Evaluation of a Sparse Representation-Based Classifier For Bird Phrase Classification Under Limited Data Conditions

Lee Ngee Tan[1], *Kantapon Kaewtip*[1], *Martin L. Cody*[2],*Charles E. Taylor*[2], *Abeer Alwan*[1]

[1]Department of Electrical Engineering, [2]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, California, USA

tleengee@ee.ucla.edu, kkaewtip@ucla.edu, mlcody @ucla.edu, taylor@biology.ucla.edu, alwan@ee.ucla.edu

Abstract

This paper evaluates the performance of a sparse representation-based (SR) classifier for a limited data, bird phrase classification task. The evaluation database contains 32 unique phrases segmented from songs of the Cassin's Vireo (*Vireo cassinii*). Spectrographic features were extracted from each phrase-segmented audio file, followed by dimension reduction using principal component analysis (PCA). A performance comparison to the nearest subspace (NS) and support vector machine (SVM) classifiers was conducted. The SR classifier outperforms the NS and SVM classifiers, with a maximum absolute improvement of 3.4% observed when there are only four tokens per phrase in the training set.

Index Terms: bird phrase classification, limited data, sparse representation, L_1 minimization.

1. Introduction

Automated recognition of bird sounds is desirable, among other reasons, for measuring biodiversity from sound recordings [1] and for studying behavior of vocalizing species [2]. These methods will gain importance as more attention is directed toward "soundscape ecology" generally [3] and a more refined understanding of bird communication, specifically. Automated bird song classification has already found use for species identification [4–8], individual recognition [9], and classification of particular syllables or phrases expressed by birds with complex vocabularies [10–12]. Much of this research has been reviewed by [1] and [2].

Classifying particular bird calls or song elements becomes especially challenging in those cases where the song repertoire is diverse, comprising large numbers of variant syllables or phrases; for example some species have been observed with thousands of distinct phrases in their lexicons [13]. In our experience the frequencies at which individual bird song elements are observed often resembles a "Zipf curve", where a few phrases are heard many times, but the majority of phrases are rare. Communication in other species, including humans, typically follows this same relation [14]. A premium is thus placed on the ability of automated classifiers to correctly classify bird song phrases based on limited training data.

The amount of data available to train models for phrase classification can be limited due to the opportunistic nature of bird vocalization collection in geographical locations of interest. Although Hidden Markov Models (HMMs) [10] and neural networks [11] have been shown to work well for acoustic unit recognition in bird vocalization applications, they typically require a substantial amount of data to train their parameters. In this study, a sparse representation-based (SR) classifier is used for a limited data, bird phrase classification task. The work is inspired by high recognition accuracies for a 100-subject face recognition task, with only 7 training images per subject [15]. The SR classifier seeks to represent the test vector by a sparse linear combination of training vectors, which is found by solving a L_1 minimization problem. Since it is a relatively new classification methodology, this SR classifier has not been used in bird vocalization recognition tasks. This paper evaluates the performance of the SR classifier for bird phrase classification using features derived from spectrographic images. Its performance is compared to two other classifiers: the nearest subspace (NS) classifier [15, 16], and the support vector machine (SVM) classifier [17, 18], which have also demonstrated good classification accuracies with limited training data.

2. Sound Data

Song fragments (phrases) for classification were obtained from recordings of Cassin's Vireo (*Vireo cassinii*). Only the males of this species give full songs. The song has been described as "... a jerky series of burry phrases, separated by pauses of 1 s. Each phrase is made up of 2 to 4 notes [syllables], with song often alternating between ascending and descending phrases ..." The "song [is] repeated tirelessly, particularly when [the singing male is] unpaired" [19].

