Audio Classification of Bird Species: a Statistical Manifold Approach

Forrest Briggs, Raviv Raich, and Xiaoli Z. Fern
School of EECS, Oregon State University, Corvallis, OR 97331-5501
{briggsf, raich, xfern}@eecs.oregonstate.edu

Abstract

Our goal is to automatically identify which species of bird is present in an audio recording using supervised learning. Devising effective algorithms for bird species classification is a preliminary step toward extracting useful ecological data from recordings collected in the field. We propose a probabilistic model for audio features within a short interval of time, then derive its Bayes risk-minimizing classifier, and show that it is closely approximated by a nearest-neighbor classifier using Kullback-Leibler divergence to compare histograms of features. We note that feature histograms can be viewed as points on a statistical manifold, and KL divergence approximates geodesic distances defined by the Fisher information metric on such manifolds. Motivated by this fact, we propose the use of another approximation to the Fisher information metric, namely the Hellinger metric. The proposed classifiers achieve over 90% accuracy on a data set containing six species of bird, and outperform support vector machines.

1 Introduction

Our goal is to develop algorithms that can predict which species of bird is present in an audio recording, by learning from a collection of labeled examples. Such algorithms will serve as part of a system to automatically collect bird species presence/absence data, which will provide valuable ecological information for species distribution modeling and conservation planning. Existing bird species distribution data are collected by manual surveys, which are labor intensive, and require observers trained in bird recognition [2]. Automated bird population surveys could provide vast amounts of useful data for species distribution modeling, while requiring less effort and expense than human surveys. Other applications of classifying bird sounds include reducing plane crashes caused by collisions with birds [5], and audio classification in general.

Sounds that birds make have a grammatical structure; two important levels of organization in this structure are songs and syllables. Syllables are single distinct utterances by a bird and serve as the basic building blocks of bird song [3]. A song consists of a series of syllables arranged in a particular pattern. In this study, our goal is to classify bird species from an interval of sound (containing one or more syllables), which roughly corresponds to the song level of organization.

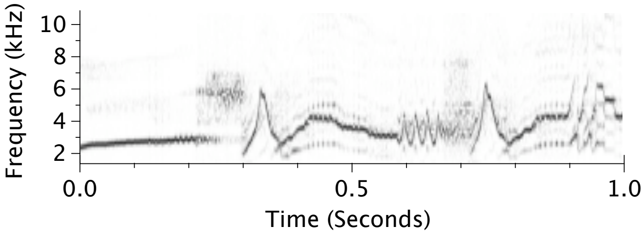

Figure 1. The spectrogram for a one-second portion of a recording of a Swainson's Thrush. Darker areas indicate higher energy at the corresponding frequency.

Audio classification systems typically begin by extracting acoustic features from audio signals. Such features often pertain to individual frames (i.e., very short segments of signal). For example, one commonly used feature is the spectrum of a signal frame, which describes the intensity of (a short segment of) the signal as a function of frequency. To apply many standard algorithms for classification, it is necessary to represent a sound, which contains multiple frames, using a fixed-length vector. To construct such a fixed-length feature vector to describe a sound as a whole, a common approach is to first identify interesting frames by segmentation, compute features for those frames, then take the average of the features over all frames [13, 23, 29]. For example, in the context of bird species recognition, a recent work by Fagerlund [13] (current state-of-the-art) averages frame-level features and applies support vector machines (SVMs).

Rather than averaging frame-level features, we represent their distributions using histograms (bag-of-codewords) defined by a 'codebook' of clustered frame-level features. Codebook based representations have been successfully applied in computer vision [7, 31, 19], and have also recently achieved success in music genre classification [26].

Our main contribution in this paper is to establish a theoretical framework that connects nearest-neighbors classifiers using histograms of features, Bayesian risk minimization, and geodesics on statistical manifolds. In particular,

- We propose a probability model for audio, then follow a Bayesian approach to derive the risk-minimizing classifier for this model. The Bayes classifier is closely approximated by a nearest-neighbor classifier using Kullback-Leibler (KL) divergence to compare feature histograms (Sec. 3);
- We explain that not only do Kullback-Leibler and the related Hellinger distance follow from a Bayesian probability model, but in the limit of a nearest-neighbor classifier, they can be thought of as approximations to geodesics using the Fisher information metric on statistical manifolds of histograms (Sec. 4);
- We experimentally compare the accuracy of nearest-neighbors using L1, L2, KL and Hellinger distances, and SVMs, with averages and histograms of frame-level features, on a data set consisting of 413 thirty-second intervals of sound from six species of bird. Results indicate that classifiers using histograms of frame-level features outperform those using averages, and that with a manifold geodesic distance between histograms, nearest neighbor can outperform SVM. Several classifiers achieve over 90% accuracy (Sec. 5).

2 Background and Related Work

We review data representation for audio classification, and related work on species identification from bird sounds.

2.1 Data Representation

Our goal is to classify a recording of bird sound as one of several species. A critical initial step toward this goal is to extract meaningful features to describe an interval of sound. This section presents our approach to constructing feature vectors to describe such intervals.

2.1.1 Basics: Signals and Spectrograms

Audio signals consists of a time-series of *samples*, which we denote as $s(t)$. It is often easier to recognize patterns in an audio signal when samples are converted to a frequency domain *spectrogram* using the Fast Fourier Transform (FFT) [3], (see Fig. 1 for an example spectrogram).

To compute a spectrogram, samples in a sound are divided into overlapping frames (Fig. 2), each of which contains a fixed number of consecutive samples. The FFT is applied to each frame to obtain the complex Fourier coefficients. The magnitudes of these coefficients are called the frame's magnitude spectrum and represent the intensity of the sound during that frame at different frequencies. A spectrogram is a plot of the spectrum for each frame in a signal.

