Feature and Dissimilarity Representations for the Sound-Based Recognition of Bird Species

José Francisco Ruiz-Muñoz[1], Mauricio Orozco-Alzate[1,2,⋆], and César Germán Castellanos-Domínguez[1]

[1] Signal Processing and Recognition Group, Universidad Nacional de Colombia Sede Manizales, km 7 vía al aeropuerto, Manizales (Caldas), Colombia
{jfruizmu,morozcoa,cgcastellanosd}@unal.edu.co
[2] Departamento de Informática y Computación, Universidad Nacional de Colombia Sede Manizales, km 7 vía al aeropuerto, Manizales (Caldas), Colombia

Abstract. Pattern recognition and digital signal processing techniques allow the design of automated systems for avian monitoring. They are a non-intrusive and cost-effective way to perform surveys of bird populations and assessments of biological diversity. In this study, a number of representation approaches for bird sounds are compared; namely, feature and dissimilarity representations. In order to take into account the non-stationary nature of the audio signals and to build robust dissimilarity representations, the application of the Earth Mover's Distance (EMD) to time-varying measurements is proposed. Measures of the leave-one-out 1-NN performance are used as comparison criteria. Results show that, overall, the Mel-ceptrum coefficients are the best alternative; specially when computed by frames and used in combination with EMD to generate dissimilarity representations.

Keywords: Automated avian monitoring, bird sounds, dissimilarity representations, feature representations.

1 Introduction

Advances in pattern recognition and digital signal processing allow the identification of bird species by their emitted sounds and, thereby, the design of automated systems for avian monitoring. In spite of those advances, biodiversity assessments have typically been carried out by visual inspection, which requires human involvement and, therefore, may be expensive and have a limited coverage. In contrast, automatic acoustic monitoring is a non-intrusive and cost-effective alternative that may provide good temporal and spatial coverages.

The simplest sounds in a bird song are called *elements* or *notes*. Several notes together in a regular pattern in a song constitute a *syllable* and, in turn, several syllables are a *song phrase* [1]. Previous studies [2–4] have shown that the sound-based recognition of bird species is suitable when considering syllables as

⋆ Mauricio Orozco-Alzate is a member of Sociedad Caldense de Ornitología (SCO), a regional ornithological society from Caldas, Colombia: http://rnoa.org/sco/

C. San Martin and S.-W. Kim (Eds.): CIARP 2011, LNCS 7042, pp. 451–458, 2011.

elementary units. Raw measurements corresponding to those elementary units have to be represented in vector spaces where classification rules can afterwards be applied. Representations for bioacoustic signals have traditionally been built by feature extraction; however, we advocate that dissimilarity representations are also a feasible option to face this problem. Furthermore, dissimilarities have the potential to build either simpler or richer representations; the later case, a richer representation, when considering for instance time-varying measurements that take into account non-stationarity. In this study, we evaluate different types of representations, including feature-based and dissimilarity-based ones, for bird sounds segmented into syllables.

Considered feature representations include the so-called *standard features* and a so-called *coarse representation of segment structure*; both of them include the syllable duration as well as features related to particular frequencies and maximum values in the frequency domain [2]. Besides, we evaluate a set of acoustical features, named in [3] and here as *descriptive features*, and the *Mel-cepstrum representation*, which is based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. Spectral analyses such as as the *Fast Fourier Transform* (FFT) and the *Parametric estimation of the Power spectral Density* (PSD) are simple initial representations to find dissimilarities between syllables. Such options to build dissimilarity representations are also considered. In addition, we calculate dissimilarities with richer spectral estimates, namely time-varying ones, that consist in dividing segments into frames and mapping each one into two-dimensional spectral or feature representations. The computation of dissimilarities between time-varying initial representations is carried out by using the *Earth Mover's Distance* (EMD), due to its usefulness to compare distributions.

The goodness of a particular representation can be roughly assessed by using measures of the leave-one-out nearest-neighbor (1-NN) performance. Such measures are commonly used as criteria when selecting features or prototypes for a representation [5]. We use them here as comparison values between the evaluated representations.

