ANIMAL BEHAVIOUR, 2000, **59**, 000–000
doi:10.1006/anbe.1999.1416, available online at http://www.idealibrary.com on IDEAL®

A procedure for an automated measurement of song similarity

OFER TCHERNICHOVSKI*, FERNANDO NOTTEBOHM*, CHING ELIZABETH HO†, BIJAN PESARAN† & PARTHA PRATIM MITRA‡

*The Rockefeller University, Field Research Center
†California Institute of Technology, Computation and Neural Systems
‡Bell Laboratories Lucent Technologies

(Received 25 May 1999; initial acceptance 24 August 1999; final acceptance 28 November 1999; MS. number: A8503)

Assessment of vocal imitation requires a widely accepted way of describing and measuring any similarities between the song of a tutor and that of its pupil. Quantifying the similarity between two songs, however, can be difficult and fraught with subjective bias. We present a fully automated procedure that measures parametrically the similarity between songs. We tested its performance on a large database of zebra finch, *Taeniopygia guttata*, songs. The procedure was an analytical framework of modern spectral analysis to characterize the acoustic structure of a song. This analysis provides a superior sound spectrogram that is then reduced to a set of simple acoustic features. Based on these features, the procedure detects similar sections between songs automatically. In addition, the procedure can be used to examine: (1) imitation accuracy across acoustic features; (2) song development; (3) the effect of brain lesions on specific song features; and (4) variability across different renditions of a song or a call produced by the same individual, across individuals and across populations. By making the procedure available we hope to promote the adoption of a standard, automated method for measuring similarity between songs or calls.

All true songbirds (order Passeriformes, suborder Oscines) are thought to develop their song by reference to auditory information (Kroodsma 1982). This can take the form of improvisation or imitation (Thorpe 1958; Marler & Tamura 1964; Immelmann 1969); both phenomena constitute examples of vocal learning because in both cases vocal development is guided by auditory feedback (Konishi 1965; Nottebohm 1968). Once sound-spectrographic analysis became available for the visual inspection of avian sounds (Thorpe 1954), the accuracy of vocal imitation among oscine songbirds became a focus of scientific interest. Some researchers were particularly interested in the choice of model or in the timing of model acquisition, others in the social context in which imitation occurred or in the brain mechanisms involved. All these approaches require a widely accepted way of describing and measuring the similarities that might exist between the song of a tutor and that of its pupil. Yet, quantifying the similarity between two songs (or calls) can be difficult and fraught with subjective bias. Most efforts at scoring song or call similarity have relied on visual inspection of sound spectrographs.

Visual scoring of song similarity can be made easier by partitioning the songs into 'syllables' or 'notes', defined as continuous sounds preceded and followed by silent intervals or by abrupt changes in frequency. The next step is to find for each of the notes of the tutor's song the best match in the pupil's song. According to the accuracy of this match, the pupil's note is assigned a numeric score. In two recent studies that used this procedure, notes for which there was a close match received a high score, those for which the match was poor or nonexistent received a low score and only notes that received high scores were said to be imitated (Scharff & Nottebohm 1991; Tchernichovski & Nottebohm 1998). It should be emphasized that imitation is always inferential and based on sound similarity as well as on other information. Clearly, the above scoring of similarity was done without the benefit of an explicit metric and the criteria for scoring similarity were arbitrary and idiosyncratic. None the less, despite these limitations, the visual approach to scoring similarity made good use of the human eye and brain to recognize patterns. This approach was satisfactory for studies aimed at establishing which songs are imitated, when model acquisition occurs, when imitation is achieved and how much of a model is learned (reviewed in Kroodsma 1982; Catchpole & Slater 1995; Zann 1996). However, song is a multidimensional phenomenon and this method is unsuitable for evaluating the components of similarity in a quantitative manner. A quantitative, automated scoring of similarity based on a

Correspondence: O. Tchernichovski, The Rockefeller University Field Research Center, Millbrook, NY 12545, U.S.A. (email: tcherno@rockvax.rockefeller.edu).

clear rationale and well-defined acoustic features would not only improve the quality of our measurements but also facilitate comparisons between results obtained by different laboratories.

Previous attempts to automate the analysis of song similarity have not gained general acceptance. Clark et al. (1987) suggested a sound-spectrographic cross-correlation as a way to measure the similarity between song notes: correlation between the spectrograms of the two notes was examined by sliding one note on top of the other and choosing the best match (the correlation peak). This method was later used for studying intraspecific variation of song learning in white-crowned sparrows, *Zonotrichia leucophrys* (Nelson et al. 1995). However, measures based on the full spectrogram suffer from a fundamental problem: the high dimensionality of the basic features. Cross-correlations between songs can be useful if the song is first partitioned into its notes and if the notes compared are simple, but even in this case mismatching a single feature can reduce the correlation to baseline level. For example, a moderate difference between the fundamental frequencies of two complex sounds that are otherwise very similar would prevent us from overlapping their spectrograms.

The cross-correlation approach, as mentioned above, requires, as a first step, that the song be partitioned into its component notes or syllables. This, in itself, can be a problem. Partitioning a song into syllables or notes is relatively straightforward in a song such as that of the canary, *Serinus canaria* (Nottebohm & Nottebohm 1978), in which syllables are always preceded and followed by a silent interval. Partitioning a song into syllables is more difficult in the zebra finch, *Taeniopygia guttata*, whose song includes many changes in frequency modulation and in which diverse sounds often follow each other without intervening silent intervals. Thus, the problems of partitioning sounds into their component notes and then dealing with the complex acoustic structure of these notes compound each other. In the present study we describe a procedure that addresses both of the above difficulties. It achieves this by reducing complex sounds to an array of simple features and by implementing an algorithm that does not require that a song be partitioned into its component notes.

Our approach is not the first one to grapple with these problems. Nowicki & Nelson (1990) first suggested an analytical approach to song comparisons using a set of 14 acoustic features for categorizing note types in the black-capped chickadee, *Poecile atricapillus*. Here too, partitioning of the song into its component notes was required although this method was not used to score the overall similarity between the songs of two birds. A similar analytical approach to the characterization of sounds was also used for bat communication calls, in a study that searched for neuronal correlates of different acoustic features (Kanwal et al. 1994; Esser et al. 1997). Recently, new techniques have been introduced for automatically partitioning a song into its component parts (notes, chunks or motifs). Kogan & Margoliash (1998) applied techniques borrowed from automated speech recognition for recognizing and categorizing these song parts. They demonstrated that these techniques work well for automated recognition of song units in the zebra finch and in the indigo buntings, *Passerina cyanea*. A robust automatic categorization of units of vocalization is an important step towards an objective scoring of similarity; however, the problem of scoring song similarity was not addressed.

To solve this latter problem, Ho et al. (1998) developed an analytical framework for the automated characterization of the vocalizations of a songbird. Their approach is based upon a robust spectral analysis technique that identifies those acoustic features that have good articulatory correlates, based on in vitro observations and theoretical modelling of sound production in an isolated syrinx (Fee et al. 1998). The acoustic features that Ho et al. (1998) chose to characterize zebra finch song are represented by a set of simple, unidimensional measures designed to summarize the multidimensional information present in a spectrogram. A procedure for scoring similarity, based on such an analytic framework has two advantages. (1) It enables the examination of one acoustic feature at a time, instead of having to cope with the entire complexity of the song of two birds. A distributed and then integrated assessment of similarity across different features promotes stability of scoring. (2) It also has the potential to evaluate how each of the chosen features emerges during development and is affected by different experimental manipulations.

