Juang, 1993; Young *et al.*, 1995) implemented in HTK finds the maximum-likelihood estimates (MLE) of unknown parameters (see Appendix A). In HTK the training of HMM prototypes is subdivided into several stages including initialization of the models, isolated training for each vocalization class, and finally the embedded training when the compound HMMs are trained on the set of entire songs (Appendix A). In addition, we divided the intersyllable intervals into five duration ranges (see Williams, 1990), then derived a simple probabilistic bigrammar by counting the frequency of co-occurrences of each vocalization class with each intersyllable interval range.

In our experiments, songs for training were selected at random, but such that the training set contained exemplars of all vocalization classes. Since some calls or introductory notes occurred very infrequently, and because compound HMMs are capable of accumulating more information from the entire songs (one or more of which were contained in each file), we selected the training set on a file level, specifically including files with examples of infrequently encountered elements. These procedures tended to increase the size of the HTK training sets. Because HTK and LCSR do not have a mechanism for selecting an optimal training (or template) set, we did not obtain precise estimates for the size of the training (template) set or other parameters that may influence recognition accuracy. This would require a more statistically oriented approach such as cross validation (e.g., McLachlan, 1992). Thus direct comparison of the size of the training sets we used for HTK and number of templates we used for LCSR is somewhat difficult to interpret, although general trends were observed (see Sec. III).

3. Recognition

Trained HMMs along with the song bigrammars were used for recognition. In HTK the recognition is performed by a dynamic programming algorithm called token passing (Young *et al.*, 1989) which represents an adaptation of the Viterbi algorithm (Viterbi, 1967; Forney, 1973; Kogan, 1996) for continuous speech recognition. Interestingly, the computational scheme employed in the token passing algorithm is also the one-stage synchronized search used in LCSR, but modified for the HMM setting.

C. Database

Oscine passerine birds learn their vocalizations by reference to auditory feedback (Thorpe, 1958; Konishi, 1965). This accounts for the complexity and individual variety in many species' songs (Catchpole and Slater, 1995), and provides examples of complex bioacoustic signals to be recognized. Both LCSR and HTK were evaluated with continuous recordings of stereotyped songs of four zebra finches (bu41, gr43, gr46, yl49) as well as stereotyped and plastic songs of four indigo buntings (ib7, ib5, ib13, ib20). These recordings are representative of different types of vocalizations (broad/narrow band, stereotyped/plastic) and different conditions (low noise/noisy). To facilitate assessment of the performance of the recognizers, each call and syllable in the data files was labeled (i.e., assigned to a vocalization class and assigned an onset and offset). Some data files were scored completely manually, others were prepared with LCSR then manually reviewed. Many of the labeled vocalizations were carefully segmented and extensively analyzed in the context of behavioral or neurophysiological experiments. In some cases, labels were cross verified in blind conditions. Thus the data base is likely to be highly reliable. The overall number of syllables and calls analyzed as well as approximate number of songs per animal can be found in Table I.

Zebra finches produce a repeated sequence of broadband vocalizations often with a prominent harmonic structure. Their songs consist of introductory notes/calls followed by one or more motifs (Sossinka and Böhner, 1980). Motifs tend to be highly stereotyped, consisting of ordered sequences of two or more syllables which typically are 20–200 ms in duration separated by brief (5–50 ms) silent intervals. Zebra finch calls may be learned or innate (Zann, 1996). Compared to syllables, calls may have a more variable structure and amplitude, be shorter in duration, and can occur in a broader range of contexts (temporal sequences of vocal elements). The number of different vocalization classes in the database were ten for zebra finch bu41, 14 for gr46, 13 for gr43, and 17 for yl49.

Indigo buntings typically produce sequences of repeated (doublets, triplets, etc.) narrow-band syllables often followed by one or several spectrally broadband syllables in their advertisement songs (Thompson, 1970). Indigo buntings sing two types of songs, stereotyped songs and plastic songs (Margoliash *et al.*, 1991). The plastic songs can be distin-

TABLE I. Database content for evaluation of LCSR and HTK.

Zebra finch				Indigo bunting			
Bird	Type[a]	Level[b]	Total (N)	Bird	Type[a]	Level[b]	Total (N)
bu41	S	file	59	ib7	S	song	154
		syl	1115			syl	1061
gr43	S	file	54		P	song	61
		syl	1131			syl	582
gr46	S	file	56	ib5	S	song	100
		syl	1700			syl	975
yl49	S	file	114		P	song	70
		syl	3732			syl	1619
				ib13	S	song	100
						syl	1199
					P	song	49
						syl	833
				ib20	S	song	100
						syl	1209
					P	song	56
						syl	692

[a]S stands for stereotyped song, P stands for plastic song.
[b]A ''file'' consists of 1–5 zebra finch's songs and usually corresponds to a recording of a song bout; ''syl'' stands for syllable.

guished from stereotyped songs by their low amplitude, larger syllable repertoire, more variable syllable structure and order, and inclusion of high-frequency ''squeaky'' notes and other transients, incompletely formed sounds and indistinct sounds typically not found in the advertisement songs (Margoliash *et al.*, 1991, 1994). The indigo buntings vocalizations consisted of 13 song element classes for ib7, 13 for ib5, 22 for ib13, and 16 for ib20, for both stereotyped and plastic songs.

For each bird, there was a large range in the frequency of expression of different elements of the vocal repertoire. Commonly, there were hundreds of exemplars of each stereotyped syllable and call class, but occasionally a syllable or call class was only represented by tens of exemplars. In addition, some call/note classes occurred even less frequently. For example, element $\{s\}$ of gr43 was recognized only once, and calls $\{k\}$ and $\{m\}$ of yl49 were recognized only once and twice, respectively. Automated recognition of such infrequently occurring elements was difficult for both recognizers. For HTK, the small sample size tended to preclude adequate training. For LCSR, just a few templates were often sufficient to recognize a class (avoiding deletion errors), but the inclusion of such templates also increased the

TABLE II. Dependence of HTK performance on parametrization for zebra finch gr43. (target-rate=6 ms, window-size=18 ms, 10 coefficients).