2 Methods

The design of a bird sound recognition system includes, at least, the following three stages: preprocessing, representation and performance evaluation. The first one consists in the segmentation of continuous records, whose objective is to detect intervals —according to the energy signal— where there are sounds emitted by birds. Consequently, it is assumed that bird sounds are located in signal regions with high energy levels. Steps of the segmentation stage are: computation of the energy signal, estimation of an energy threshold, search of syllables (regions having energies above a threshold), and the application of a criterion of deletion and merging of very short segments.

Regarding the second stage —representation— several methods of feature-based representations, commonly used in bioacoustics and bird sound classification, are compared in this paper. In addition, the dissimilarity-based approach is proposed as an alternative for representation. The last stage —performance evaluation— is carried out, as indicated at the end of Sec. 1, i.e. by using the leave-one-out 1-NN performance.

2.1 Feature Representations

Standard features: Segments are characterized by using four features as proposed in [2]; namely minimum and maximum frequencies, temporal duration and maximum power.

Coarse representation of segment structure: The following eleven variables, originally proposed in [2], are used as features for each segment: minimum and maximum frequencies, temporal duration and frequency of maximum power in eight non-overlapping frames.

Descriptive features: This set includes both temporal and spectral features. Segments are divided into overlapping frames of 256 samples with 50% overlap. For each frame, the following features are estimated: spectral centroid, signal bandwidth, spectral roll-off frequency, spectral flux, spectral flatness, zero crossing rate and short time energy. Feature vectors for classification are composed by mean and variance values of the feature trajectories along the frames. Frequency range (minimum and maximum frequencies), segment temporal duration and modulation spectrum (position and magnitude of the maximum peak in the modulation spectrum) are calculated from the entire segment. Therefore, 19 features are calculated with this method as proposed in [3].

Mel-cepstrum representation: Mel-frequency cepstral coefficients (MFCCs) are a feature representation method commonly used in many audio classification problems, e.g. in speech recognition. Mel-frequency scale is derived from the human perceptual system. Such systems in birds are not the same but exhibit similar characteristics; therefore, MFCCs have also been used in birdsong recognition [3, 4]. The first 12 MFCCs, the log-energy and the so-called delta and delta-delta coefficients are obtained for each frame. Their mean values along the frames are used as features, as proposed in [3].

2.2 Dissimilarity Representations

A dissimilarity representation consists in building vectorial spaces where coordinate axes represent dissimilarities —typically distance measures— to prototypes. In these spaces, classifiers can be built. In a full dissimilarity matrix, prototypes are all the elements available in a particular dataset. The matrix is often symmetric and must be real and have zero diagonal. “Dissimilarity representations can

be derived in many ways, e.g. from raw (sensor) measurements such as images, histograms or spectra or, from an initial representation by features, strings or graphs" [5]. Considering that the analysis of signal properties is usually done in the frequency domain, we have calculated the spectrum for each signal by using two different approaches: FFT and PSD. Dissimilarity representations have been then computed by pointwise distances between spectra.

Dissimilarity representations, derived as described above, suppose that spectral behavior is similar in the entire segment. In order to obviate such an assumption, we also use representations that change over time (time-varying). In such a way, the acoustic space for each sound segment is efficiently covered [6]. Time-varying representations are computed by dividing sound segments into frames and converting each one to either a spectral or a feature representation. Feature sets measured for each frame were: 1) spectrogram, also known as *short time Fourier transform* (SFT); 2) PSD by using the Yule Walker method; 3) selected descriptive features (spectral centroid, signal bandwidth, spectral roll-off frequency, spectral flux, spectral flatness, zero crossing rate and short time energy); and 4) the Mel-cepstrum representation as explained in Sec. 2.1. Sets 3) and 4) must be standardized because features are not in same scale.

Measuring dissimilarities between representations: In the case of equally-sized representations (e.g. FFT or PSD) a classical measure, as the Euclidean distance, can be used. Conversely, the Euclidean distance can not be directly applied to time-varying representations. To overcome this difficulty, in this study we have used the EMD. Due to space constraints, we are not able to provide a description for this distance measure; see [7] and [6] for further implementation details.

