# A METHOD FOR BIRD SONG SEGMENTATION AND PAIRWISE DISTANCE MEASURE OF SYLLABLES AND SONGS

L. Ranjard Bioinformatics Institute and School of Biological Sciences, University of Auckland, Auckland, New Zealand
H.A. Ross Bioinformatics Institute and School of Biological Sciences, University of Auckland, Auckland, New Zealand

## 1 INTRODUCTION

The songs of Passeriformes bird species are learned rather than innate [**Error! Reference source not found.**]. 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; firstly a bird hears and memorizes a song from another bird, and secondly it tries to imitate the song as accurately as possible. This imitation is not perfect, therefore songs evolve through generations as small changes occur. Thus, it appears that in certain cases, bird song is a culturally transmitted trait and therefore these songs can be considered as a collection of memes [**Error! Reference source not found.**]. Using the framework of meme evolution, an analogy of molecular evolution, song evolution involves deletion, substitution and repetition of syllables. In studying birdsong evolution, the first critical step is to convert a continuous song into a sequence of discrete syllables, which can then be classified as memes. Particular care is required when defining the memes as this definition is crucial for this work. Previous studies suggest that the syllable should be considered as the meme unit [**Error! Reference source not found.**, **Error! Reference source not found.**-**Error! Reference source not found.**, **Error! Reference source not found.**]. Here we present bioacoustics methods which have been developed for encoding bird songs as a sequence of discrete syllables. A pairwise syllable distance measure, calculated on the basis of cepstrum analysis and dynamic time warping, will be also introduced. One way to perform sequence comparison is to use alignment algorithms that minimize a distance function between a pair of sequence [**Error! Reference source not found.**]. They are used in the field of speech processing and molecular biology. In this paper, alignment algorithms are adapted and applied to calculate a distance between a pair of syllables or between a pair of songs. This measure is proportional to the number of operations required to transform one sequence into another. By this method, it has been possible to draw basic trees using classic tree-building algorithms, such as UPGMA or Neighbour-Joining. Particular attention has been given to the penalty cost of the different operations involved: insertion, deletion, substitution, compression and expansion. The trees show gradual and ordered relationships between the songs. Results obtained from the encoding of songs belonging to a sub-species of White-crowned Sparrow (*Zonotrichia leucophrys pugetensis*) show how the classification from this approach corresponds to those defined by classic approaches.

## 2 METHODS

Classic approaches in bioacoustics use basic signal features and/or human knowledge for both detection and segregation of bioacoustics signals from noise [**Error! Reference source not found.**]. Although former technical limitations are non longer restrictive [**Error! Reference source not found.**], only a few song analysis tools are inspired by automatic speech processing [**Error! Reference source not found.**] and are fully automatic [**Error! Reference source not found.**]. When expert knowledge is used in song analysis, it can introduce subjectivity which can impair the reproducibility of bird song studies. However modern speech processing techniques offer high accuracy in sound analysis for speech recognition or speaker verification [**Error! Reference source not found.**, **Error! Reference source not found.**]. Different measures of word distance have been

developed [**Error! Reference source not found.**] and they can be extended to provide syllable distance measures. On one hand, animal vocalizations are simpler that 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 therefore the meaningful features of the song may remain undetected. Initially, syllable boundaries are found by segregating the syllables from the background noise using feature-specific waves. The waves are smoothed and merged to define a switch function. Syllables are defined as regions in the song that are characterised by high amounts of power, low Wiener entropy values and a certain continuity in fundamental frequency and autocorrelation of the signal. Analysing the derivative of this function is an effective method to find the syllables. In order to compare fragments of sounds, it is possible to use frequency band filtering and then cepstrum coefficients. The mel-frequency bands have been defined for human ears and therefore there is little support to use it for bird vocalizations. If such a frequency band would exist for birds it would be very dependant on the bird species under study. Nevertheless, some previous studies successfully worked this way [**Error! Reference source not found.**]. This study indicates that a correct song recognition could be obtain using only 12 mel-frequency cepstral coefficients, but there 16 cepstral coefficients have been used as they seem to give a sufficient accuracy for segregating syllables. It has been shown [**Error! Reference source not found.**] that most birds are able to distinguish between two sounds separated by only 1 or 2ms. Therefore, considering a sampling frequency of 44.1kHz, a frame of 512 samples with 50% overlap has been used. This window size is much smaller than that used by Trawicki [**Error! Reference source not found.**] and therefore provides sufficient precision in the dynamic of the song.

