Unsupervised bird song syllable classification using evolving neural networks

Louis Ranjard[a] and Howard A. Ross[b]
Bioinformatics Institute, School of Biological Sciences, University of Auckland, Auckland 1142, New Zealand

(Received 4 December 2007; revised 5 May 2008; accepted 6 March 2008)

Evolution of bird vocalizations is subjected to selection pressure related to their functions. Passerine bird songs are also under a neutral model of evolution because of the learning process supporting their transmission; thus they contain signals of individual, population, and species relationships. In order to retrieve this information, large amounts of data need to be processed. From vocalization recordings, songs are extracted and encoded as sequences of syllables before being compared. Encoding songs in such a way can be done either by ear and spectrogram visual analysis or by specific algorithms permitting reproducible studies. Here, a specific automatic method is presented to compute a syllable distance measure allowing an unsupervised classification of song syllables. Results obtained from the encoding of White-crowned Sparrow (*Zonotrichia leucophrys pugetensis*) songs are compared to human-based analysis. © *2008 Acoustical Society of America.*
[DOI: 10.1121/1.2903861]

PACS number(s): 43.60.Np, 43.60.Lq, 43.80.Ka [JAS] Pages: 4358–4368

I. INTRODUCTION

Song birds (Passeriformes) have learned, rather than innate, songs.[1] The learning occurs when the bird is immature and in some cases, learning continues throughout a bird's life. This learning process is divided into two steps; first, a bird hears and memorizes a song from another bird, and second, it tries to imitate the song as accurately as possible. This imitation is not perfect, and so songs evolve through generations as small changes occur. From one generation to the next, a particular component of a song can be kept or disappear because no bird would have copied it. The more frequently this component occurs, the greater chance it has to be copied. Therefore, it is possible to define a neutral model of evolution of songs.[2] If two populations with a common origin are isolated, one can expect that the songs of each will accumulate modifications independently and then, with a sufficient number of generations, become significantly different. Being able to detect those changes can help us to infer population histories and relationships.

Here, bioacoustic methods developed for encoding bird songs as sequences of discrete syllables are presented. First, syllables are extracted from segmented song recordings and their spectrograms are encoded. They are then classified using a new kind of artificial neural network which utilizes dynamic time warping for the learning stage. One way to perform sequence comparison is to use alignment algorithms that minimize a distance function between a pair of sequences.[3] This approach is popular in the field of molecular biology, in speech processing where it is commonly named dynamic time warping and also in bioacoustics.[4,5] This technique has been used because it allows us to perform unsupervised classification. Here, a pairwise syllable distance measure, calculated on the basis of mel-cepstrum coefficients dynamic time warping, will be introduced. This measure is proportional to the number of operations required to transform one sequence into another, while producing an alignment. Then, this distance is used for classifying syllables into different clusters. Self-organizing maps are a kind of neural network designed for data representation and classification.[6] During a learning stage, the network is trained with the data samples. However, this process can be very time consuming and the size of the map has to be arbitrarily chosen; therefore, modifications to the shape of the self-organizing map neural network have been developed. Evolving tree neural networks[7] are not only faster to train but also allow unsupervised clustering as they grow through a learning process, reaching a size proportional to the number of clusters required. The syllable distance introduced above is incorporated in such networks and weighted average sequences are derived from the syllable sequence alignments, supporting the learning process. Syllable classification obtained from the encoding of a small data set of songs belonging to a subspecies of White-crowned Sparrow (*Zonotrichia leucophrys pugetensis*) shows how the classification from this approach corresponds to those defined by classic approaches. This analysis suggests that such a method can be useful in analyzing larger data sets and so permit large scale studies of bird vocalizations.

II. METHODS

Classic approaches in bioacoustics use basic signal features and/or human knowledge for both detection and segregation of bioacoustic signals from noise.[8] Although former technical limitations are no longer restrictive,[8] only a few song analysis tools have been inspired by automatic speech processing[9,10] and are fully automatic.[11] When humans apply expert knowledge to song analysis, it can introduce subjec-

[a]Electronic mail: l.ranjard@auckland.ac.nz
[b]Electronic mail: h.ross@auckland.ac.nz

 0001-4966/2008/123(6)/4358/11/$23.00

tivity which can impair the reproducibility of bird song studies. Furthermore, this approach becomes very time consuming and thus is not suitable for large data sets. However, modern speech processing techniques offer high accuracy in sound analysis for speech recognition or speaker verification.[12,13] Different measures of word distance have been developed[14] and they can be extended to provide syllable distance measures. On the one hand, animal vocalizations seem to be less complex than human speech but, on the other hand, available recordings are often collected in poor acoustic conditions. Therefore, the signal-to-noise ratio can be low. Moreover, the meanings of the vocalizations remain unknown so the significant features of the song may remain undetected.

A. Song segmentation and syllable encoding

No common definition of a song note or a syllable has been set in the literature. Here, it has been considered that a syllable is a part of a song characterized by a high value of autocorrelation of the signal and with a continuity in the fundamental frequency. Syllable boundaries are found by segregating the syllables from the background noise using the autocorrelation function of the signal throughout a song recording. In a second step, the boundaries at the beginning and end of the syllables are readjusted. The spectral roll-off, defined as the frequency below which a specific percentage of the energy occurs, is analyzed to complete this step. A minimum threshold in the length of a syllable is set at 50 ms. Dooling[15] showed that birds can distinguish shorter sounds but a minimum number of samples is required to perform accurate feature extraction.

a. Autocorrelation. The cross correlation is a standard method for estimating how two signals are linearly related. In the field of music analysis, it is useful for instrument recognition.[16] For each 50% overlapping window of 128 samples under 44.1 kHz sampling rate, the maximum $c_{\max}$ of the cross correlation of the signal $s_l(n)$ of the lth analysis window is calculated as

$$c_{\max}(l) = \max \sum_{n=0}^{N-m-1} s_l(n+m)s_l^*(n), \quad \forall\, 1 \le m \le N, \quad (1)$$

where s_l^* is the complex conjugate of s_l and N is the number of samples in s_l. This function is then smoothed by calculating a moving average, thus eliminating the slight local variations, in analysis windows of 512 samples with 50% overlap. Then, the potential syllables are characterized as the sections of this function greater than a specific threshold. This threshold is set as the median value of the function.

b. Frequency roll-off 60%. The spectral roll-off point ro is defined as the frequency value at which 60% of the signal energy is contained below in the spectrum.[17] Windows of 512 samples with 50% overlap are used to compute this value through the signal. The spectral roll-off point is higher during a syllable emission than during noise because bird songs are generally high-pitched and field recordings contain low frequency background noise. Slabbekoorn *et al.*[18] showed that birds can adjust the frequency of their vocalizations depending on the environmental noise. Therefore, it can be helpful for segregating the signal from the noise.

$$\mathrm{ro}(l) = f_{\mathrm{ro}}, \quad (2)$$

where l defines a window and f_{ro} is computed from the spectrum X of this window as

$$\sum_{k=1}^{f_{\mathrm{ro}}} X(k) = 0.60 \sum_{k=1}^{f_{\mathrm{Nyquist}}} X(k), \quad (3)$$

where f_{Nyquist} is half the sampling frequency. Each boundary is replaced by the local minima of the roll-off function, in a window starting 50 ms before and ending 50 ms after the time point found after analyzing the autocorrelation function.

