Dissimilarity-based classification for bioacoustic monitoring of bird species

José Francisco Ruiz-Muñoz,
Mauricio Orozco-Alzate
Universidad Nacional de Colombia
Sede Manizales

September 29, 2011
Motivation

- Estimating birds populations is a common rate to assess ecosystem health.
- In the Colombian mountains there is a rich variety of bird species and several inventories of them have been performed.
Motivation

- Estimating birds populations is a common rate to assess ecosystem health.

- In the Colombian mountains there is a rich variety of bird species and several inventories of them have been performed.
Motivation

- Estimating birds populations is a common rate to assess ecosystem health.
- In the Colombian mountains there is a rich variety of bird species and several inventories of them have been performed.
Permanent programs for monitoring populations have hardly been implemented because carrying out assessments by visual and in situ inspections implies an expensive process that includes long and exhausting journeys, and is spatially and temporally restricted.
Communication and computer technologies allow the implementation of automated solutions for biodiversity monitoring [Brandes, 2008].
Automatic systems: a solution

A number of previous efforts have been already undertaken in order to design and implement this type of systems. Their goal has been to develop software and hardware aimed at recording, transmitting, analyzing and identifying acoustical data.
The system is roughly composed by the following stages:

1. **Raw data acquisition:** sound signals are registered and digitized by using sensors.
2. **Preprocessing:** sound segments are selected.
3. **Representation:** signals are either transformed or measured to extract discriminant features from them.
4. **Classification:** discriminant functions are built.

This study focuses in stages 3 and 4.
Automatic identification system

The system is roughly composed by the following stages:

1. **Raw data acquisition**: sound signals are registered and digitized by using sensors.
2. **Preprocessing**: sound segments are selected.
3. **Representation**: signals are either transformed or measured to extract discriminant features from them.
4. **Classification**: discriminant functions are built.

This study focuses in stages 3 and 4.
The system is roughly composed by the following stages:

1. **Raw data acquisition**: sound signals are registered and digitized by using sensors.

2. **Preprocessing**: sound segments are selected.

3. **Representation**: signals are either transformed or measured to extract discriminant features from them.

4. **Classification**: discriminant functions are built.

This study focuses in stages 3 and 4.
The system is roughly composed by the following stages:

1. **Raw data acquisition**: sound signals are registered and digitized by using sensors.

2. **Preprocessing**: sound segments are selected.

3. **Representation**: signals are either transformed or measured to extract discriminant features from them.

4. **Classification**: discriminant functions are built.

This study focuses in stages 3 and 4.
The system is roughly composed by the following stages:

1. **Raw data acquisition**: sound signals are registered and digitized by using sensors.
2. **Preprocessing**: sound segments are selected.
3. **Representation**: signals are either transformed or measured to extract discriminant features from them.
4. **Classification**: discriminant functions are built.

This study focuses in stages 3 and 4.
Automatic identification system

The system is roughly composed by the following stages:

1. **Raw data acquisition**: sound signals are registered and digitized by using sensors.
2. **Preprocessing**: sound segments are selected.
3. **Representation**: signals are either transformed or measured to extract discriminant features from them.
4. **Classification**: discriminant functions are built.

This study focuses in stages 3 and 4.
Automatic identification system

- $x(t)$: analogous signal
- $x(n)$: digitized signal
- $x$: feature representation of the signal
- $d$: dissimilarity representation of the signal
- $\omega_k$: label of class
Dissimilarity-based classification consists in building classifiers in dissimilarity spaces (see [Pekalska & Duin, 2005]). One representation set \( R \) with \( n \) prototypes \( p_i \) is defined as

\[
R = \{ p_1, p_2, \ldots, p_n \}. \tag{1}
\]
Dissimilarity-based classification consists in building classifiers in dissimilarity spaces (see [Pekalska & Duin, 2005]). One representation set \( R \) with \( n \) prototypes \( p_i \) is defined as

\[
R = \{ p_1, p_2, \ldots, p_n \} .
\]
Dissimilarity-based classification consists in building classifiers in dissimilarity spaces (see [Pekalska & Duin, 2005]). One representation set $R$ with $n$ prototypes $p_i$ is defined as

$$R = \{p_1, p_2, \ldots, p_n\}.$$ (1)
An object $x$ is represented as a vector of the dissimilarities computed between $x$ and the prototypes from $R$:

$$D(x, R) = [d(x, p_1), d(x, p_2), ..., d(x, p_n)] \quad (2)$$

For a training set $T$ of $N$ objects, a classifier, e.g. k-nearest neighbor classifier $knnC$, linear Bayesian classifier $lbc$ or quadratic Bayesian classifier $qbc$, can be built.
Dissimilarity-based classification