All recordings were obtained between 23 April and 8 June, 2010 near Volcano, California (38˚29′04"N, 120˚38′04"W) in a mixed conifer-oak forest at approximately 800 m elevation. Two males on two different territories (approximately 200 m apart) were recorded. Phrase repertoires of the two males were similar, though not identical. Songs were recorded in 13 bouts in separate WAV files (16-bit, mono) at a sampling rate of 44.1 kHz, using a Marantz PMD 670 with a Telinga parabolic reflector and a Sennheiser omni-directional microphone. Manual annotation was performed using Praat (http://www.fon.hum.uva.nl/praat) to note the phrase class, and the start and end times of each phrase in the song. The phrases were categorized into one of the 65 phrase classes based on both visual examination of their spectrograms and auditory recognition. Each phrase segment or token is extracted from the original WAV file based on its start and end times, and saved in a separate WAV file. The number of tokens in each phrase class ranges from 1 to 69, and only 32 phrase classes have at least 10 tokens, while 29 of the remaining 33 classes have four tokens or less. In this paper, classification experiments were performed using data from the 32 phrase classes only, which contains a total of 1022 tokens. The 1022 tokens made up 6.3 min of audio data, with each

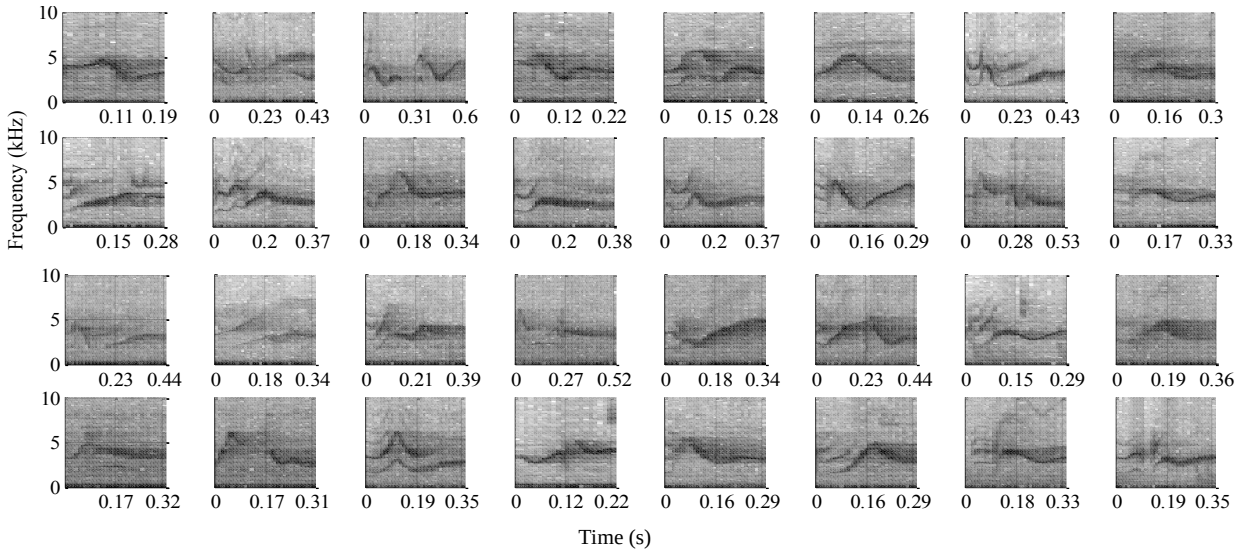

Fig. 1. Spectrograms of the 32 phrase classes from our Cassin's Vireo database

token between 0.17 to 0.97 s long. Fig. 1 shows the example spectrograms of these 32 phrase classes. The recordings and annotations for this study are available online at http://taylor0.biology.ucla.edu/al/bioacoustics/.