## 2.1 Filtering

In [**Error! Reference source not found.**], the average audibility curve shows that the greatest sensitivity of birds is between 2 and 3kHz. The limits of the birds' auditory sensitivity are between 0.5kHz and 6kHz. However, the frequency bandwidth to be filtered is strongly dependent on the species and the environment. In this study, a finite impulse response filter was applied in order to keep frequencies between 2kHz and 12kHz. A wavelet denoising approach is also performed using the Wavelab Toolbox routines [**Error! Reference source not found.**, **Error! Reference source not found.**] via a Symmlets wavelet filter [**Error! Reference source not found.**].

## 2.2 Segmentation

For each song, a set of features is computed. [**Error! Reference source not found.**] provides a complete list of sound features usable for sound characterisation. Whereas some of these features are based on psycho-acoustical studies and therefore are human-specific, some are relevant to bird song analysis. It is suggested that the spectral roll-off, defined as the frequency below which a specific percentage of the energy occurs, is applicable to bird songs. The minimum length of a syllable used in this study is 1024 samples (under 44.1kHz sample rate). Dooling [**Error! Reference source not found.**] showed that birds can distinguish shorter sounds but a minimum number of samples is required to perform accurate analysis. Song features include:

### 2.2.1 Amplitude Envelope

The amplitude envelope $s_{rms}$ is computed with the root mean square technique using a 50% overlapping window of 128 samples,

$$s_{rms}(l) = \sqrt{\frac{1}{t_2 - t_1} \sum_{x=t_1}^{t_2} s^2(x)}, \tag{1}$$

where $l$ defines a window between time $t_1$ and $t_2$ of the signal $s$. This amplitude envelope is then artificially amplified by squaring it and leveling all sample points situated above the maximum value of the original amplitude envelope.

### 2.2.2 Wiener Entropy

The chronux toolbox [**Error! Reference source not found.**] is a set of Matlab routines for neural data analysis. This library provides algorithms to perform multitaper analysis of sound signals [**Error! Reference source not found.**]. This allows a better resolution of the spectrum analysis by avoiding the use of predefined time windows. Wiener entropy has previously been applied to the study of bird songs [**Error! Reference source not found.**]. The Wiener entropy $s_{ent}$ is defined as the ratio between the geometric mean and the arithmetic mean of the power of the signal computed on the multitaper spectrum,

$$s_{ent}(l) = -\log \frac{\exp\left(\frac{1}{f_{Nyquist}} \sum_{f=1}^{f_{Nyquist}} \ln X(f)\right)}{\frac{1}{f_{Nyquist}} \sum_{f=1}^{f_{Nyquist}} X(f)}, \tag{2}$$

where $l$ defines a window and $X$ is the spectrum of the signal $s$.

### 2.2.3 Autocorrelation

The cross-correlation $s_{cor}$ is a standard method for estimating how two signals are linearly related. In the field of music analysis it is useful for instrument recognition [**Error! Reference source not found.**]. For each 50% overlapping window of 1024 samples, the maximum of the cross-correlation is calculated as

$$s_{cor}(l) = \max \sum_{n=0}^{N-m-1} s(n+m)s^*(n), \qquad \forall\, 1 \leqslant m \leqslant N, \tag{3}$$

where $s^*$ is the complex conjugate of $s$, $N$ is the number of samples of $s$ and $l$ defines a window.

### 2.2.4 Frequency roll-off 60%

The spectral roll-off point $s_{ro}$ is defined as the frequency value at which 60% of the signal energy is contained below in the spectrum [**Error! Reference source not found.**]. 50% overlapping windows of 512 samples are used to compute this value through the signal. This value is higher during a bird song emission than for the noise because bird songs are generally high-pitched. Therefore, it can be helpful for segregating the signal from the noise.