Parametrization[a]	% Correct	% Accuracy	S[b]	D[b]	I[b]
FBANK_E	40.4	36.5	738	708	95
FBANK_EDA	89.7	76.6	192	59	316
	92.2	80.6	75	13	131
LPC_EDA	82.3	74.9	280	156	172
	81.3	74.0	149	63	83
LPCEPST_EDA	88.6	82.5	164	112	149
	91.4	86.1	48	49	60
LPREFC_EDA	81.3	78.8	233	220	62
	82.8	80.2	103	92	29
MELSPEC_EDA	81.0	68.3	317	145	308
	89.7	62.9	97	20	303
MFCC_EDA	90.4	83.9	108	124	160
	92.3	88.7	30	57	41
MFCC_EDAZ	81.6	45.6	432	15	874

[a]First row tests were conducted on the set of vocalizations as manually scored; second row tests with the combined set of calls or syllables (see text).
[b]D≡deletion error, S≡substitution error, I≡insertion error. % correct ignores syllable level insertions, and song level insertions and deletions. % accuracy accounts for all types of errors.

TABLE III. HTK and LCSR recognition performance on zebra finches vocalizations (LCSR, window-size/step=25.6/12.8[a] ms, HTK, target-rate=6 ms, window-size=18 ms).

Tool/Bird	Level	% Correct[b]	% Accuracy[b]	% S[b]	% D[b]	% I[b]	Train/Test[c]
HTK/gr43	file	25.9, 53.7§		74.0, 46.2			27/54
HTK	syl	92.4, 97.9§	90.1, 96.1§	6.4, 1.4§	1.2, 0.7§	2.3, 1.8§	668/1131
HTK	file	35.2§		64.8§			14/54
HTK	syl	89.4, 97.0§	85.8, 94.6§	2.1§	0.8§	2.4§	383/1131
LCSR	syl	75.8, 92.7§	62.3, 80.0§	21.0, 4.2§	3.2, 3.2§	13.4, 12.6§	5(70)
LCSR	syl	84.6, 95.8§	81.8, 92.5§	11.9, 0.8§	3.4, 3.4§	2.8, 2.8§	(130)
HTK/gr46	file	17.9		83.1			16/56
HTK	syl	96.6	92.4	3.2	0.2	4.1	513/1700
HTK	file	16.0		84.0			13/56
HTK	syl	95.9	92.4	3.7	0.3	3.4	485/1700
LCSR 12.8/3.2	syl	92.5	82.1	5.5	2.0	10.3	5(75)
HTK/bu41	file	49.1		50.9			20/59
HTK	syl	96.5, 97.5§	95.0, 95.5§	3.1, 1.0§	0.4, 1.5§	1.5, 2.0§	439/1115
LCSR	syl	97.4, 98.6§	79.7, 95.2§	2.5, 0.7§	0.1, 0.7§	17.7, 3.4§	5(60)
LCSR	syl	98.8	95.5	0.9	0.3	3.3	(93)
HTK/yl49	file	4.4		95.6			18/114
HTK	syl	87.1	83.6	10.8	2.0	3.5	653/3732
LCSR 12.8/12.8	syl	89.7	45.4	8.3	2.0	44.3	5(78)
LCSR 12.8/12.8	syl	92.9	78.2	5.8	1.2	14.7	(156)

[a]The default FFT window-size/step, otherwise shown in the first column.
[b]§ is for combined set of calls or syllables (see text).
[c]HTK: Number of vocalizations in training/test set. Number in training set does not include following ''silent'' intervals. LCSR: first number is typical number of templates per syllable; number in parentheses is the total number of templates.

frequency of substitution and especially insertion errors elsewhere.

D. Representation and preprocessing

1. Template creation

As described by Anderson *et al.* (1996), the most significant preprocessing step in using LCSR is the creation of templates for the song units, background noises, and a variety of ''silent'' intervals to be identified from the input stream. In the current study, templates were manually selected based on visual inspection of the spectrograms of individual bird songs with some attempt made to represent the variation within each vocalization class (call or syllable), as well as cage noises and ''silent'' intervals. As a rule, the templates were chosen from the beginning, middle, and end of an entire set of recordings. In initial experiments the number of templates within each class was limited to 3–7, which in some cases has provided accurate recognition while limiting computational overhead (see Anderson *et al.*, 1996). Under challenging conditions (short confusing calls, cage noises, noisy recordings), this number of templates was insufficient for accurate recognition and we increased the number of templates by choosing confusing elements as additional templates (see below).

2. Preprocessing and parametrization

In our experiments with LCSR presented in Tables III and IV, the bird song signals were represented by discrete Fourier transforms. 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. To optimize the LCSR performance, the parameters of the FFT size and frame rate, thresholds for minimum durations of elements to be recognized, etc., were experimentally selected for each bird individually. The amplitude of the acoustic signals was not normalized, because it was thought that the signal amplitude might provide valuable information for separation of song elements. To achieve better LCSR performance on indigo buntings' stereotyped songs, the input representation was enhanced by zeroing all feature vector components of magnitude <1 s.d. above the mean value (see Anderson *et al.*, 1996). This procedure, however, requires prior knowledge regarding these songs.

HTK facilitated analysis of the applicability to bird songs of different types of representations common in speech recognition work. Thus we compared the influence of six different types of parametrizations on HTK performance. These included linear predictive coding (LPC), LPCepstral, LPCreflection, mel-frequency cepstral, log mel-filter bank channel, and linear mel-filter bank channel (Rabiner and Juang, 1993; Deller *et al.*, 1993). These parametrizations were also evaluated with different classifiers: _E (energy), _D and _A (first and second derivatives), and _Z (cepstral mean normalization). The results of these experiments on songs of zebra finch gr43 are presented in Table II. As one can see, the performance of HTK for linear parametrizations (LPC, LPCreflection, linear mel-filter bank), decreased in all cases but by a variable amount compared to nonlinear cepstral parametrizations. Among the latter, the best results were achieved for mel-frequency cepstral coefficients (MFCC). For this reason we used the MFCC parametrization for all other experiments (e.g., Tables III and IV). Classifiers _E, _D, and _A improved the performance but _Z did not.

TABLE IV. HTK and LCSR recognition performance on indigo bunting vocalizations (LCSR, window-size=25.6/3.2, HTK, target-rate=10 ms, window-size=25 ms).

