2-1-2010

Acoustic Censusing Using Automatic Vocalization Classification and Identity Recognition

Kuntoro Adi
Santa Dharma University

Michael T. Johnson
Marquette University, michael.johnson@marquette.edu

Tomasz S. Osiejuk
Adam Mickiewicz University

Published version. *Journal of the Acoustical Society of America*, Vol. 127, No. 2 (February 2010): 874-883. DOI.

Acoustic censusing using automatic vocalization classification and identity recognition

Kuntoro Adi
Santa Dharma University, Mrican, Yogyakarta 55002, Indonesia

Michael T. Johnson[a)]
Speech and Signal Processing Laboratory, Marquette University, 1515 West Wisconsin Avenue, Milwaukee, Wisconsin 53201-1881

Tomasz S. Osiejuk
Department of Behavioural Ecology, Adam Mickiewicz University, Umultowska 89, 61–614 Poznań, Poland

(Received 29 June 2009; revised 16 October 2009; accepted 16 November 2009)

This paper presents an advanced method to acoustically assess animal abundance. The framework combines supervised classification (song-type and individual identity recognition), unsupervised classification (individual identity clustering), and the mark-recapture model of abundance estimation. The underlying algorithm is based on clustering using hidden Markov models (HMMs) and Gaussian mixture models (GMMs) similar to methods used in the speech recognition community for tasks such as speaker identification and clustering. Initial experiments using a Norwegian ortolan bunting (*Emberiza hortulana*) data set show the feasibility and effectiveness of the approach. Individually distinct acoustic features have been observed in a wide range of animal species, and this combined with the widespread success of speaker identification and verification methods for human speech suggests that robust automatic identification of individuals from their vocalizations is attainable. Only a few studies, however, have yet attempted to use individual acoustic distinctiveness to directly assess population density and structure. The approach introduced here offers a direct mechanism for using individual vocal variability to create simpler and more accurate population assessment tools in vocally active species.
 [DOI: 10.1121/1.3273887]

PACS number(s): 43.60.Bf, 43.80.Ka [WWA] Pages: 874–883

I. INTRODUCTION

Individually distinct acoustic features have been observed in a wide range of vocally active animal species, for example, cetaceans (Janik, 2000), bats (Masters *et al.*, 1995), and primates (Butynski *et al.*, 1992). There is strong evidence to suggest that individual identification from vocalizations is possible in many species, just as it is in humans, and that many of the state-of-the-art techniques for robust human speaker identification and clustering (Reynolds and Rose, 1995; Tranter and Reynolds, 2006) can be applied equally well to animal vocalizations.

Within birds, the presence of vocal individuality has been shown in the European bitterns (*Botaurus stellaris*) and Black-throated divers (*Gavia arctica*) (Gilbert *et al.*, 1994), American woodcock (*Scolopax minor*) (Beightol and Samuel, 1973), Australian kingfishers (*Dacelo novaeguineae*) (Saunders and Wooler, 1978), and Tawny owls (*Strix aluco*) (Galeotti and Pavan, 1991). Birds use vocal differences to identify other members of their species nearby and to identify individual birds in their immediate vicinity. They have been shown to use vocalizations in recognizing their mates, their parents, and in differentiating between neighbors and strangers (Holschuh, 2004). While a wide variety of approaches has been used to count and monitor bird populations within a species (Peake and McGregor, 2001), most of those approaches do not use individual vocal variability or require the identification of individual birds.

For rare or elusive species that are hard to monitor or to mark visually, the possibility of recognizing individuals by their vocalizations may provide a useful census tool, e.g., Saunders and Wooler, 1978 and Gilbert *et al.*, 1994, but only a few researchers have used individual variation to assess population structure, abundance and density, seasonal distribution and trends, or impact of human-made noise on animals (Mellinger and Barlow, 2003). Peake and McGregor (2001) employed a statistical Pearson-correlation approach to identify corncrake (*Crex crex*) vocal individuality and to estimate numbers of individuals in species. Holschuh (2004) used discriminant function analysis (DFA) to explore vocal individuality of the saw-whet owl (*Aegolius acadicus*) to monitor its habitat quality. Terry and McGregor (2002) suggested neural network models to monitor and census male corncrake species, using a backpropagation and probabilistic network to re-identify the members of the known population (monitoring task) and a Kohonen network to count a population of unknown size. In cetaceans, there are several examples of the use of vocalizations in assessment, including sperm whales (*Physeter macrocephalus*), humpback whales (*Megaptera novaeangliae*), Cuvier's beaked whales (*Ziphius cavirostris*), and harbor porpoises (*Phocoena phocoena*)

[a)]Author to whom correspondence should be addressed. Electronic mail: mike.johnson@marquette.edu

(Barlow and Taylor, 1998; Mellinger and Barlow, 2003; Marques *et al.*, 2009), where combined visual and acoustic methods have significantly improved the population estimate.

Numerous quantitative approaches for analyzing vocal individuality exist. Otter (1996) was able to differentiate individual birds through a series of nested Analysis of Variance (ANOVA) Holschuh (2004) did the same using DFA. In marine mammals, Buck and Tyack (1993) utilized Dynamic Time Warping (DTW) to classify 15 bottlenose dolphin (*Tursiops truncatus*) signature whistles into five groups. The research presented here adapts a well established automatic human speech recognition framework to the task of acoustic censusing, i.e., estimating the abundance of animals in a specified survey area. Previous and current studies show the feasibility of a hidden Markov model (HMM) method to automatically classify ortolan bunting song-types, to identify individual birds (Trawicki *et al.*, 2005; Adi, 2008; Adi *et al.*, 2008), to identify individual African elephants (*Africana Loxodonta*) (Clemins *et al.*, 2005), and to cluster beluga whale (*Delphinapterus leucas*) repertoires (Adi, 2008). The proposed framework is based on HMMs and Gaussian mixture models (GMMs), both commonly used in the speech processing community to perform speech recognition and speaker identification and verification. This approach has advantages in its flexibility and robustness to duration and temporal alignment differences between training and testing examples. An integration of several techniques for supervised and unsupervised classification is proposed and combined with the mark-recapture approach to estimate animal abundance.

Following this introduction, Sec. II gives an overview of the study population, introduces two protocols for estimating the number of animals in a population, and discusses the methods behind the key individual subtasks. Section III presents the experimental results, and Sec. IV finishes with overall conclusions.

II. METHODS

A. Study population: Ortolan bunting (*Emberiza hortulana*)

The ortolan bunting is a migratory passerine bird distributed from Western Europe to Mongolia (Cramp and Perrins, 1994). They winter in Africa. The species inhabits open agricultural areas, raised peat bogs, clear-cut forest on poor sand, and cleared farmland and forest burn (Dale and Hagen, 1997).