In order to represent the syllables, it is possible to use frequency band filtering and then calculate the mel-cepstrum coefficients as well as the first-order and second-order delta features. This feature extraction approach has a great success in human speaker recognition applications.[19] The mel-frequency bands have been defined for human ears and therefore there is little support for their use with bird vocalizations. It has been decided to switch the frequency band filters in order to get more sensitivity in high frequencies where most bird vocalization occurs. However, specific bandwidth filter banks should be defined as species specific. More precisely, the first of the 26 filterbank channels starts at 1000 Hz and the last one terminates at 22.05 kHz. Under a sampling frequency of 44.1 kHz, a Hamming window of 128 samples with 50% overlap has been used for computing the spectra and the signal had first-order pre-emphasis applied using a coefficient of 0.97. Twenty coefficients were calculated and the $C0$th cepstral parameter was used as the energy component. This number of coefficients has been chosen based on empirical results (not shown). Two frames before and two frames after the current one were used to estimate the first- and second-order temporal derivatives. Energy normalization was implemented by subtracting the maximum value of the energy and adding 1.0. The cepstral coefficients were rescaled by liftering the cepstra using a coefficient of 22 so that they have similar magnitudes. This window size is much smaller than that used in previous work, for example, Trawicki,[20] and therefore provides sufficient precision to analyze the changes in the frequency content of each syllable. Therefore, each syllable is represented by a sequence of vectors, where each vector is composed of 63 coefficients, the mel-cepstrum coefficients, and the delta features, of consecutive overlapping windows.

B. Syllable distance measure D_s

Different approaches exist to measure a distance between two sounds; some use feature extraction, others spectral cross correlation.[21] There are also different approaches[3,14,22] for performing sequence comparisons and dynamic time warping. An interesting advantage of dynamic time warping is that it allows the computation of an average from the alignment of a pair of variable-length sequences. Moreover, slight changes in the rate of sound emissions are tolerated while computing the alignment. These techniques consist of finding the optimal alignment between the two sequences of vectors using dynamic programming. This in-

volves computing an edit distance that is proportional to the minimum of the sum of operation costs required to transform one sequence into the other one. The values of the operation costs are proportional to a distance measure between mel-cepstrum coefficient vectors. Many different distances have been implemented for speech processing.[12] Here, the Euclidean distance has been used for each pair of vectors in the sequences to be aligned. Let $X=x_1 \ldots x_N$ and $Y=y_1 \ldots y_M$ be two vector sequences to be compared. Five edit operations are considered:

(a) Substitution $S(v,w)$ defines the cost associated with the substitution of the vector v for the vector w. This cost is the vector distance described above.
(b) Insertion $I(v)$ defines the cost associated with the insertion of the vector v. This cost is set as half the average of the substitution cost.
(c) Deletion $D(v)$ defines the cost associated with the deletion of the vector v. This cost is set as half the average of the substitution cost.
(d) Compression $C(vw,x)$ defines the cost associated with the compression of the vectors vw into the vector x. It is defined as the mean of the substitution cost of the vector v for x and the substitution cost of the vector w for x.
(e) Expansion $E(v,wx)$ defines the cost associated with the expansion of the vector v into the vectors wx. It is defined as the mean of the substitution cost of the vectors w for v and the substitution cost of the vector x for v.

Then, a graph of size $G(a,b)$ with $1 \leqq a \leqq N$ and $1 \leqq b \leqq M$ is computed as

$$G(a,b) = \min \begin{cases} G(a-1,b) + I(x_a) \\ G(a,b-1) + D(y_b) \\ G(a-1,b-1) + S(x_a,y_b) \\ G(a-1,b-2) + C(x_a,y_{b-1}y_b) \\ G(a-2,b-1) + E(x_{a-1}x_a,y_b). \end{cases} \quad (4)$$

This graph is used to calculate the distance between a pair of syllables $D_s=G(N,M)/N+M$. Then, tracing the warping path back allows us to find the weighted average sequence of vectors as defined in Refs. 3 and 23 performing a time interpolation. In each step in the warping path, a weighted vector is computed

$$c_k = qx_a + (1-q)y_b \quad (5)$$

and

$$c_k = 0.5qx_a + 0.5qx_{a-1} + (1-q)y_b, \quad (6)$$

in the case of a compression or expansion. For deletions and insertions, the vector used in the average sequence is simply the one conserved in the alignment. The corresponding time point is the average between those of both sequences

$$t_k = |qt_a + (1-q)t_b|. \quad (7)$$

In a last step, all vectors belonging to the same time point are averaged in order to obtain a continuous time sequence.

C. Self-organizing neural network

Evolving trees are a variant of self-organizing maps[6,7] which allow the treatment of large amounts of data. An advantage of this approach is that it produces a lower dimensional representation of the data set hierarchically ordered, and this hierarchy can be used for the identification of clusters in the data set. Therefore, it is a suitable approach for unsupervised classification problems. Moreover, the learning time of the network is considerably smaller than for a self-organizing map as it is not necessary to compare each input vector with each network's weight matrix. Furthermore, the network is able to grow in order to reach a size suitable for the classification of a specific data set.