3 Experimental Results

We performed a set of experiments on a dataset of raw field recordings taken at *Reserva Natural Río Blanco* in Manizales, Colombia. The sampling frequency of the recordings is 44.1 kHz. The dataset is composed by a total of 595 syllables distributed per species as follows[1]: *Grallaria ruficapilla* (GR, 33), *Henicorhina leucophrys* (HL, 64), *Mimus gilvus* (MG, 66), *Myadestes ralloides* (MR, 58), *Pitangus sulphuratus* (PS, 53), *Pyrrhomyias cinnamomea* (PC, 36), *Troglodytes aedon* (TA, 33), *Turdus ignobilis* (TI, 74), *Turdus serranus* (TS, 78), *Xiphocolaptes promeropirhynchus* (XP, 46) and *Zonotrichia capensis* (ZC, 54).

Evaluation for each representation was assessed by using measures of the leave-one-out 1-NN performance. For each representation, a confusion matrix is reported. In addition, the following performance measures per class are presented: True Positive rate (TP), False Positive rate (FP), Accuracy (ACC) and F1 score. Results are shown in Tables 1-4.

[1] Scientific names are indicated together with a pair (Abbreviation, Number of syllables).

Table 1. Results for feature representations

(a) Standard features

		Predicted											Total	TP	FP	ACC	F1
		GR	HL	MG	MR	PS	PC	TA	TI	TS	XP	ZC					
Actual	GR	21	0	0	3	1	3	0	0	1	4	0	33	63.64	1.60	96.47	66.67
	HL	1	35	8	2	4	0	7	1	2	3	1	64	54.69	4.33	91.26	57.38
	MG	0	6	28	4	10	4	5	2	3	1	3	66	42.42	5.67	88.57	45.16
	MR	1	1	3	39	3	5	0	3	3	0	0	58	67.24	2.98	94.12	69.03
	PS	0	2	4	2	36	1	2	1	3	2	0	53	67.92	4.98	92.61	62.07
	PC	2	1	4	2	0	24	0	0	0	3	0	36	66.67	2.50	95.63	64.86
	TA	0	6	3	0	2	0	21	0	0	0	1	33	63.64	3.02	95.13	59.15
	TI	0	0	1	1	0	0	1	67	3	1	0	74	90.54	2.30	96.81	87.58
	TS	1	5	5	2	4	1	0	4	52	1	3	78	66.67	3.87	92.27	69.33
	XP	2	1	1	0	1	0	0	0	1	39	1	46	84.78	2.91	96.13	77.23
	ZC	2	1	1	0	2	0	2	1	4	1	40	54	74.07	1.66	96.13	77.67
	Total	30	58	58	55	63	38	38	79	72	55	49	595				

Total accuracy = 67.56%

(b) Coarse representation of segment structure

		Predicted											Total	TP	FP	ACC	F1
		GR	HL	MG	MR	PS	PC	TA	TI	TS	XP	ZC					
Actual	GR	27	0	3	0	0	0	0	0	3	0	0	33	81.82	0.71	98.32	84.38
	HL	1	45	2	2	3	0	4	1	2	2	2	64	70.31	2.07	94.96	75.00
	MG	0	1	43	2	2	3	1	0	11	2	1	66	65.15	4.73	91.93	64.18
	MR	0	0	1	41	1	6	0	1	7	1	0	58	70.69	2.98	94.45	71.30
	PS	2	0	6	1	40	0	0	0	0	4	0	53	75.47	2.21	95.80	76.19
	PC	0	0	1	6	0	25	1	0	1	1	1	36	69.44	2.50	95.80	66.67
	TA	0	5	1	0	3	1	22	0	0	0	1	33	66.67	1.78	96.47	67.69
	TI	0	1	2	1	0	0	0	65	3	2	0	74	87.84	0.77	97.82	90.91
	TS	0	2	6	2	0	3	1	2	60	1	1	78	76.92	5.61	92.10	71.86
	XP	0	0	1	1	1	0	0	0	1	42	0	46	91.30	2.37	97.14	83.17
	ZC	1	2	2	1	2	1	3	0	1	0	41	54	75.93	1.11	96.81	81.19
	Total	31	56	68	57	52	39	32	69	89	55	47	595				

