相関行列（距離行列）を用いた音声の分類・判別手法について

百瀬　浩　（国）農研機構　中央農業研究センター　虫・鳥獣害研究領域

はじめに

あなたが、野外で見知らぬ鳥のさえずりを録音したとします。それが既知の種のさえずりのどれに近いのか、知りたい場合などがあると思います。あるいは、ある種類の鳥が色々なさえずりのパターンを持っていたとして、それを分類（タイプ分け）したい場合とかがあるかもしれません。本稿では、そんな時に使える手法として、異なる音声どうしの相関（Cross Correlation）を用いて分類や判別を行う手法を解説します。ただし、私は統計については素人なので、手法そのものの解説は行わず、結果を出すための方法についてだけ述べます。統計については、教科書を読むなり、ここに出てくるキーワードでググるなりして、各自お勉強をお願いします。

必要なツールほか

まず、前提条件として、今回は音声の分析、特に相関の計算に、有料のソフト（コーネル大学のRaven Pro 1.5[1)]を使用します。価格は400ドル位で、学生割引とかもあるようです。相関行列を自動で計算してくれる機能（音声どうしの相関の計算を総当たりでやってくれる機能）が、有料ソフトにしかついていないためです（残念！！）。ほかは無料の統計ソフトR[2)]を使うだけですので費用はかかりません。ちなみに、Rでも同様の計算はできるのですが、総当たりの計算、という部分にプログラミングが必要で、いろいろ面倒なので、楽をしてRavenProを使うことにしました。

今回使うソフト（RavenProとR）の使い方については、長くなるので最低限の説明のみ行います。例えば入手やインストール方法などは御自分でお調べ下さい。後、表計算ソフトも使いますが、これはまあテキストエディターでも代用できます（CSVファイルを自分で書く！）。

まずは王道のやり方から

さて、異なる音声どうしを比較しよう！という場合、通常考えつくアプローチとして、各音声について、色々な要素（例えば最低周波数、最高周波数、持続時間とかですかね…）を計測した多変量データを作り、これを用いて統計的な処理をする、というやり方があるでしょう。まず、これをやってみましょう。サンプルのファイル（BushSampleSongs.wav：ウグイスのさえずり9回分の音声）を、RavenProで開いてみてください。

百瀬　浩（2016）日本鳥学会　鳥学通信（ブログ版）2016.09.06 掲載
http://ornithology-japan.sblo.jp/

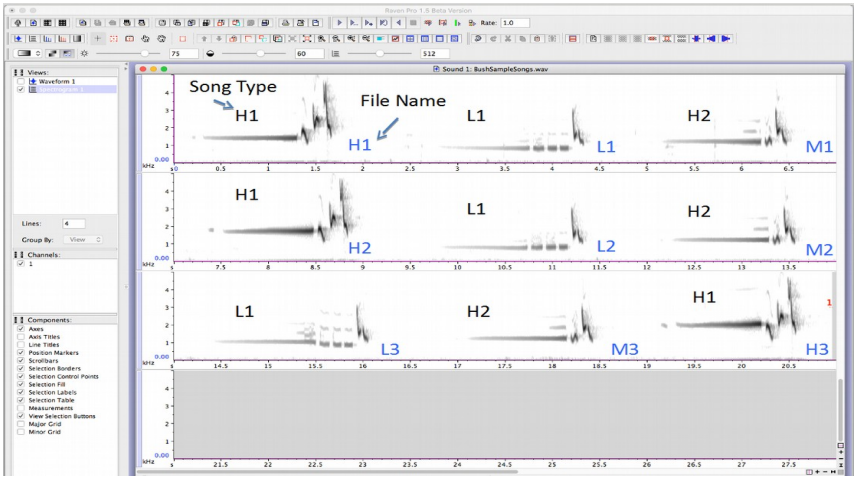


図１　サンプル用ファイル

すこし調整すると、図1のような画面になりす。具体的には、画面上の方の明るさなどを変えるツールバーで明るさを 75、コントラストを 80、DFT のサイズを規定値の倍の 512 に、左の方の Views というパネルで Waveform のチェックをはずし、Lines に4にしてください。それから、メニューの View→Configure View Axes を選んで、図2のように設定してください。