Tool	Bird/Type	Level	% Correct	% Accuracy	% S	% D	% I	Train/Test[a]
HTK	ib7/S	song	88.3		11.7			26/154
HTK		syllable	98.4	97.8	1.4	0.2	0.6	217/1061
LCSR		syllable	99.2	97.8	0.8	0.0	1.4	3(45)
HTK	P	song	65.6		35.4			29/61
HTK		syllable	98.5	94.76	0.5	1	3,7	341/582
LCSR		syllable	94.5	83.7	3.1	2.4	10.8	3(45)
HTK	ib5/S	song	67.7		26			32/68
HTK		syllable	93.2	92.7	2.8	4	0.4	318/975
LCSR	12.8/3.2	syllable	97.4	97.3	0.9	1.6	0.1	5(113)
HTK	P	song	21.3		78.7			23/47
HTK		syllable	83.4, 95.0§	76.5, 90.7§	3.2§	1.7§	4.4§	644/1619
LCSR	12.8/3.2	syllable	75.4, 86.5§	66.6, 77.7§	3.4§	10.1§	8.8§	5(113)
HTK	ib13/S	song	72, 76§		27,34§			15/100
HTK		syllable	98.8, 99.4§	97.2, 97.5§	0.1§	0.4§	1.9§	192/1199
LCSR		syllable	80.2	22.6, 78.8§	12.1	7.8	57.6,2.2§	7(161)
HTK	P	song	24.5, 44.9§		55.1§			26/49
HTK		syllable	90.9, 95.6§	89.1, 94§	3.3§	1.1§	1.5§	460/833
LCSR		syllable	62.0	−29.6, 53.3§	20.0	18.0	91.6,8.7§	7(161)
HTK	ib20/S	song	55.0		45.0			26/100
HTK		syllable	96.3	93.8	2.9	0.8	2.4	402/1209
LCSR	12.8/3.2	syllable	92.9	87.0	1.8	5.3	5.9	5(52)
HTK	P	song	64.3, 83.9§		16.1§			18/56
HTK		syllable	98.7, 99.2§	95.8, 98.6§	0.5§	0.3§	0.7§	251/692
LCSR		syllable	90.9	85.0	0.6	8.5	5.9	5(52)

[a]For LCSR, the same templates were used for analysis of stereotyped and plastic song. See Tables II and III for all symbols.

Hence, we also used _EDA as well for all other experiments.

The LPC representation was also explored for LCSR. We tested the short-time modified representation based on linear predictive coding (LPC) applied to the double autocorrelation function which de-emphasizes the amplitude of the signal. We also tested modification of the Euclidean distance for the FFT feature vectors to LPC cepstrum, lifted cepstrum, weighted likelihood ratio distances, as well as their linear combinations (Shikano and Itakura, 1991). Those methods were suggested by Kobatake and Matsunoo (1994) for improving recognition of words degraded by white noise. However, as with the results for HTK, in general the alternate representations also decreased the performance of the LCSR system.

We investigated the influence of HTK performance of the number of coefficients in the parametrizations, and the influence of target-rate (t-r) (the output sample rate which determines the period of the parameter vector) and window size (w-s) (segment of waveform used to determine each parameter vector). Our experiments have shown that the use of more than ten coefficients does not improve HTK performance. We also observed that the influence of target-rate and window-size was species dependent. The best results for zebra finches were achieved at t-r=6 ms and w-s=18 ms, whereas for indigo buntings best results were achieved at t-r=10 ms and w-s=25 ms. These differences probably arise because zebra finch vocalizations are broadband whereas indigo bunting vocalizations are narrow band. Based on these results, for most of our experiments we have used t-r=6/10 ms and w-s=18/25, depending on species, with the number of cepstral coefficients equal to 10.

E. Scoring

The performance of each recognizer was compared against a baseline of manually labeled vocalizations drawn from the database using auxiliary programs for scoring evaluation. The programs match each of the recognized and reference label sequences. Once the alignment of the two label files was found, the number of substitution (misrecognition) (S), deletion (D), and insertion (I) errors was calculated. The percentage correct (% correct) was calculated as 100% times the ratio of correctly recognized elements to a total number of song elements. This measure ignores insertion errors. To take insertion errors into account, the percentage accuracy (% accuracy) was computed as % correct minus percentage of insertion errors. The % correct on the song/file level was computed only for HTK as the percentage of completely correct recognized songs/files.

Although the main principles of scoring as applied to both recognizers are the same, they differ somewhat in detail. LCSR scoring allows one to constrain the discrepancy in the boundaries of the matched segments. In the LCSR tests reported here, the maximum time discrepancy (δ) was 100 ms. HTK scoring achieves a similar effect by performing optimal string matching (Sankoff and Kruskal, 1983) making the penalty for a substitution error (10) higher than penalties for insertion and deletion errors (7). We noted increasing δ beyond 100 ms hardly affected the numbers or distribution of LCSR errors. This implies that for δ=100 ms, the LCSR

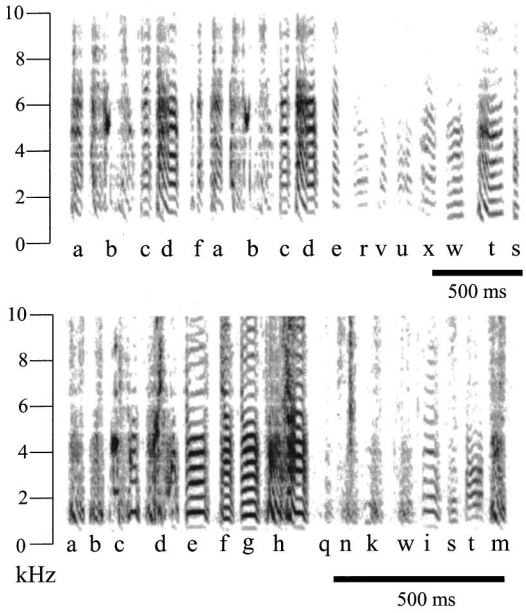

FIG. 1. Stereotyped song and additional vocalizations of zebra finch gr43 (top panel) and of zebra finch yl49 (bottom panel). The vocalization classes are represented for gr43 and all but two for yl49. ''*n*'' for yl49 is a cage noise.

score was determined by the sequence of elements and was therefore equivalent to the HTK score.

II. RESULTS OF EVALUATION

Both techniques were tested against continuous, unsegmented recordings of all animals in the database. Summaries of these tests are presented in Table III (for zebra finches) and Table IV (for indigo buntings). Unless otherwise noted, test sets included training (template) sets.

A. Comparison of recognition for zebra finch songs

Both LCSR and HTK exhibited almost perfect performance in recognition for relatively clean recordings of highly stereotyped parts of the vocalizations (e.g., major syllables of song motifs), but were vulnerable to noise and to variability in vocalizations. Performance also suffered during separation of short duration or low amplitude calls from noises such as cage hopping, bill wiping, etc. We observed that for adequate performance under noisy conditions, LCSR required many more templates for representing the variety of different cage noises and short calls which are not the part of the song than templates for stereotyped notes and syllables that comprise the bulk of song.