An object $x$ is represented as a vector of the dissimilarities computed between $x$ and the prototypes from $R$:

$$D(x, R) = [d(x, p_1), d(x, p_2), ..., d(x, p_n)] \quad (2)$$

For a training set $T$ of $N$ objects, a classifier, e.g. k-nearest neighbor classifier $knnC$, linear Bayesian classifier $lbc$ or quadratic Bayesian classifier $qbc$, can be built.
An object $x$ is represented as a vector of the dissimilarities computed between $x$ and the prototypes from $R$:

$$D(x, R) = [d(x, p_1), d(x, p_2), \ldots, d(x, p_n)]$$  \hspace{1cm} (2)$$

For a training set $T$ of $N$ objects, a classifier, e.g. k-nearest neighbor classifier $knn$, linear Bayesian classifier $lbc$ or quadratic Bayesian classifier $qbc$, can be built.
The 1-D transform contains frequency information about the acoustic signal. In this study, it was computed by using the following:

- Fast Fourier Transform ($FFT$)
- Parametric estimation of the Power Spectral Density ($PSD$)
The 1-D transform contains frequency information about the acoustic signal. In this study, it was computed by using the

- Fast Fourier Transform (FFT)
- Parametric estimation of the Power Spectral Density (PSD)
The 1-D transform contains frequency information about the acoustic signal. In this study, it was computed by using the

- Fast Fourier Transform (FFT)
- Parametric estimation of the Power Spectral Density (PSD)
However, in the FFT and the PSD, time localizations of the events are not specified in the spectrum.
However, in the FFT and the PSD, time localizations of the events are not specified in the spectrum.
Representations of bioacoustic signals
Two dimensional (2-D) representations

In order to include the spectral changes over time, 2-D transforms are also used. In particular, a 1-D transform for a short portion of the signal is multiplied by a nonzero and sliding window function.
Representations of bioacoustic signals
Two dimensional (2-D) representations

In the case of FFT, such a 2-D extension is known as Short Time Fourier Transform (STFT).
Dissimilarities between 1-D transforms are computed by applying the Euclidean distance.
Euclidean distance can not be used for 2-D transforms due to the asymmetry. In that case we use the Earth Mover’s Distance (EMD) [Rubner et al., 2000], that is applicable for distributions:

- 1-D representations along the transforms are clustered using algorithm of $k$-means. Centroids and number of objects from clusters are the input parameters to estimate EMD.
Euclidean distance can not be used for 2-D transforms due to the asymmetry. In that case we use the Earth Mover’s Distance (\textit{EMD}) [Rubner et al., 2000], that is applicable for distributions:

- 1-D representations along the transforms are clustered using algorithm of \textit{k}-means. Centroids and number of objects from clusters are the input parameters to estimate \textit{EMD}.
Outline

1. Introduction
2. Conceptual setup
3. Experimental setup
4. Results
5. Conclusion
6. Future work
A set of birdsong recordings of eleven species were used in the experiments.
Raw field recordings were taken at Reserva Natural Río Blanco in Manizales, Colombia.

- The sampling frequency of the recordings is 44.1 kHz.
- Bird sounds are detected in raw data using segmentation techniques based on the energy [Härmä, 2003].
- The data set contains a total of 595 syllables distributed per species as shown in Table 1.
Raw field recordings were taken at Reserva Natural Río Blanco in Manizales, Colombia.

The sampling frequency of the recordings is 44.1 kHz.

Bird sounds are detected in raw data using segmentation techniques based on the energy [Härmä, 2003].

The data set contains a total of 595 syllables distributed per species as shown in Table 1.
Raw field recordings were taken at Reserva Natural Río Blanco in Manizales, Colombia.

The sampling frequency of the recordings is 44.1 kHz.

Bird sounds are detected in raw data using segmentation techniques based on the energy [Härmä, 2003].

The data set contains a total of 595 syllables distributed per species as shown in Table 1.
Dataset

- Raw field recordings were taken at Reserva Natural Río Blanco in Manizales, Colombia.
- The sampling frequency of the recordings is 44.1 kHz.
- Bird sounds are detected in raw data using segmentation techniques based on the energy [Härmä, 2003].
- The data set contains a total of 595 syllables distributed per species as shown in Table 1.
Dataset

Table: 1. Birds species in the current study.