The syllable distance measure D_s defined above is used in order to hierarchically find the best matching unit in the evolving tree. Starting from the root of the network, a data sample is aligned with every child neuron weight matrix, choosing the closest one to go to the next level until a leaf neuron is reached. In this case, this neuron is defined as the best matching unit. The learning process is carried out in two stages. First, the data set is explored as the network grows quickly, and second the network is fine-tuned by pulling the neurons closer to the data samples which they aim to classify. Therefore, it is important to slow down the growing and learning rate of the network as it is trained on the sample data set. During an epoch, every data sample is used just once. For each syllable in the data set, the closest neuron, or best matching unit, is found in the network. Then, its weight matrix is updated by aligning the vector sequences and computing a weighted average sequence. Each neuron has a hit counter which is incremented every time it is chosen as a best matching unit. If this counter goes beyond the splitting threshold, the neuron is subdivided and child neurons are created. The number of new neurons created depends on the number of leaves specified. A link measure distance has been used to compute the distance between neurons in the network. This distance is computed as the number of edges separating two neurons in the network, using an implementation of the depth-first search algorithm. The neighborhood function of the neural network is defined as

$$h(c(t),i) = \alpha(t)\exp\left(\frac{-d(c(t),i)^2}{2\sigma(t)^2}\right), \quad (8)$$

where $d(c(t),i)$ is the distance between the neuron weight matrix $c(t)$ and the sample i, $\alpha(t)$ is the learning rate defined as

$$\alpha(t) = \max \begin{cases} \alpha(0)\exp\left(\frac{-t^2}{(0.75T)^2}\right) \\ \alpha_{\min}, \end{cases} \quad (9)$$

where T is the total number of epochs. $\sigma(t)$ is the neighborhood size at epoch t, defined as

$$\sigma(t) = \sigma(0)\exp\left(\frac{-t}{0.5T}\right). \quad (10)$$

The neighborhood function $h(c(t),i)$ determines the weight q used in the calculation of weighted average sequences. The

TABLE I. Values for the parameter of the network training. The left column shows the values used by default and the right column shows the selected ones, after estimation.

	Default values	Selected values
Number of epoch	5	5
Splitting threshold	50	20% of data set size
Neighborhood strength	3	1
Initial leaf number	2	3
Final leaf number	2	2
Gamma	0.95	0.90
Initial learning rate	0.90	0.90
Final learning rate	0.01	0.01

number of children can also be set to decrease linearly after each epoch, through the learning process

$$n(t+1) = \max\begin{cases} n(t) - \dfrac{n(t)}{T} \\ n_{\min}. \end{cases} \tag{11}$$

Moreover, a factor γ affects the counter of each neuron in the network, slowing down the expansion of the network[24]

$$\text{count}(c(t+1)) = \gamma \text{count}(c(t)). \tag{12}$$

D. Parameter selection

Two different features of the final neural network are used to assess the quality of the training process. First, the mapping precision is defined as the average distance between a data sample and the best matching unit in the tree. The second one is the tree size, which means the number of neurons of the neural network at the end of the learning process. The goal of unsupervised clustering is to find the best dimensionality reduction of a given data set which is directly linked to this value. In their experiments, Pakkanen *et al.*[24] used a splitting threshold of 60, implying the creation of four new neurons, to analyze a data set of 1000 vectors during ten epochs. In another experiment, they used a splitting threshold of 50, the creation of three new neurons after each split and Ref. 25 used different values of those parameters. In this study, several values of these parameters were tested in order to understand better how they can affect the final characteristics of the network. For those assessments, a data set of 200 White-crowned Sparrow syllables was employed. The default parameters values used are shown in Table I. To assess the quality of the neural network, the size of the tree and the mapping precision were evaluated for different values of the parameters.

The longer the network is trained the greater the number of neurons which will reside in the final network. The mapping precision will be better as each neuron would have been averaged with more data samples, summarizing fewer samples. As expected, the size of the final network linearly increased with the number of epochs, Fig. 1(a). However, the mapping precision did not behave in a similar way and decreased in two different stages. Mapping precision rapidly decreased at first, but in a second stage, it decreased more slowly. Therefore, a small number of epochs is sufficient to train a network. The best value for this parameter, given a specific data set, is not proportional to the number of data samples. It is actually related to the number of clusters in the data set, and more precisely the number of examples of each given cluster. Indeed, a network trained on a data set comprising only few clusters but with a great number of samples in each of them will reach a high mapping precision quicker than if it is trained with a more heterogeneous data set. The number of expected clusters is not known; therefore, it is not obvious how to choose a correct value for the number of epochs. In most of the experiments, a value of 5 was used.

The value of the splitting threshold will determine the rate at which the network grows. With increasing values for the splitting threshold, the size of the tree underwent an exponential decay, Fig. 1(b). At the same time, the mapping precision increased exponentially, corresponding to a higher average distance between a data sample and its best matching unit. After a stage of rapid decrease in the tree size and increase in the mapping precision lasting until a threshold of around 50, those values changed more slowly. The point of change occurs when the splitting threshold is approximately 20% of the data sample size.

The neighborhood strength defines the distance from the best matching unit, expressed as number of neurons, until the point at which the update will have an effect in the network. A high value means that an important region of the tree will be updated at each learning step. A value of 0 indicates that the classification is similar to *kmeans* clustering as pointed out by Ref. 24. The tree size was not sensitive to this parameter (data not shown). Considering the mapping precision, small values, a distance between 1 and 3, returned the best results. A value of 1 means that only the mother neuron in the tree network will be updated. Nevertheless, this could be different with larger networks. This parameter could also be expressed as a function of the size of the network. However, a value of 1 has been chosen in most of the experiments.

The number of leaves created at each learning step is expected to have a direct effect on the size of the final tree. It is less obvious what effect it will have on the mapping precision. Assessments, performed using a constant leaf number throughout the learning process, showed that the best results were obtained with three new leaves created at each growing step, Fig. 1(c). With higher values, assessments did not show any improvement in the mapping precision, even if the learning process required more calculation. In fact, to find the best matching unit for a given data sample, its distance to every neuron weight matrix is computed at each level in the network. Therefore, it is effective to use small values for this parameter.

Another set of assessments was performed using a decreasing number of leaves. Different initial values of the number of leaves were tested with a constant final number of leaves (data not shown). With a greater initial number of leaves, it is expected that the network will first organize itself to explore the data set while growing quickly, but at the end of the learning process each neuron will only need to be pulled closer to the data samples with little requirement for growth. This resulted in smaller networks but with similar mapping precision and thus potentially a smaller number of