$$s_{ro}(l) = f_{ro}, \tag{4}$$

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

$$\sum_{f=1}^{f_{ro}} X(f) = 0.60 * \sum_{f=1} X(f). \tag{5}$$

A function combining those features is defined as

$$s_{sw}(l) = s_{rms}(l) + s_{ent}(l) + s_{cor}(l) + s_{ro}(l). \quad (6)$$

This function reflects the changes occurring through the song between two states, signal and noise. We will then refer to it as the *switch* function. An analysis of its first derivative is performed in order to separate the signal from the noise. $s_{sw}$ is first smoothed in order to get rid of short time variations, then the main variations in the song can be detected by an increasing or a decreasing section of the function, i.e. the first derivative has a positive value during a minimum number of 1024 samples. Each section is analysed in order to characterise it as a syllable, ignore it or merge it with the previous or the next section. This step involves the computation of few discrete features. Let $S_a$ and $S_b$ be two consecutive sections, $S_a$ is between time samples $t_1$ and $t_2$ and $S_b$ between time samples $t_2$ and $t_3$, then those features are

- the maximum values $M$ of each section

$$M_a = \max s_{sw}(x) \qquad \forall\, t_1 \leqslant x \leqslant t_2,$$

$$M_b = \max s_{sw}(x) \qquad \forall\, t_2 \leqslant x \leqslant t_3,$$

- an index $I$ reflecting the variation of the *switch* function between the two sections

$$I_a = \frac{|s_{sw}(t_1) - M_a|}{|s_{sw}(t_2) - M_a|},$$

$$I_b = \frac{|s_{sw}(t_2) - M_b|}{|s_{sw}(t_3) - M_b|},$$

- an index $C_{sw}$ reflecting the continuity of the *switch* function

$$C_{sw} = \max \begin{cases} |M_b - s_{sw}(t_2)| \\ |M_a - M_b| \\ |M_a - s_{sw}(t_2)| \end{cases},$$

- an index $C_{ro}$ reflecting the continuity of the frequency based on the roll-off function

$$C_{ro} = \max \begin{cases} |\frac{1}{L}\sum_{x=t2-L}^{t2} s_{ro}(x) - s_{ro}(t_2)| \\ |\frac{1}{L}\sum_{x=t2-L}^{t2} s_{ro}(x) - \frac{1}{L}\sum_{x=t2}^{t2+L} s_{ro}(x)| \\ |\frac{1}{L}\sum_{x=t2}^{t2+L} s_{ro}(x) - s_{ro}(t_2)| \end{cases} \text{, with } L = 1024$$

A set of thresholds is then used to label a section as a syllable or to merge two successive sections. Here is a description of the algorithm

for $n = 1 \rightarrow$ number of sections
  if $M_n > 0.3$ && $I_n > 1.3$
    section $n$ is labelled as a candidate syllable
    if $M_{n+1} > 0.3$ && $C_{sw} < 0.3$ && $C_{ro} < 0.3$
      section $n$ is merged with section $n+1$
    endif

endif
end

Then, the syllable boundaries are adjusted by analysing the autocorrelation of the signal in each syllable. A threshold is defined as $t = 0.01 * \max(s_{cor}(syllable))$ and the syllable beginning and end are move to the first and the last sample point above that value.

## 2.3 Syllable distance measure $D_s$

One of the most popular techniques used for speech analysis, for example speaker identification [**Error! Reference source not found.**], involves the calculation of the cepstral coefficients from the spectrum. A syllable is represented by a sequence of vectors, where each vector is a composed of the 16 cepstrum coefficients of consecutive overlapping windows. The number of 16 coefficients has been chosen based on empirical results. Many different distance measures have been implemented for speech processing [**Error! Reference source not found.**]. [**Error! Reference source not found.**] proposed a weighted distance calculated from those coefficients. The weight for each coefficient is proportional to the variance of the coefficient along both syllables. Using these weights, a distance between a pair of vectors can be calculated. The distance between two vectors is

$$dist(x,y) = \sum_{n=1}^{V} w(n) * (x(n) - y(n))^2, \tag{7}$$

where $V$ is the number of cepstral coefficients per vector and $w(n)$ is the inverse of the variance of the cepstral coefficient. [**Error! Reference source not found.**, **Error! Reference source not found.**, **Error! Reference source not found.**] present different approaches for performing sequence comparisons and more precisely the use of dynamic time-warping. Those techniques consist in finding the optimal alignment between two sequences of vectors using dynamic programming. This involves 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. Let $X = x_1 ... x_N$ and $Y = y_1 ... y_M$ be two vector sequences to be compared. Five edit operations are considered

