A Syllable-Level Probabilistic Framework for Bird Species Identification

Balaji Lakshminarayanan, Raviv Raich, and Xiaoli Fern
School of EECS, Oregon State University, Corvallis, OR 97331-5501
{lakshmba, raich, xfern}@eecs.oregonstate.edu

***Abstract*—In this paper, we present new probabilistic models for identifying bird species from audio recordings. We introduce the independent syllable model and consider two ways of aggregating frame level features within a syllable. We characterize each syllable as a probability distribution of its frame level features. The independent frame independent syllable (IFIS) model allows us to distinguish syllables whose feature distributions are different from one another. The Markov chain frame independent syllable (MCFIS) model is introduced for scenarios where the temporal structure within the syllable provides significant amount of discriminative information. We derive the Bayes risk minimizing classifier for each model and show that it can be approximated as a nearest neighbour classifier. Our experiments indicate that the IFIS and MCFIS models achieve 88.26% and 90.61% correct classification rates, respectively, while the equivalent SVM implementation achieves 86.15%.**

***Keywords*-Probabilistic modeling; audio classification; Bayesian inference; bird species identification;**

I. Introduction

Human speech recognition systems are generally based on models that characterize the vocabulary and grammar of a particular language. The notion of vocabulary and grammar is usually ambiguous for sound signals other than human speech. However, bird vocalization is a good example of a class of natural sounds where we can expect to find an underlying vocabulary and inherent grammatical structure [1]. Bird recordings usually contain a structured pattern of brief sounds from a species-specific vocabulary. Those brief sounds are usually called elements or syllables [2]. Since these syllables are characteristic of a particular species, probabilistic models that treat these syllables as the basic building blocks lend themselves naturally to species identification.

The process of identifying the constituent syllables present in a recording is known as segmentation. After segmentation, each syllable is represented by a suitable choice of feature vector. These feature vectors then act as inputs to a classification algorithm for identifying the particular bird species. Different feature representations and machine learning methods have been applied for bird species identification in the literature [1], [3]–[6]. In [4], the authors used dynamic time warping (DTW) to compare the input spectrograms with a predefined set of templates. In [5], the authors used neural networks and multivariate statistical techniques in conjunction with a set of temporal and spectral features. In [6], the authors used wavelet coefficients along with self organizing map (SOM) and multilayer perceptron (MLP). In [1], the authors compared three different feature representations (sinusoidal model, Mel-cepstrum model, descriptive parameters) by evaluating their performance with different classification algorithms based on DTW, Gaussian mixture model (GMM), Hidden Markov model (HMM). In [3], the author used a decision tree based classifier with support vector machine (SVM) at each node.

Even though different machine learning algorithms have been applied for bird species identification, there has been little work on the development of probabilistic models specific to bird vocalization. Probabilistic models enable Bayesian inference and help in identifying the interesting characteristics of data [7]. Probabilistic models have been successfully applied in other domains, for instance, the Dirichlet-multinomial model has been applied for classification as well as clustering of documents [8] and the Hierarchical Dirichlet process hidden Markov model (HDP-HMM) has been applied for word segmentation [9].

Bird vocalization is analogous to document classification in that the probability of syllables (words) depends on the particular species (topic). Hence, we can build probabilistic models for bird vocalization in a similar manner to those developed for document classification. There has been little previous work in probabilistic modeling of syllables for bird species identification. There are different ways to characterize a syllable in terms of its frame-level features. In [1], the authors computed the spectral peak for each frame within the syllable and used analysis by synthesis overlap add (ABS/OLA) algorithm to parametrize the variation in spectral peak within a syllable. Another approach is to compute the features of each frame within the syllable and use the average of these frame-level features to represent the syllable. HMM-based approaches model syllables as states and each frame within the syllable as an observation and are quite similar to our approach.

In this paper, we characterize each syllable as a probability distribution and treat the feature representation of each frame within a syllable to be observations from that particular syllable distribution. Modeling each syllable with a probability distribution allows us to employ Bayesian inference techniques for bird species identification at the syllable level. In this paper, we make the following contributions: we introduce the independent syllable model and consider two models of treating the frames within a syllable, namely independent frame independent syllable (IFIS) model and Markov chain frame independent syllable (MCFIS) model. We derive the Bayes-risk minimizing classifiers for each case and show that they can be approximated by the nearest neighbor with the distance metric being the appropriate divergence. For both IFIS and MCFIS models, we consider the special case where the frame level features are assumed to follow a multivariate Gaussian distribution. We experimentally evaluate the accuracy of the proposed classifiers and features using cross-validation on a data set consisting of 426 thirty-second recordings of six species of birds, from the Cornell Macaulay library. In our experiments, IFIS and MCFIS models are able to achieve correct classification rates of about 88.26% and 90.61% respectively compared to the 86.15% achieved by the equivalent SVM implementation.