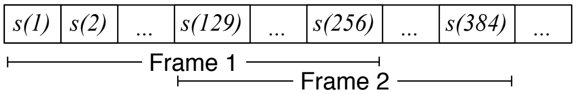

Figure 2. An audio signal is made up of samples, which are divided into overlapping frames.

2.1.2 Frame-Level Features

Many features for audio classification describe individual frames of a signal. In this section, we describe three features that we used in our experiments.

Spectrum Density. The magnitude spectrum of a frame can be normalized to form a probability distribution. If the magnitude spectrum is $(|c_1|, \ldots, |c_l|)$, where l is the number of elements in a spectrum, then the spectrum density is $f(i) = \frac{|c_i|}{\sum |c_i|}$. We can directly use $(f(1), f(2), \ldots, f(l))$ as the feature vector describing a frame.

Mean Frequency and Bandwidth. Consider the spectrogram shown in Fig. 1; each vertical slice represents the spectrum of one frame of sound. Bird sounds are usually concentrated at a few frequencies; we can see this phenomenon as horizontal strips in the spectrogram. This suggests that it is possible to condense the information contained in the spectrum density into just two values: the mean frequency and the bandwidth of the spectrum. The mean frequency of a frame indicates the vertical position of the strip, while bandwidth describes the width of the strip. Specifically, the mean frequency is $f_c = \int x f(x) dx$, and bandwidth is $BW = \sqrt{\int (x - f_c)^2 f(x) dx}$.

Mel-Frequency Cepstral Coefficients. Mel-frequency cepstral coefficients [9] (MFCCs) are one of the most widely used features for audio classification. The idea is to first compute Mel-frequency coefficients (MFCs), which are like the magnitude spectrum, but in units of mels rather than Hz (mels correspond more closely with human perception of pitch [30]). MFCs are computed by applying a collection of triangular filters to the magnitude spectrum; the MFCs are the response of each filter. The filters are evenly spaced in the mel scale. MFCCs are the result of applying the discrete cosine transform (DCT) to the log of the MFCs.

2.1.3 Aggregating Frame-Level Features

An interval contains a large number of frames, which can be aggregated to produce a single fixed-length feature vector.

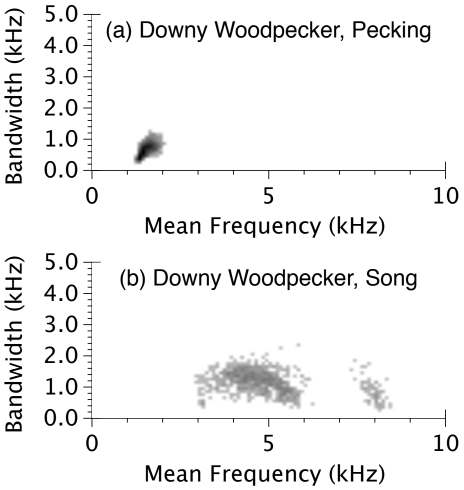

Figure 3. 2D histograms of frame mean frequency and bandwidth from two different intervals of audio recordings of the Downy Woodpecker.

A common approach that has been used in syllable classification is to average frame-level features [13, 23, 29]. However, by averaging, significant information about the distribution of features is lost, which can be problematic when the distribution of features in an interval is multimodal. For example, Fig. 3(b) shows the distribution of the features (mean-frequency and bandwidth) of the frames from a 30-second recording of a downy woodpecker (approximated by a 5000-bin histogram). In this case, the distribution is clearly multimodal and its mean will actually be in an area of relatively low probability, making it a poor representation for the overall distribution. We observed that such multimodality is common for bird sound. This observation suggests that aggregation schemes that can capture multimodality in feature distributions may be more successful than averages (our experimental results support this idea; Sec. 5.6). Inspired by the use of codebooks for image classification [7, 31, 19], and recent work in music genre classification [26], we consider aggregating frame-level features by representing their distributions with histograms.

Low-Dimensional Feature Histograms. Given an interval (i.e., a set of frames), each of which is described by a d-dimensional feature vector, a natural way to represent the interval is to use the probability distribution of features in this interval. This distribution can be approximated by a d-dimensional histogram, where dimension i is discretized into k_i bins, leading to a total of $\prod_{i=1}^{d} k_i$ bins. Note that since the total number of bins grows exponentially with d, this method can only be applied for small values of d. The vector of frequencies for each histogram bin can be used as a feature vector for classification.

Codebook Feature Histograms. The simple binning approach does not work for higher dimensional frame-level features such as spectra or MFCCs — we would need an infeasible number of bins to cover these high-dimensional spaces. Instead, we take a ‘codebook’ approach [26] to constructing histograms for high-dimensional features, which amounts to using non-uniform bins. A codebook is a collection of k codewords, each of which is a feature vector that is considered as representative in the feature space. There is one bin associated with each codeword. Given an interval (i.e., a set of frames each described by a feature vector) and a codebook, to compute a feature for the interval, assign each frame to its closest codeword, then count the number of frames assigned to each codeword. The vector of counts, normalized by dividing by their sum, gives the final feature vector, which is a histogram.

2.2 Related Work

Bird species can be classified using features extracted from audio recordings. A common approach to bird species classification is to identify distinct syllables, then construct feature vectors for those syllables and apply a standard classifier such as nearest neighbor or support vector machines to predict the species for each syllable [29, 13, 16, 12, 23, 22, 27]. Song-level species prediction has also been investigated using Hidden Markov Models [21, 29], Gaussian Mixture Models [29], based on comparisons of syllable-pair histograms [28], or nearest-neighbor classifiers using a feature constructed by aggregating syllable features [23].