The automated procedure we present here is based on the analytical approach suggested by Ho et al. (1998). We tested this procedure on a large database of zebra finch songs, including the songs of pairs of birds known to have had a tutor–pupil relation. The formal description of the song features that we measured, the spectral analysis techniques used and the rationale for using them appear in Ho et al. (1998). We describe the new technique in a manner that, we hope, will be useful and accessible to biologists. We then present the computational frame of our procedure and focus on the meaning and the limitations of the computational steps. Finally, we test the procedure and present a few examples that demonstrate its power. We have incorporated the procedure (including all the graphical tools presented in this article) into a user-friendly Microsoft Windows™ application, available at no charge for purposes of studying animal communication (excluding human) from O. Tchernichovski. We are aware that our procedure is sensitive to the nature of the sounds compared; researchers that wish to use it may have to modify it to maximize its usefulness in species whose sounds are very different from those of the zebra finch. However, we hope that our program will promote the adoption of an automated standard for measuring vocal imitation in birds.

METHODS

Song Recording

We recorded female-directed songs (Morris 1954; reviewed in Jarvis et al. 1998) in a soundproof room. A female and a male were placed in two adjacent cages. An

omnidirectional Shure™ 33-1070C microphone was placed just below the perch used by the female so that the male sang facing the microphone. Songs were digitally recorded using Goldwave™ sound recorder software at a frequency of 44 100 Hz and at an accuracy of 16 bits.

Glossary of Terms and Units of Analysis

The following terms characterize the kinds of spectral analysis done by our algorithm and thus allow us to calculate the four sound features that we used to quantify song similarity.

Song notes

A song note is a continuous sound (Price 1979; Cynx 1990) bordered by either a silent interval or an abrupt transition from one frequency pattern (e.g. a stack of harmonically related frequencies) to a different one (e.g. a frequency vibrato or a pure tone).

Song motifs

A song motif is composed of dissimilar notes repeated in fixed order.

Fourier transformation

Fourier transformation (FT) transforms a short segment of sound to the frequency domain. The FT is implemented algorithmically using the fast Fourier transformation technique (FFT).

Time window

The time window is the duration of the segment of sound upon which FFT is performed, in our case 7 ms. The time window determines both time and frequency resolution of the analysis. In this study 307 samples of sound pressure were obtained during the 7-ms period, which corresponds to a frequency resolution of 287 Hz. The next window starts 1.4 ms after the beginning of the previous one and therefore has an 80% overlap. The spectrogram is a sequence of spectra computed on such windows, typically represented as an image where power is represented on a scale of grey ranging from white to black. Because frequency resolution is finite, the spectrogram does not capture a 'pure sine wave' but represents 'frequency' as a 'trace'. The width of this trace is, in our case, 287 Hz.

Multitaper spectral analysis

Multitaper (MT) methods are a framework for performing spectral analysis (Thomson 1982). In particular, they produce spectral estimates that are similar but superior to the traditional spectrogram. Multitaper methods also provide robust estimates of derivatives of the spectrogram as well as a framework for performing harmonic analysis (detection of sine waves in a broadband noisy background). This technique is described in Percival & Walden (1993).

Spectral derivatives

Spectral derivatives are derivatives of the spectrogram in an appropriate direction in the time–frequency plane. These derivatives can be estimated using MT spectral methods (Thomson 1990, 1993). The derivatives have the same resolution as the spectrogram and are not artificially broadened. Here we use them for tracking frequency traces in the spectrogram. As one cuts across a horizontal frequency trace, from low to high, there is a sharp increase in power, then a plateau, then a decrease in power. The frequency derivatives for the same cut are first positive and then negative, passing through zero at the peak power location. A useful property of these derivatives is that they show a sharp transition from positive to negative values, providing a contour that is more accurately defined than the frequency trace. If the frequency trace is not horizontal, then the direction of maximum change in power is not in the frequency axis, but rather at an angle to both time and frequency axes. To capture the direction of maximal power change in the frequency trace, it is then natural to take a directional derivative perpendicular to the direction of frequency modulation. The directional derivative is easily computed as a linear combination of the derivatives in the time and frequency directions, and may be thought of as an edge detector in the time–frequency plane. We find the derivatives spectrogram an excellent means of visualizing the spectral information in a song.

We illustrate the above procedure in Fig. 1 using two examples. Figure 1a presents an MT spectrogram of a note, and Fig. 1b presents the directional time–frequency derivatives of the same note. The arrows below the time axis in Fig. 1 indicate the angle of the derivatives. As shown, this angle is perpendicular to the direction of frequency modulation. As a result of this edge detector technique, zero crossings (transitions from black to white in the middle of frequency traces) are equally sharp in the modulated and in the unmodulated portions of a note.

Peak frequency contours

Peak frequency contour is defined by the zero crossings of successive directional derivatives. Figure 1c presents the frequency contours as red lines and this constitutes a parametric representation of the sound analysed. It contains less information than the original sound but this information can be analysed more readily. By simplifying the song to a series of frequency contours we have excluded all information about absolute power. So, for example, the representation of the note with many harmonics shown in Fig. 1c shows all harmonics with equal emphasis, although it is clear from Fig. 1a that some harmonics were louder than others.

Features Used to Characterize and Compare Songs

Wiener entropy

Wiener entropy is a measure of randomness that can be applied to sounds (Ho et al. 1998), as shown in Figs 2a and 3a. It is a pure number, that is, it is unitless. On a