3. Feature Extraction

3.1. Spectrographic feature computation

Since spectrograms contain discriminative information that aids phrase annotation, we derived the features for this bird phrase classification task explicitly from phrase-segmented spectrograms. Although spectrographic features are generally not noise-robust without additional signal processing, they yield reasonable classification accuracies for our task because the majority of segmented phrase tokens in our database has a high signal-to-noise ratio (SNR). The sampling rate was first reduced to 20 kHz because little energy is found above 10 kHz. Since every phrase token has a variable file duration, to generate a spectrographic feature vector of the same dimension for each token, a file-duration-dependent frame shift is used to compute its spectrogram. This file-duration-dependent frame shift is calculated as shown in Eq. (1), to ensure that the spectrogram of each file always contains 64 frames in time. The *round*(.) operator rounds off the input argument to the nearest integer. The starting sample index for frame t is denoted by S_t , while D and W denote file duration and frame length in number of samples, respectively. This translates to a frame shift of between 3 to 16 ms for the phrase durations present in the database. A frame length of 20 ms is used, thus W = 400 samples.

$$S_t = round\left(1+\frac{D-W}{63}t\right), \quad t = 0,1,...,63. \tag{1}$$

A 512-point FFT is computed at each frame, and the values are converted to decibels (dB) units. It was observed that most of the bird phrase acoustic energy fall within 1.5 and 6.5 kHz, hence only the 128 FFT bins corresponding to this frequency range in the spectrogram are retained. Next, this 128-by-64 spectrographic image is normalized so that it takes values between 0 and 255 (inclusive), as shown in Eq. (2), where $X_{norm}(k,t)$ and $X(k,t)$ are the post- and pre-normalized pixel values, respectively, with frequency bin index k, and frame index t. $X_{\min}$ and $X_{\max}$ are the corresponding minimum and maximum values of the pre-normalized spectrogram.

$$X_{norm}(k,t) = \frac{255\left(X(k,t)-X_{\min}\right)}{X_{\max}-X_{\min}} \tag{2}$$

Finally, the normalized spectrographic image is reshaped into a 8192-by-1 feature vector by concatenating the column vectors in $X_{norm}(k,t)$.

3.2. Dimension reduction via PCA

The dimension of the feature vector is then reduced by doing a principal component analysis (PCA) on the training set. Mean subtraction is not performed before computing the eigenvectors because this yields higher phrase classification accuracies for all classifiers evaluated. For our performance evaluation, the feature dimension, d is reduced to 32, 50 and 128, corresponding to image resizing factors of 1/16, 1/12 and 1/8, respectively. This is to investigate if the performance of the classifiers would exhibit different trends when d is varied. The dimensionally reduced feature vectors are subsequently normalized to unit length.

4. SR, NS, and SVM Classifiers

The following sections describe the SR, NS, and SVM classifiers implementation used in our experiments.

4.1. Sparse representation-based (SR) classifier

The sparse representation-based classifier performs phrase classification through representing the test feature vector, b, by a sparse linear combination of feature vectors or exemplars present in the training set, as shown in Fig. 2. This sparse linear combination can be found by solving for a sparse vector x via the L_1 minimization convex optimization problem defined in Eq. (3), where each column in matrix A, contains one exemplar (corresponding to one token) from the training set. After the sparse representation is found, the residual vector, r_i between the original test feature vector and various vectors, each reconstructed using the sparse training exemplars