To classify syllables or songs, most prior work relies on segmentation of the input audio into syllables [29]. As such, the classifier accuracy can be strongly dependent upon accurate segmentation [12]. A standard approach to segmentation is to compute the energy of each frame, then adaptively compute a threshold that separates syllables from background noise [29, 13, 25, 25]. It is difficult to obtain reliable segmentation using this method in recordings with low signal-to-noise ratio. In this paper, we use a simple approach to detect a set of interesting frames within the signal that correspond to bird sound, and do not require that they precisely match syllable time-boundaries (Sec. 5).

Audio classification in general has been widely studied, with applications to human speech and music being the most common. Our work is closely related to recent work by Seyerlehner et al. [26] on music genre identification. They follow a codebook approach to constructing audio feature histograms (Sec. 2.1.3), and use a nearest-neighbor classifier with L1 distance to classify these features. However, it is not obvious why a nearest-neighbor classifier is ideal for classifying histograms of features, or which distance measures are the best for comparing histograms. In this paper, we show that the Bayes optimal classifier for a probability model for audio is closely related

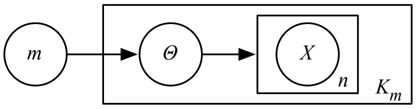

Figure 4. The plate diagram for the Interval-IID model.

Table 1. Notations

Variable	Description
m	class label (bird species)
n	number of interesting frames in an interval
θ	frame feature histogram parametrization
x_i	ith test frame feature vector
x_{ik}^t	ith training frame feature vector for the k training interval
X	frame feature vector collection for an interval $X = [x_1, \ldots, x_n]$
X_k^t	frame feature vector collection for the kth training interval
y_k^t	class label associated with the kth training interval
K	number of training intervals
K_m	number of training intervals from class m
$P(m)$	class prior probability
$p(\theta\|m)$	class-conditional histogram probability
$p_{x\|\theta}(x\|\theta)$	interval-conditional frame-feature probability
$p_{X\|\theta}(X\|\theta)$	interval-conditional features probability
$p_{X\|m}(X\|m)$	class-conditional features probability

to nearest-neighbor classifiers using histograms of features with appropriate distance measures.

3 Probability Model for Sound

In the following section, we present a theoretical justification for the frame-level feature histogram representation through a probability model, namely the Interval-IID model, and show that the corresponding Bayes risk-minimizing classifier can be approximated by a nearest-neighbor classifier with KL divergence.

3.1 The Interval-IID model

The Interval-IID model follows the graphical representation in Fig. 4. The model suggests that to generate an interval, we first determine its class label m based on the class prior $p(m)$. Given m, we then generate an interval-specific parameterization θ based on $p(\theta|m)$, which parameterizes the the frame feature distribution $p_{x|\theta}(x|\theta)$ of that interval. Given θ, we then generate n independent and identically distritbuted (i.i.d.) frame feature vectors x_i based on $p_{x|\theta}(x|\theta)$ (thus the name Interval-IID, i.e., frames are i.i.d. within an interval).

Given an observed interval represented by its collection of frame features $X = [x_1, x_2, \ldots, x_n]$, using Bayes rule, we write its class-conditional probability as

$$p_{X|m}(X|m) = \int p_{X|\theta}(X|\theta)p(\theta|m)d\mu(\theta), \qquad (1)$$

where θ is a parametrization determining the interval-conditional feature distribution $p_{x|\theta}(x|\theta)$ and m denotes the class label. Here we marginalized over the interval-conditional features probability model $p_{X|\theta}(X|\theta)$ according to the class conditional histogram parametrization distribution $p(\theta|m)$. As the Interval-IID model name suggests, conditioned on θ, the frame-level features are assumed i.i.d., and hence $p_{X|\theta}(X|\theta)$ can be written as a product of the marginal distributions of each frame-level feature:

$$p_{X|\theta}(X|\theta) = \prod_{i=1}^{n} p_{x|\theta}(x_i|\theta), \qquad (2)$$

where x_i denotes the feature vector of the ith frame. Substituting (2) into (1), the class conditional model for the Interval-IID model is given by

$$p_{X|m}(X|m) = \int \prod_{i=1}^{n} p_{x|\theta}(x_i|\theta)p(\theta|m)d\mu(\theta). \qquad (3)$$

The integral w.r.t. θ here applies to the product of marginal probabilities. By writing p as $e^{\log p}$ and replacing the integral with the expectation notation, (3) becomes

$$p_{X|m}(X|m) = E_\theta\left[e^{\sum_{i=1}^{n} \log p_{x|\theta}(x_i|\theta)}|m\right]. \qquad (4)$$

To express $p_{X|m}(X|m)$ in (4) in terms of the Kullback-Leibler (KL) divergence, start by introducing the following terms:

$$\begin{aligned} \theta^* &= \arg\max_\theta \frac{1}{n}\sum_{i=1}^{n} \log p_{x|\theta}(x_i|\theta), && (5) \\ \tilde{H}(\theta^*) &= -\frac{1}{n}\sum_{i=1}^{n} \log p_{x|\theta}(x_i|\theta^*), && (6) \\ D(\theta^*, \theta) &= \frac{1}{n}\sum_{i=1}^{n} \log \frac{p_{x|\theta}(x_i|\theta^*)}{p_{x|\theta}(x_i|\theta)} && (7) \end{aligned}$$

Note that θ^* is the maximum-likelihood estimator of θ. Using (5-7) and the observation that $\sum_{i=1}^{n} \log p_{x|\theta}(x_i|\theta) = -n(\tilde{H}(\theta^*) + D(\theta^*, \theta))$, we rewrite (4) as

$$p_{X|m}(X|m) = e^{-n\tilde{H}(\theta^*)} E_{\theta|m}\left[e^{-nD(\theta^*,\theta)}\right]. \qquad (8)$$