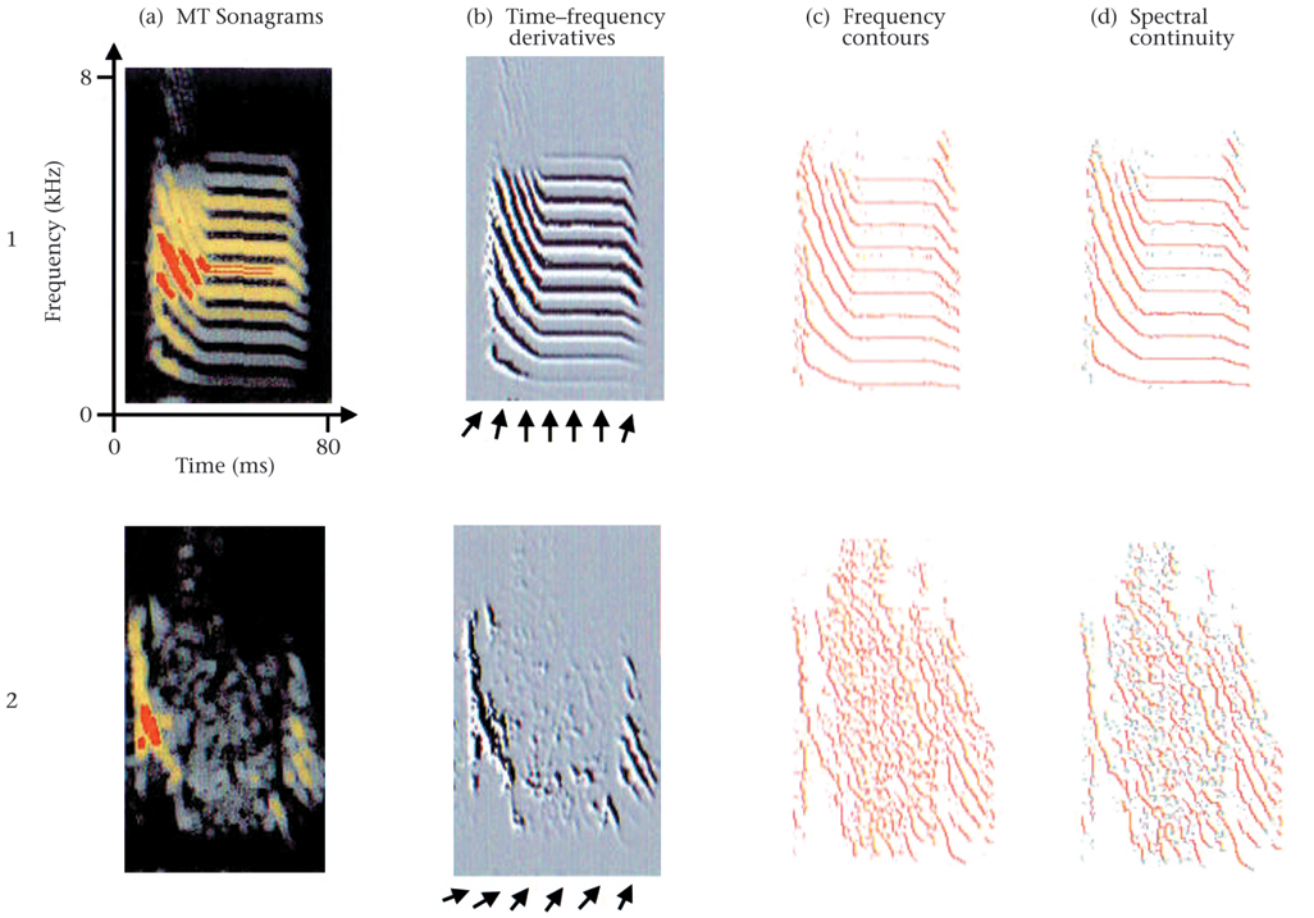

Figure 1. Computation of spectral derivatives and frequency contours of a note (example 2) and a song chunk (example 3). (a) Multitaper sound spectrograph improves the definition of frequencies. This technique allows us to approximate the spectral derivatives as shown in (b), where the light areas represent an increase of power, and the dark areas, a decrease of power. The arrows below the *X* axis in (b) indicate the direction of the derivatives presented. We chose the direction that maximizes the derivatives and hence the sharp transition between white and black at the middle of each frequency trace. This allows us to accurately locate frequency peaks of modulated and unmodulated frequencies as shown in (c). The red lines in (d) correspond to continuous frequency contours and the grey lines indicate discontinuous contours. Spectral derivatives are used in the analysis of pitch, frequency modulation and spectral continuity.

scale of 0–1, white noise has an entropy value of 1 and complete order; for example, a pure tone has an entropy value of 0. To expand the dynamic range, the Wiener entropy is measured on a logarithmic scale from 0 to minus infinity (white noise: log(1)=0; complete order: log(0)=minus infinity). The Wiener entropy of a multiharmonic sound depends on the distribution of the power spectrum: if narrow (the extreme of which is a pure tone), the Wiener entropy approaches minus infinity; if broad, it approaches zero. The amplitude of the sound does not affect its Wiener entropy value, which remains virtually unchanged when the distance between the bird and the microphone fluctuates during recording. Yet, the entropy time series (or curve) of a song motif is negatively correlated with its amplitude time series. This is because noisy sounds tend to have less energy than tonal sounds. A similar phenomenon has also been observed in human speech, where unvoiced phonemes have low amplitude. Wiener entropy may also correlate with the dynamic state of the syringeal sound generator, which shifts between harmonic vibrations and chaotic states (Fee et al. 1998). Such transitions may be among the most primitive features of song production and maybe of song imitation.

Spectral continuity

Spectral continuity estimates the continuity of frequency contours across time windows, as illustrated in Figs 1d, 2b, 3b. Frequency contours are mostly continuous in example 1 shown in Fig. 1d, but not in the more complex set of notes in example 2. It is clear from Fig. 1d that the noisier a sound, the lower its spectral continuity score and the higher its Wiener entropy. Importantly, although both measures are related to 'noise', they are measured orthogonally to each other: Wiener entropy is measured on the *Y* axis, spectral continuity is measured on *X* axis. Although at their extremes, Wiener entropy and spectral continuity are correlated, there is a broad middle range in these two measures where one does not predict the other. For example, adding more and more harmonics to a sound would not change spectral continuity but would increase Wiener entropy value.

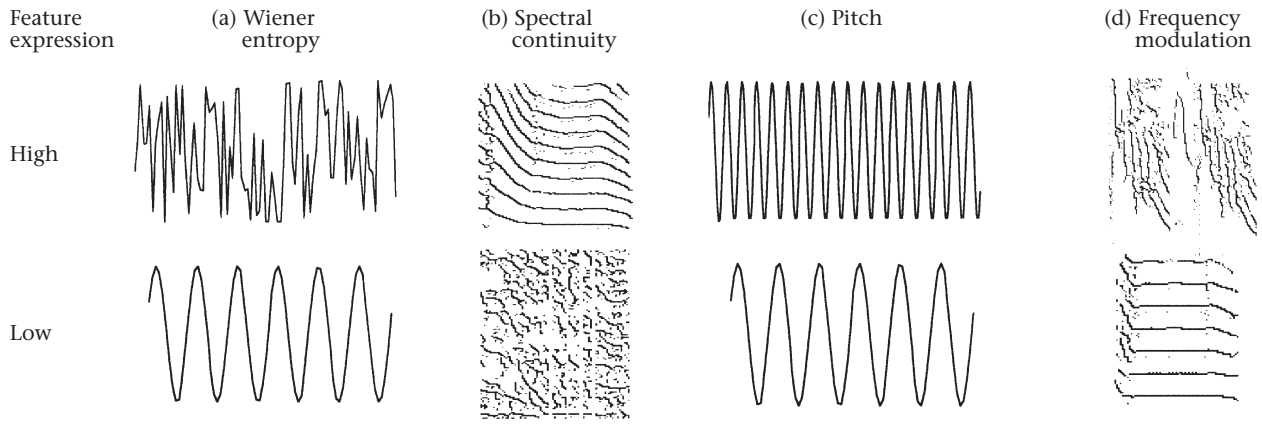

Figure 2. (a) Wiener entropy (a measure of randomness) is high when the waveform is random, and low when the waveform is of pure tone. (b) The spectral continuity value is high when the contours are long and low when the contours are short. (c) Pitch is a measure of the period of the sound and its value is high when the period is short and low when the period is long. (d) Frequency modulation is a measure of the mean slope of frequency contours.