Ortolan buntings are classified as an endangered species (Steinberg, 1983; Dale, 2001a) and have shown a major population decline both in individual numbers and in their distribution. In Finland, Vepsalainen *et al.* (2005) studied their population density changes and environment associations in years 1984–2002. They observed a population crash between 1990 and 1993, resulting in a 54% reduction in population density, with a total density reduction of 72% between years of 1984–2002. The Norwegian ortolan bunting, meanwhile, currently numbers approximately 150 singing males and has shown decline over the past 50 years as well. In years 1996–2000 the decline rate was 8% per year (Dale, 2001a). The decline is most likely related to female-biased dispersal away from the population which results in many unpaired males and low population productivity (Dale, 2001b; Steifetten and Dale, 2006). Recently, it was revealed that ortolan bunting males are able to discriminate vocally between neighbors and strangers based on single song derived from the repertoire of a particular male (Skierczyński *et al.*, 2007).

The ortolan bunting vocalizations being examined for this study were collected from County Hedmark, Norway in May of 2001 and 2002 (Osiejuk *et al.*, 2003). The male vocalizations were recorded on 11 out of 25 sites within an area of about 500 km^2. The total number of males in the covered area of the years 2001 and 2002 was about 150. Individual identity was determined for a high percentage of the vocalizations, based on visual observation of individuals using wing markings. For the purposes of this study the 2001 data are used as training data ("marking" data in the mark-recapture protocol) for building song-type and individual identity classification models and for determining repertoire statistics, while the 2002 data are used as test data ("recapture" data) for classification and overall censusing evaluation. It is also used to demonstrate the method of overall abundance estimation using an acoustic mark-recapture model, although because the data were not collected originally for that purpose and does not meet the time and locality guidelines for a mark-recapture study, the mark-recapture model estimates given here are only illustrative.

The ortolan bunting has a relatively simple song and small repertoire size of typically 2–3 song-types for each individual. Individual repertoires are relatively stable but do change some over time. However, in the Norwegian population males use on average 4.2 song-types and have a repertoire size of between 1 and 24 (Łosak, 2007). Song frequencies are in a range between 1.9 and 6.7 kHz. As described by Osiejuk *et al.* (2003), these ortolan vocalizations were recorded between 04:00 and 11:00 am by using an HHB PDR 1000 professional Digital Audio Tape (DAT) recorder. All recordings were transferred to a PC using 48 kHz/16 bit sampling.

The ortolan bunting has a relatively simple repertoire, with songs described by syllable, song-type, and song-variant. A syllable, the minimal unit of song production, is described using letter notation, as illustrated in Fig. 1. Syllables are grouped together into patterns, with each general pattern called a song-type and each unique song called a song-variant. For example, *b* and *c* are examples of syllables, *cb* is an example of a song-type pattern consisting of one or more instances of a *c* syllable followed by one or more instances of a *b* syllable, and *cccccb* is an example of a specific song-variant within the *cb* song-type, consisting of exactly five repetitions of a *c* syllable followed by exactly one repetition of a *b* syllable. Although syllables of the same type differ in length and frequency between individuals, they have the same basic spectrogram shape. Figure 2 illustrates an example of this song-variant for two different individuals in

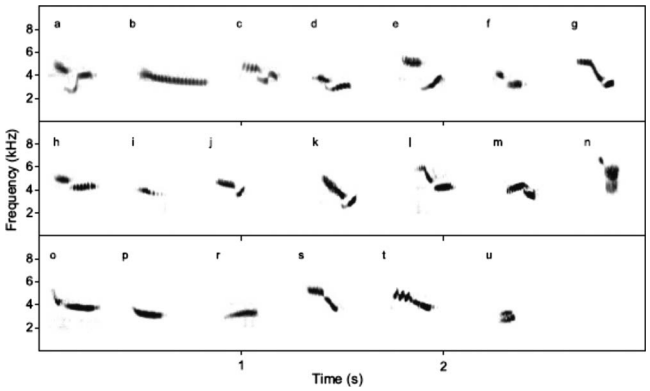

FIG. 1. Ortolan bunting syllables (after Osiejuk *et al.*, 2003).

this data set. In total, the data sets used here include 63 different song-types and 234 different song-variants composed of 20 different syllables.

B. Overview of population assessment protocols

This research proposes two separate population assessment protocols. Both protocols assume single species data, which may require the preprocessing step of species classification (in most cases a somewhat simpler task than song-type and identity classification). In the simpler single-pass protocol illustrated in Fig. 3, a single acoustic data set is used for analysis, with training data limited to enough song-type labeled repertoire examples that classification models can be built, typically five to ten examples of each song-type. In this case song-type classification and individual clustering methods can be used to estimate the total number of individuals within each song-type, and repertoire statistics can then be used to give an overall population estimate within that data set, with confidence intervals.

In the more complex mark-recapture protocol illustrated in Fig. 4, two acoustic data sets are collected in accordance with a closed population mark-recapture survey design. The first set is labeled with song-type and individual identity and used to build song-type and identity classification models.

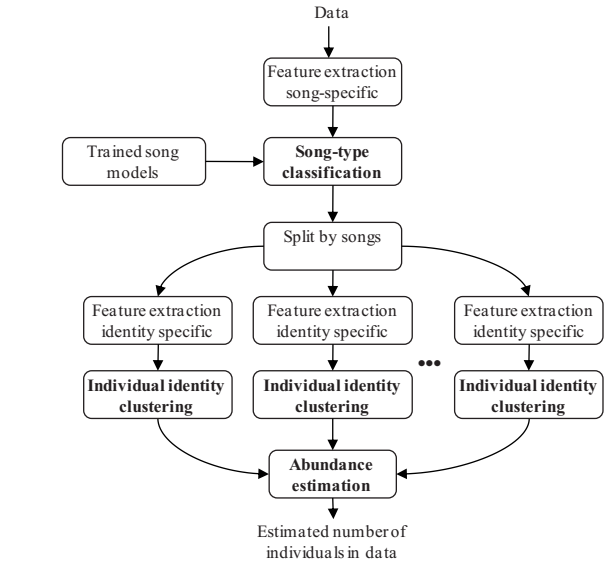

FIG. 3. Single pass protocol, resulting in an estimate of the number of individuals present in the acoustic data set under analysis. Required subtasks include song-type classification (Sec. II C 1), individual clustering (Sec. II C 2), and abundance estimation weighted by repertoire statistics (Sec. II C 3).

The second data set then uses those models to automatically split vocalizations into song-types using song-type classification, cluster individuals within each song type using identity classification, and then match up the individuals in each set to obtain the number of "marked" and "recaptured" animals for total population estimation.