II. Problem statement

Our objective is to identify bird species based on audio recordings. We have a collection of recordings of bird sounds, each of which is labeled with a particular species. The recordings differ in their duration, so they are split into equal-length intervals. The task is to learn an acoustic model for each species based on these training set intervals so that we can correctly classify a test interval. Analogous to human speech where syllables are characteristic of a particular language, there are some 'syllables' that are characteristic of a particular bird species. Hence, it makes more sense to treat these syllables as the basic building blocks and develop probability models for each syllable. It is common practice to divide syllables further into frames, where each frame corresponds to the sound in a very short span of time. The frames can then be represented by features such as: mean frequency, spectral bandwidth, short time energy, zero crossing rate, Mel frequency cepstral coefficients (MFCC) and energy. More formally, a syllable $\mathbf{x}(i)$ consisting of n_i frames can be viewed as a sequence of observations, i.e., syllable $\mathbf{x}(i) = [x_1(i), x_2(i), \ldots, x_{n_i}(i)]$ where observation $x_j(i)$ corresponds to the feature vector representation of the j^{th} frame in the i^{th} syllable. The duration of the syllables is characteristic of a particular species, hence the number of frames within a syllable n_i is included as part of the model. Each recording can now be viewed as a sequence of syllables where the number of syllables within a fixed interval of time depends on the particular species. Mathematically, the data in an interval of sound may be represented as $\mathcal{D} = [\mathbf{x}(1), n_1, \ldots, \mathbf{x}(N), n_N, N]$ where N represents the number of syllables in a fixed interval. In this paper, we build a generative probability model for the syllables produced by each species (from the labeled training examples) and extend this probabilistic model to build Bayes-optimal classifiers for bird species identification.

Table I
Nomenclature

Variable	Description
m	Class index (bird species)
N	Number of syllables present in a fixed interval
θ_i	Syllable parametrization vector of i^{th} syllable
n_i	Number of frames in i^{th} syllable (length)
$\mathbf{x}(i)$	i^{th} syllable features, $[x_1(i)x_2(i)\cdots x_{n_i}(i)]$
$x_j(i)$	Feature representation of the j^{th} frame (i.e., j^{th} observation) in i^{th} syllable
$N_m^t(l)$	Number of syllables present in the l^{th} training interval from class m
N_m^t	Total number of training syllables from class m, $\sum_l N_m^t(l)$
K_m^t	Number of training set intervals from class m
K^t	Total number of training set intervals
$\mathcal{D}$	Data in an interval $[\mathbf{x}(1), n_1, \ldots, \mathbf{x}(N), n_N, N]$

Subscript m indicates class m, Superscript t indicates training set

III. Independent syllable model

We start with the syllable independence assumption common to both of our models. Figure 1 explains how an interval of recording may be generated using the independent syllable model. We assume that the syllables present in a recording are independent and identically distributed (i.i.d.) i.e., we select both a syllable parametrization vector θ_i (for the frame-level features) and length n_i for each syllable independent of other syllables. If we denote the class (species) by m, we have

$$P(\theta_1, n_1, \ldots, \theta_N, n_N, N|m) = P(N|m) \prod_{i=1}^{N} P(\theta_i, n_i|m),$$

where $P(\theta_i, n_i|m)$ represents the joint distribution of (θ_i, n_i) for class m and $P(N|m)$ represents the probability for the number of syllables per interval for class m. The syllable parametrization vector θ_i is a hidden parameter and cannot be observed directly. Instead, we observe the frame level features of the frames constituting that particular syllable, i.e., information about θ_i is available only through the observations $\mathbf{x}(i) = [x_1(i), x_2(i), \ldots, x_{n_i}(i)]$. The likelihood of a syllable $\mathbf{x}(i)$ can be found by marginalizing over the hidden variable θ_i as in

$$P(\mathbf{x}(i), n_i|m) = E_{\theta_i}\Big[P(\mathbf{x}(i)|\theta_i, n_i)\Big] P(n_i|m). \qquad (1)$$

Assuming that we have a collection of all the syllables produced by the different species of birds, the independent syllable model is quite similar to the 'bag-of-words' model used for document classification [8]. The probability of each syllable (word) depends on the particular species (topic) and hence, our inference method is quite similar to that employed in document classification. However, we include the length of the syllable n_i and the number of syllables present within an interval of recording N as part of our model. Even though

$$\underbrace{(\theta_1, n_1)}_{\text{Syllable 1}}, \underbrace{(\theta_2, n_2)}_{\text{Syllable 2}}, \dots$$

$$\underbrace{\{[x_1(1), \dots, x_{n_1}(1)], n_1\}}_{\text{Syllable 1}}, \underbrace{\{[x_1(2), \dots, x_{n_2}(2)], n_2\}}_{\text{Syllable 2}}, \dots$$