By the definition of θ^* in (5), we have that $D(\theta^*, \theta) \geq 0$ for all θ and is zero for $\theta = \theta^*$. We proceed with the specific case in which the features are discretized into L non-intersecting bins defined by the sets A_l. Hence, we represent the class-conditional distribution of frame-level features using histograms. Each frame-level feature x_i can fall into one of the histogram bins $\{A_1, \ldots, A_L\}$ with probability $\{\theta_1, \ldots, \theta_L\}$, respectively. The vector $\theta = [\theta_1, \ldots, \theta_L]^T$ is a probability mass function (or a histogram), i.e., $\sum \theta_l = 1$ and $\theta_l \geq 0$. The interval-conditional probability model for a frame-level feature is given by

$$p_{x|\theta}(x|\theta) = \prod_{l=1}^{L} \theta_l^{I(x \in A_l)}, \tag{9}$$

where $I(\cdot)$ is the indicator function which takes the value one if its argument is true and zero otherwise. We would like to point out that when $p_{x|\theta}(\cdot|\theta)$ is given by (9) and $p(\theta|m)$ is the Dirichlet distribution, then (3) becomes the Dirichlet-Multinomial model, which is also referred to as Polya distribution [24] or the Dirichlet compound multinomial (DCM) model [11]. This model is often used as a topic model in text document classification. One criticism concerning the choice of Dirichlet prior is the limited capability of representing multimodal priors [34]. Our experience with bird sounds suggests that the probability model $p(\theta|m)$ is indeed multimodal; as Fig. 3 shows, frame-level feature histograms for the same species differ between intervals.

For $p_{x|\theta}(x|\theta)$ given by (9), we have

$$\begin{aligned} \theta_l^* &= \hat{p}_l, & (10) \\ \tilde{H}(\theta^*) &= H(\theta^*) & (11) \\ D(\theta^*, \theta) &= D_{kl}(\theta^* \| \theta), & (12) \end{aligned}$$

where $\hat{p}_l = \frac{1}{n}\sum_{i=1}^{n} I(x_i \in A_l)$ is the lth empirical histogram bin probability estimate based on the observed feature collection $X = [x_1, x_2, \ldots, x_n]$, $H(p) = -\sum_{l=1}^{L} p_l \log p_l$ is the entropy associated with a multinomial parameterized by p (the vector of bin probabilities of a histogram), and

$$D_{kl}(\theta^* \| \theta) = \sum_{l=1}^{L} \theta_l^* \log \frac{\theta_l^*}{\theta_l}$$

is the Kullback-Leibler (KL) divergence between a multinomial parameterized by θ^* and another parameterized by θ. Substituting (10)-(12) into (8), we have

$$p_{X|m}(X|m) = e^{-nH(\hat{p})} E_\theta\big[e^{-nD_{kl}(\hat{p}\|\theta)}|m\big]. \tag{13}$$

This form for $p_{X|m}(X|m)$ acts as the likelihood component in the Bayes risk minimizing classifier in the following section. Moreover, it highlights the role of the KL divergence in optimal Bayesian classification for the problem at hand.

3.2 Bayes Risk Minimizing Classifier

We start with a brief review of the Bayes risk minimization approach to classification [14]. The probability of error for a given classification rule $\hat{m}(X)$ is

$$Pr(\text{error}) = \sum_{m=1}^{M} P(\hat{m}(X) \neq m|m)P(m). \tag{14}$$

The classification rule that minimizes the error in (14) is

$$\hat{m} = \arg\max_m \; p_{m|X}(m|X). \tag{15}$$

This rule is also referred to as maximum a-posteriori (MAP), as it assigns a decision based on the highest class probability given the set of observations. Using Bayes rule $p_{m|X}(m|X) = p_{X|m}(X|m)P(m)/p_X(X)$ and the fact that $p_X(X)$ is constant w.r.t. to the class variable m, yields an equivalent form to the MAP classifier:

$$\hat{m} \;=\; \arg\max_m \; \log P(m) + \log p_{X|m}(X|m). \tag{16}$$

After replacing the likelihood $P_{X|m}(X|m)$ with (13), the MAP classification rule (16) for the Interval-IID model in (3) is

$$\begin{aligned} \hat{m} \;=\; & \arg\max_m \; \log P(m) \\ & + \log E_\theta\big[\exp\big(-nD_{kl}(\hat{p}\|\theta)\big)|m\big]. \end{aligned} \tag{17}$$

Note that since $H(\hat{p})$ in (13) is independent of m, it is not incorporated into (17). It is equivalent to replace the maximization with minimization and divide by n

$$\hat{m} = \arg\min_m \; -\frac{1}{n}\log\Big(E_\theta\big[e^{-nD_{kl}(\hat{p}\|\theta)}|m\big]P(m)\Big). \tag{18}$$

With no exact knowledge about $P(m)$ and the PDF $p(\theta|m)$ used to compute the expectation $E_\theta[\cdot|m]$, we propose estimating these quantities from the training data.

3.3 Training

To describe the training process, we start by explaining the format of the training data. Each interval k in the training data contains n training features $X_k^t = [x_{1k}^t, x_{2k}^t, \ldots, x_{nk}^t]$ and is associated with a class label y_k^t. We assume that K training intervals are available, i.e., $(X_1^t, y_1^t), (X_2^t, y_2^t), \ldots, (X_K^t, y_K^t)$. We denote training variables using the superscript t notation.