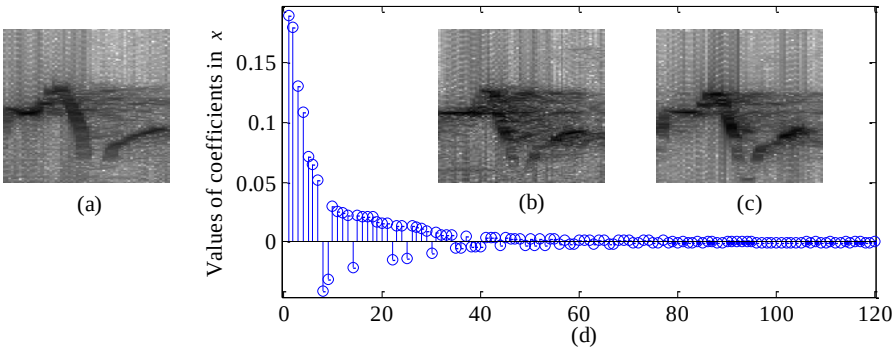

Fig. 2. Testing phrase token expressed by a sparse linear combination of training tokens. (a) Spectrogram of the test phrase token. (b) and (c) spectrograms of the two training phrase tokens from the same class as (a) that respectively correspond to the first and second largest coefficients in the sparse vector x plotted in (d). The spectrograms shown are the cropped and normalized spectrographic features prior to dimension reduction. The coefficients of x in (d) had been sorted in order of decreasing absolute value.

from a specific phrase class i, are computed. The class that yields the r_i with the smallest norm is the output of the SR classifier, O_{SR}. This classification algorithm, which is summarized in Eqs. (3) – (5), follows "Algorithm 1" described in [15] for their face recognition application. The L_1-MAGIC MATLAB toolbox [20] is the L_1 solver used to perform the L_1 minimization in Eq. (3), with $\varepsilon = 0.05$. The $\delta_i(x)$ function zeros the non-zero coefficients in x corresponding to training exemplars that do not belonging to class i.

$$\min \|x\|_1 \quad \text{subject to} \quad \|Ax - b\|_2 \le \varepsilon \tag{3}$$

$$r_i = b - A\delta_i(x) \quad \text{for} \quad i = 1, \ldots, 32. \tag{4}$$

$$O_{SR} = \arg\min_i \|r_i\|_2 \tag{5}$$

4.2. Nearest subspace (NS) classifier

The classical nearest subspace classifier [16] works by finding the class subspace that best represents the test vector, b, as described in Eqs. (6) – (8). First, the orthonormal basis of each phrase class is found by performing singular value decomposition (SVD) on a matrix A_i. Each column in A_i contains one feature vector from class i in the training set. The first n left singular vectors, $u_{1, \ldots, n}$ in the orthogonal matrix U_i form the orthonormal basis, P_i for the subspace of class i. The value of n is the number of training tokens in that class, since n is always less than the dimension of the feature vector in our limited training data experimental setup. The norm of the P_i subspace-projected b is used as a similarity measure between b and class i. The output of the NS classifier, O_{NS} is the class that yields the maximum similarity value, as shown in Eq. (8).

$$\text{SVD:} \quad A_i = U_i \Sigma_i V_i^T \tag{6}$$

$$P_i = \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix} \tag{7}$$

$$O_{NS} = \arg\max_i \left\| P_i^T b \right\|_2 \tag{8}$$

4.3. Support vector machine (SVM) classifier

The multi-class SVM classifier is implemented using the LIBSVM [21], which uses a one-against-one decomposition strategy. The selected kernel function is the Gaussian radial basis function (RBF). The classifier for each training set is trained using a five-fold cross-validation to search for an optimal pair of regularization factor (a.k.a soft margin parameter), C and the RBF parameter, γ. The search points of C and γ used for cross- validation are $[2^{-1}, 2^0, \ldots, 2^7]$ and $[2^{-4}, 2^{-3}, \ldots, 2^5]$, respectively.

5. Results and Discussions

To evaluate the performance of the classifiers under limited data conditions, we varied the amount of data for training, and ran experiments with n = 4 to 7 training tokens from each of the 32 classes, while all remaining $1022 - 32n$ tokens are used for testing purposes. The phrase classification accuracy is defined as the percentage of the total number of test tokens that are classified correctly. For each case (corresponding to a particular # of training token, n and feature dimension, d), the results tabulated in Table 1 were obtained from averaging five independent experimental results in which training tokens were randomly selected. In general, the phrase classification accuracies improve with increasing n and d for all three classifiers.