Figure 1. The independent syllable model

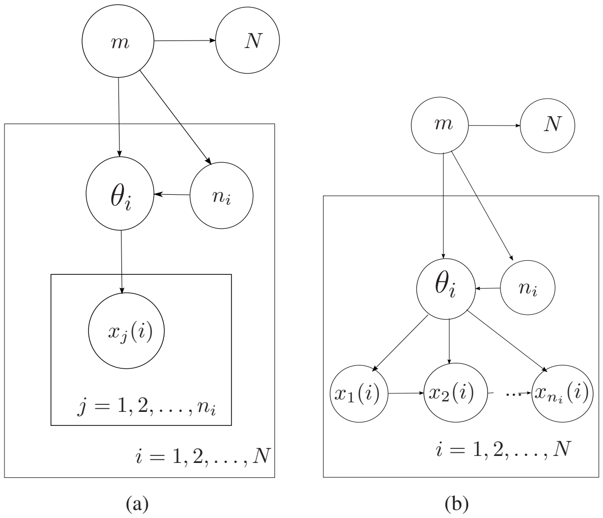

Figure 2. Graphical models of (a) the IFIS model and (b) the MCFIS model

two birds may vocalize in the same frequency range, the length of the syllable and number of syllables present within an interval are usually species-dependent and hence provide valuable information for identifying the different species of birds. Note that the independent syllable model does not specify how the frame level features within a syllable are generated in Eq. (1). Next, we consider two models for $P(\mathbf{x}(i)|\theta_i, n_i)$ namely IFIS and MCFIS.

A. IFIS model

In this model, each syllable is assumed to be an i.i.d. sequence of observations, i.e., $\mathbf{x}(i) = [x_1(i), x_2(i), \dots, x_{n_i}(i)]$ where each observation is drawn independently according to $p_x(x|\theta_i)$. Based on the graphical model for the IFIS model in Fig. 2(a), the likelihood that an interval belongs to class m is given by

$$P(\mathcal{D}|m) = \tag{2}$$
$$P(N|m) \prod_{i=1}^{N} E_{\theta_i}\Big[\prod_{j=1}^{n_i} p_x(x_j(i)|\theta_i)\Big| n_i, m\Big] P(n_i|N, m).$$

where $E_{\theta_i}[\cdot|n_i, m]$ represents marginalization over θ_i. The logarithm of Eq. (2) can be written as

$$\log P(\mathcal{D}|m) = C(\mathbf{X}) + \log P(N|m)$$
$$+ \sum_{i=1}^{N} \log \int_{\theta} e^{-n_i \hat{D}_{kl}(p_x(\cdot|\hat{\theta}) \| p_x(\cdot|\theta))} dF(\theta, n = n_i|m), \tag{3}$$

where we have used integral form of expectation and $\hat{\theta}$, $C(\mathbf{X})$ and $\hat{D}_{kl}$ are defined as

$$\begin{aligned} \hat{\theta} &= \arg\max_{\theta} \sum_{j=1}^{n_i} \log p(x_j(i)|\theta), \\ C(\mathbf{X}) &= \sum_{i=1}^{N} \sum_{j=1}^{n_i} \log p_x(x_j(i)|\hat{\theta}), \\ \hat{D}_{kl}(p_x(\cdot|\hat{\theta}) \| p_x(\cdot|\theta)) &= \frac{1}{n_i} \sum_{j=1}^{n_i} \log \frac{p_x(x_j(i)|\hat{\theta})}{p_x(x_j(i)|\theta)}. \end{aligned} \tag{4}$$

Due to space limitations, we omit the lengthy derivations leading to Eq. (3), and refer to [10] instead. By definition, $\hat{D}_{kl}$ is a non-negative quantity. $\hat{\theta}$ for a syllable can be obtained by maximizing the likelihood of frame level features conditioned on the other parameters, i.e., $P(\mathbf{x}(i)|n_i, N, m)$. We can also interpret $\hat{D}_{kl}$ as a sample estimate of the KL divergence $D_{kl}(p_x(\cdot|\hat{\theta}) \| p_x(\cdot|\theta))$. The log-likelihood model in Eq. (3) is very general in that it does not specify the precise form of the frame probability model $p_x(\cdot|\theta)$. Next, we will present a special case where the frame probability model is assumed to follow a Gaussian distribution.

Gaussian IFIS: Here, we assume that the frame probability model $p_x(x|\theta)$ follows a Gaussian distribution, i.e.,

$$p_x(x|\theta) = \frac{1}{\sqrt{\det 2\pi C}} e^{-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)}, \tag{5}$$

where $\theta = (\mu, C)$, i.e., the syllable parametrization vector is specified by the mean and covariance of the multivariate Gaussian distribution. $p_x(\cdot|\hat{\theta})$ is characterized as $p_x(\cdot|\hat{\mu}, \hat{C})$, where $\hat{\mu}$ and $\hat{C}$ are computed from the test data. $\hat{D}_{kl}$ can be written as follows [10]