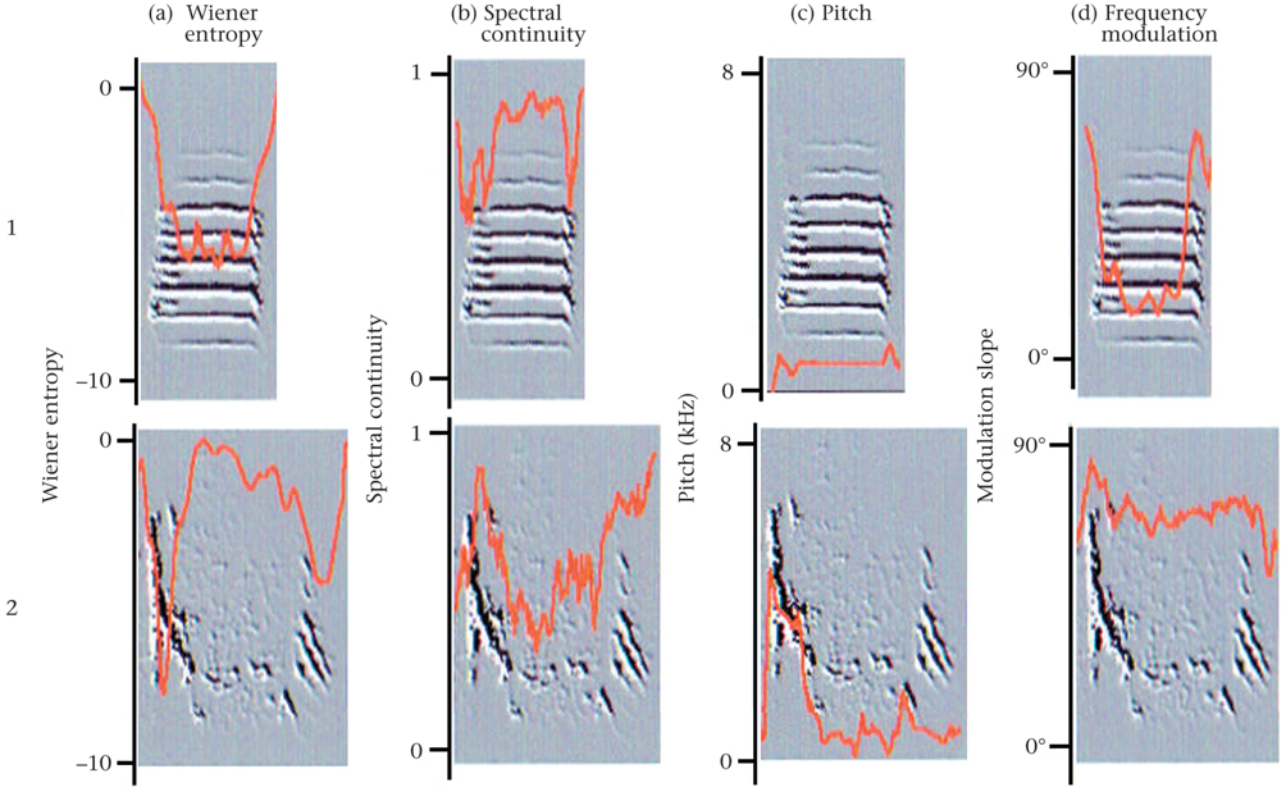

Figure 3. Values, in red, are presented for each of the four features in two examples of song chunks. The grey traces correspond to the time frequency of the sounds represented. Each sound has a unique combination of feature values. Note that values of different features may show independent changes (compare for example the curves of pitch and frequency modulation in example 2).

Continuity is defined according to the time and frequency resolution of the analysis. Sounds are examined across a grid of time and frequency pixels (each pixel 1.4 ms × 43 Hz). If a contour continues across five consecutive pixels (pixels that have at least one common corner), it crosses a section of 7 ms × 215 Hz, approximately the resolution of analysis, and is defined as continuous. A consecutive pixel can belong to the next consecutive time window or to the same window, but not to the previous window. On a scale of 0–1, continuity is 1 when all the frequency contours of a time window are continuous and 0 when none of the contours is continuous. Figure 3b presents examples of the continuity measurement.

Pitch

Pitch is determined by the period of a sound (Fig. 2c) and is a very important song feature. It is not always easy

to measure pitch. In the simplest situation, that of a pure tone, the frequency of this tone is its pitch. In sounds with many harmonics, pitch is the fundamental frequency, as defined by the separation between successive harmonics, and the median difference between consecutive frequency contours is our estimate of harmonic pitch. However, several situations can occur that require an explanation. In some cases, a harmonic contour in a stack of harmonics has been suppressed; this is rarely a problem because unless there are many missing harmonics the median measure remains unchanged. In addition, noisy sounds that do not offer clear stacks of harmonics (e.g. example 2 in Fig. 1d) can also occur. Careful inspection, however, reveals that there is an embedded structure of frequency contours that, for any time window, tends to show a periodic relation; in this case as well, the median gives a robust estimate of this embedded periodic relation. Figure 3c shows examples of pitch measures in the two above situations. The sound in Fig. 3c, example 2 is the same one as in Fig. 1, example 2. The third situation, and the most troublesome, is when the frequency contours in a same harmonic stack include more than one family of harmonics, suggesting two independent sound sources. In this case the median difference between successive frequency contours is not an ideal solution. It would be useful to have an algorithm that distinguished between single- and double-source sounds and treated each source separately, but ours does not do this. Sounds in which two separate sound sources can be inferred from the simultaneous occurrence of at least two families of unrelated harmonics are probably relatively rare in adult zebra finch song, but we have not seen a quantitative estimate of their incidence.

Frequency modulation

Frequency modulation is computed as described above for spectral derivatives (also see Fig. 3d). It is defined as the angle of the directional derivatives as shown in Fig. 2d.

RESULTS

The Computational Frame

The problem of defining song units

A zebra finch song motif consists of discrete notes that are often imitated in chunks of variable size (Williams & Staples 1992). Partitioning a motif into its component notes would seem, therefore, the obvious first step for scoring imitation. However, pupils can transform elements of a tutor's song in many different ways: they can merge and split notes or modify them in such a way that sharp transitions of frequency structure are replaced by a smooth transition and so forth. For an automatic procedure, recognizing homologous notes can be very difficult. Yet, if a note-based procedure fails to recognize such transformations it may, as a result, underestimate the similarity between two songs. We chose, therefore, to avoid any partitioning of the song motif into component 'notes'. Instead, we examined each time window for similarity, throughout the songs, omitting silent intervals. This approach allowed us to detect discrete 'segments of imitation' that typically emerge from the analysis. The technique of extracting a similarity score from a set of features that vary in time is described below and summarized as a sequence of steps in the Appendix.

Integration of the Song Measures

Each time window of a tutor's song is represented by measurements of four features: Wiener entropy, spectral continuity, pitch and frequency modulation. Each of these features has different units and different statistical distributions in the population of songs studied. To arrive at an overall score of similarity, we transformed the units for each feature to a common type of unit that could be added. One can transform the units of pitch, for example, from Hertz to units of statistical distances. In a certain population of songs, two measurements of pitch may be 3 standard deviations away from each other and so forth (although in practice, we did not use units of SD but 'median absolute deviation' from the mean). These normalized measures can then be integrated (see Appendix). We scaled measures based on their distribution in a sample of 10 different songs. Because the distribution of features may vary between populations (e.g. pitch is distributed differently in wild and domestic zebra finches; Zann 1996), a new normalization may be desirable before starting on new material to prevent a distortion of comparisons or an unintended change of a measure's weight.

A Method for Reducing Scoring Ambiguity

For the sake of simplicity, we demonstrate how one measure, pitch, performs when comparing an artificial tutor–pupil pair of songs that show perfect similarity. First we singled out a particular time window of the 'tutor's' song and compared its measures to those of each window in the 'pupil's' song. Ideally, there would be only one good match in the pupil's song. We repeated this procedure for each window of the tutor's song (see Fig. 4a). The resulting matrix spans all possible combinations of pairs of 'tutor' and 'pupil' windows. The difference in pitch between each pair of windows is encoded into a colour scale. In this case there is a marked tendency for the strongest similarity between pairs of windows to show as a red diagonal line. In practice, however, similar pitch values are seldom restricted to a unique pair of windows of the tutor and pupil's song. Different windows often share similar patterns of power spectrum. Therefore, even when all four measures are taken into account, there are likely to be several windows in the pupil's song that show close similarity to a specific window of the tutor's song. Therefore, scoring similarity between songs on the scale of a single window is hopeless, as is comparing pictures one pixel at a time.