Total accuracy = 75.80%

(c) Descriptive features

		Predicted											Total	TP	FP	ACC	F1
		GR	HL	MG	MR	PS	PC	TA	TI	TS	XP	ZC					
Actual	GR	33	0	0	0	0	0	0	0	0	0	0	33	100.00	0.36	99.66	97.06
	HL	0	54	1	0	0	0	7	1	1	0	0	64	84.38	1.88	96.64	84.38
	MG	1	2	58	1	3	0	0	0	1	0	0	66	87.88	1.51	97.31	87.88
	MR	0	0	2	54	0	0	0	1	0	0	1	58	93.10	0.93	98.49	92.31
	PS	0	0	0	0	52	0	0	0	1	0	0	53	98.11	0.74	99.16	95.41
	PC	0	0	0	1	0	34	1	0	0	0	0	36	94.44	0.00	99.66	97.14
	TA	0	5	2	0	0	0	26	0	0	0	0	33	78.79	1.42	97.48	77.61
	TI	0	0	1	0	0	0	0	72	1	0	0	74	97.30	0.38	99.33	97.30
	TS	0	3	1	1	1	0	0	0	70	1	1	78	89.74	0.77	97.98	92.11
	XP	0	0	0	0	0	0	0	0	0	46	0	46	100.00	0.18	99.83	98.92
	ZC	1	0	1	2	0	0	0	0	0	0	50	54	92.59	0.37	98.99	94.34
	Total	32	64	65	59	56	34	34	74	74	47	52	595				

Total accuracy = 92.27%

(d) Mel-cepstrum representation

		Predicted											Total	TP	FP	ACC	F1
		GR	HL	MG	MR	PS	PC	TA	TI	TS	XP	ZC					
Actual	GR	33	0	0	0	0	0	0	0	0	0	0	33	100.00	0.00	100.00	100.00
	HL	0	58	0	0	0	0	5	0	0	1	0	64	90.63	1.69	97.48	88.55
	MG	0	0	66	0	0	0	0	0	0	0	0	66	100.00	0.38	99.66	98.51
	MR	0	1	0	54	1	0	1	0	0	0	1	58	93.10	0.19	99.16	95.58
	PS	0	0	0	0	53	0	0	0	0	0	0	53	100.00	0.37	99.66	98.15
	PC	0	0	0	1	0	34	0	0	1	0	0	36	94.44	0.18	99.50	95.77
	TA	0	5	0	0	0	0	28	0	0	0	0	33	84.85	1.25	97.98	82.35
	TI	0	0	1	0	0	0	1	70	0	2	0	74	94.59	0.38	98.99	95.89
	TS	0	3	1	0	0	1	0	2	69	1	1	78	88.46	0.58	97.98	92.00
	XP	0	0	0	0	1	0	0	0	2	43	0	46	100.00	0.18	99.83	98.92
	ZC	0	0	0	0	0	0	0	0	0	0	54	54	100.00	0.37	99.66	98.18
	Total	33	67	68	55	55	35	35	72	72	47	56	595				

Total accuracy = 94.45%

Table 2. Results for dissimilarity representations computed from 1-D spectra

(a) Spectra computed by FFT

		Predicted											Total	TP	FP	ACC	F1
		GR	HL	MG	MR	PS	PC	TA	TI	TS	XP	ZC					
Actual	GR	20	4	1	0	1	0	1	1	1	0	4	33	60.61	1.42	96.47	65.57
	HL	0	28	6	3	1	4	7	6	3	1	5	64	43.75	7.91	86.89	41.79
	MG	1	5	35	1	7	1	0	7	4	4	1	66	53.03	7.18	88.40	50.36
	MR	0	2	0	33	2	0	1	6	5	0	9	58	56.90	5.03	91.26	55.93
	PS	1	1	3	5	35	1	0	4	2	0	1	53	66.04	3.87	93.45	64.22
	PC	1	6	1	1	0	17	1	1	6	0	2	36	47.22	3.76	93.28	45.95
	TA	1	6	4	0	2	1	13	0	1	2	3	33	39.39	1.78	94.96	46.43
	TI	1	5	7	3	1	3	0	40	10	3	1	74	54.05	7.49	87.73	52.29
	TS	2	5	8	4	5	7	0	10	29	5	3	78	37.18	8.12	84.71	38.93
	XP	0	5	8	0	1	1	0	2	6	21	2	46	45.65	2.73	93.28	51.22
	ZC	1	3	0	10	1	3	0	2	4	0	30	54	55.56	5.73	90.76	52.17
	Total	28	70	73	60	56	38	23	79	71	36	61	595				