$$\hat{D}_{kl}\big(p_x(\cdot|\hat{\mu}, \hat{C}) \| p_x(\cdot|\mu, C)\big) = \frac{1}{2}\Big(\log \frac{\det C}{\det \hat{C}} \tag{6}$$
$$+ \mathrm{tr}(C^{-1}\hat{C} - I) + (\hat{\mu} - \mu)^T C^{-1}(\hat{\mu} - \mu)\Big),$$

where the RHS of Eq. (6) equals the true KL divergence between two Gaussian distributions $\mathcal{N}(\hat{\mu}, \hat{C})$ and $\mathcal{N}(\mu, C)$ [11]. For the Gaussian case, the sample based estimate of the KL-divergence is shown to be equal to the actual KL-divergence between the estimated observation PDF and the model PDF. However, this may not be the case with other distributions. Note that if the test syllable is very similar to the training syllable, i.e., $\hat{\theta} \approx \theta$, the value of $\hat{D}_{kl}$ in Eq. (6) approaches zero, thus maximizing the log-likelihood in Eq. (3).

B. MCFIS model

In the previous model, we considered each syllable to be an i.i.d. sequence of observations. Doing so, we ignored any temporal structure within a syllable. For instance, the i.i.d. assumption does not capture the gradient increase or decrease in mean frequency between successive frames within a syllable. A simple method to incorporate temporal structure would be to model each syllable as a Markov chain of observations i.e., $p(x_j(i)|x_{j-1}(i), x_{j-2}(i), \ldots, x_1(i), \theta_i) = p(x_j(i)|x_{j-1}(i), \theta_i)$. Assuming that the first observation is generated according to a probability distribution $p_x(\cdot|\theta)$ and denoting the conditional distribution by $p_x(x_j(i)|x_{j-1}(i), \theta_i)$, the likelihood of the i^{th} syllable can be written as product of the likelihood of the first frame and the conditional likelihood of the remaining frame-level features, i.e.,

$$P(\mathcal{D}|m) = P(N|m) \prod_{i=1}^{N} \Big(P(n_i|N, m) \tag{7}$$

$$\cdot E_{\theta_i} \Big[p_x(x_1(i)|\theta_i) \prod_{j=2}^{n_i} p_x(x_j(i)|x_{j-1}(i), \theta_i) \Big| n_i, m \Big] \Big).$$

For tractability, we assume that $p_x(x_1(i)|\theta_i)$ follows an uniform distribution. Since $p_x(x_1(i)|\theta_i)$ is uniformly distributed, it is irrelevant for classification and hence we proceed with the conditional likelihood for the MCFIS model given by

$$P(\mathcal{D}|m) = P(N|m) \prod_{i=1}^{N} \Big(P(n_i|N, m)$$

$$\cdot E_{\theta_i} \Big[\prod_{j=2}^{n_i} p_x(x_j(i)|x_{j-1}(i), \theta_i) \Big| n_i, m \Big] \Big). \tag{8}$$

The logarithm of Eq. (8) can be written as follows [10]

$$\log P(\mathcal{D}|m) = C(\mathbf{X}) + \log P(N|m) \tag{9}$$

$$+ \sum_{i=1}^{N} \log \int_{\theta} e^{-(n_i-1)\hat{D}_{kl}(\hat{\theta}\|\theta)} dF(\theta, n_i|m),$$

where we have used the integral form of expectation and $\hat{\theta}$, $C(\mathbf{X})$ and $\hat{D}_{kl}$ are defined as

$$\begin{aligned} \hat{\theta} &= \arg\max_{\theta} \sum_{j=2}^{n_i} \log p_x(x_j(i)|x_{j-1}(i), \theta), \\ C(\mathbf{X}) &= \sum_{i=1}^{N} \sum_{j=2}^{n_i} \log p_x(x_j(i)|x_{j-1}(i), \hat{\theta}), \qquad (10) \\ \hat{D}_{kl}(\hat{\theta}\|\theta) &= \frac{1}{n_i - 1} \sum_{j=2}^{n_i} \log \frac{p_x(x_j(i)|x_{j-1}(i), \hat{\theta})}{p_x(x_j(i)|x_{j-1}(i), \theta)}. \end{aligned}$$

$\hat{\theta}$ for a syllable can be obtained by maximizing the likelihood $P(\mathbf{x}(i)|n_i, N, m)$. Note that this likelihood involves Markov dependency between successive frame level features in the MCFIS model. In the MCFIS case, we can interpret $\hat{D}_{kl}(\hat{\theta}\|\theta)$ as a sample estimate of the KL divergence between the conditional distributions $p(x_j(i)|x_{j-1}(i), \hat{\theta})$ and $p(x_j(i)|x_{j-1}(i), \theta)$.