The solution is to compare intervals consisting of several windows. If such intervals are sufficiently long, they will contain enough information to identify a unique

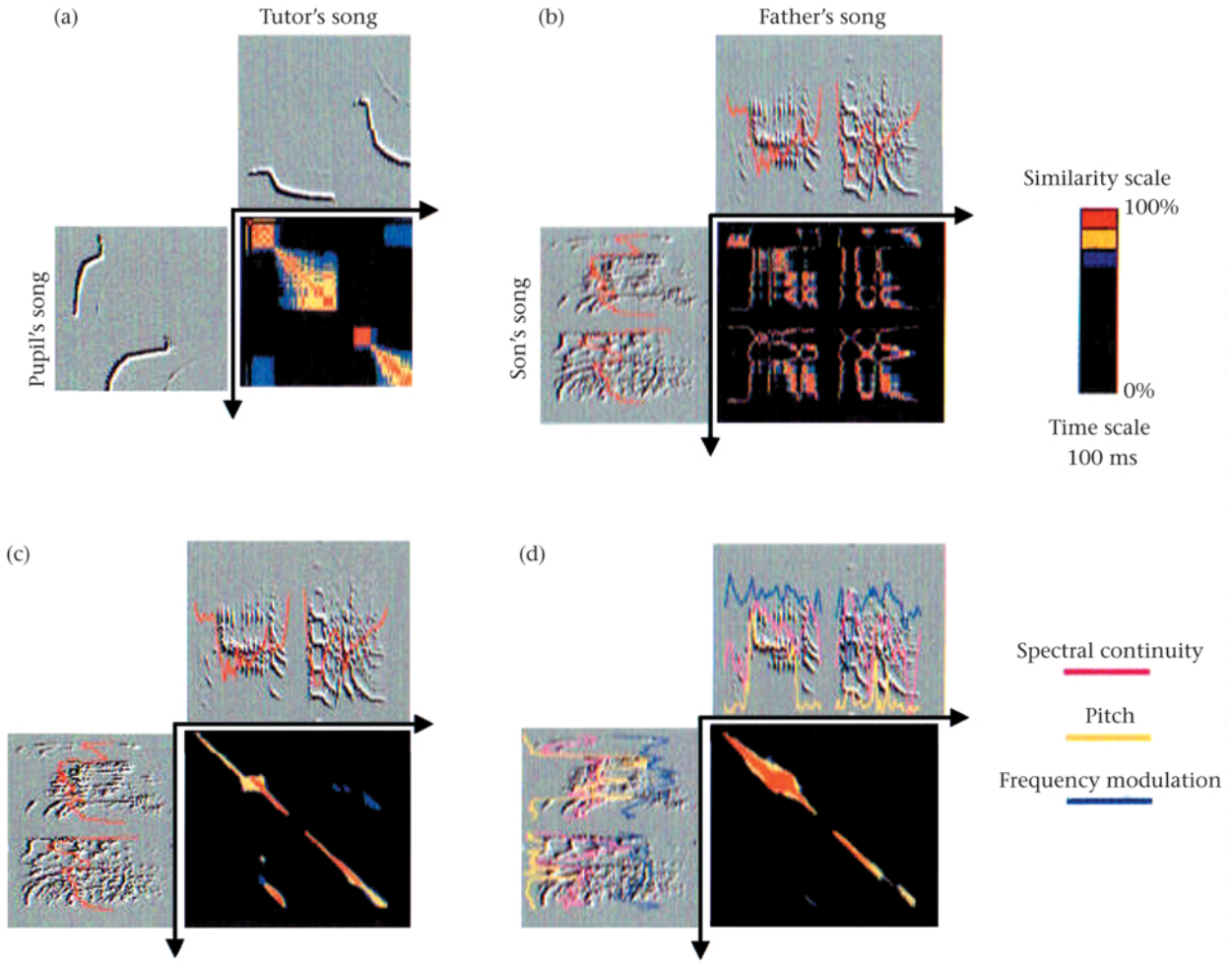

Figure 4. Similarity measure improves as comparisons include longer intervals and more features. (a) Similarity matrix between identical, artificial sounds. Because each of these simple sounds has a unique pitch, the similarity matrix shows high similarity values (indicated in red) across the diagonal and low values elsewhere. Comparing complex sounds would rarely give such result. As shown in (b), although the songs are similar, high similarity of Wiener entropy values are scattered. (c) Ambiguity is reduced when we compare Wiener entropy curves between 50-ms intervals. (d) A combined similarity matrix between 50-ms intervals across features. High similarity values are now restricted to the diagonal, indicating that each of the notes of the father's song was imitated by his son in a sequential order. Similarity scale: 0–70% (black), 71–80% (blue), 81–90% (yellow), 91–100% (red). The coloured curves overlaying the time–frequency derivative in (d) correspond to spectral continuity, pitch and frequency modulation (see colour code).

song segment. Yet, if the intervals are too long, similarities that are real at a smaller interval size may be rejected and that would reduce the power of analysis. We found empirically that comparisons using 50-ms intervals, centred on each 7-ms time window were satisfactory. Perhaps not surprisingly, the duration of these song intervals is on the order of magnitude of a typical song note. Figure 4b–c illustrates this approach, in this case for measures of Wiener entropy. This time, we compared the song of a father to the song of his son. The two birds were kept together until the son reached adulthood. Figure 4b presents the similarity of Wiener entropy values between 7-ms windows of the father and the son's songs. As expected, the result was ambiguous. Figure 4c presents the similarity of Wiener entropy values, this time between 50-ms intervals of the same songs. As indicated by the diagonal red line, narrowing the definition of the similarity measurement eliminates most of the ambiguity. Measuring similarity in 50-ms intervals across all four measures (Fig. 4d) was in this case sufficient for identifying a unique diagonal line, which reflects that the two songs being compared were very similar. Our final score of similarity combined the two scales: the 'large scale' (50 ms) is used for reducing ambiguity, while the 'small scale' (7 ms) is used to obtain a fine-grained quantification of similarity (see below).

Setting a similarity threshold

We compared windows and their surrounding intervals (the large, 50-ms scale) in the tutor's song with windows and their surrounding intervals in the pupil's song. We accepted the hypothesis of similarity between the two windows when a critical 'similarity threshold' was met. Setting the threshold for this decision was critical, because of the danger of making false rejections. A positive decision at this stage is not final, because it does not

guarantee that the two sounds compared offer the best possible match. The final step in the procedure involves choosing the best match from all possible alternatives, as explained below.

We took a statistical approach to our setting of the similarity threshold. We recorded songs from 20 unrelated zebra finches from our colonies at the Field Research Center and paired them randomly. We then compared all 50-ms time windows in one song with all of the 50-ms windows in the other song and calculated the distribution of similarity values of the entire sample. We used this statistical distribution to arrive at a probability curve that assigns a probability value for each measured difference between two 50-ms intervals. We set a threshold that would accept only similarity values that were likely to occur by chance alone with a probability equal to or lower than 1%. The selection of this *P* value for a similarity threshold can be tailored to the needs of a particular study. For example, a more liberal similarity threshold is required when searching for the first evidence of similarity between the emerging song of a juvenile and its putative tutor. As mentioned earlier, although our 'categorical decision' to label two sounds as 'similar' is based on 'large-scale' similarity, the actual similarity value is based only on the 'small-scale' similarity.

The final similarity score

For each pair of time windows labelled as 'similar' for two songs being compared, we calculated the probability that the goodness of the match would have occurred by chance as described above. We are left, then, with a series of *P* values, and the lower the *P*, the higher the similarity. For convenience we transform these *P* values to 1-*P*; therefore, a 99% similarity between a pair of windows means that the probability that the goodness of the match would have occurred by chance is less than 1%. In this case, 99% similarity does not mean that the features in the two songs being compared are 99% similar to each other. In practice and because of how our thresholds were set, songs or sections of songs that get a score of 99% similarity tend, in fact, to be very similar.