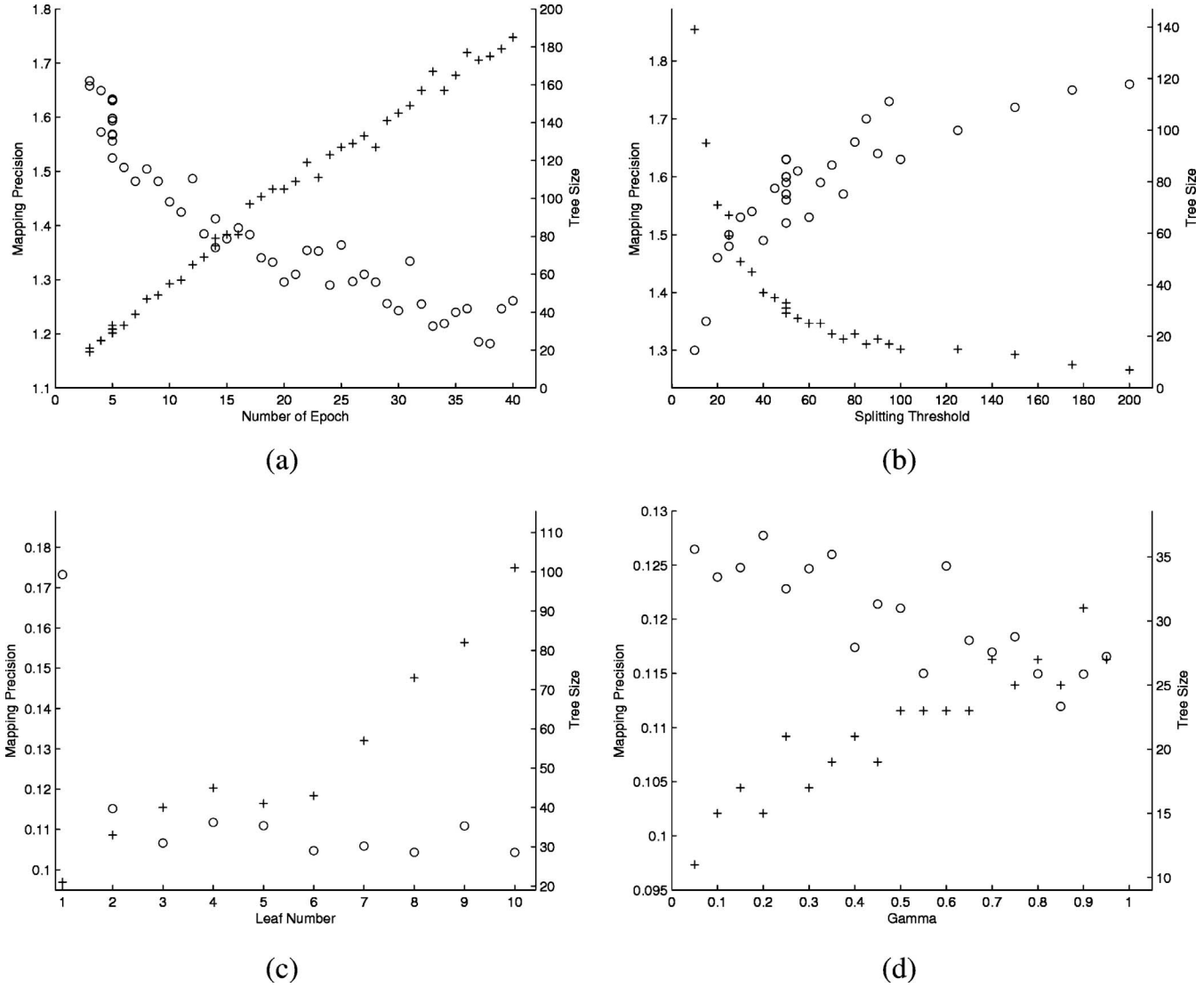

FIG. 1. Neural network size (+) and precision (○) are affected by varying values of the learning parameters: the number of epochs (a), the splitting threshold (b), the number of leaves (c), and gamma (d). The mapping precision is the average distance between the data samples and the best matching units in the network; therefore, lower values correspond to higher precision.

cluster centers. All assessments returned a better precision than the one obtained with a constant leaf number of 2, except for an initial value of 4. Surprisingly, with values greater than 8, the precision worsened while the final network was larger. More assessments were performed with decreasing number of leaves throughout the learning process. Different values for the initial and the final number of leaves were tested (data not shown). It appeared that initial numbers between 7 and 10 return the best mapping precision with an increasing network size. A small final number of leaves at the final stage of the learning process is shown to give the best results. Overall, the best results were obtained with a value decreasing from 7 to 3, 7 to 1, 8 to 3, and 8 to 1. A ratio around 2.5 between the initial and final numbers produced the best precision. More computing will be required for training the network, with higher initial number of leaves, because of the way the search for best matching units is performed. Therefore, these values have to be minimized and an initial value of 3 was selected.

Another way to slow the growth of the tree is to multiply the hit counter of each neuron by a factor gamma after the end of every epoch. In assessments using a constant number of leaves (2) at each split, larger values of the gamma factor reduced the mapping precision and increased the network size, Fig. 1(d). However, a better mapping precision can be obtained by varying the number of leaves, but with final trees of greater size.

The learning rate will define how strongly the neurons are pulled toward the data sample at every network update. This rate also decreased through the learning process in order to allow the network to first explore the data set and then assess it accurately. Although different initial values of the learning rate did not affect the final result significantly, better precision is returned with small final learning rates (data not shown). As a result of these assessments, the set of selected parameter values is presented in Table I and were used for other analysis.

It could be asked how well these results can be extrapolated to data sets of greater size. In particular, the first stages of learning, while the network is small, require the algorithm

TABLE II. Dialects are defined in Ref. 26 on the basis of the different types of simple syllables. The corresponding cluster numbers obtained with neural network classification are given for simple and complex syllables. Clusters in bold uniquely define a dialect. Cluster 27 contains one occurrence of a syllable belonging to the dialect B and the four complex syllables of the dialect E.

	Complex syllable		Simple syllable	
Dialect	Syllable type	Cluster	Syllable type	Cluster
A	22	11, 16, **30**, 31	1	**10**, 21, **23**, 24, **25**
B	12, 13	11, **12**, **27**, 31	2a	**4**, **19**, 22, 24, **26**
C	21	**6**, **8**, **20**, 31	7	**13**, 21, 22
D	n/a	**1**	8, 9	**3**, **32**
E	24	**27**	Absent	Absent

to be able to conquer the data set and thus to grow at a rate proportional to its size. Consequently, the parameters affecting this section, the initial leaf number and the neighborhood strength, should be expressed as a function of the data set size.

The final step of the syllable classification is performed by matching the data samples extracted from recordings to the neurons of the network. Each neuron defines a cluster center. An extra cluster merging step can be added to the classification. This step involves computing the pairwise distance between neurons and then merging the corresponding clusters if this distance falls below a specific threshold. It is also possible to apply a limit on the distance between a data sample and its closest neuron. Neurons of different depths define different levels of classification, basically obtaining clusters of different sizes and precision.