In general, a template set with five elements per class was sufficient to achieve high % correct scores of LCSR performance, but this measure ignores the relatively large number of insertion errors typically caused by mislabeling cage noises as short calls. Thus such small template sets were typically inadequate for achieving high accuracy. For example, for gr43 vocalizations, which included five short duration calls that were relatively confusing (see Fig. 1), LCSR accuracy was only 62.3% (accuracy) when tested with five templates for each class of elements. Increasing to 26 the number of templates for noises which were confused with calls or other vocalizations helped to reduce the number of insertion errors by a factor of 5. To reduce the number of substitution errors caused by misclassified elements of the motif, the number of templates for the introductory note $\{a\}$ was increased to 18 and for syllable $\{d\}$ to 9, and the number of templates for short confusing calls $\{r\}$ and $\{u\}$ were increased to 26 and 14, respectively. These changes completely eliminated the 35 substitution errors for $\{a\}$, eliminated two of three substitution errors for $\{d\}$, and reduced substitution errors for $\{r\}$ from 107 to 50, and for $\{u\}$ from 59 to 41. These changes also slightly increased the substitution errors for other calls, however. Another approach to resolving this problem which may have applicability for some experimental designs is to combine the confusing calls into one class. For gr43, combining confusing calls into one class improved LCSR accuracy to 80.0% for five templates per class and to 92.5% for the larger set of templates.

The same strategy was used to improve LCSR performance for zebra finches bu41 and yl49. For example, by selecting nine additional templates for confusing cage noises, and 24 additional templates for two confusing calls and one syllable, we dramatically improved LCSR performance for bu41, achieving a better score than achieved by HTK. For yl49, selecting 16 additional templates for confusing cage noises and 62 additional templates for the most confusing syllables $\{a,b,f,g\}$ and calls $\{i,s,w\}$, we dramatically improved the accuracy from 45.4% to 78.2%, with only modest increases in correctness (89.7% to 92.9%). The decreased insertion error was largely attributed to a drop in the number of insertions (from 1450 to 190) for a low-amplitude, short duration call $\{q\}$ that was manually labeled only three times out of 3732 vocalizations. The additional templates also helped to reduce the number of substitution and deletion errors for almost all elements. However, the number of insertions for some elements even increased, for example, for confusing calls $\{s\}$ (from 0 to 195), and $\{w\}$ (from 97 to 119). Further attempts to improve the performance for yl49 by increasing the template size for confusing elements did not give uniform results.

In most experiments, HTK exhibited better performance in accuracy than did LCSR, although the correctness of LCSR was higher for bu41 and yl49 using the expanded template sets (Table III). The primary reason for HTK's superior performance was that typically it produced fewer insertion errors than did LCSR. This effect was most noticeable for noisy recording as for yl49. On the other hand, LCSR could make fewer substitution and deletion errors, (e.g., bu41 and yl49), which partially offset the improvement with HTK. For comparison on the most challenging set of recordings (yl49) the number of errors for LCSR/HTK were as follows: 218/406 (substitution), 44/74 (deletion), 550/113 (insertion). The impact of the song elements on the number and type of errors was also quite different. For example, most of the substitution (and therefore deletion) errors for HTK were caused by confusing an introductory note $\{a\}$ and syllables of the motif with each other, and to a lesser extent from misclassifying calls within call classes, whereas for LCSR errors were mostly caused by confusing call $\{i\}$ with syllable $\{g\}$, and to a lesser extent by confusing introductory

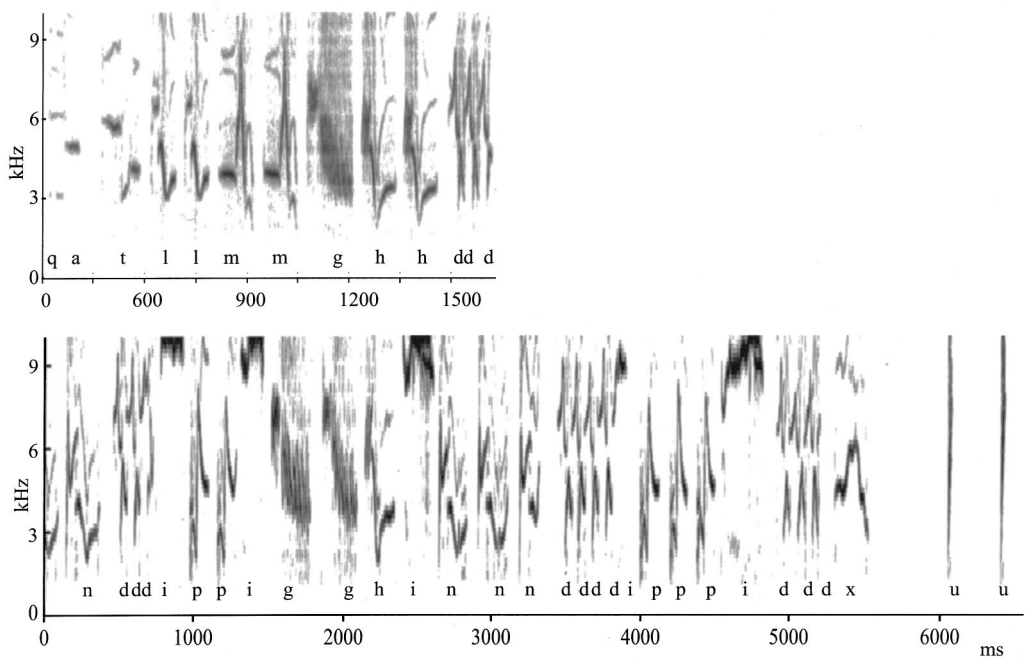

FIG. 2. Stereotyped song (top panel) and plastic song (bottom panel) for ib5, showing all 13 vocalization classes. Notes ''i'' of plastic songs are ''squeaky'' notes (see Margoliash *et al.*, 1991).

note $\{a\}$ with calls $\{w\}$, $\{s\}$, and syllable $\{b\}$. The percentage of insertion errors for LCSR/HTK were distributed across the song elements as follows: low amplitude calls $\{s\}$, $\{q\}$, and $\{w\}$: 35.5/34.6, 35.8/0.0, and 21.6/0.0, respectively; introductory note $\{a\}$: 1.3/10.5; and syllables $\{e\}$: 0.4/16.5, and $\{h\}$: 0.0/9.7.