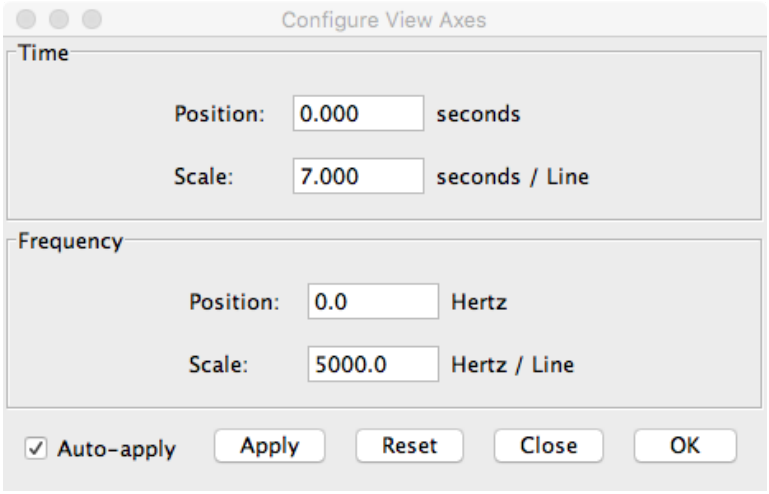


図２　RavenPro の View Axes 設定

できましたか？　さて、この状態でマウスをスペクトログラム上に持っていくとウィンドウの一番下に時間、周波数などが表示されるので、各さえずりの周波数などを計測してみます。ここでは少し手抜きをして、スペクトログラムから直接周波数を測りますが、正式にやる場合は、上から2行目のツールにある「Selection Spectrum View」などを使って、各時点でのスペクトラムを表示させ、スペクトログラム（声紋）ではなく スペクトラム（周波数特性を示すグラフ）上で周波数を計測するようにしてください。

できた計測結果が表 1 です。表の一番左側の FileName はファイル名で、図1に青い字で記入してあります（後でこの名前の音声ファイルをつくります）。ちなみに、黒い字で記入してある Song Type は私が分類したさえずりタイプの名称です[3)]。例えば、図 1 の9回のさえずりの中で、先頭から3，6，8番目のさえずりは、正式には H2 型（H 型の中で2番目に高い声）なのですが、ここでは H2 の代わりに M という名前を付けて、M1、M2、M3 というファイル名にしてあります。

表 1　9回のさえずりの計測値

FileName	cf_freq	n_cf_note	fm_min_freq	fm_max_freq	n_fm_note	fm_freq_range
H1	1400	2	1125	4500	3	3375
L1	825	4	1050	2900	2	1850
M1	1200	1	1075	2925	2	1850
H2	1700	2	1225	4500	3	3275
L2	775	4	1125	2800	2	1675
M2	1225	1	1025	2925	2	1900
L3	925	4	1150	2900	2	1750
M3	1200	1	1050	2875	2	1825
H3	2000	2	1250	4500	3	3250

表 1 の各項目を簡単に説明しておきます。

FileName	ファイル名
cf_freq	CF 部分（ホーホケキョのホーの部分）の周波数
n_cf_note	CF 部分のノートの数（ホーなら1）
fm_min_freq	FM 部分で使われた最低周波数
fm_max_freq	FM 部分で使われた最高周波数
n_fm_note	FM 部分のノートの数（ホケキョなら3）
fm_freq_range	FM 部分の最低～最高周波数の差