E. Validation with White-crowned Sparrow songs

To test whether an automated approach could reproduce a human classification, this method was applied to a set of songs from a species with well-documented dialects. A set of 17 songs, recorded between 1998 and 2000, was chosen to represent the diversity of song types sung by a subspecies of White-crowned Sparrow (*Zonotrichia leucophrys pugetensis*), in the west of the United States of America. Previous studies[26,27] have characterized dialects in this region by manually grouping the syllables of the same type. Human judges chose features to define each dialect and classified songs in accordance with these features. In Refs. 26 and 27, the visual classification is performed in two steps. First, syllables are classified into four main types: whistle, buzz, complex syllables, and simple syllables, depending on their spectrogram main shape and the position at which they occur in a song. Second, syllables belonging to the complex syllable and simple syllable types are classified into subtypes. It is this second classification which yields the definition of dialects. These 17 songs were visually classified into five dialects using the catalog of syllables of Nelson,[26] Table II. The complex syllable of dialect E was not found but it is visually different enough to constitute a new complex syllable type. To classify the syllables using the method described here, the rules introduced above for parameter selection were used. All syllables of this data set were classified by training a network, with the parameter values given above. After a neuron merging step, the network obtained was composed of 71 neurons with a mapping precision of 0.086. All data samples were classified and a threshold of 0.05 was used on the distance between weight matrices for cluster merging.

III. RESULTS

Using the method described here, the set of White-crowned Sparrow songs was found to contain 170 syllables in 32 clusters. The segmentation of the songs was performed by analyzing the autocorrelation and the roll-off of the songs as explained before. However, the threshold of the autocorrelation function has been manually modified in some cases, for the purpose of obtaining a consistent segmentation of the songs. Figure 2 shows an example of the spectrogram derived from one of those songs as well as the limits between syllables after song segmentation. When human experts classified the data set by visual inspection, 128 syllables were identified. The reason for this difference is that whistles and complex syllables have sometimes been subdivided by the segmentation algorithm, as can be noticed in Fig. 2. The resulting segments contain a continuous trace in the spectrogram separated from each other by short gaps of low frequency roll-off. By visual inspection, 28 of the 32 clusters are consistent with the four main syllable types (buzz, whistle, complex, and simple syllables). Figure 3 shows the spectrograms of the different data samples being part of clusters 4, 5, and 8. These three clusters illustrate characteristic cases. The presence of strong background noise, Fig. 3(a), has an influence on the clustering. Syllables are clustered in accordance with their dominant frequency, Fig. 3(b), and the variation in the shape of their spectrograms, Fig. 3(c).

This study differs from earlier studies in that all syllables are computationally classified at the same time rather than being classified in two steps. The neural network approach identified a larger number of clusters than had been

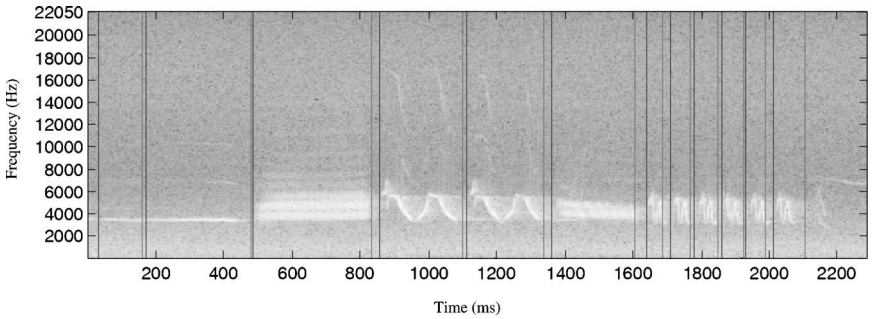

FIG. 2. Spectrogram of a song (Oak Harbor, WA, 1998, dialect D), showing the boundaries between syllables. The vertical axis is frequency in Hz and the horizontal axis is time in ms. This song consecutively contains two whistle syllables, a buzz syllable, two complex syllables, another buzz syllable, and finally six simple syllables which constitute an end trill.