The reason for the high level of substitution errors, which also often involved insertions and deletions, is that HTK (and LCSR) does not have a good mechanism for discriminating similar sounds. Therefore use of a larger number of exemplars per syllable in the training set did not guarantee better performance on the test set. For example, for gr43, only 16 exemplars for syllable $\{f\}$ were sufficient to obtain 100% accuracy for that syllable. However, 135 exemplars for syllable $\{a\}$ still gave five substitution and two deletion errors, and increasing the training set for $\{a\}$ to 241 exemplars still gave two substitution and one deletion errors. The same trend was noted in other experiments as well, where a large number of training exemplars did not necessarily guarantee good performance, and syllables with larger numbers of training exemplars did not necessarily have higher performance than did syllables with fewer numbers of training exemplars. For example, for yl49, 95/62 exemplars for syllables $\{a\}/\{d\}$ gave 180/18 substitution, 8/1 deletion, 2/0 insertion errors, respectively. Good performance depended more on homogeneity within the category. We did not investigate the use of discriminative procedures, but these may significantly improve error performance, and to some extent automate the selection of a representative set of templates for LCSR, or training set for HTK (see Sec. III).

B. Comparison of recognition for indigo bunting songs

The summary of LCSR and HTK performance for indigo buntings is given in Table IV. We experimented with both techniques using different numbers of templates or songs in the training sets. Both plastic and stereotyped songs were included in all training sets, and both types of songs were tested using the same training set. In general, performance at the syllable level was comparable comparing zebra finches and indigo buntings but was better for the latter at the song/file level because the finch recordings typically contained many songs per file whereas the bunting recordings typically contained only one song per file (Table I).

In general, both LCSR and HTK exhibited excellent performance for indigo bunting stereotyped songs (Table IV). LCSR exhibited marginally better performance than did HTK for stereotyped songs of one bird (ib5) (see Fig. 2). LCSR also exhibited poor performance on stereotyped songs of one bird (ib13). LCSR performance on plastic songs was also somewhat variable, ranging from excellent to poor depending on the bird (Table IV). In contrast, HTK exhibited excellent performance for stereotyped songs for all birds but one, and very good to excellent performance on plastic songs, significantly better performance than LCSR, for all birds (Table IV). The performance of HTK on plastic songs is a result of potential importance to biologists, because it demonstrates the potential of HMM-based approaches to properly represent variable vocalizations in juvenile and adult birds that learn songs (e.g., see Marler and Peters, 1982b; Margoliash *et al.*, 1994). It is noteworthy also that the robust performance of HTK was accomplished without enhancement of the input representation.

Excellent LCSR performance for bird ib7 has been previously reported by Anderson *et al.* (1996). We also achieved excellent LCSR performance for ib7 using a different set of templates (Table IV). As was noted by Anderson *et al.* (1996), LCSR performance on stereotyped and plastic indigo buntings songs was mixed, and depended on enhancement of the signal. This resulted in dramatic improvement of

performance on stereotyped songs, for example, for ib7 (from 54.6% to 97.3% in accuracy), but with some decrease in performance for plastic songs (87.6% to 82.6% accuracy) (Anderson *et al.*, 1996). We noted the same impact of enhancement on LCSR performance for other indigo bunting recordings (not shown), and therefore enhancement was used for all LCSR results given in Table IV. The opposite effects of enhancement on accuracy of recognition of high-amplitude stereotyped songs and low-amplitude plastic songs is another example of LCSR sensitivity to amplitude (hence S/N ratio) in the input representation. The slight decrease in LCSR performance on plastic songs probably occurs because the enhancement reduces not only the noise signal but the low-amplitude plastic song signal as well.

The poor performance of LCSR on ib13 resulted from a large number of insertion errors. Analysis of the confusion matrix (not shown) revealed that 665 out of 682 insertion errors for stereotyped songs, and 663 out of 733 insertion errors for plastic songs, were caused by a buzzy syllable $\{z\}$ which was present only four times in the manually scored plastic songs and absent from the stereotyped songs. As shown in Table IV, elimination of this sound, which could be considered an indistinct vocalization in the sense of Margoliash *et al.* (1991), improves accuracy of LCSR from 22.6% to 78.8%. HTK still performed better on ib13, however, without eliminating syllable $\{z\}$. In contrast, on stereotyped songs of ib5, HTK exhibited slightly worse performance than the excellent performance (97.3% accuracy) of LCSR. Interestingly, the high-amplitude stereotyped song (but not the low-amplitude plastic song) recordings of this bird were badly contaminated by aliasing noise, which created some discrepancies within the signal classes that trained HTK. In contrast, aliasing was removed when the stereotyped songs were enhanced prior to processing by LCSR. For ib5, most of the HTK errors resulted from a deletion (22 of 26 deletion errors) of part of a triplet or quadruplet of narrow-band syllables $\{d\}$. The typical error arose from substitution of a pair of $\{d\}$ syllables by another narrow-band syllable. Thus HTK is also susceptible to producing a relatively large number of insertion or deletion errors caused by one or more syllables recorded under noisy conditions, but on a much smaller scale than is LCSR.

To understand how the overlapping of the training and testing sets may influence the results of recognition, we also tested the performance of HTK under two extreme cases, when the training and testing sets were identical and when they did not overlap. For example, for identical sets HTK performed for ib5 as follows: for stereotyped songs 93.4% correct and 93.4% accuracy; for plastic songs 87.7 (95.2)% correct and 84.0 (94.1)% accuracy on the entire (or combined) set of syllables. Results of recognition for the same bird for nonoverlapping training and testing sets are slightly lower but generally very similar (Table IV). Thus since HTK learns a statistical model of the data, the absence of overlapping training and testing sets does not significantly degrade HTK performance, especially for stereotyped songs. One should note the differences when comparing the effects on LCSR performance of including templates in the test set. If a constituent in the test set which is also a template has not been improperly segmented by a prior error, LCSR is guaranteed to match that constituent to the template.

III. DISCUSSION

Bird songs represent relatively complex animal vocalizations, but such complexity is also known for the vocalizations of nonavian species. Our results highlight the main differences, strengths, and limitations of DTW-based and HMM-based approaches to recognition of bird song elements, and by extension to animal vocalizations in general. These conclusions serve to outline some important features which reliable animal recognition systems should inherit from the current recognizers.