<table>
<thead>
<tr>
<th>Latin name</th>
<th>Abbreviation</th>
<th>Syllables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grallaria ruficapilla</td>
<td>GR</td>
<td>33</td>
</tr>
<tr>
<td>Henicorhina leucophrys</td>
<td>HL</td>
<td>64</td>
</tr>
<tr>
<td>Mimus gilvus</td>
<td>MG</td>
<td>66</td>
</tr>
<tr>
<td>Myadestes ralloides</td>
<td>MR</td>
<td>58</td>
</tr>
<tr>
<td>Pitangus sulphuratus</td>
<td>PS</td>
<td>53</td>
</tr>
<tr>
<td>Pyrrhomyias cinnamomea</td>
<td>PC</td>
<td>36</td>
</tr>
<tr>
<td>Troglodytes aedon</td>
<td>TA</td>
<td>33</td>
</tr>
<tr>
<td>Turdus ignobilis</td>
<td>TI</td>
<td>74</td>
</tr>
<tr>
<td>Turdus serranus</td>
<td>TS</td>
<td>78</td>
</tr>
<tr>
<td>Xiphocolaptes promeropirhynchus</td>
<td>XP</td>
<td>46</td>
</tr>
<tr>
<td>Zonotrichia capensis</td>
<td>ZC</td>
<td>54</td>
</tr>
</tbody>
</table>
Experiments

- Spectral (FFT and PSD) and time-varying (STFT and PSD time-varying) representations were calculated from dataset.

- Dataset was divided into train and test (70% by training and 30% by testing) and the selected prototype set consists of 60 random syllables.

- The classifiers built and tested were knnc, lbc and qbc; besides for the two last ones, a regularization of 0.01 was used.

The process was repeated 40 times.
Experiments

- Spectral (FFT and PSD) and time-varying (STFT and PSD time-varying) representations were calculated from dataset.

- Dataset was divided into train and test (70% by training and 30% by testing) and the selected prototype set consists of 60 random syllables.

- The classifiers built and tested were knnc, lbc and qbc; besides for the two last ones, a regularization of 0.01 was used.

The process was repeated 40 times.
Experiments

- Spectral (FFT and PSD) and time-varying (STFT and PSD time-varying) representations were calculated from dataset.

- Dataset was divided into train and test (70% by training and 30% by testing) and the selected prototype set consists of 60 random syllables.

- The classifiers built and tested were knnc, lbc and qbc; besides for the two last ones, a regularization of 0.01 was used.

The process was repeated 40 times.
Experiments

- Spectral \( (FFT \text{ and } PSD) \) and time-varying \( (STFT \text{ and } PSD \text{ time-varying}) \) representations were calculated from dataset.

- Dataset was divided into train and test (70% by training and 30% by testing) and the selected prototype set consists of 60 random syllables.

- The classifiers built and tested were \textit{knn}, \textit{lbc} and \textit{qbc}; besides for the two last ones, a regularization of 0.01 was used.

The process was repeated 40 times.
The following performance measures were considered:

- The percentage of samples that were correctly assigned to the class with respect to the number of samples belonging to that class, known as true positive (TP) rate.

- The percentage of samples that were incorrectly labelled as belonging to each class respect to the number of samples belonging to the other classes, also known as false positive (FP) rate.
The following performance measures were considered:

- The percentage of samples that were correctly assigned to the class with respect to the number of samples belonging to that class, known as true positive (TP) rate.

- The percentage of samples that were incorrectly labelled as belonging to each class respect to the number of samples belonging to the other classes, also known as false positive (FP) rate.
The following performance measures were considered:

- The percentage of samples that were correctly assigned to the class with respect to the number of samples belonging to that class, known as true positive (TP) rate.
- The percentage of samples that were incorrectly labelled as belonging to each class respect to the number of samples belonging to the other classes, also known as false positive (FP) rate.
A remarkable result is the one shown in the experiment using STFT for representation and qbc for classification, where percentage average of TP is $94 \pm 7.05$ and FP is $7.17 \pm 10.73$.

An optimal result is obtained in this experiment for classification of *Grallaria ruficapilla* (TP of 100% and FP of $0.16 \pm 0.99\%$) using STFT representation and a qbc.
A remarkable result is the one shown in the experiment using *STFT* for representation and *qbc* for classification, where percentage average of TP is $94 \pm 7.05$ and FP is $7.17 \pm 10.73$.

An optimal result is obtained in this experiment for classification of *Grallaria ruficapilla* (TP of 100% and FP of $0.16 \pm 0.99\%$) using *STFT* representation and a *qbc*. 
A strange case is that results obtained for classes *Turdus serranus* and *Turdus ignobilis*, when using *lbc*, are poor but; in contrast, good performances were obtained when using *knncc*.
Results shows that a general problem is the high dispersion in the performance evaluation, that is observable by values of standard deviation.