この表をテキスト形式で保存したものが song_keisoku.csv です。これを統計ソフトRに読み込んでみましょう。Rのコンソールに次のコマンドをコピーしてリターンキーを押し、実行してみてください。

```
x <- read.table("d:/sound/mds_sample/song_keisoku.csv", sep=",",
header=TRUE, row.names=1)
```

これは、ウィンドウズ版のRでやった場合の例で、実際には d:/sound/… の部分は皆様の実行環境に合わせて変えていただく必要があります。ウィンドウズ版の場合はメニューの ファイル → 作業ディレクトリの変更 マック版の場合は、メニューのその他 → 作業ディレクトリの変更 を選んでファイルが入っているフォルダーの場所を指定しておけば、ファイル名を書くだけで済みます。

```
x <- read.table("song_keisoku.csv", sep=",", header=TRUE,
row.names=1)
```

ここまでできたところで、コンソールに x と入力するとデータそのものを、 summary(x) と入力するとデータの概略を見ることができます。

さて、この部分は本題ではないので、タイプが明確に分かれることだけ確認しておきましょう。そこで、今読み込んだデータを、主成分分析により2つ位の変数にまとめて散布図を描いてみましょう。以下のコマンドを打ち込んでみてください。

```
pc <- prcomp(x, scale=TRUE) #先ほど読み込んだxに対して主成分分析を行います。
summary(pc) #同じく主成分分析の結果概要を表示します。
```

出力はこんな感じです。第 2 主成分（PC2）まででデータの変動の 94%を説明できますね。

```
Importance of components:
                          PC1  PC2  PC3  PC4  PC5  PC6
Standard deviation     2.11 1.10 0.55 0.20 0.03 0.00
Proportion of Variance 0.74 0.20 0.05 0.01 0.00 0.00
Cumulative Proportion  0.74 0.94 0.99 1.00 1.00 1.00
```

次に、主成分1と2をデータとして取り出します。

```
pc1 <- pc$x[,1]
pc2 <- pc$x[,2]
```

これを散布図にしてみましょう。

```
plot(pc1, pc2, type="n")
text(pc1, pc2, rownames(x))
```

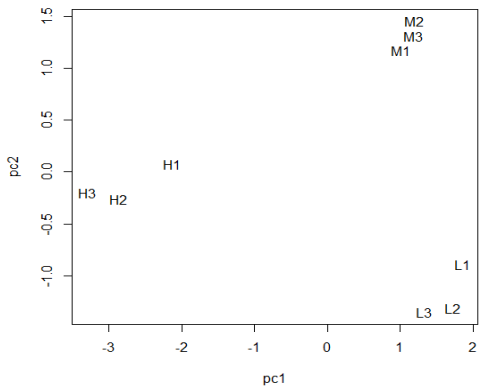


図 3　ウグイスさえずりタイプの分類結果
（第 1、第 2 主成分による散布図）

このように、H（H1 型）、M（H2 型）、L（L1 型）がそれぞれまとまり、3つのグループに分かれること

が確認できたと思います。

さて、ここからが本題…

前項のやり方は実に王道ではあるのですが、例えば構造が複雑な音声の場合、どこを測るのかも簡単に決められない、といった場合があります。また、面倒な手順抜きにざっくり比較できればいいや、というような場合は、前項のような要素に分けた計測をしないで、音声全体、というか音声を解析した声紋（スペクトログラム）のパターンそのものを使った分析方法があります。いわば、音声を二次元データである画像として扱い、画像どうしの類似度を数値化するようなイメージです。

そのやり方は、音声のセットに対して、そのすべての組み合わせについて相関を計算したマトリックスを作って、それを元に分析を行うというものです。相関の値は、音声どうしの類似度のようなものと考えてください。まったく同じ音声どうしだと相関が1に、正反対の音声、画像でいえばすべて黒の画像と白の画像のようなもの、であれば相関が0になります。