A. Sensitivity and discriminative power of LCSR and HMMs

1. Template and training set sizes

As a rule, LCSR performance was determined by the discriminative power of the templates. In LCSR, since all features of sound spectrograms are weighted equally, more variable sounds required more templates for recognition. Thus a mere five templates per class were in general sufficient for reliable recognition of stereotyped syllables and calls of zebra finches, and only three templates sufficed for reliable recognition of stereotyped and even many of the plastic syllables of indigo buntings. On the other hand, further increase of the template size for these types of sounds did not give significant improvement in performance (see also Anderson *et al.*, 1996). The more variable types of sounds such as some calls and syllables, especially those of short duration and low amplitude, and cage noises, often required tens of carefully selected templates for proper discrimination. This limitation of LCSR has been previously noted by Anderson *et al.* (1996), who obtained an accuracy of 98.1% for the songs of a zebra finch by combining confusing calls into one class, increasing the number of templates for those calls to 13, and by using 27 templates for noises and nine for silent intervals.

The size of the training sets we used for HTK was usually much larger. However, as with the template set, simply increasing the training set size did not reliably lead to better performance. From 20 to 40 exemplars per syllable or call class were sufficient to obtain near-maximal performance, and subsequent increases in the training set resulted in only very modest improvements in recognition accuracy. Our actual training set was typically large, because we trained HTK on entire files (continuous recordings), and selected several files to provide sufficient exemplars for infrequently encountered vocalization classes. Training on continuous records is a natural approach for an HMM-based system.

2. Parametrization and normalization

The performance of the recognizers depends on the parametrization and amplitude normalization of the signals. Parameters which can affect LCSR performance are the resolution of FFT which determines the window and temporal step sizes. The choice of parameters depends on the specifics of a bird's vocalizations. In general, adjacent time windows

have to overlap at least 50%. A smaller step size (especially for stereotyped vocalization of indigo buntings) may reduce the number of deletions but increase the number of insertions (Anderson *et al.*, 1996). Conflicting effects on performance are also obtained in some cases when manipulating the minimum permitted syllable duration. In considering the number of templates chosen or the degree of window overlap, it is also important to remember that computational load increases as a square of the number of input patterns and template time frames.

The problem of efficient and parsimonious representation of acoustic signals has not received much attention in bird song research. Our experiments with various representations of bird songs have shown that nonlinear parametrizations are better than linear ones (e.g., LPC). Because linear representations take into account only the second-order dependencies in time (correlations or covariances) of the signal, they give smoothed averages of the signals. Therefore linear parametrizations are better suited for quasilinear signals such as in speech (Gidas and Murua, 1996), but are less appropriate for the rapid transients and nonlinearities which are common in bird songs. In contrast, nonlinear parametrizations are better at modeling bird songs because they can better represent the higher-order dependences of the signals (see Bell and Sejenowski, 1996). It is reasonable to expect that nonlinear parametrizations derived specifically for bird songs will give even better results.

Amplitude normalization of signals produced mixed results. On one hand, LCSR without normalization often gave results that were more correct but less accurate than HTK with normalization. This means that LCSR made less substitution and deletion errors which in HTK were caused by misrecognition of motif elements of relatively high amplitude. On the other hand, HTK made much fewer insertion errors than LCSR, especially for noisy recordings which caused LCSR to confuse (insert) low-amplitude short calls instead of noises. Interestingly, our experiments with LCSR using the double autocorrelation function (for LPC), which attempts to de-emphasize signal amplitude, often resulted in fewer insertion errors but at a cost of greatly increased substitution and deletion errors. Therefore normalization may be useful for recognition of low-amplitude sounds in relatively noisy recordings (e.g., as with field recordings), but may worsen the recognition of prominent song elements (as often achieved in laboratory recordings).

3. HMM structure

We evaluated increasing or decreasing the number of states and number of mixtures, and changing left-to-right models to recurrent models. In general, we observed that reduction in the number of states, for example, from five to four, improved the performance of recognition for shorter duration sounds. Reduction of the number of mixtures of distributions per state, for example, from four to two, typically worsened performance. Thus this procedure could potentially be considered in order to reduce the number of estimated parameters only in cases where a very limited amount of training data were available. We observed insignificant changes in performance with recurrent models. We conclude that using HTK, the default structure is best suited for modeling the longer duration (more stationary) bird song syllables and calls. For the shorter duration calls and notes it is often useful to reduce the number of states (which also has the effect of reducing the number of exemplars required for training). There is no mechanism within HTK, however, to select an optimal structure for a given acoustical pattern.

We also attempted to improve the discrimination of confusing syllables and calls using models having different number of states, as well as using recurrent models. These experiments did not give consistent results. It is noteworthy that easily confused sounds shared similar morphologies, that is they appeared as classes within a larger group. This can explain why improvement in discrimination of some sounds within such a group also resulted in an increase in the number of substitution errors for other sounds of the group. This represents another significant limitation of HTK. For HTK, training the statistical models for each song element is based on the maximum-likelihood estimation within the group but not across the groups, and therefore often lacks sufficient discriminative power to distinguish confusing transient calls and introductory notes of bird songs. In addition, in traditional HMMs the transition probabilities of the underlying Markov chains have smaller effects on the model than do the observation probabilities. Taking into account that the observation probabilities are assumed to be conditionally independent explains our result that HTK models are much better suited for representing more stationary rather than more transient bird song sounds (cf. Bourlard *et al.*, 1994; Gidas and Murua, 1996).

B. Further development of LCSR and HMMs for automated analyses of bird vocalizations

Both LCSR and HTK can be successfully used to automate segmentation (labeling) of bird song elements of nonoverlapping continuous vocalizations recorded in the laboratory. However, both techniques produce sufficient errors that may limit their utility in daily laboratory use. Here we suggest that extracting more biologically meaningful information from bird song recordings may significantly improve performance.

1. Discriminative features extraction

One major drawback of both techniques is the lack of a method to find discriminative features from the given signals. This limitation results in many obvious errors of song element misrecognition. For both LCSR and HTK, this limitation can be addressed by the discriminative training method (Chang and Juang, 1993; Juang *et al.*, 1997). This approach weighs more heavily those features which distinguish between classes. For example, subtly distinct budgerigar calls were separated using DTW applied to the frequencies of the two most intense peaks of each call with a specially designed scoring function of frequency tolerance (Ito *et al.*, 1996). The specific model used by Ito *et al.* (1996) was designed for budgerigar calls, however, and presumably is not generally valid for all birds. Therefore it would be valuable to have a technique that finds such features, which can potentially be of significant biological interest, with a

minimum of prior knowledge. For HMMs, one attractive approach may be to use the so-called hybrid HMM/ANN models, which switch the emphasis from observation probabilities to transition probabilities (Bourlard *et al.*, 1994). These hybrid models use an artificial neural network as an input to an HMM which are then also discriminatively trained.