There are four distinct tasks needed to implement these protocols: song-type classification, identity clustering, identity cluster matching, and abundance estimation. Song-type classification trains repertoire models using labeled data and then classifies unknown data. Identity clustering groups data, within one specific song-type, to find the number of clusters representing unique individuals. Identity matching then matches those identity clusters in the unlabeled data to a specific known individual in the labeled training set, using a speaker verification model. The final task is the population

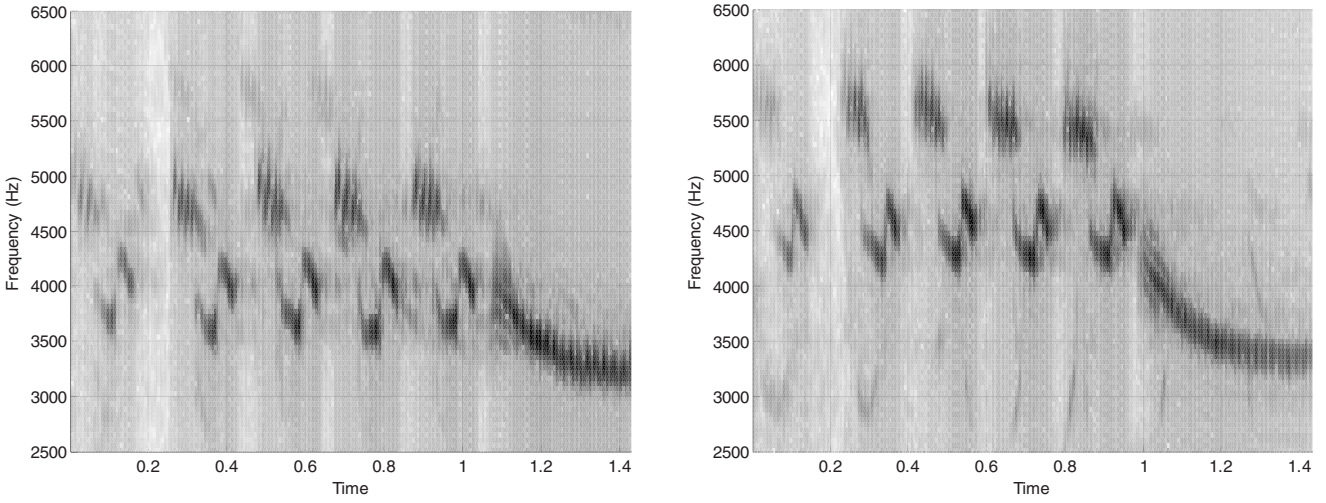

FIG. 2. Examples of song-type *cb*, song-variant *cccccb*, from two different individuals. Note the similarity in basic pattern but difference in timing and mean frequency level.

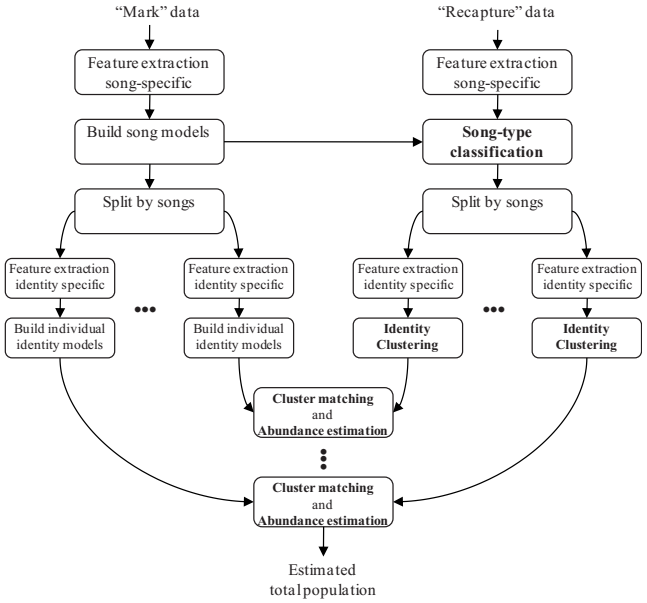

FIG. 4. Mark-recapture protocol, resulting in a net population estimate. The data set on the left represents the "mark" data set used for building both song-type models and individual identity models. The data set on the right represents the "recapture" data set, which is separated into songs using song-type classification (Sec. II C 1), separated by individual using individual clustering (Sec. II C 2), matched across data sets to identify "recaptures" using cluster matching (Sec. II C 3), after which a mark-recapture abundance estimation model (Sec. II C 4) is used to estimate total population.

assessment itself, which is accomplished in the single-pass protocol using a weighted averaging of sub-population estimates within each song-type and is accomplished in the mark-recapture protocol using established maximum-likelihood methods for abundance estimation, based on the estimate of individuals in the two data sets plus how many were present in both.

Section II C discusses in more detail the separate tasks involved in the above scenarios, including song-type classification, individual identity clustering, individual cluster matching, and population size estimation.

C. Individual sub-task methodology

1. Song-type classification using HMMs

Song-type classification is implemented using HMMs (Clemins and Johnson, 2003; Trawicki *et al.*, 2005). Classification features are based on Greenwood function cepstral coefficients (GFCCs) (Clemins, 2005; Clemins and Johnson, 2006), normalized to minimize individual vocal variability. Specifically, a 39-element feature vector is calculated consisting of the 12 GFCCs plus energy, appended with first and second derivatives. The waveforms are first Hamming windowed with frame-size of 3 ms and overlap of 1.5 ms. The Greenwood frequency warping constants are calculated using 26 filter banks spaced across the ortolan bunting hearing range of $f_{\min}$ 400 Hz to $f_{\max}$ 7400 Hz (Edwards, 1943) for each frame.

Song-type classification is implemented using syllables as the base unit. Each syllable is represented by a 15-state HMM to capture the temporal pattern of that syllable. The

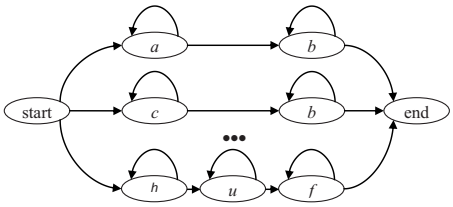

FIG. 5. Song-type classification language model constraints.

corresponding HMMs are connected together for training and recognition using a song-type language model which constrains the output to a valid song-type. Figures 5 and 6 illustrate the language model and waveform-to-HMM matching process.

The HMMs are trained using the Baum–Welch algorithm (Baum *et al.*, 1970), a maximum likelihood estimation (MLE) method based on expectation maximization. Classification on new data is accomplished using the Viterbi algorithm (Forney, 1973) to identify the most likely syllable sequence given the waveform. All HMM tasks for these experiments were implemented using the Cambridge University HMM Toolkit (HTK) version 3.2 (Cambridge University Engineering Department, 2002).