Gaussian MCFIS: Here, we consider the frame probability model in Eq. (7) to be Gaussian, i.e., $\theta = (\tilde{\mu}, \tilde{C})$,

$$\begin{bmatrix} x_j(i) \\ x_{j+1}(i) \end{bmatrix} \sim \mathcal{N}(\tilde{\mu}, \tilde{C}),\ \tilde{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix},\ \text{and}\ \tilde{C} = \begin{bmatrix} C_{11} & C_{12} \\ C_{12}^T & C_{22} \end{bmatrix}.$$

Note that in this case, the syllable parametrization vector θ includes the cross-covariance between consecutive frame level features C_{12}. For a multivariate Gaussian distribution, the conditional distribution $p_x(x_j(i)|x_{j-1}(i), \theta)$ is also Gaussian

$$p_x(x_j(i)|x_{j-1}(i), \theta) = \frac{1}{\sqrt{\det 2\pi C_c}} e^{-\frac{1}{2}(x_j(i) - \mu_{j|j-1}) C_c^{-1} (x_j(i) - \mu_{j|j-1})}, \tag{11}$$

where $\mu_{j|j-1} = \mu_2 + C_{12}^T C_{11}^{-1}(x_{j-1}(i) - \mu_1)$ and $C_c = C_{22} - C_{12}^T C_{11}^{-1} C_{12}$. If we assume the distribution of frame level features to be stationary within a syllable, we have $\mu_1 = \mu_2 = \mu$, $C_{22} = C_{11}$, and the model is characterized by $\theta = (\mu, C_{11}, C_{12})$. Here, we select $\hat{\theta}$ by maximizing the conditional likelihood as $\hat{\theta}_{ML} = (\hat{\mu}_{ML},\ \hat{C}_{11_{ML}},\ \hat{C}_{12_{ML}})$. For a test syllable $\mathbf{x}(i) = [x_1(i),\ x_2(i),\ \ldots,\ x_{n_i}(i)]$, the conditional ML solutions for $\hat{\mu}$, $\hat{C}_{11}$, and $\hat{C}_{12}$ are given by [10]

$$\begin{aligned} \hat{\mu}_{ML} &= (I - \hat{M})^{-1}(\hat{\mu}_2 - \hat{M}\hat{\mu}_1) \\ \hat{C}_{11_{ML}} &= \sum_{k=0}^{\infty} \hat{M}^k \hat{C}_c (\hat{M}^k)^T \\ \hat{C}_{12_{ML}} &= \hat{C}_{11_{ML}} \hat{M}^T \end{aligned}$$

where $\hat{M} = \hat{C}_{12}^T \hat{C}_{11}^{-1}$, $\hat{\mu}_1 = \frac{1}{n_i-1} \sum_{j=1}^{n_i-1} x_j(i)$, $\hat{\mu}_2 = \frac{1}{n_i-1} \sum_{j=1}^{n_i-1} x_{j+1}(i)$, $\hat{C}_{11} = \frac{1}{n_i-1} \sum_{j=1}^{n_i-1} (x_j(i) - \mu_1)(x_j(i) - \mu_1)^T$, $\hat{C}_{12} = \frac{1}{n_i-1} \sum_{j=1}^{n_i-1} (x_j(i) - \mu_1)(x_{j+1}(i) - \mu_2)^T$, $\hat{C}_c = \hat{C}_{22} - \hat{C}_{12}^T \hat{C}_{11}^{-1} \hat{C}_{12}$, and $\hat{C}_{22} = \frac{1}{n_i-1} \sum_{j=1}^{n_i-1} (x_{j+1}(i) - \mu_2)(x_{j+1}(i) - \mu_2)^T$. During the training phase, the model parameters μ, C_{11}, C_{12} can be estimated in a similar fashion in terms of $\hat{\mu}_1^t, \hat{\mu}_2^t, \hat{C}_{11}^t, \hat{C}_{12}^t, \hat{C}_{22}^t$. Substituting these ML estimates in Eq. (10), $\hat{D}_{kl}$ for MCFIS model can be written in the following form [10]

$$\hat{D}_{kl}(\hat{\theta}\|\theta) = \frac{1}{2} \Big(\log \frac{\det \hat{C}_c^t}{\det \hat{C}_c} + \mathrm{tr}(\hat{C}_c^{-1^t} \hat{C}_c - I)$$

$$+ \mathrm{tr} \Big[\hat{C}_c^{-1^t} (\hat{M}^t - \hat{M}) \hat{C}_{11} (\hat{M}^t - \hat{M})^T \Big]$$

$$+ (\Delta\hat{\mu}_2 - \hat{M}^t \Delta\hat{\mu}_1)^T C_c^{-1} (\Delta\hat{\mu}_2 - \hat{M}^t \Delta\hat{\mu}_1) \Big), \tag{12}$$

where $\Delta\hat{\mu}_2 = \hat{\mu}_2 - \hat{\mu}_2^t$ and $\Delta\hat{\mu}_1 = \hat{\mu}_1 - \hat{\mu}_1^t$. Note that if the test syllable is very similar to the training syllable, i.e., $\hat{\theta} \approx \theta$, the value of $\hat{D}_{kl}$ in Eq. (12) approaches zero, thus maximizing the log-likelihood in Eq. (9).