相関（類似度）行列の計算

では、実際に RavenPro を使って相関の計算をバッチで実行してみます。先ほどの図1の画面上で、マウスを使って、9個のさえずりのFM部分を選択（左上をクリックして右下までドラッグ）して、コピー（Edit メニューから Copy または ctrl + c）し、File メニューから New → Sound Window を選択してペースト（Edit メニューから Paste または ctrl + v）、そのまま File メニューから “SAVE SOUND xx AS...”を選択して、新しい FM 部分だけのファイルを作って下さい。こうして作ったファイルが H1～3.aif、M1～3.aif、L1～3.aif の9個のファイルです。*.wav ではなく、*.aif になっていますが、どちらでも動きますのでご心配なく。これを、何でも良いのですが、例えば「FM_Notes」のような一つのフォルダーに入れておいてください。

さて、ここからが RavenPro の本領発揮です。Tools メニューから Batch Correlator…を実行して下さい。次の図4のような画面になります。

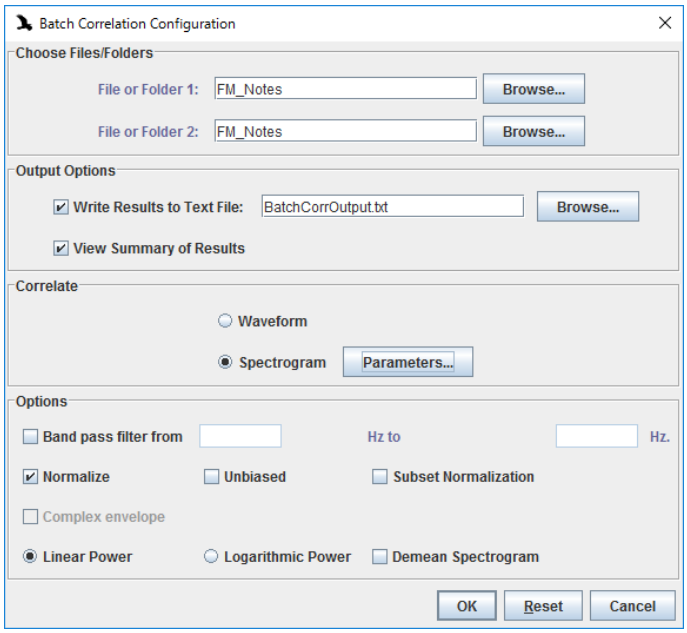


図4　Batch Correlator…　の設定

一番上の Choose Files/Folders のところでは、Browse ボタンを押して、先ほど9個のファイルを格納したフォルダーを選んでください。次の Output Options では、結果を出力するフォルダーとファイル名を指定します。その下の Correlate では、「Spectrogram」を選択して、隣の Parameters… ボタンを押し、次の図5のように設定して OK を押してください。