Table 1. *Phrase classification accuracies (%) using varying number of training token per phrase class, n and feature dimension, d. The highest value for each test case is boldfaced.*

n	Classifier	d		
		32	50	128
4	SR	**85.24**	**85.75**	**86.85**
	NS	81.77	82.31	83.76
	SVM	82.41	82.19	83.24
5	SR	**86.68**	**87.70**	**88.24**
	NS	84.18	84.99	85.78
	SVM	84.64	84.66	85.50
6	SR	**88.07**	**88.55**	**89.25**
	NS	86.22	86.75	87.66
	SVM	86.70	87.32	87.30
7	SR	**88.65**	**89.57**	**90.06**
	NS	87.12	87.82	88.65
	SVM	87.54	88.07	87.90

The bird phrase classification accuracies achieved by the SR classifier are consistently the highest compared to the NS and SVM classifiers in all cases. The McNemar test [22] was performed to evaluate the statistical significance of SR classifier's performance against the NS and SVM classifiers. A p-value of less than 0.03 is obtained for all the cases tabulated above, indicating that the improvement in bird phrase classification accuracies of the SR classifier over the other two classifiers is of statistical significance. The

performance gain of the SR classifier becomes more significant as n decreases, which implies that the performance of the SR classifier is less dependent on the amount of training data compared to the NS and SVM classifiers. One reason for the good performance of the SR classifier is that spectrograms of phrases from the same class are generally very similar to one another for this database, except for time differences in the silence interval before and after a phrase onset and offset due to variations in specifying phrase boundaries during the manual annotation process.

The SR classifier tends to yield good performance over a wider range of conditions compared to the nearest neighbor and NS classifiers [15], because the L_1 minimization algorithm selects the smallest subset of training exemplars that best represents the test vector, so it is not dependent solely on the nearest exemplar, nor on the subspace spanned by all the training exemplars of a class. However, this also means that the SR classifier would be more sensitive to classification errors due to outliers or mislabeled exemplars in the training set, since the SR classifier might misclassify a test vector that is in close proximity to a single or a neighborhood of outlier(s) or mislabeled exemplar(s). For the SVM classifier, if the outliers are not one of the support vectors, its decision boundary would not be affected by these outliers. On the other hand, when the classes in training data are non-separable by the SVM, the SR classifier might have an advantage over the SVM classifier, depending on similarity between the test vector and the training exemplars from the same class. This is the most likely scenario that led to the SR classifier's superior performance over the SVM's.

6. Conclusion and Future Work

The SR classifier that had reported good face recognition performance was applied to limited data, bird phrase classification. On our 32-class, Cassin's Vireo phrase database, the SR classifier obtains the highest classification accuracies compared to NS and SVM classifiers. An increasing performance gain over the latter two classifiers was observed with a decreasing number of tokens per phrase used for training. This suggests that the SR classifier is a promising technique for bird phrase classification when the amount of training data is limited.

In this paper, bird vocalization recordings with high SNRs were used to focus on the classifiers' performance evaluation under limited data conditions. However, field recordings tend to contain overlapping bird songs and other acoustic interferences, e.g. wind, rustling leaves, etc. Since the classification accuracy of the SR classifier is highly dependent on the similarity between the training and test sets from the same class, denoising and robust feature extraction will be important to yield similar good performance with the SR classifier when we proceed to a more challenging database.

7. Acknowledgements

This study was supported in part by National Science Foundation Award No. 0410438. We thank George Kossan for his assistance with phrase identification.

8. References

[1] Brandes, T. S., "Automated sound recording and analysis techniques for bird surveys and conservation," *Bird Conservation International*, vol. 18 , pp. S163–S173, 2008.

[2] Mennill, D. J., "Individual distinctiveness in avian vocalizations and the spatial monitoring of behavior," *Ibis*, vol. 153, pp. 235–238, 2011.