IV. Classification and Training

We consider the Bayes risk minimizer of the probability of error for classification. Hence, our classifier is the maximum-a-posteriori (MAP) rule [7]:

$$\hat{m} = \arg\max_{m} P(\mathcal{D}|m)P(m), \tag{13}$$

which is equivalent to the maximization of the posterior $P(m|\mathcal{D})$. We proceed with a log version of the MAP rule:

$$\hat{m} = \arg\max_{m} \log P(\mathcal{D}|m) + \log P(m).$$

Next, we proceed with the evaluation of the MAP criterion for the IFIS and MCFIS models.

A. IFIS model

To obtain the MAP criterion for the IFIS model, Eq. (3) is substituted into Eq. (14), yielding

$$\begin{aligned}\max_{m} \quad & \log P(m) + \log P(N|m) \\ & + \sum_{i=1}^{N} \log \int e^{-n_i \hat{D}_{kl}(p_x(\cdot|\hat{\theta}) \| p_x(\cdot|\theta))} p(\theta, n_i|m) d\mu(\theta). \end{aligned} \tag{14}$$

Typically, the models $p(\theta, n|m)$, $P(N|m)$, $P(m)$ in Eq. (14) are not available. We propose to estimate them from training samples in a non-parametric fashion. To estimate $p(\theta, n|m)$, we follow the kernel density estimation approach. Since only a small number of samples are available for a given n (or potentially zero), we employ smoothing via the kernel $q(n|n(k,m))$ in our estimator:

$$\hat{p}(\theta, n|m) = \frac{1}{N_m^t} \sum_{k=1}^{N_m^t} q(n|n(k,m)) \delta(\theta - \theta(k,m)), \tag{15}$$

where N_m^t denotes number of training syllables from class m and $\theta(k,m)$, $n(k,m)$ respectively denote the syllable parametrization vector and length of the k^{th} training syllable from class m. The estimator $\hat{p}(\theta, n|m)$ is essentially a weighted average of all the training syllables from class m where the weight $q(n|n(k,m))$ accounts for the syllable length similarity. Next, we estimate the class prior probability via the following ratio of counts

$$\hat{P}(m) = \frac{K_m^t}{K^t}, \tag{16}$$

where K_m^t denotes the number of training set intervals from class m and K^t denotes the total number of training set intervals. Finally, we estimate the class conditional probability for the number of syllables within an interval using the kernel density estimator

$$\hat{P}(N|m) = \frac{1}{K_m^t} \sum_{j=1}^{K_m^t} q_k(N|N(j,m)), \tag{17}$$

where $q_k(\cdot|\cdot)$ is the kernel and $N(j,m)$ denotes the number of syllables in the j^{th} training interval from class m. Substituting these estimated models $\hat{p}(\theta, n|m)$, $\hat{P}(N|m)$, $\hat{P}(m)$ into Eq. (14), we obtain the following MAP criterion [10]

$$\begin{aligned}\min_{m} \quad & -\log \hat{P}(m) - \log \hat{P}(N|m) + N \log \frac{N_m^t}{N^t} \\ & + \sum_{i=1}^{N} n_i d((\hat{p}_{x_i}, n_i) \| (\theta^{(1,i,m)}, n^{(1,i,m)})) \\ & - \sum_{i=1}^{N} \log(1 + \sum_{k=2}^{N_m^t} e^{-n_i \partial d((\hat{p}_{x_i}, n_i) \| (\theta^{(1,i,m)}, n^{(1,i,m)})}) \end{aligned} \tag{18}$$

where $d((\theta_1, n_1) \| (\theta_2, n_2))$ measures a divergence between the syllable parametrization vector and length of one syllable to those of another by

$$d((\theta_1, n_1) \| (\theta_2, n_2)) = \hat{D}_{kl}(\theta_1 \| \theta_2) + d_q(n_1, n_2), \tag{19}$$

where $d_q(n_1, n_2)$ is a non-negative divergence for comparing syllables lengths and is given by

$$d_q(n_1, n_2) = \frac{1}{n_1} \log \frac{q(n_1|n_1)}{q(n_1|n_2)}. \tag{20}$$

Also, $\partial d((\hat{p}_{x_i}, n_i) \| (\theta^{(k,i,m)}, n^{(k,i,m)}))$ in (18) is given by

$$\begin{aligned}\partial d((\hat{p}_{x_i}, n_i) \| (\theta^{(k,i,m)}, n^{(k,i,m)})) = & \\ d((\hat{p}_{x_i}, n_i) \| (\theta^{(k,i,m)}, n^{(k,i,m)})) & \\ -d((\hat{p}_{x_i}, n_i) \| (\theta^{(1,i,m)}, n^{(1,i,m)})). & \end{aligned}$$