- substitution $S(v,w)$ defines the cost associated with the substitution of the vector $v$ for the vector $w$. This cost is the weighted distance described above.
- insertion $I(v)$ defines the cost associated with the insertion of the vector $v$. This cost is set as half the average of the substitutions costs.
- deletion $D(v)$ defines the cost associated with the deletion of the vector $v$. This cost is set as half the average of the substitutions costs.
- compression $C(vw,x)$ defines the cost associated with the compression of the vectors $vw$ into the vector $x$. It is defined as the sum of the substitution cost of the vectors $v$ for $x$ and the substitution cost of the vectors $w$ for $x$ divided by two.
- expansion $E(v,wx)$ defines the cost associated with the expansion of the vector $v$ into the vectors $wx$. It is defined as the sum of the substitution cost of the vectors $w$ for $v$ and the substitution cost of the vectors $x$ for $v$ divided by two.

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

$$G(a,b) = \min \begin{cases} G(a-1,b) + I(x_a) & \text{case 1} \\ G(a,b-1) + D(y_b) & \text{case 2} \\ G(a-1,b-1) + S(x_a, y_b) & \text{case 3} \\ \frac{1}{2} G(a-1,b-2) + C(x_a, y_{b-1} y_b) & \text{case 4} \\ \frac{1}{2} G(a-2,b-1) + E(x_{a-1} x_a, y_b) & \text{case 5} \end{cases} \quad (8)$$

## 2.4 Song distance measure

From this pairwise syllable distance measure, it is possible to compute a syllable distance matrix. Having each song encoded as a sequence of syllables, a classic alignment algorithm can be used in order to compare them [**Error! Reference source not found.**]. Only three basic operations are allowed: insertion, deletion and substitution. Substitution cost is the syllable distance measure introduced above and the insertion and deletion costs are set as half the average of the substitution cost [**Error! Reference source not found.**]. From this alignment of length $L$, the number $N$ of positions where the aligned syllables have a distance below a specific threshold is counted; this threshold has been arbitrarily chosen ($1.2$) and more analysis is required in order to refine this definition. The distance between a pair of songs is then defined as

$$d = \frac{L - N}{L}$$

which results in a value between $0$ and $1$. This setting can have strong consequences on the construction of a distance tree from a distance matrix built this way.

# 3 RESULTS

This method has been applied to a set of songs from a sub-species of White-crowned Sparrow (*Zonotrichia leucophrys pugetensis*), recorded in the Puget Sound region. Previous studies [**Error! Reference source not found.**, **Error! Reference source not found.**] have characterised dialects in this region but they were delineated using subjective human measurements. Human judges chose features to define each dialects and classified songs accordance to these features. The aim of this study was to see if similar results can be analytically obtained, thereby avoiding any human bias and thus permitting a reproducible song classification. A Neighbor-Joining tree has been built from the distance matrix of this set of songs (Figure1). Different threshold values have been used for this analysis. As pointed out above, the shape of the tree is significantly altered by the syllable distance measure threshold used. However, when different values for this threshold are used (unpublished data), songs belonging to the same birds cluster together.

# 4 CONCLUSION

We have introduced a method showing that modern sound analysis techniques could be successfully used in the study of bird songs. The songs were aligned and compared by counting the number of shared syllables. Basically, this can be viewed as counting the number of memes shared by a pair of songs. This approach allow us to cluster songs belonging to individual birds but does not group songs belonging to similar dialects as defined in previous studies [**Error! Reference source not found.**, **Error! Reference source not found.**]. There are multiple reasons which help account for the differences. Firstly, the songs for this study have only been chosen because of the

region where they were recorded, not because they are known to belong to a specific dialect. Secondly, those dialects have been defined by considering some specific parts of the song. As previously suggested [**Error! Reference source not found.**] some parts of songs may reflect population's relationships while others are likely to evolve independently. The method applied in this paper considers the song in a whole. Therefore, this can reduce the weight of the signal relevant for dialect identification.

## ACKNOWLEDGEMENTS

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 New Zealand Government funding, administered by the Royal Society of New Zealand. The computations were performed using the Matlab® programming environment.