2. Individual identity clustering using GMMs and deltaBIC analysis

The clustering of vocalizations according to individual identity focuses on accurately estimating the number of unique individuals, each of which is represented by a cluster. Because the number of individuals is completely unknown and some individuals may vocalize very few times, this task is much more difficult than that of song-type clustering. The approach used here is based on a method called deltaBIC (BIC denotes Bayesian information criterion) analysis, commonly used in human speech recognition systems for speaker diarization, the task of associating dialog segments in a conversation with specific individuals (Trantor and Reynolds, 2006). DeltaBIC analysis (Ajmera and Wooters, 2003) uses GMMs rather than HMMs as a model for each individual.

The deltaBIC method is based on differential values of the BIC as the number of clusters is increased. The BIC value itself is a similarity measure between two probability

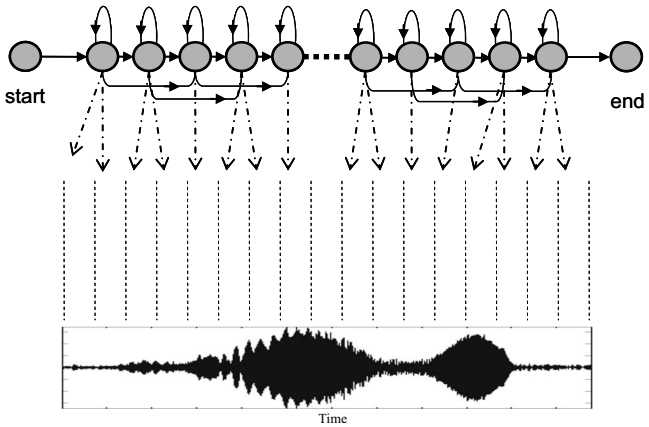

FIG. 6. Illustration of HMM to waveform matching for classification.

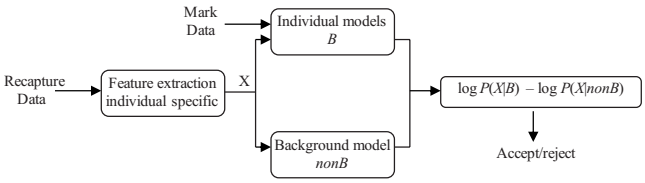

FIG. 7. Individual verification process for individual-to-cluster matching.

density functions, here GMMs. The process starts with over-clustering of the data sets and iteratively merges clusters and retrains a new cluster until no pair of clusters is left with a positive deltaBIC distance measure. The deltaBIC measure is given by

$$\text{deltaBIC} = \sum_{X \in D} \log p(X|\theta) - \sum_{X \in D_1} \log p(X|\theta_1) - \sum_{X \in D_2} \log p(X|\theta_2), \quad (1)$$

where x represents the feature vectors for each frame and D_1, θ_1 and D_2, θ_2 represent the two clusters and cluster parameters being considered for merging. When the clustering process is complete, the remaining number of clusters is used as an estimate of the number of individuals.

Features for individual bird clustering consist of a 39-element GFCC feature vector similar to that used for song-type classification, except implemented without mean and variance normalization in order to preserve individually specific vocal characteristics. The individual bird models are 15 mixture GMMs, implemented in HTK using a single-state HMM with a GMM observation model. The modeling and clustering process is always done on data that consist of a single song-type, i.e., individual identity clustering is always performed after song-type classification or clustering, so that differences in feature characteristics can be reliably associated with individual differences rather than vocalization differences.

3. Identity cluster matching using speaker verification models

To implement an acoustic mark-recapture protocol, it is necessary to match the identity clusters from the recapture data set to known individuals in the marking data set in order to find the overlap, or "recaptures," between the sets. The process is similar to the HMM-based classification as done for song-types, but is implemented using un-normalized GFCC features as with the individual clustering, and also adds a verification step to allow for classifying a vocalization or group of vocalizations as unknown. This is accomplished using a basic likelihood-ratio speaker verification approach, as used in the field of human speaker recognition. The process is implemented separately for each song-type, and then individual results are globally combined. Figure 7 gives an overview of this task.

To implement this, 15-state HMMs (one for each specific song-type) are created for each known individual in the training set. In addition, a verification model called a universal background model is created for each song-type across all individuals in the data set. A likelihood-ratio test is then implemented using all vocalizations in a cluster to discriminate whether the vocalizations in that cluster come from a specific known individual or represent a new unseen individual. The threshold of this accept/reject decision can be varied to control the degree of confidence required to verify that the cluster vocalizations match a known individual.

4. Abundance estimation

a. Single-pass protocol: Data set population estimation using song distribution statistics. The single-pass protocol is able to estimate the number of individuals in the data set under analysis based on a known repertoire distribution for each song-type. Within each song-type, an overall local population estimate is obtained from the estimated number of individuals within that song-type, found through identity clustering, combined with knowledge of how many individuals within the population make that particular song. The final estimated population is then the estimated number of birds within a song type group divided by the percentage of birds that typically use that song type.

These population estimates are then combined using an average or a weighted average according to overall song-type occurrence, leading to a final data set estimate as well as a variance that represents margin of estimation error due to dissimilarity in individual estimates.

If repertoire information regarding the percentage of individuals who make specific song-types is unknown, it is not possible to directly combine the population estimates for each song-type into an overall local population estimate. However, in this case upper and lower bounds on the population can still be established, with a lower bound equal to the maximum number of individuals in any one category (an implicit assumption that all individuals make that song-type) and an upper bound equal to the sum of individuals in each category (an implicit assumption that no individuals make more than one type of song).

b. Abundance estimation using a mark-recapture model. The mark-recapture protocol addresses the bird abundance estimation problem using the MLE framework of a standard mark-recapture model. A two sample mark-recapture involves one session of catching and marking, and another session of recapturing. In the context of this study, catching refers to recording bird vocalizations in an initial session, marking refers to labeling the vocalizations with specific identities, and recapturing refers to recording a second set of vocalizations and acoustically comparing identities.

The process of labeling and recapture or re-labeling involves the tasks of supervised recognition, unsupervised clustering, and identity cluster matching discussed in sections II C I, II C 2, and II C3. The previous steps, therefore, provide the number of individuals (u_1) in one data set, the estimated number of individuals (u_2) in the second data set, and the estimated number of individuals present in both data sets (m_2).