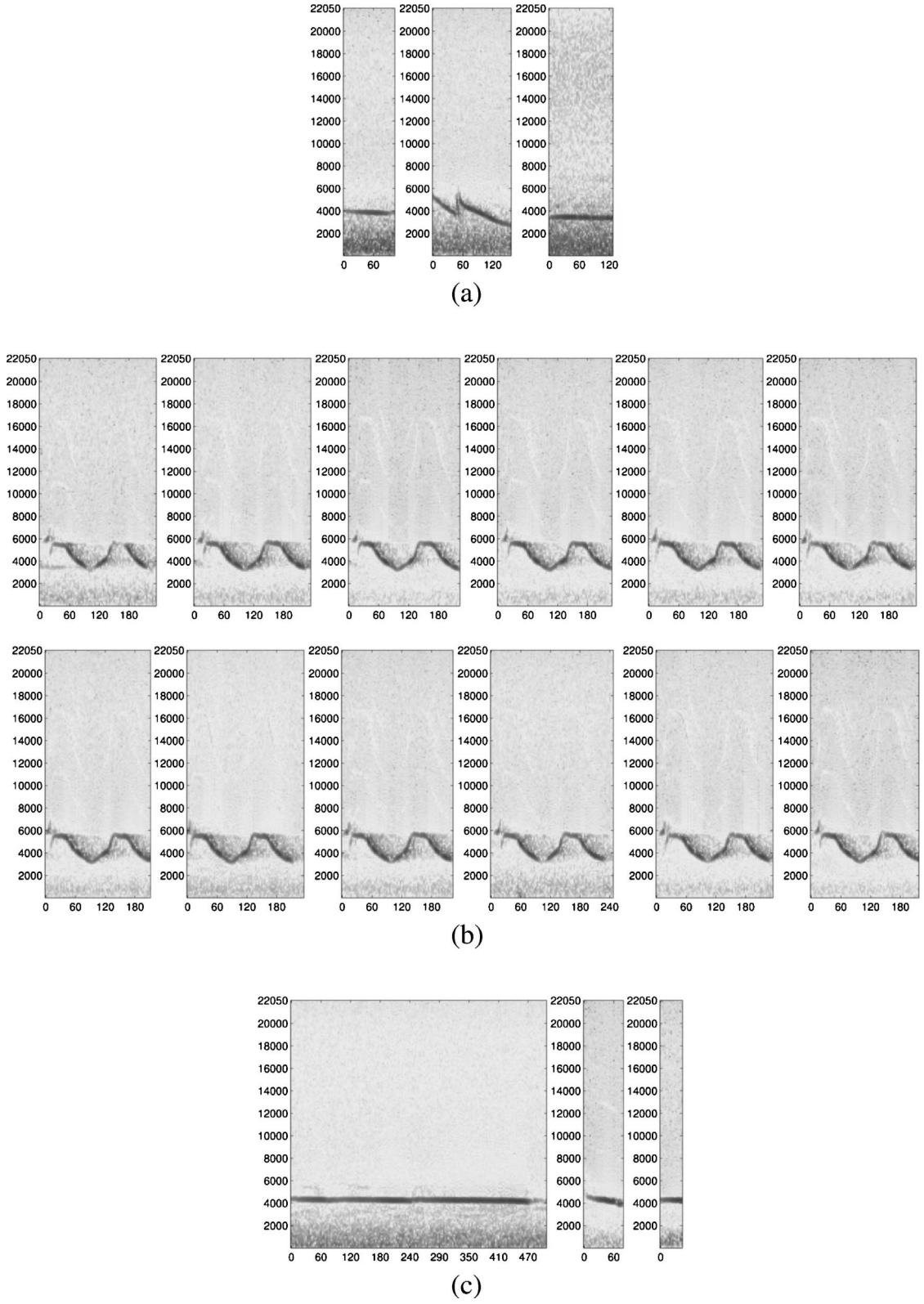

FIG. 3. Syllable spectrograms of the members of three different clusters. The first cluster (a), number 4, is affected by the presence of strong noise. Cluster 5 (b) appears to be consistent and Cluster 8 (c) comprises syllables sharing similar dominant frequencies. Axes are the same as in Fig. 2.

previously recognized so that a given syllable type corresponded to several clusters. For example, the simple syllable type 9 corresponds to the clusters 3 and 32, Table II. Moreover, several clusters contain only one representative. For two cases, the presence of strong noise has influenced the classification. This was particularly true when syllables had a short duration, Fig. 3. When syllables are very short, there is less information after sound encoding for performing the comparisons and the syllables are classified according to their dominant frequency rather than their spectral shape. Whistle syllables were classified according to their dominant frequency and buzz syllables according to their spectral

TABLE III. Occurrence of syllable types per localities, i.e., the number of times each syllable type is found in each different location. For each syllable cluster, the first level of visual classification is given, B for buzz, W for whistle, SS for simple syllable, and CS for complex syllable. The neural network cluster number is shown as well. The locations are ordered from north to south. Oak Harbor being the furthest north.

Neural net	18	16	30	10	25	23	20	13	7	12	19	26	2	3	9	5	32
Visual	W	CS	CS	SS	SS	SS	CS	SS	CS	CS	SS	SS	B	SS	W	CS	SS
Location[a]																	
OH													6	27	6	12	8
PC												1					
N										1	5						
Y									1								
F							1	3									
B		2	2	2	6	1											
PO	1																
Neural net	1	17	6	29	8	21	23	31	14	11	15	27	24	28	4		
Visual	W	W	B-CS	W	W-CS	SS	SS	CS	W	CS-W	B	CS	SS	B	W-SS		
OH														6	1		
PC								1	1	1	1	1	2		2		
N					1		1			2	1						
Y			3	1	1	9		3	1								
F	1	2	3	1			4	4									
B	2	1	3	1	1	2		4	2	1			2				
PO				2					1	1	1	4		3			

See Ref. 26 for location. OH, Oak Harbor; PC, Pacific City; N, Newport; Y, Yachats; F, Florence; B, Bandon; PO, Port Orford.

shape. Complex syllables, when they had been extracted in a similar way as in Refs. 26 and 27, were classified according to their geographical origin, for example, cluster 5, Fig. 3(b). However, when the segmentation process divided those syllables in short different parts, the alignment algorithm clustered the syllables with similar dominant frequency.

A unique set of clusters was identified for each of the visually specified syllable types used to define a dialect (Table II). The only exception was cluster 27 where a complex syllable, unique in the data set, was clustered with the four complex syllables of the dialect E. Similarity among dialects A, B, and C is shown by the shared possession of clusters from both the complex and simple syllables. Dialects D and E, on the other hand, had only unique syllable clusters.

Table III shows that each location possessed particular syllable types, which can be a consequence of the presence of dialects. However, some syllable types were present at different locations, showing regional patterns, for example, clusters 1, 17, 6, 29, and 8, while others did not, Table III. To test for regional patterns, the average geographic pairwise distance between the locations of every occurrence were calculated for each cluster. Then, a test statistic was defined as the average pairwise distance over all clusters. In the presence of dialect, this distance is expected to be smaller than if syllable clusters were randomly distributed among locations. Indeed, the number of shared syllables between the songs of any two locations should decrease with the geographical distance separating them. Over all syllables, the observed test statistic was less than that obtained for a random distribution ($N=100\,000$ replicates, data not shown). When each main group of syllables is examined individually, the strongest differences were for simple and complex syllables. A slight difference was apparent for whistle syllables, but in the case of buzz syllables, no significant difference between the statistic calculated with observed occurrences and that derived from random occurrences of syllables was noticeable.

Moreover, the alignment distances between the simple syllables reflect this geographic pattern, Fig. 4. The pairwise distance matrix between the most frequent syllables produced at the end of the songs for each location was used to build a hierarchical nearest neighbor tree. These syllables are repeated during the end trill of the songs, except for the songs of Port Orford which lack this part. This tree groups the locations sharing the same dialects, both Pacific City and Newport, and Yachats and Florence, together. However, it also suggests that some parts of the syllables are shared between dialects. The beginning of the syllables produced at Yachats and Florence looks similar to the terminal syllables of Bandon, and the end part resembles the syllables of Pacific City and Newport. Therefore, some syllables could be hybrids, created by the union or concatenation of syllables. Another explanation could be that these syllables have evolved as the consequence of the deletion of a part in the original syllables. The structure of the spectrogram of the terminal syllable at Oak Harbor are very different from the others, as can be seen in the tree.

IV. DISCUSSION

The segmentation of bird songs into syllables and the subsequent classification of those syllables have largely been based on human perception and aesthetics.[28] How birds perceive and parse songs is not known. The aim of this study has been to apply analytic methods to this problem, to extract reproducible classifications for large data sets.