Our procedure requires that there be a unique relation between a time window in the model and a time window in the pupil. Yet, our technique allows that more than one window in the pupil song will meet the similarity threshold. The probability of finding one or more pairs of sounds that meet this threshold increases with the number of comparisons made and so, in some species at least, the duration of the pupil's song will influence the outcome. When a window in a tutor's song is similar to more than one window in the pupil's song, the problem is how to retain only one pair of windows. Two types of observations helped us make this final selection: the first is the magnitude of similarity, the second one is the length of the section that met the similarity criterion. Windows with scores that meet the similarity threshold are often contiguous to each other and characterize discrete 'sections' of the song. In cases of good imitation, sections of similarity are interrupted only by silent intervals, where similarity is undefined. Depending on the species, a long section of sequentially similar windows (i.e. serial sounds similar in the two songs compared) is very unlikely to occur by chance, and thus the sequential similarity we observed in zebra finches was likely the result of imitation. Taken together, the longer the section of similarity and the higher the overall similarity score of its windows, the lower the likelihood of this having occurred by chance. Therefore, as described below, the overall similarity that a section captures has preeminence over the local similarity between time windows.

To calculate how much similarity each section captured we used the following procedure. Consider for example, a tutor's song of 1000 ms of sound (i.e. excluding silent intervals) that has a similarity section of 100 ms with the song of its pupil, and the average similarity score between windows of that section is 80%. The overall similarity that this section captures is therefore:

$$80\% \times 100 \text{ ms}/1000 \text{ ms}=8\%.$$

We repeated the procedure for all sections of similarity. Then, we discarded parts of sections that showed overlapping projections, either on the tutor or on the pupil's song (see Fig. 5). Starting from the section that received the highest overall similarity score (the product of similarity × duration, as shown above), we accepted its similarity score as final and removed overlapping parts in other sections. We based the latter decision on the overall similarity of each section and not on the relative similarity of their overlapping parts. We repeated this process down the scoring hierarchy until all redundancy was removed. The remainder was retained for our final score of similarity.

We demonstrate the results of this procedure for excellent song imitation (Fig. 5a), partial imitation (Fig. 5b) and unrelated songs (Fig. 5c).

Testing the procedure

Figure 6a presents the similarity scores of songs produced by birds that were housed as juveniles singly with their father, a condition that promotes accurate imitation (Tchernichovski & Nottebohm 1998). For comparison, we scored similarity between the songs of birds that were raised in different breeding rooms of our colony. As shown, similarity scores were much higher when comparing the songs of a tutor and its pupil than when comparing the songs of two randomly chosen individuals. We next wanted to determine whether the procedure could detect subtle differences in the completeness of an imitation. For this we used the effect of fraternal inhibition (Tchernichovski & Nottebohm 1998): when several pupils are kept together with a single tutor, imitation completeness is reduced in some pupils but not in others. Because this effect works even when pupils are kept in separate cages, we constructed an arena of 10 cages around a central cage as shown in Fig. 6b. We placed an adult tutor in the middle cage, and in each of the 10 peripheral cages we placed a single 30-day-old pupil that had not been exposed to male zebra finch song from day 10 onwards. The 10 pupils and the tutor were kept in the arena until the pupils were 100 days old, at which time we recorded the songs of the tutor and the pupils. A

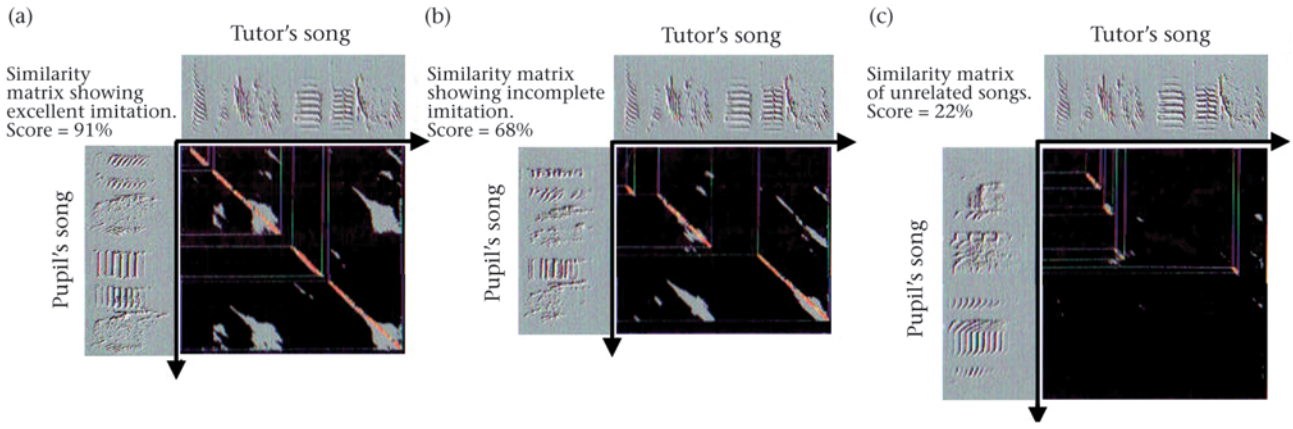

Figure 5. Similarity scores for three pairs of tutor–pupil songs. Each pair shows a different degree of similarity. The grey traces in the black panel represent sections of similarity that met the similarity threshold but were rejected in the final analysis. These sections of similarity were rejected because their projections on either the tutor's or the pupil's song overlapped with sections of higher similarity. The similarity values of sections that passed this final selection are encoded into colour code as in Fig. 4, where similarity is measured on a small scale, across time windows. The thin, blue lines connect the beginning and end of corresponding sections of similarity.

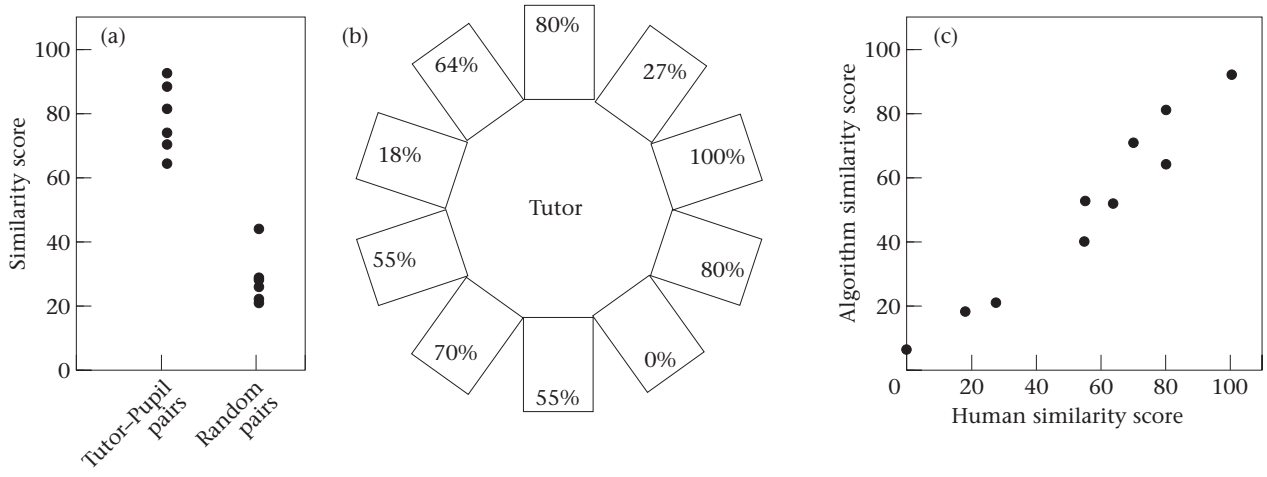

Figure 6. (a) Song similarity scores computed by the automated procedure in cases of tutor–pupil pairs and in random pairs. (b) Ten pupils were kept singly in 10 cages as shown. The tutor was kept in the central cage. Human similarity scores are presented for each pupil. (c) The correlation between human and automated procedure scores for the 10 pupils shown in (b) (r=0.91, P<0.01).

human then scored, as in Tchernichovski & Nottebohm (1998), the percentage of the tutor notes for which the pupils produced a close match. The results of the human scores are presented in Fig. 6b. As expected, imitation was highly variable. Figure 6c presents the correlation between the human (visually guided) score and automated scores of similarity. As shown, the correlation was high (r=0.91).