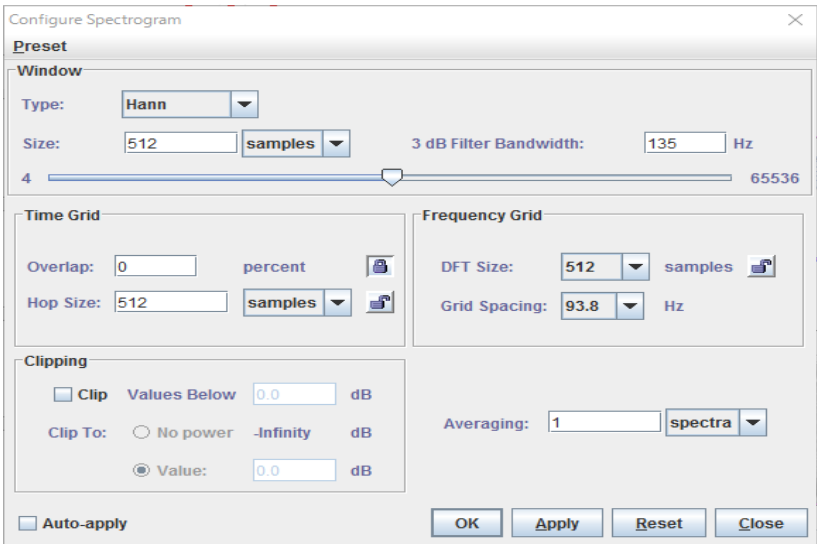
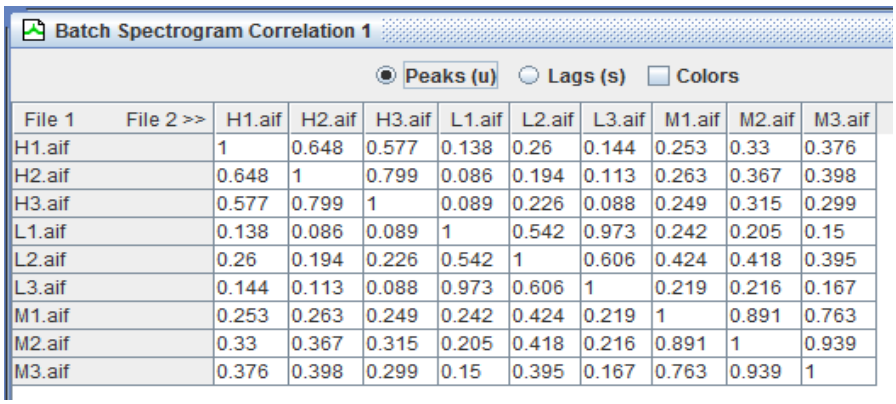


図5　Spectrogram Parameters の設定

さてさて、先ほどの Batch Correlator…　の設定画面に戻って OK を押すと計算が始まり、計算が終わると、次の図 6 のような画面が表示されます。これが、最初にいった音声どうしの総当たりの相関行列（類似度のマトリックス）というわけです。



Batch Spectrogram Correlation 1

Peaks (u) Lags (s) Colors

File 1 File 2 >>	H1.aif	H2.aif	H3.aif	L1.aif	L2.aif	L3.aif	M1.aif	M2.aif	M3.aif
H1.aif	1	0.648	0.577	0.138	0.26	0.144	0.253	0.33	0.376
H2.aif	0.648	1	0.799	0.086	0.194	0.113	0.263	0.367	0.398
H3.aif	0.577	0.799	1	0.089	0.226	0.088	0.249	0.315	0.299
L1.aif	0.138	0.086	0.089	1	0.542	0.973	0.242	0.205	0.15
L2.aif	0.26	0.194	0.226	0.542	1	0.606	0.424	0.418	0.395
L3.aif	0.144	0.113	0.088	0.973	0.606	1	0.219	0.216	0.167
M1.aif	0.253	0.263	0.249	0.242	0.424	0.219	1	0.891	0.763
M2.aif	0.33	0.367	0.315	0.205	0.418	0.216	0.891	1	0.939
M3.aif	0.376	0.398	0.299	0.15	0.395	0.167	0.763	0.939	1

図 6　Batch Correlator…　の出力結果

このデータをRに読み込む必要がありますので、先ほど指定した出力結果のファイル（画面の例では BatchCorrOutput.txt）をテキストエディターで開いて、前半部分の　Batch Correlation Peaks (u) だけを残した次のようなファイルを作って下さい。