To train the classification rule in (18), we replace $P(m)$ and $E[\cdot|m]$ through their sample estimates

$$\hat{m} = \arg\min_m \; \frac{-1}{n}\log\Big(\sum_{k=1}^{K} I(y_k^t = m)e^{-nD_{kl}(\hat{p}\|\hat{\theta}^{(k)})}\Big), \tag{19}$$

where k is the interval number, K is the total number of training intervals, y_k^t is the class label for the kth training interval, and $\hat{\theta}^{(k)}$ is a histogram estimated from the kth training interval given by

$$\hat{\theta}_l^{(k)} = \frac{1}{n}\sum_{i=1}^{n} I(x_{ik}^t \in A_l), \tag{20}$$

where x_{ik}^t is the ith feature vector from the kth interval. With a slight abuse of notations, we rewrite (19) as

$$\hat{m} = \arg\min_m \; -\frac{1}{n}\log\Big(\sum_{k=1}^{K_m} e^{-nD_{kl}(\hat{p}\|\hat{\theta}^{(k,m)})}\Big), \tag{21}$$

where the $\hat{\theta}^{(k,m)}$'s are the sorted version of the $\hat{\theta}^{(k)}$s from class m such that $D_{kl}(\hat{p}\|\hat{\theta}^{(1,m)}) \leq D_{kl}(\hat{p}\|\hat{\theta}^{(2,m)}) \leq \ldots D_{kl}(\hat{p}\|\hat{\theta}^{(K_m,m)})$, and K_m is the number of training intervals for the mth class. Using the ordered training class histograms, we reorganize (21) as

$$\begin{aligned}\hat{m} \;&=\; \arg\min_m \; D_{kl}(\hat{p}\|\hat{\theta}^{(1,m)}) \qquad (22)\\ &\quad -\frac{1}{n}\log\Big(1+\sum_{i=2}^{n_m} e^{-n(D_{kl}(\hat{p}\|\hat{\theta}^{(i,m)})-D_{kl}(\hat{p}\|\hat{\theta}^{(1,m)}))}\Big).\end{aligned}$$

We refer to (22) as the Interval-IID MAP classifier. While equivalent to (21), (22) provides insight into the relation between Bayes risk-minimization, nearest-neighbor classifiers, and manifold geodesics. Identifying the training intervals with their feature histograms, and the test interval with its feature histogram, the first term on the RHS of (22) is a KL divergence based nearest neighbor rule in histogram space. Note that if the KL distance to points other than the first nearest neighbor $D_{kl}(\hat{p}\|\hat{\theta}^{(i,m)})$ is sufficiently larger than the distance to the first nearest neighbor $D_{kl}(\hat{p}\|\hat{\theta}^{(1,m)})$ then the second term on the RHS of (22) becomes negligible, and (22) is simply a nearest neighbor classifier using KL divergence.

4 Nearest Neighbors on Statistical Manifolds

The connection between optimal Bayes classification and the histogram KL nearest neighbor rule leads us to extend the approach to nearest neighbor classification on histograms. Note that a collection of probability models (i.e, histograms) can be regarded as a manifold. Denote a model by $p(X|\theta)$ or in short by $p(\cdot|\theta)$. The collection of models given by

$$\mathcal{M} = \left\{p(\cdot|\theta) \mid \theta \in \Theta \in \mathbb{R}^d\right\}, \tag{23}$$

is a d-dimensional statistical manifold if there exist a one-to-one smooth mapping between θ to $p(\cdot|\theta)$. In the geometric approach to statistical models [20], one can measure the geodesic distance between two histograms by using the Fisher information metric (FIM) as the Riemannian metric

$$D_F(p(\cdot|\theta), p(\cdot|\theta')) = \min_{\substack{\theta(\cdot),\\ \theta(0)=\theta,\\ \theta(l)=\theta'}} \int_0^l \sqrt{\dot{\theta}(t)^T \mathcal{I}(\theta(t))\dot{\theta}(t)}dt, \tag{24}$$

where $\mathcal{I}(\theta)$ is the Fisher information matrix given by

$$\mathcal{I}_{ij}(\theta) = E\Big[\frac{d\log p(x|\theta)}{d\theta_i}\frac{d\log p(x|\theta)}{d\theta_j}\Big]. \tag{25}$$

The FIM is considered a natural metric for statistical manifolds as it reflect the capability to discriminate between probability models from their samples.

To generalize the nearest neighbor approach discussed in the previous section in the context of statistical manifolds, we consider a geodesic nearest neighbor rule using $D_F(p(\cdot|\theta), p(\cdot|\theta'))$ defined in (24). As the precise form of the manifold is unavailable, an exact computation of the geodesic distance $D_F(p, p')$ is impossible. Since the nearest neighbor approach prompts us to calculate short geodesic distances, local approximations of $D_F(p, p')$ can be used instead. For two close probability models $p \to p'$ it is known [20] that $\sqrt{2D_{kl}(p\|p')} \to D_F(p, p')$. The KL divergence provides a computable approximation to the FIM manifold geodesic distance.

Note that other approximations for the FIM are available (e.g., certain Ali-Silvey divergences, and specifically, Hellinger divergence). In this paper, we use the Hellinger divergence given by

$$D_H(p,q)^2 = \sum(\sqrt{p_i} - \sqrt{q_i})^2, \tag{26}$$

which is a metric as opposed to the KL divergence. The approximation of the FIM using Hellinger distance for close models is $2D_H(p, p') \to D_F(p, p')$ [20].

For the purpose of comparison, we experimentally evaluate nearest neighbor classifiers using L1 and L2 distances as well as KL and Hellinger. L2 is the standard Euclidean distance, which is widely used, but not theoretically justified for the comparison of probability distributions. L1 is fairly common for comparing probability distributions. It is a member of the Ali-Silvey family, but due to non-differentiability, it is not an approximation to the FIM. However, it is related to Hellinger by the inequality, $\frac{1}{2}D_H(p,q)^2 \leq D_{L1}(p,q) \leq D_H(p,q)$ [8]. This relation between L1 and Hellinger hints at why classifiers using these distances achieve similar results (Sec. 5.6).