Total accuracy = 50.58%

(b) Spectra computed by PSD

		Predicted											Total	TP	FP	ACC	F1
		GR	HL	MG	MR	PS	PC	TA	TI	TS	XP	ZC					
Actual	GR	31	0	1	0	0	0	0	0	0	1	0	33	93.94	0.18	99.50	95.38
	HL	0	48	3	1	1	0	5	0	1	1	4	64	75.00	3.39	94.29	73.85
	MG	0	4	44	1	2	0	1	3	6	5	0	66	66.67	2.65	93.95	70.97
	MR	0	1	0	46	1	2	0	2	5	0	1	58	79.31	0.93	97.14	84.40
	PS	0	1	0	0	41	2	0	1	3	4	1	53	77.36	2.03	96.13	78.10
	PC	0	0	0	0	0	34	1	0	1	0	0	36	94.44	0.89	98.82	90.67
	TA	0	6	2	1	0	0	23	1	0	0	0	33	69.70	1.60	96.81	70.77
	TI	0	0	1	1	1	0	0	71	0	0	0	74	95.95	2.11	97.65	91.03
	TS	0	2	2	1	4	1	0	3	58	7	0	78	74.36	3.29	93.78	75.82
	XP	1	2	5	0	2	0	1	1	1	33	0	46	71.74	3.28	94.79	68.04
	ZC	0	2	0	0	0	0	1	0	0	0	51	54	94.44	1.11	98.49	91.89
	Total	32	66	58	51	52	39	32	82	75	51	57	595				

Total accuracy = 80.67%

Table 3. Results for dissimilarity representations derived from:

(a) SFT

		Predicted											Total	TP	FP	ACC	F1
		GR	HL	MG	MR	PS	PC	TA	TI	TS	XP	ZC					
Actual	GR	33	0	0	0	0	0	0	0	0	0	0	33	100.00	0.00	100.00	100.00
	HL	0	58	1	0	0	0	5	0	0	0	0	64	90.62	1.51	97.65	89.23
	MG	0	2	59	0	0	0	0	1	3	1	0	66	89.39	0.76	98.15	91.47
	MR	0	1	0	52	0	0	1	0	4	0	0	58	89.66	0.00	98.99	94.55
	PS	0	0	0	0	50	0	1	0	1	1	0	53	94.34	0.18	99.33	96.15
	PC	0	0	0	0	0	36	0	0	0	0	0	36	100.00	0.18	99.83	98.63
	TA	0	5	0	0	0	0	26	0	1	1	0	33	78.79	1.60	97.31	76.47
	TI	0	0	1	0	0	0	1	72	0	0	0	74	97.30	0.19	99.50	97.96
	TS	0	0	0	0	0	1	0	0	76	1	0	78	97.44	1.74	98.15	93.25
	XP	0	0	2	0	0	0	0	0	0	44	0	46	95.65	0.73	98.99	93.62
	ZC	0	0	0	0	1	0	1	0	0	0	52	54	96.30	0.00	99.66	98.11
	Total	33	66	63	52	51	37	35	73	85	48	52	595				

Total accuracy = 93.78%

(b) Time-varying PSD

		Predicted											Total	TP	FP	ACC	F1
		GR	HL	MG	MR	PS	PC	TA	TI	TS	XP	ZC					
Actual	GR	33	0	0	0	0	0	0	0	0	0	0	33	100.00	0.00	100.00	100.00
	HL	0	53	2	0	0	0	7	0	2	0	0	64	82.81	3.01	95.46	79.70
	MG	0	1	56	0	0	0	1	0	4	4	0	66	84.85	1.89	96.64	84.85
	MR	0	2	0	47	1	0	2	0	5	0	1	58	81.03	0.37	97.82	87.85
	PS	0	1	1	0	45	0	0	1	3	2	0	53	84.91	0.55	98.15	89.11
	PC	0	1	0	0	0	33	1	0	1	0	0	36	91.67	0.54	98.99	91.67
	TA	0	5	0	0	0	0	27	0	0	1	0	33	81.82	1.96	97.14	76.06
	TI	0	0	2	1	1	1	0	68	1	0	0	74	91.89	0.77	98.32	93.15
	TS	0	1	2	1	0	1	0	3	64	6	0	78	82.05	3.68	94.45	79.50
	XP	0	1	3	0	0	0	0	0	2	40	0	46	86.96	2.37	96.81	80.81
	ZC	0	4	0	0	1	1	0	0	1	0	47	54	87.04	0.18	98.66	92.16
	Total	33	69	66	49	48	36	38	72	83	53	48	595				