Given the variables u_1, m_2, and u_2, the likelihood of population estimate is computed using the Jolly–Seber (Seber, 1982) equation

TABLE I. Song-type recognition confusion matrix. Rows represent correct song-type categories, and columns represent the classifications made by the algorithm so that entries along the diagonal represent correct classifications (accuracy of 89.6%).

	ab	*c*	*cb*	*cd*	*eb*	*ef*	*gb*	*guf*	*h*	*hb*	*hd*	*huf*	*jd*	*kb*
ab	1561	3	24	1	0	0	1	0	0	5	0	0	0	9
c	0	53	1	11	0	0	1	0	2	0	0	7	0	0
cb	1	2	706	11	0	0	85	1	0	2	0	2	1	1
cd	0	0	9	434	0	0	0	0	0	0	0	0	5	0
eb	1	1	4	0	384	11	1	0	0	0	0	0	0	0
ef	0	0	0	0	0	57	0	0	0	0	0	0	0	1
gb	3	1	19	1	0	0	320	5	0	3	0	0	0	37
guf	0	0	2	0	0	0	1	130	0	0	0	6	0	0
h	0	32	44	0	0	0	3	0	138	27	0	0	0	17
hb	0	0	0	0	0	0	0	0	0	32	0	0	0	1
hd	0	0	0	0	0	0	0	0	0	5	8	3	0	0
huf	0	2	7	41	0	0	0	1	0	2	22	285	0	1
jd	0	0	1	2	0	0	0	2	0	0	0	0	47	0
kb	0	1	0	0	0	0	0	0	0	0	0	0	0	87

$$L(N,p) = \prod_{s=1}^{2} \binom{U_s}{u_s} p^{u_s}(1-p)^{U_s-u_s} \binom{M_s}{m_s} p^{m_s}(1-p)^{M_s-m_s} \tag{2}$$

as a function of the unknown variables N (population) and p (the probability of recapture). The N and p where the likelihood function achieves its maximum value is the maximum likelihood estimate of the population N.

III. EXPERIMENTAL RESULTS

Before system integration, the individual task components are evaluated separately, including song-type classification, individual identity clustering, and identity cluster matching. Following this each of the two protocols is evaluated separately.

A. Evaluation of separate sub-tasks

1. Song-type classification

Speaker independent song-type classification experiments are performed across 14 of the most common song-types. Each song-type contains multiple song-variants. As described previously, the experiment uses the 2001 data set for training and the 2002 data set for evaluation (experiments with those designations reversed yielded similar results).

Results are summarized in the classification confusion matrix in Table I. Individual song-type classification accuracy ranges from a low of 50% (song-type *hd*, which had only 16 examples, 5 of which were confused for type *hb*) to a high of 98.8% (song-type *kb*), with an average overall accuracy of 89.6%. Nearly all errors were made between two-syllable song-types where one syllable was the same and the other was closely related (often syllables *b* and *d*). Full confusion matrices and more detailed analysis of these results are available in Adi (2008). It is worth noting that perfect song-type separation is not necessary to do the larger task of counting individuals. Error in song-type classification tends to lead to an upward bias in the ensuing step of individual identity recognition and clustering. This can be compensated for by downward calibration of the final population estimation methods, although no such adjustment has been made in these experiments.

2. Individual identity clustering

The number of individuals in the test set was estimated using identity clustering via the deltaBIC analysis discussed in Sec. III A 1. Results of the deltaBIC clustering are shown in Fig. 8 (only a subset of the calls are shown for readability), with final estimated number of clusters, true number of individuals, and error shown in Table II. Error ranges from 2.9% to 50% and is notably lower for the more common song-types.

3. Identity cluster matching

An example of identity cluster matching is given in Table III for the *cb* song-type. Identity clusters from the target data are shown in each row, with known individuals from the training data in each column. The number in each cell represents the "acceptance value" from the likelihood-ratio

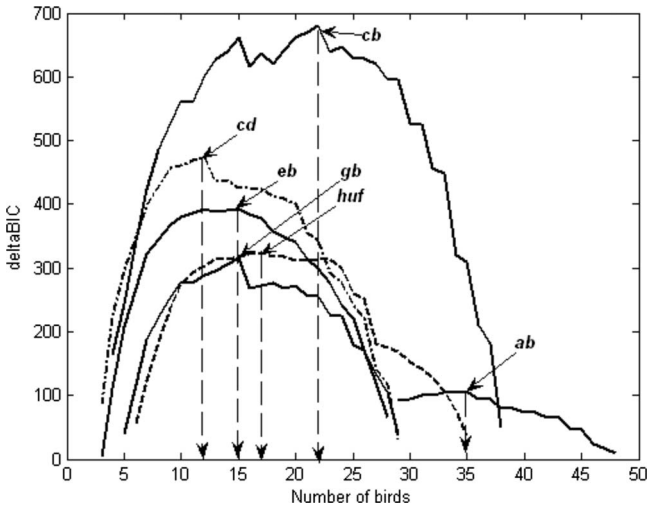

FIG. 8. DeltaBIC analysis curves for selected song-types. The peak value in each curve is used as the estimated number of individuals for that song-type.

TABLE II. Comparison of deltaBIC cluster estimates to known populations.

Song-type	Estimated number of individuals for song	Known number of individuals for song	Percent error
ab	35	34	2.9
cb	22	20	10.0
cd	12	9	33.3
eb	15	12	25.0
ef	3	2	50.0
gb	15	13	15.4
guf	6	4	50.0
huf	17	20	15.0
kb	9	7	28.6

test, with a positive value indicating that the cluster matched more closely to a specific individual than the background model. Here we have arbitrarily chosen a threshold of 1.0 as a cutoff point, which results in a match to two different individuals in the original set, noted b3 and b11. The actual number of individuals present in both the training and test data sets is 3 for a matching error of 33%.

B. Population estimation evaluation

1. Single-pass protocol results

In the single-pass protocol, the goal is to estimate the number of individuals within each song-type, and then use those estimates to arrive at an overall estimate of the population present within the data set. To do this, the calls are first separated by song-type and clustered for individual identity using the deltaBIC approach, and then the number of clusters is used to arrive at a local population estimate.

The song type-specific estimates for the test data were shown previously in Table II. These can be used to project overall local population estimates by dividing the song-type estimate by the percentage of individuals in the population that make each song. These separate estimates can then be combined through an average or weighted average (to emphasize the more frequent song-types which give more accurate estimates), as shown in Table IV.

TABLE IV. Single pass protocol estimates of number of individuals in data set.

Song	Proportion of individuals for each song (%)	Estimated number of individuals for each song	Estimated number of total individuals (true value=81)
ab	51.8	35	67.6
cb	26.8	22	82.1
cd	21.4	12	56.1
eb	12.5	15	120.0
ef	5.3	3	56.6
gb	26.8	15	56.0
guf	8.9	6	67.4
huf	10.7	17	158.9
kb	5.3	9	169.8
Average ± stdev (error)			92.7 ± 45 (14.4%)
Weighted average			79.0 ± 31 (2.4%)

The results indicate that the projected population estimates tend to be biased on the high side, likely due to addi-

TABLE III. Song-type *cb* identity matching results (overlap=2 individuals, *b*3 and *b*11).