Our measurements suggest that in many cases the automated procedure makes similar decisions to those made by a human observer. In addition, we gain analytic power to ask questions that might have been difficult to answer without the procedure. For example, in the experiment represented in Fig. 6b, the similarity between tutor and pupil for each of the 10 birds was closely paralleled by the duration of each pupil's song. That is, a pupil that got a score of 55% had a song that was approximately 55% as long as that of the tutor (Tchernichovski & Nottebohm 1998). Apparently, the difference in completeness of imitation was explained by how many notes were present in the pupil's song. We then used our procedure to determine whether the quality of imitation for each of the notes sung was related to the completeness of imitation. In this case, incompleteness of imitation was not correlated with accuracy of imitation (r<0.1). Thus, for those notes sung, the match with the tutor's notes was equally good whether just a few or all of the notes were imitated.

Performances of the procedure

The software is relatively easy to master. To compare two sounds, a user must first digitize the sounds and store them in a data file. The software will then extract features from the two songs (a process that takes approximately 3 s of analysis per 1 s of sound analysed, using a 500-MHz

Pentium PC). The user can then outline the corresponding parts of the songs for which a similarity score is desired (e.g. the whole song or parts thereof). Scoring similarity between a pair of songs that each last 1 s, takes about 10 s; for a pair of 2-s songs, scoring will take approximately 40 s, and so forth. Only the memory resources of the computer limit the overall duration of the comparison.

Our procedure can also be used to enhance, rather than replace visual inspection. It allows the user to alternate between different representations of the sound: sonagram, spectral derivatives and frequency contours (as in Fig. 1). The user can outline each note and type comments, while the software generates a data file that transparently combines the visual inspection with a summary of the objective features of each note.

DISCUSSION

We presented a procedure that uses four simple, unidimensional acoustic features to measure the similarity between two sounds. The measurements for each of the features are integrated into a global similarity score. Thus, one of the more novel aspects of this new approach to score similarity in natural sounds is that it has an explicit and reliable metric. Even subtle differences between sounds can be quantified and compared, and it is possible to track, in quantitative terms, the small daily changes that occur during song development.

Our initial motivation for developing our procedure was the need to have an easy, reliable and fast method to score song imitation. However, the procedure also may be used for scoring similarity between unlearned sounds. The measures provided by the procedure allow for a standardization that will also make it easier to describe signal variability during development, in adulthood and between members of a population. Such a measure has been lacking in studies of development as well as in studies examining the consequence of various hormonal or neurological interventions. We chose the features because those features are thought to bear a close relation to the articulatory variables involved in sound production (Ho et al. 1998).

Our algorithm for measuring the similarity between songs used several parameter values that can be altered without changing the conceptual framework. The software that is available allows for changing these parameter values. We are aware that the parameters used will be determined, to some extent, by the properties of the sounds compared and by the nature of the questions asked. Similarly, the weight assigned to each sound feature and its contribution to the final index of similarity can be altered. In the current report, we gave equal weight to measures from all four features analysed, but this need not be so. We also used an arbitrary criterion for deciding what was the 'similarity threshold', which also can be modified. Fine tuning of the algorithm will reflect not just the properties of the sounds compared and the questions asked but, in time, will also reflect the experience of many users. But even if the parameter values that we used are proven to be suboptimal, they are stated and have a quantitative reality. To this extent, they differ from the unstated and unexplainable idiosynchracies that have often permeated our way of talking about similarities between animal vocalizations. But even with these limitations, we hope that others will find our approach useful for scoring the similarity between animal sounds. This, in turn, should allow for a more rigorous and quantitative approach to the study of vocal learning, vocal imitation and vocal communication.

Acknowledgments

We thank Marcelo Magnasco, Boris Shriaman, Michael Fee and Thierry Lints for their useful comments. Supported by NIMH 18343, the Mary Flagler Cary Charitable Trust and the generosity of Rommie Shapiro and the late Herbert Singer. The research presented here was described in Animal Utilization Proposal No. 93161, approved September 1998 by The Rockefeller Animal Research Ethics Board.

References

Catchpole, C. K. & Slater, P. J. B. 1985. *Bird Song*. Cambridge: Cambridge University Press.

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

Cynx, J. 1990. Experimental determination of a unit of song production in the zebra finch (*Taeniiopygia guttata*). *Journal of Comparative Psychology*, **104,** 3–10.

Esser, K. H., Condon, C. J., Suga, N. & Kanwal, J. S. 1997. Syntax processing by auditory cortical neurons in the FM–FM area of the mustached bat *Pteronotus parnellii*. *Proceedings of the National Academy of Sciences, U.S.A.*, **94,** 14 019–14 025.

Fee, M., Pesaran, B., Shraiman, B. & Mitra, P. 1998. The role of nonlinear dynamics of the syrinx in birdsong production. *Nature*, **395**, 67–71.

Ho, C. E., Pesaran, B., Fee, M. S. & Mitra, P. P. 1998. Characterization of the structure and variability of zebra finch song elements. *Proceedings of the Joint Symposium on Neural Computation*, **5,** 76–83.

Immelmann, K. 1969. Song development in the zebra finch and in other estrilded finches. In: *Bird Vocalization* (Ed. by R. A. Hinde), pp. 61–74. Cambridge: Cambridge University press.

Jarvis, E. D., Scharff, C., Grossman, M. R., Ramos, J. A. & Nottebohm, F. 1998. For whom the bird sings: context-dependent gene expression. *Neuron*, **21,** 775–788.

Kanwal, J. S., Matsumura, S., Ohlemiller, K. & Saga, N. 1994. Analysis of acoustic elements and syntax in communication sounds emitted by mustached bats. *Journal of the Acoustical Society of America*, **96,** 1229–1254.

Kogan, J. A. & Margoliash, D. 1998. Automated recognition of bird song elements from continuous recording using dynamic time warping and hidden Markov models: a comparative study. *Journal of the Acoustical Society of America*, **103,** 2185–2196.

Konishi, M. 1965. The role of auditory feedback in the control of vocalization in the white-crowned sparrow. *Zeitschrift für Tierpsychologie*, **22,** 770–783.

Kroodsma, D. E. 1982. Learning and the ontogeny of sound signals in birds. In: *Acoustic Communication in Birds* (Ed. by D. E. Kroodsma & E. H. Miller), pp. 11–23. New York: Academic Press.

Marler, P. & Tamura, M. 1964. Culturally transmitted patterns of vocal behavior in sparrows. *Science*, **146,** 1483–1486.

Morris, D. 1954. The reproductive behaviour of the zebra finch (*Taeniopygia guttata*) with special reference to pseudofemale behaviour and displacement activities. *Behaviour*, **6,** 271–322.

Nelson, D. A., Marler, P. & Pallerone, A. 1995. A comparative approach to vocal learning: intraspecific variation in the learning process. *Animal Behaviour*, **50,** 83–97.

Nottebohm, F. 1968. Auditory experience and song development in the chaffinch, *Fringilla coelebs*. *Ibis*, **110,** 549–568.