5 Experiments

In this section, we describe the experimental setup used to measure the accuracy of the proposed methods for bird species classification, and to compare with SVMs [13]. We consider various frame-level features (mean frequency and bandwidth, MFCCs, and spectral density), interval-level features (averages vs. histograms), and metrics for nearest-neighbor classification (L1, L2, KL, and Hellinger). We also empirically verify that a nearest-neighbor classifier using Kullback-Leibler closely approximates the Interval-IID MAP classifier (22) as suggested in Sec. 3.

5.1 Data

We have 1.13 GB of recordings from the Cornell Macaulay library, of 6 species: Black Throated Blue Warbler, Hermit Warbler, Downy Woodpecker, Swainson's Thrush, Western Tanager, and Winter Wren. All of these recordings are at least 30 seconds long, and most are less than 10 minutes. We divide each recording into intervals of 30 seconds, resulting in 413 intervals. Our goal is to classify these intervals according to species.

The recordings were collected over several decades, mainly in the western United States. Most are made using a directional microphone in the field. The amount of noise in the recordings varies widely. In addition to static and wind, some recordings contain cars sounds, human speech, and other non-bird sounds. We manually removed most portions of sound with human voices. Although each recording is labeled with just one species, some recordings contain multiple birds, sometimes of different species; usually the loudest bird present corresponds to the label for the recording. The sampling frequency for all recordings is 44.1 kHz. The audio data is stored as mono-channel WAV files.

5.2 Preprocessing

Section 2.1 covers the process of converting a sequence of samples from an audio interval into interval-level features. We proceed by further elaborating on the specific details of our experimental setup.

When dividing a signal into frames, we use 256 samples per frame, and successive frames overlap by 50%. To reduce noise and decrease processing time in later stages, we discard the lowest 8 and highest 64 elements of each frame's spectrum, leaving 56 elements from the original 128 (equivalent to removing all sound below 1.378 kHz and above 10.852 kHz).

Instead of syllables, we detect a subset of interesting frames (which are more likely to contain bird sound) in an interval. To find these interesting frames, we compute the total magnitude of each frame, $\sum |c_i|$, and retain only the 10% of frames with highest total magnitude in all subsequent calculations. Note that the total magnitude is similar to, but not the same as the energy of a frame.[1]

Our implementation of MFCCs (Sec. 2.1.2), is based on the description provided by Ganchev et al. [15] of the MFCCs computed in the Cambridge Hidden Markov Models Toolkit (for MATLAB), known as HTK [33]. We use 24 filters,[2] resulting in 24 MFCs, then take only the first 12 elements of the output of the DCT as the frame-level feature.

For constructing 2D histograms, we divide the range of values for mean frequency and bandwidth into square bins 100 Hz wide, with 100 bins on the mean frequency axis, and 50 bins for the bandwidth axis (for a total of $50 \times 100 = 5000$ bins, covering a range of 0 Hz to 10,000 Hz for f_c and 0 Hz to 5000 Hz for BW). There is one element in the feature vector for each histogram bin, so this representation results in a 5000-dimensional feature vector.[3]

5.3 Clustering for Codebooks

For constructing codebooks, we apply the k-means++ clustering algorithm [1] to the frame-level features from a training data set. Note that there are several hundred-thousand frames to cluster in our data set. To speed-up codebook construction, we follow a two-staged clustering proceedure suggested by Seyerlehner et al. [26]. In particular, we first cluster features within each 30-second interval, then cluster the resulting cluster centers to obtain the final codewords. In the first stage of clustering, the feature vectors are either the spectrum density, or MFCCs for the interesting frames. In the second stage, the examples are the cluster centers from the first stage. We use $k = 10$ clusters for the first stage and $k = 100$ for the second. Thus, the final interval-level features constructed using this method are 100-dimensional. In our preliminary experiments, this approach to clustering yielded an order-of-magnitude speedup over clustering all frame-level features at once, because the first stage of clustering does not need to be repeated in each fold of cross-validation.

5.4 Classifiers

There are many combinations of frame-level features and methods of aggregating them. The combinations we consider in this study are: averages of f_c and BW, spectrum

[1] Parseval's theorem states that the energy of a frame can be computed from its spectrum via the formula $E = \sum |c_i|^2$, were c_i is the ith FFT coefficient.

[2] In our implementation, the filters span a range of frequencies from $f_{low} = 1000Hz$ to $f_{high} = 22050Hz$. Following an exact implementation of the filters described by Ganchev et al. [15], we got aliased triangle filters because some were narrower than a single spectrum bin, which caused artifacts in the MFCs. To fix this problem, we numerically integrate the triangle filter function over the range of each bin. Many other implementations of MFCCs work with lower sampling frequencies [15], so we suspect this problem is related to working with sound sampled at 44.1 kHz, as well as our choice of values for f_{low} and f_{high}.

[3] We apply Laplace smoothing to the histogram estimation by starting with a count of 1 for each bin.

Frame Feature	**Representation**	**Classifier**	**% correct**
f_c, BW	Average	NN-L1	42.85
f_c, BW	Average	NN-L2	42.85
MFCCs	Average	NN-L1	81.11
MFCCs	Average	NN-L2	81.11
MFCCs	Average	SVM	84.50
Spectrum Density	Average	NN-L1	79.42
Spectrum Density	Average	NN-L2	81.35
Spectrum Density	Average	SVM	84.75
f_c, BW	2D Histogram	**Interval-IID MAP**	87.40
f_c, BW	2D Histogram	**NN-Kullback-Leibler**	87.40
f_c, BW	2D Histogram	**NN-Hellinger**	88.13
f_c, BW	2D Histogram	NN-L1	86.44
f_c, BW	2D Histogram	NN-L2	83.05
MFCCs	Codebook	NN-L1	84.41 ± .89
MFCCs	Codebook	NN-L2	83.49 ± .62
MFCCs	Codebook	**NN-Kullback-Leibler**	85.42 ± .62
MFCCs	Codebook	**NN-Hellinger**	86.59 ± .50
MFCCs	Codebook	SVM	87.17 ± .58
Spectrum Density	Codebook	NN-L1	92.54 ± .44
Spectrum Density	Codebook	NN-L2	88.28 ± .99
Spectrum Density	Codebook	**NN-Kullback-Leibler**	90.70 ± .40
Spectrum Density	Codebook	**NN-Hellinger**	92.10 ± .27
Spectrum Density	Codebook	SVM	88.14 ± .58