		Known individuals from training set												
		b1	b2	b3	b4	b5	b6	b7	b8	b9	b10	b11	b12	b13
Test set cluster	c1	−6.1911	−4.8384	0.2598	−4.8071	−4.8664	−1.7287	−1.4306	−2.2258	−7.4660	−6.6986	−2.0243	−3.5749	−4.4040
	c2	−3.1360	−2.2224	**1.0759**	−3.5349	−3.7813	−2.4249	−1.1016	−2.9164	−5.1645	−3.5034	0.0294	−4.0007	−4.9570
	c3	−1.2931	−4.0089	−0.9428	−7.4533	0.5253	−5.3972	−5.5260	−2.3992	−7.3673	−6.3985	0.7298	−2.5238	−7.1605
	c4	−1.7191	−4.5681	0.1981	−7.0703	0.3497	−5.6525	−5.4115	−3.0178	−6.9223	−5.9974	**1.1603**	−3.5598	−7.2065
	c5	−3.7740	−3.0045	−0.9646	−3.2053	−2.8519	0.1312	−1.1781	−1.4876	−4.4815	−4.4405	−2.4751	−3.5888	−1.6194
	c6	−4.1879	−3.8325	−2.2324	−2.5864	−1.4738	−2.3203	−1.7527	−1.4110	−4.8890	−5.3197	−1.9983	−2.9655	−2.9405
	c7	−4.0524	−3.7983	−0.1179	−1.9972	−3.6972	−3.7014	0.0847	−1.5080	−5.6273	−6.7853	−1.0934	−3.2686	−4.7191
	c8	−1.6285	−1.0683	−2.5166	−3.7007	−1.7817	−2.0835	−2.3154	−1.9211	−4.3459	−4.5966	−0.9438	−3.1081	−2.8231
	c9	0.4203	−3.9631	−3.2250	−6.2160	−0.4092	−4.9076	−3.0624	−3.7383	−5.5757	−4.7141	0.8304	−4.6439	−6.6938
	c10	−6.1358	−4.7771	0.5048	−2.9390	−5.8793	−0.0643	−0.9827	−2.1916	−7.4385	−6.1976	−2.9593	−5.2991	−4.8353
	c11	−6.0416	−5.6546	−4.8061	−1.5755	−1.3735	−3.8146	−1.4443	−1.3929	−6.0052	−7.2179	−4.3271	−1.7333	−4.4404
	c12	−3.1280	−3.3624	−0.9467	−3.5546	−3.8155	−2.6841	−1.1312	−3.4759	−1.5937	−4.5810	−1.0143	−4.8189	−4.4083
	c13	−5.5911	−4.8649	0.2842	−6.4990	−4.7271	−4.2375	−2.6389	−1.0592	−7.7188	−6.7524	−0.0592	−3.8018	−6.1107
	c14	−6.4177	−4.8428	0.1138	−5.4141	−4.8679	−3.0886	−2.9369	−2.7645	−6.2715	−7.2595	−0.6921	−4.7532	−5.7999
	c15	−5.5897	−6.7359	0.1789	−3.8895	−4.4871	−1.3617	−2.4674	−4.0671	−6.9669	−1.7588	−3.1082	−6.2582	−5.5221
	c16	−3.1855	−5.3260	0.6257	−4.6915	−3.3572	−2.0442	−1.1992	−3.6259	−5.3410	−1.8490	−2.0575	−5.4783	−5.9606
	c17	−5.0415	−6.5233	−2.2900	−1.9306	−4.1976	−4.9813	−1.2144	−3.6470	−4.0607	−7.0770	−2.6779	−6.1275	−5.9435
	c18	−5.8263	−6.6272	−4.4404	−2.7344	−0.0812	−2.7804	−2.1755	−0.2400	−7.9769	−7.7126	−4.6130	0.1282	−4.1389
	c19	−1.8148	−1.8384	−1.4369	−3.8289	−1.8039	−3.3647	−1.9384	−3.0285	−4.5124	−3.7390	**1.1541**	−3.6040	−4.5305
	c20	−0.9310	−3.3875	−4.1278	−3.6522	0.4725	−4.5987	−1.8132	−1.5103	−4.6590	−5.4152	−1.8644	−3.8411	−4.7288
	c21	−3.6516	−4.8461	−0.7648	−2.9481	−3.1399	−3.7093	−0.2985	−1.6854	−4.7744	−6.5691	−1.6969	−3.7253	−4.4330
	c22	−4.6355	−4.2748	0.4865	−2.9744	−3.8810	−3.0470	−0.1036	−1.6721	−6.2508	−7.0564	−1.7838	−3.6584	−5.1683

tional non-individual variability present in the separate song-type clusters due to the song-type classification error (which as seen previously is typically in the 5%–15% range). There is also significant variance in the projected estimates, which occurs due to the original error in the per-song estimates combined with errors in the song distribution statistics. The song distribution statistics can have significant impact, especially for infrequently occurring song-types whose population estimates are amplified by a large factor when projecting to an overall estimate. For the data given here, there were, in fact, substantial differences between the training set call distribution used for statistics and the test set used for population estimates. For example, song-type *huf* had a distribution of just over 10% in the first data set, but that increased to nearly 25% in the second set. This resulted in a larger number of individuals in the test set for that song-type, but the results were projected using the original statistics, so that even though the per-call population estimate error was only 15%, the projected population error was over 90%. However, since changes in repertoire statistics by necessity go both ways, with some increasing and others decreasing, this source of error tends to balance itself out.

Weighting the individual population estimates toward the more frequent song-types not only significantly decreases the error, but also tends to lower the variance. It is also possible to compensate for the upward bias with a global adjustment factor (which would likely be species dependent but not data set dependent), but that has not been done here.

2. Mark-recapture protocol illustrative results

In the mark-recapture protocol, illustrated in Fig. 4, song-type classification and speaker identification are performed on the recapture data set to separate the calls by type and remove vocalizations from known individuals in the training set. Following this, identity clustering is performed on the remaining vocalizations and used to determine the total number of new individuals, and the mark-recapture method is used to estimate overall population abundance. Each song may be used to determine an abundance estimate, and the results combined to get a single final estimate.

As discussed previously, the two ortolan data sets used here were not collected in a specific mark-recapture protocol, but instead are representative data sets over two successive breeding seasons across multiple locations, too separated in time and location to be used for a mark-recapture abundance survey. Thus the results given here for this scenario should be considered illustrative of the method rather than true estimates, and for this reason cannot be compared to overall abundance numbers obtained through other survey methods on this population.