Nottebohm, F. & Nottebohm, M. 1978. Relationship between song repertoir and age in the canary, *Serinus canarius*. *Zeitschrift für Tierpsychologie*, **46,** 298–305.

Nowicki, S. & Nelson, D. 1990. Defining natural categories in acoustic signals: comparison of three methods applied to chick-a-dee' call notes. *Ethology*, **86,** 89–101.

Percival, D. B. & Walden, A. T. 1993. *Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques*. Cambridge: Cambridge University Press.

Price, P. 1979. Developmental determinants of structure in zebra finch song. *Journal of Comparative and Physiological Psychology*, **93,** 260–277.

Scharff, C. & Nottebohm, F. 1991. A comparative study of the behavioral deficits following lesions of various parts of the zebra finch song system: implication for vocal learning. *Journal of Neuroscience*, **11,** 2896–2913.

Slepian, D. & Pollak, H. O. 1961. Prolate spheriodal wave-functions Fourier analysis and uncertainty. *Bell Systems and Technology Journal*, **40,** 43–63.

Tchernichovski, O. & Nottebohm, F. 1998. Social inhibition of song imitation among sibling male zebra finches. *Proceedings of the National Academy of Sciences, U.S.A.*, **95,** 8951–8956.

Thomson, D. 1982. Spectrum estimation and harmonic analysis. *Proceedings of the Institute of Electrical and Electronics Engineers*, **70,** 1055–1096.

Thomson, D. 1990. Quadratic-inverse spectrum estimates: applications to palaeoclimatology. *Philosophical Transactions of the Royal Society of London, Series A*, **332,** 539–597.

Thomson, D. 1993. Non-stationary fluctuations in stationary time series. *Proceedings of the International Society of Optical Engineering*, **2027,** 236–244.

Thorpe, W. H. 1954. The process of song-learning in the chaffinch as studied by means of the sound spectrograph. *Nature*, **173,** 465.

Thorpe, W. H. 1958. The learning of song patterns by birds, with special reference to the song of the chaffinch, *Fringella coelebs*. *Ibis*, **100,** 535–570.

Williams, H. & Staples, K. 1992. Syllable chunking in zebra finch (*Taeniopygia guttata*) song. *Journal of Comparative Psychology*, **106,** 278–286.

Zann, R. E. 1996. *The Zebra Finch: Synthesis of Field and Laboratory Studies*. New York: Oxford University Press.

Appendix

Computational steps to construct a similarity score between songs starting from a set of one-dimensional features time series

1. Obtain the distribution of each feature in a sample of n different songs (say N=10). Scale the units of each feature to the absolute median difference from its mean.

2. Measure sound features for every time window of tutor and pupil's songs and scale them.

3. Compute short-scale Euclidean distances across time windows of tutor and pupil's songs. Let $\mathbf{L}$ ($M \times N$) be a rectangular matrix where M is the number of time windows in tutor's song and N is the number of time windows in pupil's song. For a pair of windows a and b of a tutor and pupil's song, respectively, our estimate of the small-scale distance (D_s) between the two sounds is the Euclidean distance between the scaled sound features $f_1, f_2, \ldots, f_n$, namely:

$$D_s(a,b) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i(a) - f_i(b))^2}.$$

Specifically for our four features: pitch (p), frequency modulation (FM), Weiner entropy (W) and spectral continuity (C):

$$D_s(a,b) = \frac{1}{2}\sqrt{(p_a - p_b)^2 + (FM_a - FM_b)^2 + (W_a - W_b)^2 + (C_a - C_b)^2}$$

and the matrix $\mathbf{L}$ is defined as

$$\mathbf{L}_{i,j} = D_s(i,j) = 1 \ldots M,\ j = 1 \ldots N$$

4. Compute long scale distances across (say, 50 ms) intervals of several time windows. Let $\mathbf{G}$ ($M \times N$) be a rectangular matrix where M is the number of intervals in tutor's song and N is the number of intervals in pupil's song. Each interval is composed of a sequence of T time windows, each centred on the corresponding window in the $\mathbf{L}$ matrix (edge effects are neglected). Our estimate of the large-scale distance (D_l) between the two sounds is the mean Euclidean distance between features of corresponding time windows within the intervals. For two intervals A and B of a tutor and pupil's song, respectively, consisting of time windows A_t, B_t (t=1 . . . T)

$$D_l(A,B) = \sqrt{\frac{1}{T}\sum_{t=1}^{T} D_s(A_t, B_t)^2}$$

and the matrix $\mathbf{G}$ is defined as

$$\mathbf{G}_{i,j} = D_l(i,j)\ i = 1 \ldots M,\ j = 1 \ldots N$$

Note that D_l (A, B) is sensitive to the order of time windows within each interval. That is, features are compared only across time windows of the same sequential order.

5. Transformation of the entries of the matrices $\mathbf{L}$ and $\mathbf{G}$ from Euclidean distances to P values: based on the distribution of D_s and D_l across 10 unrelated songs, plot the cumulative distributions of D_s and D_l and use the plots to transform Euclidean distances to P values, $P(\mathbf{G}_{i,j})$ and $P(\mathbf{L}_{i,j})$.

6. Setting a threshold for rejection of similarity hypothesis. Construct a matrix $\mathbf{S}$ of similarities as follows:

$$\mathbf{S}_{i,j} = \begin{cases} 1 - P(\mathbf{L}_{i,j}) & \text{if } P(\mathbf{G}_{i,j}) < 0.01 \\ 0 & \text{otherwise,} \end{cases}$$

that is,

$$\mathbf{S}_{i,j} = [1 - P(\mathbf{L}_{ij})]\theta\{P_{Th} - P(\mathbf{G}_{ij})\}$$

where $P_{Th}=0.01$ is the threshold probability. Note that large-scale P values are used for the threshold, but when the similarity hypothesis is accepted, the small-scale P values of the local matrix **L** are used for estimating its magnitude. For convenience we transform these P values to $1-P$ so that a high P value now refers to a high similarity and vice versa.

7. Define continuous sections of similarity by grouping each pair of cells $\mathbf{S}_{i,j}$, $\mathbf{S}_{k,i}$ that fulfil the following conditions:

$$\mathrm{ab}_s(i-k)\leq 1,\ \mathrm{ab}_s(j-l)\leq 1,\ \mathbf{S}_{ij}>0,\ \mathbf{S}_{k,l}>0$$

that is, we group neighbouring cells of significant similarity value. We now have a series of sections, each represented by two features: its dimensions and its partial similarity value. The dimensions of a section B are defined by a bounding rectangle that just encloses a continuous portion of **S**. The partial similarity value of a section estimates the proportion of the tutor's song that is accounted for by that section. For a tutor's song with a total of n intervals, the partial similarity value of a section B is.

$$P_s=\frac{1}{n}\sum_i \max_j(\mathbf{S}_{i,j})|(i,j)\in \mathrm{B}$$

8. Elimination of redundancy. Sort sections from high to low values of P_s. Starting from the interval of highest partial similarity value, clip all overlapping parts of other intervals (i.e. set those entries in **S** to zero) and recompute their partial similarity values according to step 7.

9. The final similarity score is the sum of partial similarities of all sections. In many cases, however, it may be of benefit to adjust the partial similarity value of a section so as to give higher values to sections of longer duration. This is particularly important when examining imitation of song models of variable duration. For example, such an adjustment for the P_s values of a section S bounded by a rectangle of dimensions $A\times B$ of the song **G** ($M\times N$) could be

$$P_s\rightarrow P_s*\frac{\log\sqrt{\mathrm{A}^2+\mathrm{B}^2}}{\log\sqrt{M^2+N^2}}$$