File 1	File 2 >>	H1.aif	H2.aif	H3.aif	L1.aif	L2.aif	L3.aif	M1.aif	M2.aif	M3.aif
H1.aif	1	0.648	0.577	0.138	0.26	0.144	0.253	0.33	0.376	
H2.aif	0.648	1	0.799	0.086	0.194	0.113	0.263	0.367	0.398	
H3.aif	0.577	0.799	1	0.089	0.226	0.088	0.249	0.315	0.299	
L1.aif	0.138	0.086	0.089	1	0.542	0.973	0.242	0.205	0.15	
L2.aif	0.26	0.194	0.226	0.542	1	0.606	0.424	0.418	0.395	
L3.aif	0.144	0.113	0.088	0.973	0.606	1	0.219	0.216	0.167	
M1.aif	0.253	0.263	0.249	0.242	0.424	0.219	1	0.891	0.763	
M2.aif	0.33	0.367	0.315	0.205	0.418	0.216	0.891	1	0.939	
M3.aif	0.376	0.398	0.299	0.15	0.395	0.167	0.763	0.939	1	

このテキストファイルを表計算ソフトに読み込んで、いらない情報（例えばファイル名の .aif という部分など）を消し、CSV 形式で保存してください。こうしてできたのが、Output2.csv　ファイルです。

Rを用いた多次元尺度法および階層クラスター分析

ここまで（相関行列をつくるまで）が Raven Pro の仕事で、後は統計ソフトRでサクッと計算が行えます。Rでは、相関行列から距離行列というものを使って、多次元尺度法（Multidimentional Scaling： MDS）という分析をおこないます（参考文献 4～6）。これは、距離行列を用いて、互いに位置が近いものを近くに、遠いものを遠くに配置した分布図を作るような分析方法です。では、やってみましょう。

Rを起動し、先ほどの CSV ファイルを読み込みます。

```
data <- read.table("d:/sound/mds_sample/Output2.csv", sep=",",
header=TRUE, row.names=1)
```

この時、`read.table` の最後の2つのパラメータ、つまり `header=TRUE, row.names=1` の部分は必ずこの通りに入力して下さい。でないと、次のステップがうまく動かなくなります。

次に、読み込んだデータセットを、距離行列という形式に変換します。さきほど RavenPro で作ったのは相関行列（類似度のようなもの）でしたが、多次元尺度法 で使うのは相関係数（類似度）ではなく、その反対の距離（相違度？）なので、その変換も同時に行ってしまいます。計算式は

距離 ＝ 1 － 相関 （あるいは 相違度 ＝ 1 － 類似度）

です（計算式というほどのものではないw）。これでやっと、距離行列の完成です。Rで実際にやってみましょう。文章で書くと長いですが、Rではあっという間に計算できますよ。

```
d <- as.dist( 1 - data )   # 距離行列に変換
```

この距離行列を使って多次元尺度法を実行して、結果をプロットしてみましょう。

```
mds <- cmdscale(d) #多次元尺度法の計算
plot(mds, type="n") #結果をプロット、ただし実際の出力はまだしない
text(mds, rownames(mds)) #さきほどの場所にラベルをプロット
```

結果は、次の図7のようになります。

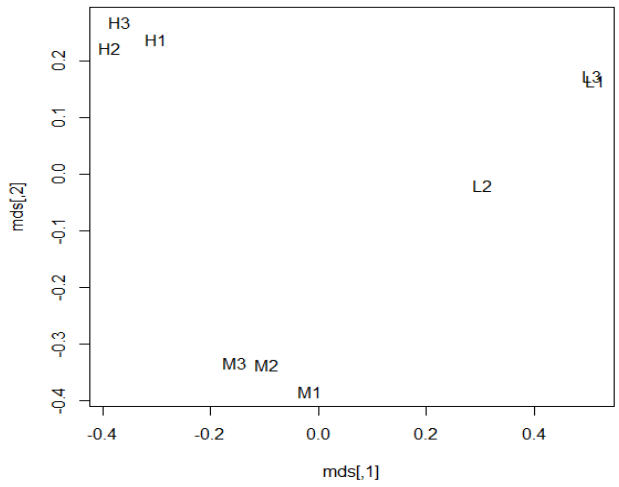


図7　多次元尺度法によるウグイスのさえずりの分類結果

最初にやった王道メソッド（多要素の計測結果を用いた多変量解析）と同様に、H（H1 型）、M（H2 型）、L（L1 型）がそれぞれまとまり、3つのグループにうまく分かれていますね。