Table 2. The accuracy of each classifier in predicting bird species based on 413 thirty-second intervals of sound. NN means nearest neighbor. The values listed for classifiers using a codebook are average accuracy over 5 trials, ± average deviation. Our proposed methods are listed in bold.

density, and MFCCs, 2D histograms of f_c and BW, and codebook histograms of spectrum density and MFCCs.

Using the above features extracted from the data described in Sec. 5.1, we compare several classification algorithms: nearest neighbor with L1, L2, KL and Hellinger distances, and the Interval-IID MAP classifier proposed in Sec. 3.2 (22), as well as support vector machines. Of these classifiers, Interval-IID Map, KL and Hellinger are our proposed methods, and the others are included for comparison.

5.4.1 Support Vector Machines

Support vector machines [6] (SVMs) are a family of algorithms for supervised classification that find a linear decision boundary by maximizing the margin between two classes. In cases where linear classification is insufficient, the kernel trick is applied to non-linearly project features into a higher dimensional space where linear separability is possible. The implementation of SVMs that we used is WLSVM [10], which integrates LIBSVM [4] into the Weka [32] machine learning system. Following Fagerlund [13], and the recommendations of Hsu, Chang and Lin [17], we use a radial basis function kernel, and optimize the SVM parameter C and the kernel parameter γ, by grid search. We evaluate the SVM at all combinations of C and γ in $\{10^{-1}, 10^0, 10^1, 10^2\}$, and report the best accuracy achieved with any set of parameters. To handle multiple classes (in our case, species), LIBSVM use the one-against-one voting scheme [18].

5.5 Cross Validation and Multiple Trials

To measure the accuracy of the proposed classifiers, we use them to predict the species in each of 413 thirty-second intervals of sound. Each classifier is trained using all of the intervals that do not come from the same recording as the interval being classified (the data set consists of longer recordings that are split into intervals). We use this setup so the classifier must identify species without already having example recordings of the individual bird being classified. Fagerlund [13] used a similar 'individual independent' setup for cross-validation.

Classifiers that use a codebook to construct feature histograms depend on a randomized clustering algorithm. To account for the randomness, we ran five trials with different random seeds, and report average accuracy, ± average deviation.

5.6 Results

Table 2 lists the accuracy of each classifier on the species recognition problem. We make the following key observations.

- Regardless of which frame-level features we use, histograms of features achieved better accuracy than averages. One possible explanation of this result is that feature distributions may be multimodal (Fig. 3(b)), so the mean alone may not be enough to discriminate between distributions from different species.

- Using the 2D histogram of f_c and BW, the Interval-IID MAP classifier (22) produced identical results to a nearest-neighbor classifier with KL divergence. This result confirms our theoretical argument that a nearest neighbor classifier using KL divergence is a close approximation to the Interval-IID MAP classifier. Accordingly, we recommend using the more efficient nearest neighbor with KL as opposed to (22), when audio data is believed to be generated as in the Interval-IID model.

- Comparing different distance functions when using histograms, we observe that L1, Hellinger and KL were generally more accurate than using L2, with the performance of Hellinger being the most robust across different settings. Interestingly, while L1 is not an approximation to the FIM, its performance is highly competitive to KL and Hellinger. For histograms of spectrum density, L1 slightly outperformed Hellinger (although not statistically significantly). This is possibly due to the close relationship between L1, Hellinger and KL, as explained in Sec. 4. Note that MFCCs are essentially a compressed version of the spectrum (from 56 elements to 12), so it is not surprising that classifiers using them are slightly less accurate than those using spectrum density.

- Despite their relative simplicity, classifiers using 2D histograms of mean frequency and bandwidth provide remarkably accurate predictions. Being able to visualize a 2D histogram as an image provides insight into the structure of bird sound (for example, we can see that interval feature histograms may be multimodal).

- Finally, we note that the proposed methods achieved accuracy similar to or better than SVMs. In particular, using codebook histograms of MFCCs, SVMs are slightly more accurate than a nearest neighbor classifier with Hellinger, although the difference is not statistically significant. On codebook histograms of spectrum density, nearest-neighbor classifiers using statistical divergence measures (i.e. L1, Kullback-Leibler and Hellinger) outperform SVM. We want to emphasize that unlike SVMs, which require significant parameter tuning, the proposed methods also offer additional advantages in terms of their simplicity and scalability, making them more usable in practice.

6 Conclusion and Future Work

In this paper, we addressed the problem of bird species classification from audio recordings. Following a Bayesian approach to classification, we introduced the interval-IID model to describe the distribution of feature vectors within an interval consisting of frames, and derived the corresponding MAP classifier. The MAP classifier suggests aggregating features into histograms and using KL nearest neighbor to classify. This connection to nearest neighbor classification on statistical manifolds led us to extended the classifier by proposing different metrics (e.g., Hellinger). To use the MAP classifier with high-dimensional frame-level features, we employ codebook histograms.

Our study suggests that **1)** using histograms of frame-level features in an audio classifier can produce better results than using averaged frame-level features **2)** nearest-neighbor classifiers using Kullback-Leibler and Hellinger distance to compare feature histograms results are competive with state-of-the-art method such as SVM and **3)** metrics appropriate for histograms such as Hellinger, KL, and L1 perform better than the Euclidean L2 metric.