The characteristics of evolving neural networks can be modified by using different values of the parameters direct-

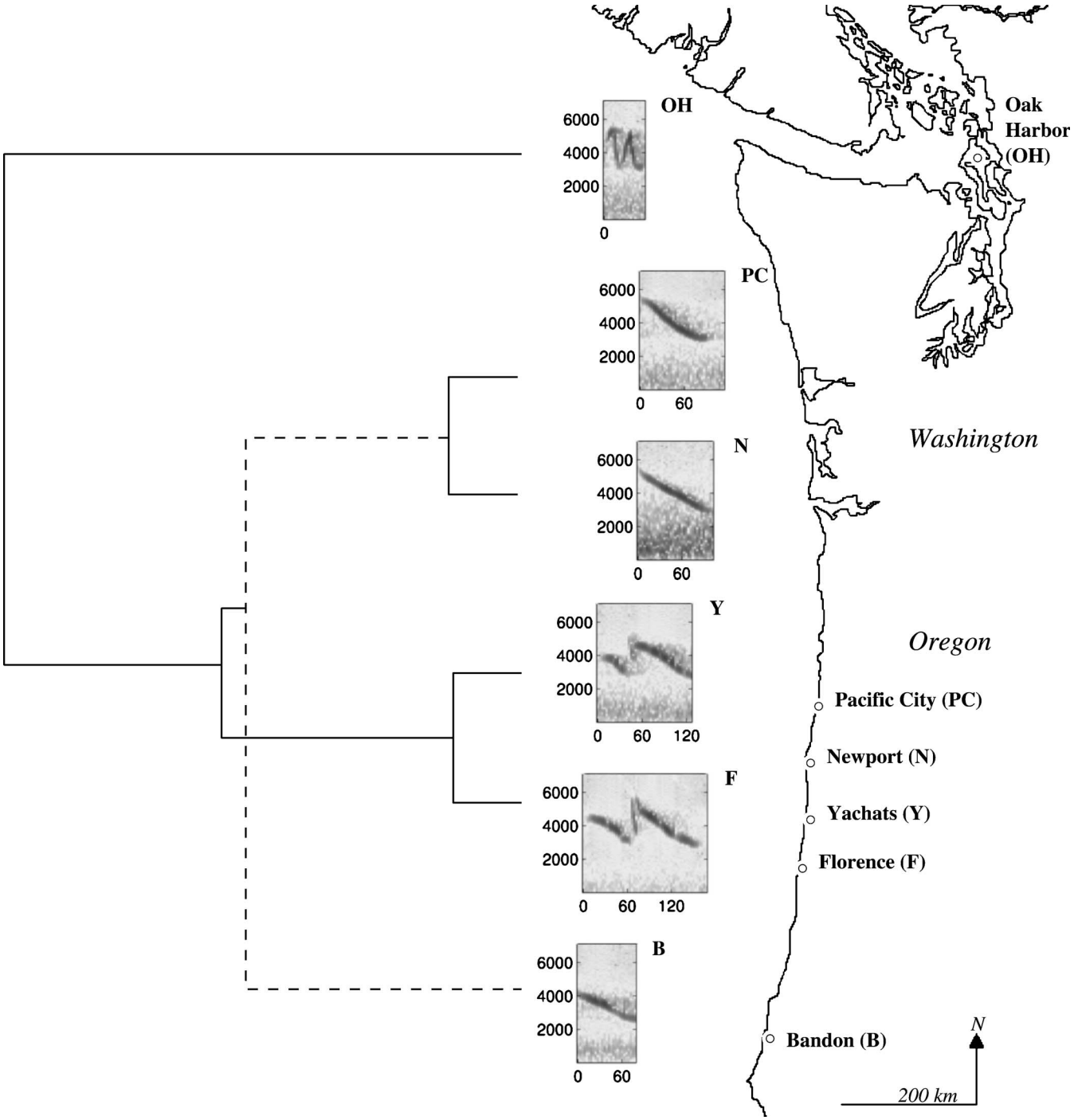

FIG. 4. Spectrograms of the most common clusters occurring at the end part of the songs. The spectrograms are ordered according to the geographic location where they were recorded. The spectrograms are positioned so that similar parts are vertically aligned in the time dimension. A nearest neighbor tree is shown on the left. Axes of the spectrograms are the same as in Fig. 2.

ing its growth. Those parameters are not independent but computer power limitations forbid an exhaustive parameter assessment. For example, the splitting threshold needs to be adjusted with the size of the tree. With a large initial number of leaves, the size of the network will increase rapidly and therefore the splitting threshold should be increased if one wants to limit the size of the final network. Single parameter assessments were conducted in order to get a better insight into their implications. The investigators may fine-tune the parameters for each specific data set depending on whether they wish to optimize the mapping precision or the size of the final network. Indeed, prior knowledge can indicate which aspect of the classification is more important. For example, the size of the network can be adjusted according to a range in the number of expected clusters.

Two main reasons can explain why the neural network classification conflicts with the visual classification in a few cases. First, the presence of strong background noise generates clusters which have been built principally on this criterion. With a visual approach, the noise is not taken into account for grouping syllables. In this work, frequencies below 1 kHz have been ignored, but the threshold could have been

increased, reducing the chance of creating spurious clusters. One could also focus on frequencies where birds have higher hearing sensitivity by excluding high frequencies and reducing the frequency bandwidth used for mel-cepstrum computation. In their work, Anderson *et al.*[4] limited the signal to 10 kHz. In this analysis, the recordings were not filtered before feature extraction and encoding of syllables. Ideally, the cepstrum coefficients should be computed with the help of a frequency scale specific to the bird species studied. Second, conflicts occur when long syllables are split into short parts during the segmentation of the songs. In this case, the information about frequency shifts throughout the syllable production is lost. The algorithm will cluster syllables differently depending on their duration. This shows the importance of the segmentation algorithm to this method. Appropriate threshold values must be chosen to obtain robust song segmentation.

The alignment distance calculation also has a strong effect. In particular, the insertion/deletion cost set for the alignment algorithm affects the importance of the syllables' duration in the distance calculation. Diminishing this cost will allow syllables of different lengths to be clustered together because of the resulting small pairwise distance. Consequently, the frequency similarities between the syllables will have greater importance. Weighted gap ends or weighted Euclidean distance between vectors are conceivable techniques for improving this distance measure. As noticed before, better results would be obtained by applying a limit to the distance between a sample and its best matching unit before affecting it to the cluster. In this way, potentially more centered clusters will be returned but at the expense of having some unclassified syllables. With larger data sets, each syllable type may occur more frequently so that each neuron would receive more hits of similar syllables in a constant number of epochs. This should improve the computation of the neuron weight matrices. An interesting aspect of this method is that it is able to retrieve small clusters containing rare syllables. This shows that the main cluster centers can be found in the data set. Therefore, it is possible to classify new recordings in order to examine how related they are to previously processed songs.

The distribution of the White-crowned Sparrow syllables, extracted from songs recorded in the west of America, strongly supports the presence of dialect. It is apparent that such culturally transmitted traits are continuously varying in space and time. The simple and complex syllables present the strongest geographical pattern but the whistle syllables seem to follow this trend too. Heterogeneity in the geographical patterns confirms that distinct positions in the White-crowned Sparrow songs evolve differently. Unique syllable clusters were found for most, but not all, dialects. A larger data set of songs and improved segmentation parameters would result in better resolution of syllable clusters and a closer match to human-based song classification. In some cases, the syllable's spectral structure seems to be a combination of different syllables, as seen in the end trill of the songs. Birds could potentially mix syllables together to produce new types. Thus, it is important to define relevant distance metric for performing comparisons independently from human biases.