距離行列を用いた階層クラスター分析

先ほど作った距離行列があると、似たものどうしをグルーピングする、階層クラスター分析も行えます。

```
plot(hclust(d))　#距離行列のデータを使って階層クラスター分析
```

結果は、次の図8のようになります。3種類のタイプがそれぞれ分かれるだけではなく、H（H1）とM（H2）が、一つのグループ（H 型）を形成してL 型と別れ、H 型の中でさらにH（H1）とM（H2）が分かれていることが良くわかると思います。

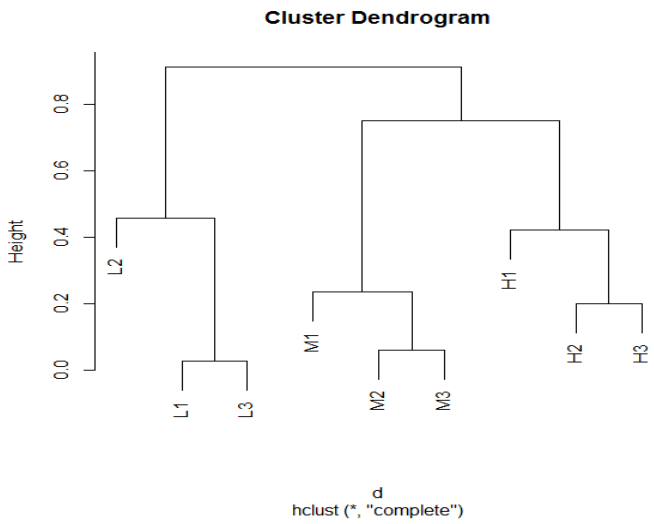


図8　階層クラスター分析によるウグイスさえずりタイプの分類結果

終わりに

この稿では、鳥のさえずりなどの音声を分類・判別する手法について説明してきました。音声を主体に説明を行いましたが、手法としては任意の多変量データに応用可能です。最初の方で説明した多変量データからは、すぐに相関行列が計算できるので、それを距離行列に変換すれば、本稿で述べた多次元尺度法などの計算が行える、というわけです。

私が以前この鳥学通信に書いた、音声関係の解説記事（参考文献7、8）もぜひ参考にして下さい。

こうした手法を活用していただき、皆様がお持ちのデータを積極的に論文などにまとめていただければと思います。ぜひ、がんばってくださいね。

参考文献

1) Cornell Lab. of Ornithology Bioacoustics Research Program (2010) Raven Pro 1.4 User's Manual http://www.birds.cornell.edu/brp/raven/Raven14UsersManual.pdf

2) R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

3) Momose, H. (1999) Structure of Territorial Songs in the Japanese Bush Warbler (*Cettia diphone*). Mem. Fac. Sci. Kyoto Univ. (*Ser. Biol.*), 16: 55-65.

4) Caroll, JD & Kruskal, JB (1978) Multidimensional scaling. Pp. 892-907 in J. Tanur & W. Kruscal (Eds.), *International Encyclopedia of Statistics*. New York: MacMillan and the Free Press.

5) Clark, CW, Marler, P. & Beeman, K (1987) Quantitative Analysis of Animal Vocal Phonology: an Application to Swamp Sparrow Song. Ethology 76: 101-115.

6) Commitee on Hearing, Bioacoustics, and Biomechanics Commission on Behavioral and Social Sciences and Education, National Research Council (1989) Classification of Complex Nonspeech Sounds. Panel on Classification of Complex Nonspeech Sounds

7) 百瀬　浩(2008) 鳥類研究者のための音声分析ガイド 日本鳥学会、鳥学通信 No.12(2)

8) 百瀬　浩(2009) 鳥類研究者のための野外録音ガイド 日本鳥学会、鳥学通信 No.24(4)