[3] Pijanowski, B. C., Villanueva-Rivera, L. J., Dumyahn, S. L., Farina, A., Krause, B. L., Napoletano, B. M., Gage, S. H., and Pieretti, N., "Soundscape Ecology: The Science of Sound in the Landscape," *BioScience*, vol. 61, pp. 203-216, 2011.

[4] Härmä, A., "Automatic recognition of bird species based on sinusoidal modeling of syllables," *IEEE Int. Conf. on Acoustic, Speech and Signal Processing (ICASSP)*, pp. 545–548, 2003.

[5] Somervuo, P. and Härmä, A., "Bird song recognition based on syllable pair histograms," in *IEEE ICASSP*, pp. 825–828, 2004.

[6] Chu, W., Blumstein, D. T., "Noise robust bird song detection using syllable pattern-based hidden Markov models," *IEEE ICASSP*, pp. 345–348, 2011.

[7] Trifa, V. M., Krischel, A. N. G., and Taylor, C. E., "Automated species recognition of antbirds in a Mexican rainforest using hidden Markov models," *J. Acoust. Soc. of Am.* (*JASA*), vol. 123, 2424–2431, 2008.

[8] Chen, Z., and Maher, R. C., "Semi-automatic classification of bird vocalizations using spectral peak tracks," *JASA*, vol. 120, pp. 2974 –2984, 2006.

[9] Kirschel, A. N. G., Cody, M. L., Harlow, Z. T., Promponas, V. J., Vallejo, E. E. and Taylor, C. E., "Territorial dynamics of Mexican Ant-thrushes *Formicarius moniliger* revealed by individual recognition of their songs," *Ibis*, vol. 153, pp. 255–268, 2011.

[10] Kogan, J. A., and Margoliash, D., "Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study," *JASA*, vol. 103, pp. 2185–2196, 1998.

[11] Ranjard, L., and Ross, H. A., "Unsupervised bird song syllable classification using evolving neural networks," *JASA*, vol. 123, pp. 4358–4368, 2008.

[12] Sasahara, K., Cody, M. L., Cohen, D., and Taylor, C. E., "Structural Design Principles of Complex Bird Songs: A Network-Based Approach," submitted.

[13] Catchpole, C. K., and Slater, P. J. B., *Bird Song: Biological Themes and Variations*. Cambridge, UK: Cambridge University Press, 2008.

[14] McCowan, B., Doyle, L. R., Jenkins, J., and Hanser, S. F., "The appropriate use of Zipf's law in animal communication studies," *Animal Behaviour*, vol. 69, pp. F1-F7, 2005.

[15] Yang A. Y., Wright J., Ma Y., and Sastry, S., "Feature selection in face recognition: A sparse representation perspective," UC Berkeley Tech Report UCB/EECS-2007-99, 2007.

[16] Laaksonen, J., "Subspace classifiers in recognition of handwritten digits", Doctoral dissertation, Helsinki University of Technology, 1997.

[17] Acevedo, M. A., Corrada-Bravoc, C. J., Corrada-Bravob, H., Villanueva-Riverad, L. J., and Mitchell, T. A, "A*utomated classification of bird and amphibian calls using machine learning: A comparison of methods,"* Ecological Informatics, Vol. 4, pp. 206–214, 2009.

[18] Zhang, L., Lin, F., and Zhang, B., "Support vector machine learning for image retrieval," *Int. Conf. on Image Processing*, pp. 721-724, 2001.

[19] Goguen, C. B., and David, R. C., "Cassin's Vireo (Vireo cassinii), The Birds of North America Online (A. Poole, Ed.)," 2002. Ithaca: Cornell Lab of Ornithology; http://bna.birds.cornell.edu/bna/species/615.

[20] Candes, E., and Romberg, J., "'L_1-Magic: Recovery of Sparse Signals via Convex Programming," 2005 http://users.ece.gatech.edu/~justin/l1magic/.

[21] Chang, C. C., and Lin, C. J., "LIBSVM : a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, 27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[22] McNemar, Q., "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, pp. 153–157, 1947.