V. CONCLUSION

Bird song syllable classification is a difficult task and the development of automatic methods acknowledges that the true classification is unknown. Indeed, bird song syllables are continuously varying characters. Retrieving cluster centers from syllable data sets can be achieved using evolving neural networks and a distance measure based on dynamic time warping. This method allows the processing of large data sets and reproducibility. Nevertheless, particular care must be given to the segmentation of the songs into syllables, to the encoding of the syllables, and to the choice of the distance measure and the parameters of the neural network learning process. It has been shown that this method is useful for the analysis of bird song evolution at different levels. First, the geographical distribution of syllables offers the opportunity to study dialects and potential population structure. Second, the distance measure gives a better insight into the fine spectrogram structure relationships between syllables.

ACKNOWLEDGMENTS

This work has been feasible thanks to the song recordings provided by the Borror Laboratory of Bioacoustics, Department of Evolution, Ecology, and Organismal Biology, Ohio State University, Columbus, OH and it has been supported by the Marsden Fund Council from Government funding, administrated by the Royal Society of New Zealand.

[1]P. Marler and M. Tamura, "Culturally transmitted patterns of vocal behavior in sparrows," Science **146**, 1483–1486 (1964).
[2]A. Lynch and A. J. Baker, "A population memetics approach to cultural evolution in chaffinch song: Meme diversity within populations," Am. Nat. **141**, 597–620 (1993).
[3]J. B. Kruskal and M. Liberman, "The symmetric time-warping problem: From continuous to discrete," in *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, edited by D. Sankoff and J. B. Kruskal (CSLI, Stanford, CA, 1999), Chap. 4.
[4]S. E. Anderson, A. S. Dave, and D. Margoliash, "Template-based automatic recognition of birdsong syllables from continuous recordings," J. Acoust. Soc. Am. **100**, 1209–1219 (1996).
[5]J. C. Brown, A. Hodgins-Davis, and P. J. O. Miller, "Classification of vocalizations of killer whales using dynamic time warping," J. Acoust. Soc. Am. **119**, EL34-EL40 (2006).
[6]T. Kohonen, "The self-organizing map," Proc. IEEE **78**, 1464–1480 (1990).
[7]J. Pakkanen, J. Iivarinen, and E. Oja, "The evolving tree—A novel self-organizing network for data analysis," Neural Processing Letters **20**, 199–211 (2004).
[8]D. W. Bradley and R. A. Bradley, "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 (CSLI, Stanford, CA, 1999), Chap. 6, pp. 189–209.
[9]M. D. Skowronski and J. G. Harris, "Acoustic detection and classification of microchiroptera using machine learning: Lessons learned from automatic speech recognition," J. Acoust. Soc. Am. **119**, 1817–1833 (2006).
[10]J. A. Kogan and D. Margoliash, "Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study," J. Acoust. Soc. Am. **103**, 2185–2196 (1998).
[11]O. Tchernichovski, F. Nottebohm, C. E. Ho, B. Pesaran, and P. P. Mitra, "A procedure for an automated measurement of song similarity," Anim. Behav. **59**, 1167–1176 (2000).

[12]S. Furui, *Digital Speech Processing, Synthesis, and Recognition*, 2nd ed. (Dekker, New York, 2000).
[13]B. Gold and N. Morgan, *Speech and Audio Signal Processing, Processing and Perception of Speech and Music* (Wiley, New York, 2000).
[14]G. Kondrak, "Phonetic alignment and similarity," Computers and the Humanities **37**, 273–291 (2003).
[15]R. J. Dooling, M. R. Leek, O. Gleich, and M. L. Dent, "Auditory temporal resolution in birds: Discrimination of harmonic complexes," J. Acoust. Soc. Am. **112**, 748–759 (2002).
[16]S. Essid, G. Richard, and B. David, "Musical instrument recognition by pairwise classification strategies," IEEE Trans. Audio, Speech, Lang. Process. **14**(4), pp. 1401–1412 (2006).
[17]M. Vacher, D. Istrate, L. Besacier, J. Serignat, and E. Castelli, "Life sounds extraction and classification in noisy environment," in SIP 2003: Fifth IASTED (The International Association of Science and Technology for Development) International Conference on Signal and Image Processing, Honolulu, HA, edited by M. Hamza, August 13–15, 2003.
[18]H. Slabbekoorn and A. den Boer-Visser, "Cities change the songs of birds," Curr. Biol. **16**(23), 2326–2331 (2006).
[19]S. Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Trans. Acoust., Speech, Signal Process. **ASSP–29**, 254–272 (1981).
[20]M. B. Trawicki, M. T. Johnson, and T. S. Osiejuk, "Automatic song-type classification and speaker identification of norwegian ortolan bunting (*Emberiza hortulana*) vocalizations," in 2005 IEEE Workshop on Machine Learning for Signal Processing (2005), pp. 277–282.
[21]S. Sharp and B. Hatchwell, "Individuality in the contact calls of cooperatively breeding long-tailed tits (*Aegithalos caudatus*)," Behaviour **142**, 1559–1575 (2005).
[22]B. J. Oommen, "String alignment with substitution, insertion, deletion, squashing, and expansion operations," Inf. Sci. (N.Y.) **83**, 89–107 (1995).
[23]P. Somervuo and T. Kohonen, "Self-organizing maps and learning vector quantization for feature sequences," Neural Processing Letters **10**, 151–159 (1999).
[24]J. Pakkanen, J. Iivarinen, and E. Oja, "The evolving tree—Analysis and applications," IEEE Trans. Neural Netw. **17**, 591–603 (2006).
[25]J. Vesanto, "Neural network tool for data mining: SOM toolbox," Technical Report, Proceedings of TOOLMET 2000 — 3rd International Symposium on Tool Environments and Development Methods for Intelligent Systems, April 13–14, University of Oulu, Oulu, Finland. http://citeseer.ist.psu.edu/388267.html.
[26]D. A. Nelson, K. I. Hallberg, and J. A. Soha, "Cultural evolution of Puget sound white-crowned sparrow song dialects," Ethology **110**, 879–908 (2004).
[27]L. F. Baptista, "Geographical variation in song and dialects of the Puget sound white-crowned sparrows," Condor **79**, 356–370 (1977).
[28]P. S. Warren, "Geographic variation and dialects in songs of the bronzed cowbird (*Molothrus aeneus*)," Auk **119**, 349–361 (2002).