Total accuracy = 86.22%

Table 4. Results for dissimilarity representations derived for frame-based:

(a) Descriptive features

		Predicted											Total	TP	FP	ACC	F1
		GR	HL	MG	MR	PS	PC	TA	TI	TS	XP	ZC					
Actual	GR	33	0	0	0	0	0	0	0	0	0	0	33	100.00	0.00	100.00	100.00
	HL	0	56	0	0	0	0	5	0	3	0	0	64	87.50	1.13	97.65	88.89
	MG	0	0	65	0	1	0	0	0	0	0	0	66	98.48	0.57	99.33	97.01
	MR	0	0	0	55	0	0	0	0	0	1	2	58	94.83	0.19	99.33	96.49
	PS	0	0	0	0	53	0	0	0	0	0	0	53	100.00	0.18	99.83	99.07
	PC	0	0	0	0	0	35	0	0	1	0	0	36	97.22	0.00	99.83	98.59
	TA	0	5	0	0	0	0	28	0	0	0	0	33	84.85	0.89	98.32	84.85
	TI	0	0	1	0	0	0	0	72	1	0	0	74	97.30	0.19	99.50	97.96
	TS	0	0	2	0	0	0	0	1	74	0	1	78	94.87	1.35	98.15	93.08
	XP	0	0	0	0	0	0	0	0	0	46	0	46	100.00	0.18	99.83	98.92
	ZC	0	1	0	1	0	0	0	0	2	0	50	54	92.59	0.55	98.82	93.46
	Total	33	62	68	56	54	35	33	73	81	47	53	595				

Total accuracy = 95.29%

(b) Mel-Cepstrum

		Predicted											Total	TP	FP	ACC	F1
		GR	HL	MG	MR	PS	PC	TA	TI	TS	XP	ZC					
Actual	GR	33	0	0	0	0	0	0	0	0	0	0	33	100.00	0.00	100.00	100.00
	HL	0	59	0	0	0	0	5	0	0	0	0	64	92.19	0.94	98.32	92.19
	MG	0	0	66	0	0	0	0	0	0	0	0	66	100.00	0.00	100.00	100.00
	MR	0	0	0	58	0	0	0	0	0	0	0	58	100.00	0.19	99.83	99.15
	PS	0	0	0	0	53	0	0	0	0	0	0	53	100.00	0.00	100.00	100.00
	PC	0	0	0	0	0	36	0	0	0	0	0	36	100.00	0.00	100.00	100.00
	TA	0	5	0	0	0	0	28	0	0	0	0	33	84.85	0.89	98.32	84.85
	TI	0	0	0	1	0	0	0	73	0	0	0	74	98.95	0.00	99.83	99.32
	TS	0	0	0	0	0	0	0	0	78	0	0	78	100.00	0.19	98.83	99.36
	XP	0	0	0	0	0	0	0	0	0	46	0	46	100.00	0.00	100.00	100.00
	ZC	0	0	0	0	0	0	0	0	1	0	53	54	98.15	0.00	99.83	99.07
	Total	33	64	66	59	53	36	33	73	79	46	53	595				

Total accuracy = 97.98%

In order to obtain an overall impression of the one-against-all subproblems, the above-reported confusion matrices were summed across all the representations. As a result, the following observations can be made: In ascending order, the total number of syllables that were erroneously assigned to each class (FP) were: 24 (GR), 59 (PC), 60 (ZC), 74 (MR), 76 (TI), 82 (PS), 85 (XP), 91 (TA), 134 (MG), 148 (HL) and 151 (TS). Similarly, the number of syllables that were erroneously assigned to other classes (FN), in ascending order, were: 33 (GR), 52 (PC), 60 (XP), 70 (TI), 72 (PS), 72 (ZC), 88 (TA), 101 (MR), 140 (MG), 146 (HL) and 150 (TS). In consequence, the easiest identification corresponds to GR and the most difficult ones are HL, MG and MR. Notice also that the most frequent confusions are those between HL and TA.