Note the use of order statistics notation $(\theta^{(1,i,m)}, n^{(1,i,m)})$ to denote the nearest neighbor for the i^{th} test syllable amongst all the training examples from class m. Consider the MAP criterion in Eq. (18) as a sum of five terms. The first term accounts for the fact that the number of intervals from different training classes might not be equal. If all classes have equal number of training intervals, the first term becomes a constant and therefore is irrelevant to the classification. The second term accounts for the fact that the number of syllables within a fixed interval is species-dependent. The last three terms correspond to the likelihood of the observations. The last term accounts for the contribution due to training syllables other than the nearest neighbor from class m. For large n_i, the last term becomes negligible. If we consider the contribution of the nearest neighbor alone, the MAP classifier reduces to

$$\begin{aligned}\min_{m} \quad & -\log \hat{P}(m) - \log \hat{P}(N|m) + N \log \frac{N_m^t}{N^t} \\ & + \sum_{i=1}^{N} n_i d((\hat{p}_{x_i}, n_i) \| (\theta^{(1,i,m)}, n^{(1,i,m)})) \end{aligned} \tag{21}$$

In the Gaussian IFIS case, $\hat{D}_{kl}$ in Eq. (19) is replaced by $\hat{D}_{kl}$ from Eq. (6). Due to the nearest-neighbor nature of the IFIS classifier in Eq. (21), the training process for IFIS involves only the computation of mean, covariance, and length of frame level features for each syllable in the training set.

B. MCFIS model

Starting with the conditional log-likelihood as defined in Eq. (9), the MAP classification rule for MCFIS can be derived in a similar fashion to the IFIS model [10]. Substituting $\hat{D}_{kl}$ as defined in Eq. (12) for $\hat{D}_{kl}$ in Eq. (19), we obtain

$$d((\theta_1, n_1)\|(\theta_2, n_2)) = \hat{D}_{kl}(\theta_1\|\theta_2) + d_q(n_1, n_2).$$

The MCFIS MAP classifier is based on a nearest neighbor rule. Hence, the training process for MCFIS involves the computation of $\hat{\mu}_1^t$, $\hat{\mu}_2^t$, $\hat{C}_{11}^t$, $\hat{C}_{12}^t$, $\hat{C}_{22}^t$, and the length of each syllable in the training set. Next, we provide a numerical evaluation of the proposed classifiers.

V. Numerical Results

In this section, we describe the experimental setup used to measure the classification error rates obtained by the proposed classifiers. We first describe the implementation details of our experimental setup and then discuss the results.

A. Implementation details

Data We used recordings from the Cornell Macaulay library, of 6 species as described in Table II. All of the recordings are 44.1 kHz PCM WAV files. We manually removed portions of recordings which contain human voices, then divide each recording into intervals of 30 seconds, resulting in 426 intervals.

Segmentation of the audio recording To compute spectrograms for the recordings, we divide sounds into frames of size = 512 samples, with 93.75% overlap between successive frames, then apply a Hamming window of the same size followed by a 512-point FFT to obtain the magnitude spectrum of the frame. To remove background noise, we consider only frequency bins in the range 1000-8000 Hz. Next, we compute the KL divergence between the normalized power spectral density (PSD) of each frame and the uniform distribution. We use the locations of local minima of the KL divergence to determine boundaries between elements. The regions within the boundaries are treated as elements and the energy of each element is computed. We then apply an adaptive thresholding algorithm similar to that used in [3]. Only the elements with energy greater than this threshold are treated as syllables and used for further processing.

Frame level features After segmentation, the next task is to compute frame level features. We choose to use mean frequency and bandwidth as the frame level features. We treat the normalized PSD as a probability density function and compute the mean and variance as our mean frequency and BW. We compute the energy of each frame and build a cumulative density function (CDF) for energy of the frames within a syllable. Only frames that fall within a 95% of the CDF are used to compute the mean and covariance matrix for the syllable. We used a Poisson probability mass function (PMF) to perform kernel density estimation in both Eq. (17) and Eq. (15).

SVM setup SVM has been successfully employed for classifying bird syllables individually [3]. To classify a sound interval, we used SVM to classify each of the individual syllables and performed a majority vote based on these individual SVM decisions. We employed the one vs. one strategy for multiclass SVM as in [3]. The Matlab SVM implementation SVM-KMToolbox [12] was used for our simulation purposes. Six features: 2 mean vectors, 3 unique entries from covariance matrix and syllable length were used and the features were normalized to lie in the range -1 to 1. Due to MATLAB memory limitations, we randomly subsampled five sets, each containing 4000 syllables, thus choosing 20,000 syllables totally from the entire training set. Five SVM classifiers were obtained by training one on each subsampled training set. The final SVM decision was computed by adding the votes of all the five SVM classifiers. To prevent issues due to unbalanced training set, we ensured that each species constitutes at least 10% of the syllables in each of the subsampled training sets. Gaussian kernel was employed and the parameters of SVM (kernel width and regularization parameter) were optimized by performing a grid search.