The individual per-call population estimates and accuracies, and the overlap determination using the identity matching technique, were presented previously in Tables II and III. From these tables, we can see that for the *cb* song-type, the number of individuals in the training set was 13 and the number of identity clusters over the test set was 22, with two individuals matched between the data sets. Given these numbers, the likelihood of the population estimate is computed using the mark-recapture formula in Eq. (2). The likelihood

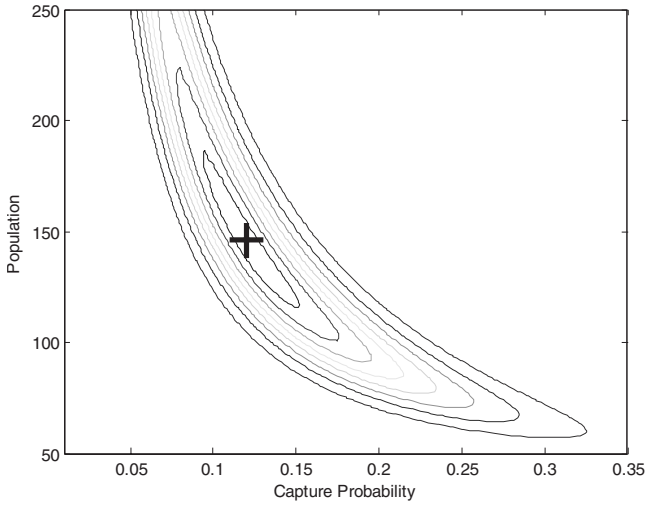

FIG. 9. Likelihood contour plot for song *cb* (ML estimate=146, $p=0.12$).

function, shown in Fig. 9, reaches a maximum value when $p=0.12$ and $N=146$, yielding a total abundance estimate of 146 individuals that use the *cb* song-type. A likelihood confidence interval (95% probability level) is then constructed using the variance of Chapman's modified estimator (Seber, 1982), with a resulting estimate of $\hat{N}=146\pm44$.

Given a properly implemented mark-recapture survey, this approach can be used to give an overall abundance estimate for individuals that make a specific song-type. Repertoire information can then be used to combine these estimates into an overall abundance estimate for the population, as described previously and illustrated in Sec. III B 3.

3. Discussion

The results presented here are intended to illustrate the feasibility of applying advanced methods for vocalization and individual classification to the task of population estimation. It is important to note that the two specific protocols addressed in this work are by no means the only possible approaches to incorporating models of individual vocal variability within an acoustic population assessment design. The concept of "acoustic mark-recapture" can, in fact, be extended to essentially any type of mark-recapture survey, and there are a wide variety of ways in which call-type or identity classification and/or clustering methods can be incorporated into these designs.

For example, within the single-pass protocol of Fig. 3, it was assumed that initial training data were available for the purposes of building the necessary song-type classification models. However, it is possible to implement this step of the process even if substantial training data are not available, using song-type clustering rather than classification as a front-end step. [This has, in fact, been implemented, for details see (Adi, 2008)]. Additionally, within the mark-recapture protocol presented in Fig. 4, it was assumed that individual identity was expertly labeled in the "marking" data set in order to build individual identity models, which negates some of the value of the automated methods, since half of the data still require manual identity analysis. By using identity clustering within both data sets, however, it is

possible to extend the mark-recapture protocol and make it more fully automated. This requires a different mechanism for matching the individuals across the two data sets, which is a focus of continuing work. A third example lies with the way in which repertoire data were used for the purpose of combining abundance estimates within individual song-types into an overall population estimate, using repertoire statistics. While the existence and knowledge of a stable repertoire allow for higher estimation accuracy as well as providing a much more accurate idea of confidence intervals for the estimate, it is not a necessary part of the protocol. In the case of most terrestrial or marine species, the vast majority of which have less sophisticated vocalization repertoires than those of song-birds, it would be sufficient and much simpler to identify one or two basic vocalization types which exist across all or nearly all of the population and to use the resulting population estimates directly.

There are number of key factors that must be considered in generalizing this kind of approach to other species, populations, and habitat regions. Because the technique relies on separation of vocalizations into call-type categories before identity analysis, it is important to have some prior knowledge of the underlying vocalization repertoire of a species as well as the repertoire's consistency across subpopulations and geographic regions. The larger and more complex the repertoire, the more difficult it is to separate call-type differences from individual differences. Species where individuals have relatively large repertoires, with a corresponding fewer number of vocalizations of each type, will thus be more difficult to accurately census. Clearly, in order for acoustic censusing techniques to be feasible, it is necessary that a significant majority of individuals within the species under study vocalize relatively often and loudly enough to allow design of a complete acoustic survey. An equally important but less problematic requirement is that individuals within the species possess sufficient individual vocal variability that they can be separated and identified, which although generally present in a very wide range of species, is certainly not guaranteed, and has not yet been carefully studied outside of terrestrial and marine mammals and songbirds (for example, in insects). This individual distinctiveness must be measurable and methods must be robust to the presence of environmental and microphone channel noise, an area which is also the focus of much research for human speaker identification and verification tasks.

There are a number of parameter settings within the classification and particularly clustering methods that may have impact on the accuracy of results. The most substantial of these currently is the threshold for matching the individuals within the marking data set to those within the recapture data set, because the number of overlapping individuals is the single biggest factor in resulting population estimate.

Overall, though, the promising initial results presented here suggest that there is great potential for the inclusion of individual vocal variability analysis into population assessment designs. This work has significant potential for further extension as well, including, in particular, application to distance sampling by using a point or line transect protocol for recording the vocalizations combined with source localization methods for determining distances to the survey line, which could yield accurate estimates of population density. In general, the underlying approach introduced here creates a means for integrating acoustics and individual vocal variability into many different tools for population assessment.

IV. CONCLUSIONS

The results from this work strongly indicate that the individual distinctiveness of vocalizations can be used to accurately estimate abundance within a data set as well as to match individuals across data sets, and furthermore that this methodology can be incorporated into a larger mark-recapture survey design for overall population assessment. We have illustrated this idea here for the ortolan bunting.

It is likely that nearly all vocally active animals have individually distinct vocalization characteristics, as has already been observed across many different species. The framework for abundance estimation presented in this paper is thus applicable to any vocally active species with a distinct vocal repertoire. This approach addresses the problem of population assessment in a new way, employing algorithms for automatic human speech and speaker recognition to estimate animal abundance. The method has advantages over physical marking techniques, as it is less invasive and is more cost and labor effective. It also has the potential to be more autonomous than current approaches to acoustic or visual surveying, since once basic repertoire models have been built local population can be estimated from any continuous recording. In general, if obtained under the larger umbrella of a well-designed mark-recapture or distance sampling survey protocol, this mechanism may allow for a substantially more accurate understanding of overall population structure and abundance on a larger scale.

ACKNOWLEDGMENT

The authors would like to thank the National Science Foundation (Grant No. IIS-0326395 "The Dr. Dolittle Project") for supporting this work.

Adi, K. (**2008**). "Hidden Markov model based animal acoustic censusing: Learning from speech technology," Ph.D. thesis, Marquette University, Milwaukee, WI.

Adi, K., Sonstrom, K., Scheifele, P., and Johnson, M. T. (**2008**). "Unsupervised validity measures for vocalization clustering," in ICASSP, Las Vegas, NV.

Ajmera, J., and Wooters, C. (**2003**). "A robust speaker clustering algorithm," in Automatic Speech Recognition and Understanding Workshop, St. Thomas, VI.