4 Discussion

In this paper, three approaches for representing bird sounds have been analyzed: 1) feature representations, 2) dissimilarity representations for signals in the frequency domain and 3) dissimilarity representations for time-varying signal transforms. Representations —for each approach— with the highest accuracies were Mel-cepstrum representation (Table 1(d)), dissimilarity representation for PSDs (Table 2(b)) and dissimilarity representations for time-varying Mel-cepstrum (Table 4(b)); respectively. The last one was the representation with the overall highest total accuracy. In general, all representations have good accuracies per class; however, in this case, accuracy is not a reliable performance measure due to the unbalanced nature of the multiclass problem, i.e. the sample size of a class

is much smaller than the combined sample size of the rest of the classes. In this case, the F1 score gives more confident results.

Mel-cepstrum and descriptive features showed a good performance in both feature and dissimilarity representations, as expected because those representations are specifically designed for audio recognition. The dissimilarity representation for PSDs, in spite of being a rather simple representation, yielded an acceptable performance with a total accuracy of 80.67%. Performances of dissimilarity representations for time-varying signal transforms are remarkable. This fact reveals the importance of taking into account the non-stationarity; which is observable by comparing results of dissimilarity representations computed for FFTs, the poorest ones with a total accuracy of 50.58%, and results of the dissimilarity representations for time-varying FFTs (SFTs) that had a total accuracy of 93.78%. In the case of PSDs, the performance also increased when deriving dissimilarities for time-varying transforms but in less proportion, with a total accuracy of 86.22%.

In summary, we conclude that Mel-cepstrum coefficients are suitable for bird sound representation, even more when dissimilarities are computed from them; i.e. when non-stationarity is taken into account. Furthemore, classifying in dissimilarity spaces derived from time-varying representations was found to be preferable instead of doing so in the corresponding 1-D representations.

Acknowledgments. This research is supported by "Programa Jóvenes Investigadores e Innovadores 2010, Convenio Interadministrativo Especial de Cooperación No. 146 de enero 24 de 2011 entre COLCIENCIAS y la Universidad Nacional de Colombia Sede Manizales" and the research program "Fortalecimiento de capacidades conjuntas para el procesamiento y análisis de información ambiental" (code Hermes-12677) funded by Universidad Nacional de Colombia.

References

1. Brenowitz, E., Margoliash, D., Nordeen, K.: An introduction to birdsong and the avian song system. Journal of Neurobiology 33(5), 495–500 (1997)
2. Acevedo, M.A., Corrada-Bravo, C.J., Corrada-Bravo, H., Villanueva-Rivera, L.J., Aide, T.M.: Automated classification of bird and amphibian calls using machine learning: A comparison of methods. Ecological Informatics 4(4), 206–214 (2009)
3. Fagerlund, S.: Bird species recognition using support vector machines. EURASIP Journal on Advances in Signal Processing 2007(1), 64–64 (2007)
4. Chou, C., Liu, P., Cai, B.: On the Studies of Syllable Segmentation and Improving MFCCs for Automatic Birdsong Recognition. In: Asia-Pacific Services Computing Conference, APSCC 2008, pp. 745–750. IEEE (2009)
5. Pękalska, E., Duin, R.P.W., Paclík, P.: Prototype selection for dissimilarity-based classifiers. Pattern Recognition 39(2), 189–208 (2006)
6. Logan, B., Salomon, A.: A music similarity function based on signal analysis. In: IEEE International Conference on Multimedia and Expo, ICME 2001, pp. 745–748 (August 2001)
7. Rubner, Y., Tomasi, C., Guibas, L.: The Earth Mover's Distance as a Metric for Image Retrieval. International Journal of Computer Vision 40(2), 99–121 (2000)