2. Bird song structure

Additional improvement in performance can be gained by learning more of the structure of the bird songs of the training set. For example, even the simplest bigrammar model significantly improved HTK performance for both stereotyped and plastic songs. This implies that the probabilistic song structure can be learned based on Markov dependencies (co-occurrences) and that longer dependencies of song elements can be approximated as a product of those Markov transitions. It would therefore be valuable to automate the learning of the song probabilistic structure, for example, using probabilistic automata with a fixed length memory such as *n*-grammars (Rabiner and Juang, 1993), or more biologically meaningful ones with a variable length memory (Ron *et al.*, 1996).

More generally, hierarchical dependences of bird songs are well established, even at the physiological level (e.g., see Yu and Margoliash, 1996). This potent source of information regarding song structure is ignored in the current models, however, often causing obvious errors (e.g., song level recognition for HTK, see Tables III and IV). This limitation can be addressed by embracing appropriate hierarchical HMMs (hHMMs) (Hihi and Bengio, 1995; Fine *et al.*, 1998), or possibly by applying dynamic programming in multiple layers (see Ney *et al.*, 1992). hHMMs would learn (and hence, describe) the hierarchy from the training examples.

3. Self-learning

Finally, a significant limitation of LCSR and HTK is that the training and template creation stages are separated from recognition. As a result, classes of elements which are not present in the training/template set will be misrecognized. Presently there are no mechanisms to automatically identify new classes and thereafter include them in the training/template set. As well as a technical problem (missing classes in a fixed repertoire), this is also a biologically important problem (as when birds acquire new vocalizations). It would be very valuable to modify the one stage DTW (token passing) algorithm in such a way as to include adaptive thresholds on the discrepancies in measures for matching the input stream against the reference (training) patterns. Such thresholds will not only prevent these algorithms from often obvious errors, but will also allow a closed form of self-learning: first, by incorporating unrecognized elements in a new training/template set; second, by adapting the thresholds to the modified training/template set. Such an algorithm would have useful applications in many areas.

IV. SUMMARY

Massive databases are common in bioacoustic studies of animal vocalizations. Automated analyses may significantly facilitate such studies. We explored the application of DTW and HMM recognition techniques to a large database of continuous bird song recordings. Our experiments highlight the advantages and limitations of the two approaches. Excellent performance was achieved under some conditions. Nevertheless, further development is needed to improve the automated recognition of short duration bird song sounds, and to improve performance under adverse conditions.

ACKNOWLEDGMENTS

We thank Sven E. Anderson and Alexander Peryshkin for help in using and modifying LCSR, Amish Dave for modifying some auxiliary code, and Han Y. Kim for extensive work preparing the vocalization database. S.E.A. provided valuable comments on the manuscript. This work was supported in part by U.S. Army Research Office (DACA88-95-C-0016) and the NIH (NS25677).

APPENDIX A: BAUM–WELCH REESTIMATION

At the initialization stage, the observation vectors are equally subdivided among the model states for each label, then the initial MLE of the mean and variance for each state j are computed as averages

$$\hat{\mu}_j=\frac{1}{T}\sum_{t=1}^{T} o_t, \quad \hat{\Sigma}_j=\frac{1}{T}\sum_{t=1}^{T}(o_t-\mu_j)(o_t-\mu_j)', \qquad \text{(A1)}$$

where the prime denotes the vector transpose. Then, the ML state sequence is found by using the Viterbi algorithm (see Appendix B), the observation vectors are reassigned to the states that were found, and Eq. (A1) is used again to get better estimates. This process is repeated until the estimates converge below some threshold value.

Because each observation vector o_t may contribute to the ML parameter values for each state j, the initial estimates are then improved by assigning each observation to every state in proportion to the probability $L_j(t)$ of the model being in that state j when the vector is observed at time t

$$\begin{aligned}\hat{\mu}_j&=\sum_{t=1}^{T} L_j(t)o_t \Big/ \sum_{t=1}^{T} L_j(t),\\ \hat{\Sigma}_{\mathbf{j}}&=\sum_{t=1}^{T} L_j(t)(o_t-\mu_j)(o_t-\mu_j)' \Big/ \sum_{t=1}^{T} L_j(t),\end{aligned} \qquad \text{(A2)}$$

where $L_j(t)=P(x(t)=j/\mathbf{O},\mathbf{H})$ and is computed by using the Forward–Backward algorithm (Rabiner and Juang, 1993). Equations (A2) are the Baum–Welch reestimation formulas for the means and covariances of an HMM. Similar formulas can be derived for reestimation of the coefficients of the mixtures and the transition probabilities. These Baum–Welch formulas are used in HTK first for reestimation of the parameters of each single HMM by training isolated labels, and then for embedded training when all HMMs are trained in parallel using the entire training set.

APPENDIX B: VITERBI ALGORITHM

Given an HMM, let $\pi_j(t)$ represent the maximum log likelihood of observing acoustic vectors $\{o_1,\ldots,o_t\}$ and being in state j at time t. Then, $\pi_j(t)$ can be efficiently computed using the following dynamic programming equation:

$$\pi_j(t)=\max_i\{\pi_i(t-1)+\log(a_{ij})\}+\log\, b_j(o_t), \qquad \text{(B1)}$$

where $\pi_1(1)=0$, and $\pi_j(1)=\log(a_{1j})+\log b_j(o_1)$, for $j\neq 1$. The MLE state sequence $\hat{X}$ is then retrieved by keeping track of the argument which maximizes Eq. (B1) for each j and $1<t\leqslant T$

$$f_j(t)=\operatorname{argmax}_i\{\pi_i(t-1)+\log(a_{ij})\}$$

by backtracking, $\hat{\imath}_t=f_{t+1}(\hat{\imath}_{t+1})$,

$$t=T-1,\ldots,1, \text{ starting from } \hat{\imath}_T=\operatorname{argmax}_i\{\pi_T(i)\}.$$

Anderson, S. E., Dave, A. S., and Margoliash, D. (**1996**). ''Automated recognition and analysis of birdsong syllables from continuous recordings,'' J. Acoust. Soc. Am. **100**, 1209–1219.

Bell, A. J., and Sejnowski, T. J. (**1996**). ''Learning the higher-order structure of a natural sound,'' Comp. Neural Sys. **7**, 261–267.

Bourlard, H., Konig, Y., and Morgan, N. (**1994**). ''REMAP: Recursive estimation and maximization of a posteriori probabilities, application to transition-based connectionist speech recognition,'' Technical Report TR-94-064, International Computer Science Institute, Berkeley, CA.