Cross Validation To measure the accuracy of the proposed classifiers, we used them to predict the species in each 30 second interval of sound. Our cross validation setup is similar to that in [1]. If the test interval belongs to recording R, all the syllables from the recording R are excluded from the training set. This ensures that the training set does not include any syllables from the individual being classified.

Table II
Comparison of correct classification rates for the IFIS, MCFIS, and SVM classifiers on the six species dataset

Species Name	Number of intervals	Number of syllables	SVM %	IFIS %	MCFIS %
Winter Wren	73	23471	100.0 ± 0.0	100.0 ± 0.0	97.26 ± 1.9
Swainsons Thrush	50	5635	46.00 ± 7.0	96.00 ± 2.7	88.00 ± 4.6
Black throated Blue Warbler	43	1411	88.37 ± 4.9	83.72 ± 5.6	81.40 ± 5.9
Black capped Chickadee	56	6845	71.43 ± 6.0	55.36 ± 6.6	73.21 ± 5.9
Downy Woodpecker	114	52075	91.23 ± 2.6	92.98 ± 2.3	96.49 ± 1.7
Western Tanager	90	12877	98.89 ± 1.1	91.11 ± 3.0	94.45 ± 2.4
Overall accuracy			86.15 ± 1.6	88.26 ± 1.5	90.61 ± 1.4

B. Comparison of classifiers with SVM

The results of our experiments are summarized in Table II. The table contains the accuracy (correct classification rate) in % for the IFIS, MCFIS models and SVM along with the standard deviation due to cross-validation. SVM achieves an overall accuracy of 86.15% while the IFIS model and MCFIS model produce an overall accuracy of 88.26% and 90.61% respectively. We observe that IFIS and MCFIS models perform better than the equivalent SVM implementation. Training SVMs is very time-intensive and requires parameter optimization. Care should be taken to ensure that subsampling is done appropriately whereas no such intensive training or parameter optimization is required for our models. In this work, we used just mean frequency and bandwidth as the features. If the different species differ in their vocalization frequency, the IFIS model will produce good results with these features. If the frequency range of vocalization is similar for two different species, but the nature of syllables is different, the MCFIS model can outperform the IFIS model. Alternatively, one could use features that capture temporal structure in conjunction with the IFIS model.

VI. Conclusion

In this paper, we introduced the independent syllable model along with two methods of aggregating frame level features within syllables, namely, the IFIS and MCFIS models. We derived the MAP classifiers for each model and showed that it can be approximated using the nearest neighbor approach. The training process is simple and involves only the computation of the syllable parametrization vectors of all the syllables in training set. We numerically evaluated the performance the classifiers proposed and show that our models provide competitive classification rates. The independent syllable model does not capture temporal structure across syllables. This can be incorporated by assuming a Markov model for the distribution of syllables within an interval. We believe that the combination of frame level and syllable level temporal structure could further improve upon the classification rates presented here.

References

[1] P. Somervuo, A. Harma, and S. Fagerlund, "Parametric representations of bird sounds for automatic species recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2252–2263, 2006.

[2] C. Catchpole and P. J. B. Slater, *Bird song: biological themes and variations*. Cambridge University Press, 1995.

[3] S. Fagerlund, "Bird species recognition using support vector machines," *EURASIP J. Appl. Signal Process.*, vol. 2007, no. 1, pp. 64–64, 2007.

[4] S. E. Anderson, A. S. Dave, and D. Margoliash, "Template-based automatic recognition of birdsong syllables from continuous recordings," *Journal of the Acoustical Society of America*, vol. 100, no. 2, pp. 1209–1219, 1996.

[5] A. L. McIlraith and H. C. Card, "Birdsong recognition using backpropagation and multivariatestatistics," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2740–2748, 1997.

[6] A. Selin, "Wavelets in recognition of bird sounds," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. 1–9, 2007.

[7] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc. Secaucus, NJ, USA, 2006.

[8] 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*. ACM New York, NY, USA, 2006, pp. 289–296.

[9] Y. W. Teh and M. I. Jordan, *Hierarchical Bayesian Nonparametric Models with Applications*. Cambridge University Press, To Appear. [Online]. Available: http://www.stat.berkeley.edu/tech-reports/770.pdf

[10] B. Lakshminarayanan, X. Fern, and R. Raich, "A Syllable-Level Probabilistic Framework for Bird Species Identification," Oregon State university, Tech. Rep., 2009, in preparation.

[11] J. R. Hershey and P. A. Olsen, "Approximating the Kullback Leibler divergence between Gaussian mixture models," in *Proc. ICASSP*, vol. 4, 2007, pp. 317–320.

[12] S. Canu, Y. Grandvalet, V. Guigue, and A. Rakotomamonjy, "Svm and kernel methods matlab toolbox," Perception Systmes et Information, INSA de Rouen, Rouen, France, 2005.