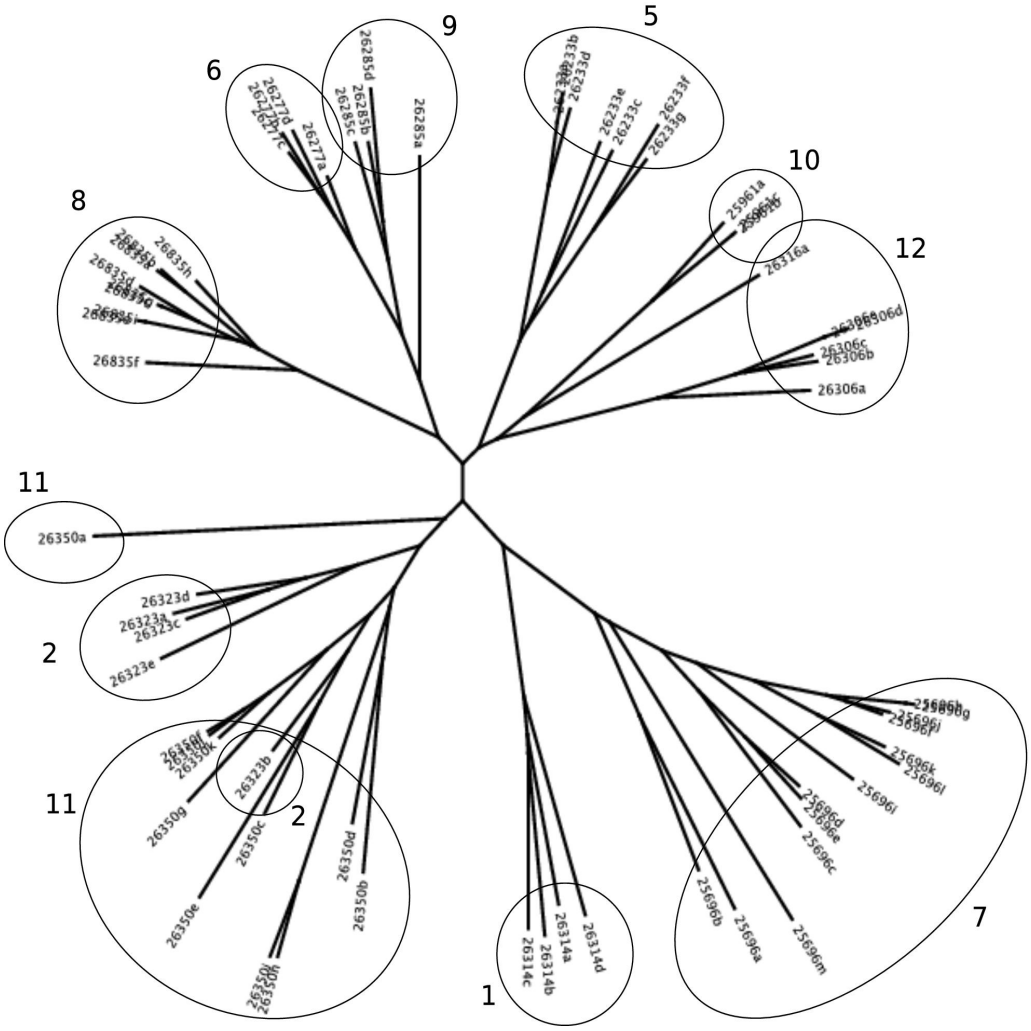


Figure 1: Neighbor-Joining tree of White-crowned songs from the Puget Sound region. Each leaf of the tree corresponds to a single song. The number in the reference of each song identifies the bird and the letter the song. Circles indicate the different dialects identified by [**Error! Reference source not found.**, **Error! Reference source not found.**] in the regions where the songs were recorded.