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, and 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.

Catchpole, C. K., and Slater, P. J. B. (**1995**). *Bird Song: Biological Themes and Variations* (Cambridge U.P., Cambridge, England).

Chang, P. C., and Juang, B. H. (**1993**). ''Discriminative training of dynamic programming based speech recognizers,'' IEEE Trans. Speech Audio Process. **1**, 135–143.

Deller, J. R., Proakis J. G., and Hansen J. H. L. (**1993**). *Discrete-Time Processing of Speech Signals* (Prentice-Hall, Upper Saddle River, NJ).

Fine, S., Singer, Y., and Tishby (**1998**). ''The hierarchical hidden Markov model: analyses and applications,'' Machine Learning (to be published).

Forney, Jr., G. D. (**1973**). ''The Viterbi algorithm,'' Proc. IEEE **61**, 268–278.

Gidas, B., and Murua, A. (**1996**). ''Stop consonant discrimination and clustering using nonlinear transformations and wavelets,'' in *Image Models (and Their Speech Model Cousins)*, edited by S. E. Levinson and L. Shepp (Springer-Verlag, New York), pp. 13–62.

Hihi, S. E., and Bengio, Y. (**1995**). ''Hierarchical recurrent neural networks for long-term dependencies,'' in *Advances in Neural Information Processing Systems 7*, edited by G. Tesauro, D. Touretzky, and T. Leen (Morgan Kaufmann), pp. 493–499.

Ito, K., Mori, K., and Iwasaki, S.-i. (**1996**). ''Application of dynamic programming matching to classification of budgerigar contact calls,'' J. Acoust. Soc. Am. **100**, 3947–3956.

Juang, B. H., Chou, W., and Lee, C. H. (**1997**). ''Minimum classification error rate methods for speech recognition,'' IEEE Trans. Speech Audio Process. **5**, 257–265.

Kobatake, H., and Matsunoo, S. (**1994**). ''Degraded word recognition based on segmental signal-to-noise ratio weighting,'' Proc. IEEE ICASSP I-425–I-428.

Kogan, J. A. (**1996**). ''Hidden Markov models estimation via the most informative stopping times for the Viterbi algorithm,'' in *Image Models (and Their Speech Model Cousins)*, edited by S. E. Levinson and L. Shepp (Springer-Verlag, New York), pp. 115–130.

Konishi, M. (**1965**). ''The role of auditory feedback in the control of vocalization in the white-crowned sparrow,'' Z. Tierpsychol. **22**, 770–783.

Kubala, J. (**1995**). ''Design of the 1994 CSR Benchmark Tests,'' in *Spoken Language Systems Technology*, Workshop (ARPA), pp. 41–46.

Larkin, R. P., Margoliash, D., Kogan, J. A., and Pater, L. L. (**1996**). ''Recognition of the utterances of terrestrial wildlife: A new approach,'' J. Acoust. Soc. Am. **99**, 2532.

Makhoul, J., and Schwartz, R. (**1995**). ''State of the art in continuous speech recognition,'' Proc. Natl. Acad. Sci. USA **92**, 9956–9963.

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

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

Marler, P., and Peters, S. (**1982a**). ''Developmental overproduction and selective attrition: New processes in the epigenesis of birdsong,'' Dev. Psychobiol. **15**, 369–78.

Marler, P., and Peters, S. (**1982b**). ''Subsong and plastic song: Their role in the vocal learning process,'' in *Acoustic Communication in Birds, Vol. 2, Song Learning and Its Consequences*, edited by D. E. Kroodsma and E. H. Miller (Academic, New York), pp. 25–50.

McLachlan, G. (**1992**). *Discriminative Analysis and Statistical Pattern Recognition* (Wiley, New York).

Ney, H. (**1984**). ''The use of a one-stage dynamic programming algorithm for connected word recognition,'' IEEE Trans. Acoust., Speech, Signal Process. **29**, 284–297.

Ney, H., Mergel, D., Noll, A., Paeseler, A. (**1992**). ''Data driven search organization for continuous speech recognition,'' IEEE Trans. Signal Process. **40**, 272–281.

Payne, R. B., Thompson, W. L., Fiala, K. L., and Sweany, L. L. (**1981**). ''Local song traditions in indigo buntings: cultural transmission of behavior patterns across generations,'' Behaviour **77**, 199–221.

Rabiner, L. R., and Juang, B. H. (**1993**). *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliffs, NJ).

Ron, D., Singer, Y., and Tishby, B. (**1996**). ''The power of amnesia: learning probabilistic automata with variable memory length,'' Machine Learning **25**, 117–149.

Sankoff, D., and Kruskal, J. B. (**1983**). *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison* (Addison-Wesley, Reading, MA).

Shikano, K., and Itakura, F. (**1991**). ''Spectrum distance measures for speech recognition,'' in *Advances in Speech Signal Processing*, edited by S. Furui and M. Sondhi (Marcel Dekker, New York), pp. 419–452.

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.

Thompson, W. L. (**1970**). ''Song variation in a population of indigo buntings,'' Auk. **87**, 58–71.

Thorpe, W. H. (**1958**). ''The learning of song patterns by birds, with especial reference to the song of the chaffinch, Fingilla coelebs,'' Ibis **100**, 535–570.

Vintsyuk, T. K. (**1971**). ''Element-wise recognition of continuous speech composed of words from a specified dictionary,'' Kibernetika **7**, 133–143.

Viterbi, A. J. (**1967**). ''Error bounds for convolutional codes and an asymptotically optimal decoding algorithm,'' IEEE Trans. Inf. Theory **IT-13**, 260–269.

Williams, H. (**1990**). ''Bird song,'' in *Neurobiogy of Comparative Cognition*, edited by R. C. Kesner and D. C. Olton (Erlbaum, Hillsdale, NJ), pp. 77–125.

Young, S., Odel, J., Ollason, D., Valtchev, V., and Woodland, P. (**1995**). *The HTK Book* (Cambridge University Technical Services Ltd., Cambridge, England).

Young, S., Russel, N. H., and Thorton, N. H. (**1989**). ''Token passing: a simple conceptual model for connected speech recognition system,'' Technical Report, Cambridge University Engineering Department.

Yu, A. C., and Margoliash, D. (**1996**). ''Temporal hierarchical control of singing in birds,'' Science **287**, 1871–1875.

Zann, R. A. (**1996**). *The Zebra Finch* (Oxford U.P., Oxford).