Barlow, J., and Taylor, B. L. (**1998**). "Preliminary abundance of sperm whales in the northeastern temperate pacific estimate from a combined visual and acoustic survey," International Whaling Commission Working Paper No. SC/50/CAWS20.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (**1970**). "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," Ann. Math. Stat. **41**, 164–171.

Beightol, D. R., and Samuel, D. E. (**1973**). "Sonographic analysis of the American woodcock's peent call," J. Wildl. Manage. **37**, 470–475.

Buck, J. R., and Tyack, P. L. (**1993**). "A quantitative measure of similarity for *tursiops truncatus* signature whistles," J. Acoust. Soc. Am. **94**, 2497–2506.

Butynski, T. M., Chapman, C. A., Chapman, L. J., and Weary, D. M. (**1992**). "Use of male blue monkey 'pyow' calls for long-term individual identification," Am. J. Primatol **28**, 183–189.

Cambridge University Engineering Department (**2002**). Hidden Markov Model Toolkit (HTK) Version 3.2.1 User's Guide, Cambridge, MA.
Clemins, P. J. (**2005**). "Automatic speaker identification and classification of animal vocalizations," Ph.D. thesis, Marquette University, Milwaukee, WI.
Clemins, P., and Johnson, M. T. (**2006**). "Generalized perceptual linear prediction (gPLP) features for animal vocalization analysis," J. Acoust. Soc. Am. **120**, 527–534.
Clemins, P. J., and Johnson, M. T. (**2003**). "Application of speech recognition to African elephant (*Loxodonta africana*) vocalizations," in 2003 International Conference on Acoustics, Speech, and Signal Processing, Hong Kong.
Clemins, P. J., Johnson, M. T., Leong, K. M., and Savage, A. (**2005**). "Automatic classification and speaker identification of African elephant (*Loxodonta africana*) vocalizations," J. Acoust. Soc. Am. **117**, 956–963.
Cramp, S., and Perrins, C. M. (**1994**). *The Birds of the Western Palearctic* (Oxford University Press, Oxford).
Dale, S. (**2001a**). "Causes of population decline in ortolan bunting in Norway," in Proceedings in the Third International Ortolan Symposium, pp. 33–41.
Dale, S. (**2001b**). "Female-biased dispersal, low female recruitment, unpaired males, and the extinction of small and isolated bird populations," Oikos **92**, 344–356.
Dale, S., and Hagen, O. (**1997**). "Population size, distribution, and habitat choice of the ortolan bunting *Emberiza hortulana* in Norway," Fauna Norv. Ser. C, Cinclus **20**, 93–103.
Edwards, E. P. (**1943**). "Hearing ranges of four species of birds," Auk **60**, 239–241.
Forney, G. D. (**1973**). "The Viterbi algorithm," Proc. IEEE **61**, 268–278.
Galeotti, P., and Pavan, G. (**1991**). "Individual recognition of male tawny owls (*Strix aluco*) using spectrograms of their territorial calls," Ethol. Ecol. Evol. **3**, 113–126.
Gilbert, G., McGregor, P. K., and Tyler, G. (**1994**). "Vocal individuality as a census tool: Practical considerations illustrated by a study of two rare species," J. Field Ornithol. **65**, 335–348.
Holschuh, C. (**2004**). "Monitoring habitat quality and condition of Queen Charlotte saw-whet owls (*Aegolius Acadicus Brooksi*) using vocal individuality," University of Northern British Columbia.
Janik, V. M. (**2000**). "Whistle matching in wild bottlenose dolphins (Tursiops truncatus)," Science **289**, 1355–1357.
Łosak, K. (**2007**). "A comparative analysis of song variation in ortolan bunting (Emberiza hortulana) from populations of different status and quality," Ph.D. thesis, Adam Mickiewicz University, Poznan, Poland.
Marques, T., Thomas, L., Ward, J., DiMarzio, N., and Tyack, P. (**2009**). "Estimating cetacean population density using fixed passive acoustic sensors: An example with Blainville's beaked whales," J. Acoust. Soc. Am. **125**, 1982–1994.
Masters, W. M., Raver, K. A., and Kazial, K. A. (**1995**). "Sonar signals of big brown bats, *Eptesicus fuscus*, contain information about individual identity, age, and family affiliation," Anim. Behav. **50**, 1243–1260.
Mellinger, D. K., and Barlow, J. (**2003**). "Future directions for acoustic marine mammal surveys: Stock assessment and habitat use," NOAA OAR Special Report No. NOAA/PMEL 2557, NOAA, La Jolla, CA.
Osiejuk, T. S., Ratynska, K., Cygan, J. P., and Dale, S. (**2003**). "Song structure and repertoire variation in ortolan bunting (Emberiza hortulana L.) from isolated Norwegian population," Ann. Zool. Fenn. **40**, 3–19.
Otter, K. (**1996**). "Individual variation in the advertisement call of male saw-whet owls," J. Field Ornithol. **67**, 398–405.
Peake, T. M., and McGregor, P. K. (**2001**). "Corncrake Crex crex census estimates: A conservation application of vocal individuality," Animal Biodiversity and Conservation **24**, 81–90.
Reynolds, D. A., and Rose, R. C. (**1995**). "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech Audio Process. **3**, 72–83.
Saunders, D. A., and Wooler, R. D. (**1978**). "Consistent individuality of voice in birds as management tool," Emu **88**, 25–32.
Seber, G. A. F. (**1982**). *The Estimation of Animal Abundance and Related Parameters* (MacMillan, New York).
Skierczynski, M., Czarnecka, K. M., and Osiejuk, T. S. (**2007**). "Neighbor-stranger song discrimination in territorial ortolan bunting *Emberiza hortulana* males," J. Avian Biol. **38**, 415–420.
Steifetten, O., and Dale, S. (**2006**). "Viability of endangered populations of ortolan buntings: The effect of a skewed sex ratio," Biol. Conserv. **132**, 88–97.
Steinberg, B. D. (**1983**). *Microwave Imaging with Large Antenna Arrays* (Wiley, New York).
Terry, A. M., and McGregor, P. K. (**2002**). "Census and monitoring based on individually identifiable vocalizations: The role of neural networks," Animal Conservation **5**, 103–111.
Tranter, S. E., and Reynolds, D. A. (**2006**). "An overview of automatic speaker diarization systems," IEEE Trans. Audio, Speech, Lang. Process. **14**, 1557–1565.
Trawicki, M. B., Johnson, M. T., and Osiejuk, T. S. (**2005**). "Automatic song-type classification and speaker identification of Norwegian ortolan bunting (Emberiza hortulana)," in IEEE International Conference on Machine Learning in Signal Processing (MLSP).
Vepsalainen, V., Pakkala, T., Piha, M., and Tiainen, J. (**2005**). "Population crash of the ortolan bunting Emberiza hortulana in agricultural landscapes of Southern Finland," Ann. Zool. Fenn. **42**, 91–107.