The classifiers in this study make predictions from intervals based on the collection of frames within the interval. A common alternative is to instead focus on individual syllables. We are working on an experimental survey of methods for classifying bird species from syllables, as well as probability models that are specialized for this purpose.

The experiments and algorithms presented here are a preliminary step toward analyzing a large (terabyte scale) data set of bird sounds that our collaborators collected in field conditions, using an array of omnidirectional microphones. We intend to apply algorithms for bird species classification to these recordings to extract information about patterns of bird activity at an unprecedented spatial and temporal resolution.

7 Acknowledgments

This work is partially funded by the IGERT program through National Science Foundation grant No. DGE 0333257, as well as the College of Engineering, Oregon State University. We would also like to thank Dr. Matthew Betts at Oregon State University for his help in acquiring the bird sound data.

References

[1] D. Arthur and S. Vassilvitskii. k-means++: the advantages of careful seeding. In *SODA '07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

[2] M. G. Betts, A. Diamond, G. Forbes, M.-A. Villard, and J. Gunn. The importance of spatial autocorrelation, extent and resolution in predicting forest bird occurrence. *Ecological Modelling*, 191(2):197 – 224, 2006.

[3] C. Catchpole, P. Slater, and N. Mann. *Bird song: biological themes and variations*. Cambridge Univ Pr, 2003.

[4] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001.

[5] Z. Chen and R. C. Maher. Semi-automatic classification of bird vocalizations using spectral peak tracks. *J Acoust Soc Am*, 120(5 Pt 1):2974–2984, November 2006.

[6] C. Cortes and V. Vapnik. Support vector networks. In *Machine Learning*, pages 273–297, 1995.

[7] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, volume 1, page 22, 2004.

[8] A. DasGupta. Asymptotic theory of statistics and probability. *International Statistical Review*, 77(1):160–161, 04 2009.

[9] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Readings in speech recognition*, pages 65–74, 1990.

[10] Y. EL-Manzalawy and V. Honavar. *WLSVM: Integrating LibSVM into Weka Environment*, 2005.

[11] C. Elkan. Clustering documents with an exponential-family approximation of the Dirichlet compound multinomial distribution. In *Proceedings of the 23rd international conference on Machine learning*, pages 289–296, 2006.

[12] S. Fagerlund. *Automatic Recognition of Bird Species by their Sounds*. PhD thesis, Helsinki University of Technology, 2004.

[13] S. Fagerlund. Bird species recognition using support vector machines. *EURASIP Journal on Advances in Signal Processing*, 2007, 2007.

[14] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.

[15] T. Ganchev, N. Fakotakis, and G. Kokkinakis. Comparative evaluation of various mfcc implementations on the speaker verification task. In *in Proc. of the SPECOM-2005*, pages 191–194, 2005.

[16] A. Härmä. Automatic identification of bird species based on sinusoidal modeling of syllables. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, volume 5, pages V–545–8 vol.5, 2003.

[17] C. Hsu, C. Chang, C. Lin, et al. A practical guide to support vector classification, 2003.

[18] C. Hsu and C. Lin. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, 2002.

[19] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *Tenth IEEE International Conference on Computer Vision, 2005. ICCV 2005*, volume 1, 2005.

[20] R. Kass and P. Vos. *Geometrical foundations of asymptotic inference*. Wiley-Interscience, 1997.

[21] J. A. Kogan and D. Margoliash. Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden markov models: A comparative study. *The Journal of the Acoustical Society of America*, 103(4):2185–2196, 1998.

[22] C. Kwan, G. Mei, X. Zhao, Z. Ren, R. Xu, V. Stanford, C. Rochet, J. Aube, and K. Ho. Bird classification algorithms: Theory and experimental results. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 289–292, 2004.

[23] C.-H. Lee, Y.-K. Lee, and R.-Z. Huang. Automatic recognition of bird songs using cepstral coefficients. *Journal of Information Technology and Applications*, 1(1):17 – 23, 2006.

[24] T. Minka. Estimating a Dirichlet distribution. *Unpublished paper available at http://research. microsoft. com/ minka*, 2003.

[25] A. Selin, J. Turunen, and J. Tanttu. Wavelets in recognition of bird sounds. *EURASIP Journal on Advances in Signal Processing*, 2007:1–9, 2007.

[26] C. Seyerlehner, G. Widmer, and P. Knees. Frame Level Audio Similarity - A Codebook Approach. In *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx-08)*, Espoo, Finland, September 1–4 2008.

[27] P. Somervuo and A. Härmä. Analyzing bird song syllables on the self-organizing map. In *Proceedings of the Workshop on Self-Organizing Maps (WSOM'03)*, Kitakyushu, Japan, Sept. 2003.

[28] P. Somervuo and A. Harma. Bird song recognition based on syllable pair histograms. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP04*, volume 5, pages 825–828, 2004.

[29] P. Somervuo, A. Härmä, and S. Fagerlund. Parametric representations of bird sounds for automatic species recognition. In *IEEE Transactinos on Audio, Speed, and Language Processing*, volume 14. IEEE Press, 2006.

[30] J. Volkmann, S. S. Stevens, and E. B. Newman. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):208–208, 1937.

[31] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *Tenth IEEE International Conference on Computer Vision, 2005. ICCV 2005*, volume 2, 2005.

[32] I. Witten and E. Frank. Data mining: practical machine learning tools and techniques with Java implementations. *ACM SIGMOD Record*, 31(1):76–77, 2002.

[33] S. Young. The hidden markov model toolkit. *Entropic Cambridge Research Laboratory, Ltd*, 2:2–44, 1995.

[34] K. Yu, S. Yu, and V. Tresp. Dirichlet enhanced latent semantic analysis. In *Conference in Artificial Intelligence and Statistics*, 2005.