- In tests considering 1-D transforms, \textit{knn} \textit{c} achieved a better performance than \textit{lbc} and \textit{qbc}.
- 2-D transforms, overall, show a better performance than 1-D ones.
- In spite that \textit{PSD} has, in general, a better performance than the one for \textit{FFT}, the overall best result was obtained for the \textit{STFT}.
- \textit{PSD} and 2-D \textit{PSD} transforms exhibit similar results.
Results shows that a general problem is the high dispersion in the performance evaluation, that is observable by values of standard deviation.

In tests considering 1-D transforms, $\textit{knn}$ achieved a better performance than $\textit{lbc}$ and $\textit{qbc}$.

2-D transforms, overall, show a better performance than 1-D ones.

In spite that $\textit{PSD}$ has, in general, a better performance than the one for $\textit{FFT}$, the overall best result was obtained for the $\textit{STFT}$.

$\textit{PSD}$ and 2-D $\textit{PSD}$ transforms exhibit similar results.
Results show that a general problem is the high dispersion in the performance evaluation, that is observable by values of standard deviation.

In tests considering 1-D transforms, \textit{knn} \textit{c} achieved a better performance than \textit{lbc} and \textit{qbc}.

2-D transforms, overall, show a better performance than 1-D ones.

In spite that \textit{PSD} has, in general, a better performance than the one for \textit{FFT}, the overall best result was obtained for the \textit{STFT}.

\textit{PSD} and 2-D \textit{PSD} transforms exhibit similar results.
Results

Overall performance

- Results shows that a general problem is the high dispersion in the performance evaluation, that is observable by values of standard deviation.
- In tests considering 1-D transforms, $knn_c$ achieved a better performance than $lbc$ and $qbc$.
- 2-D transforms, overall, show a better performance than 1-D ones.
- In spite that $PSD$ has, in general, a better performance than the one for $FFT$, the overall best result was obtained for the $STFT$.
- $PSD$ and 2-D $PSD$ transforms exhibit similar results.
Results

Overall performance

- Results show that a general problem is the high dispersion in the performance evaluation, that is observable by values of standard deviation.
- In tests considering 1-D transforms, \textit{knn} achieved a better performance than \textit{lbc} and \textit{qbc}.
- 2-D transforms, overall, show a better performance than 1-D ones.
- In spite that \textit{PSD} has, in general, a better performance than the one for \textit{FFT}, the overall best result was obtained for the \textit{STFT}.
- \textit{PSD} and 2-D \textit{PSD} transforms exhibit similar results.
Spectral (1-D) and time varying (2-D) transforms, together with dissimilarity-based classification, can be a feasible option for recognizing bird sounds.

To choose a representation or transform, it might be important to take into account non-stationarity because, according to the experimental results, a best performance is obtained in general for a 2-D transform: \textit{STFT}. 

Spectral (1-D) and time varying (2-D) transforms, together with dissimilarity-based classification, can be a feasible option for recognizing bird sounds.

To choose a representation or transform, it might be important to take into account non-stationarity because, according to the experimental results, a best performance is obtained in general for a 2-D transform: $STFT$. 
Future work

As future work, we propose to explore other types of 2-D transforms (other types of spectral or feature representations for frames) as well as other dissimilarity measures, in order to obtain both good classification performances and low computational costs.

In addition, a long-term objective is the design of a complete system for avian monitoring, including hardware and software to be deployed in a real environment.
Future work

As future work, we propose to explore other types of 2-D transforms (other types of spectral or feature representations for frames) as well as other dissimilarity measures, in order to obtain both good classification performances and low computational costs.

In addition, a long-term objective is the design of a complete system for avian monitoring, including hardware and software to be deployed in a real environment.
Resources

ARBIMON
http://arbimon.uprrp.edu/arbimonweb/

Birds photos
http://birdsmanizales.blogspot.com/
http://www.birdphotos.com
http://www.opeco.org/
http://www.animalpicturesarchive.com/
References

Automated sound recording and analysis techniques for bird surveys and conservation.  
*Bird Conservation International, 18*(S1), S163–S173.

Härmä, Aki. 2003 (apr).  
Automatic identification of bird species based on sinusoidal modeling of syllables.  

The Dissimilarity Representation for Pattern Recognition: Foundations And Applications.  
River Edge, NJ, USA: World Scientific Publishing Co., Inc.

The earth mover’s distance as a metric for image retrieval.  
Questions ?