1. References

2. M. C. Baker and J. T. Boylan. A catalog of songs syllables of indigo and lazuli buntings. *The Condor*, 97:1028–1040, 1995.
3. L. F. Baptista. Geographical variation in song and dialects of the Puget Sound white-crowned sparrows. *The Condor*, 79:356–370, 1977.
4. D. W Bradley and R. A Bradley. Application of sequence comparison to the study of bird songs. In D. Sankoff and J. B. Kruskal, editors, *Time Warps, String Edits, and Macromolecules - The Theory and Practice of Sequence Comparison*, chapter 6, pages 189–209. CSLI Publications, 1999.
5. J.C. Brown. Musical instrument identification using autocorrelation coefficients. In *Proceedings International Symposium on Musical Acoustics 1998, Leavenworth, Washington*, 1998.
6. J. Buckheit, S. Chen, D. Donoho, I. Johnstone, and J. Scargle. *About WaveLab*, January 1995.
7. R. Dawkins. *The Selfish Gene*. Oxford University Press, 1976.
8. P. de Chazal and R. B. Reilly. A comparison of the use of different wavelet coefficients for the classification of the electrocardiogram. In *Proceedings of the 14th International Conference on Pattern Recognition, September 2000*, 2000.
9. R. J. Dooling, M. R. Leek, O. Gleich, and M. L. Dent. Auditory temporal resolution in birds: Discrimination of harmonic complexes. *Journal of the Acoustical Society of America*, 112(2):748–759, August 2002.
10. S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(2):254–272, 1981.
11. S. Furui. *Digital Speech Processing, Synthesis, and Recognition*. Marcel Dekker, 2 edition, 2000.
12. B. Gold and N. Morgan. *Speech and Audio Signal Processing, Processing and Perception of Speech and Music*. John Wiley and Sons, inc., 2000.
13. G. Kondrak. Phonetic alignment and similarity. *Computers and the Humanities*, 37:273–291, 2003.
14. J. B. Kruskal and M. Liberman. The symmetric time-warping problem: from continuous to discrete. In D. Sankoff and J. B. Kruskal, editors, *Time Warps, String Edits, and Macromolecules - The Theory and Practice of Sequence Comparison*, chapter 4. CSLI Publications, 1999.
15. A. L. Lang and J. C. Barlow. Cultural evolution in the eurasian tree sparrow: Divergence between introduced and ancestral populations. *The Condor*, 99:413–423, 1997.
16. A. Lynch and A. J. Baker. A population memetics approach to cultural evolution in chaffinch song: Differentiation among populations. *Evolution*, 48(2):351–359, 1993.
17. A. Lynch and A. J. Baker. A population memetics approach to cultural evolution in chaffinch song: Meme diversity within populations. *The American Naturalist*, 141(4):597–620, 1993.
18. P. Marler and M. Tamura. Culturally transmitted patterns of vocal behavior in sparrows. *Science*, 146:1483–1486, Dec 1964.
19. P. Marler and H. Slabbekoorn. *Nature's Music: The Science of Birdsong*. Elsevier Academic Press, 2004.
20. P. Mitra. Chronux. http://mitralab.org.
21. 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.
22. J. Nerbonne, W. Heeringa, and P. Kleiweg. Edit distance and dialect proximity. In D. Sankoff and J. B. Kruskal, editors, *Time Warps, String Edits, and Macromolecules - The Theory and Practice of Sequence Comparison*, pages V–XV. CSLI Publications, 1999.
23. B. J. Oommen. String alignment with substitution, insertion, deletion, squashing, and expansion operations. *Information Sciences*, 83:89–107, 1995.
24. G. Peeters and X. Rodet. A large set of audio feature for sound description (similarity and classification) in the cuidado project. Technical report, Ircam, Analysis/Synthesis Team, 1 pl. Igor Stravinsky, 75004 Paris, France, 2004.
25. J. Podos, S. K. Huber, and B. Taft. Bird song: The interface of evolution and mechanism. *Annual Review of Ecology, Evolution, and Systematics*, 2004.

26. D. Sankoff and J. B. Kruskal, editors. *Time Warps, String Edits, and Macromolecules - The Theory and Practice of Sequence Comparison*. CSLI Publications, 1999.
27. M. D. Skowronski and J. G. Harris. Acoustic detection and classification of microchiroptera using machine learning: Lessons learned from automatic speech recognition. *Journal of the Acoustical Society of America*, 119(3):1817–1833, 2006.
28. O. Tchernichovski, F. Nottebohm, C. E. Ho, B. Pesaran, and P. P. Mitra. A procedure for an automated measurement of song similarity. *Animal Behaviour*, 59:1167–1176, 2000.
29. D. J. Thomson. Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, 70(9):1055–1096, 1982.
30. Y. Tohkura. A weighted cepstral distance measure for speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-35(10):1414–1422, October 1987.
31. 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*, pages 277–282, September 2005.
32. Vacher and Istrate. Life sounds extraction and classification in noisy environment, 2003.