Dynamic time warping and sparse representation classification for birdsong phrase classification using limited training data[a)]

Lee N. Tan[b)] and Abeer Alwan
Department of Electrical Engineering, University of California, Los Angeles 56-125B Engineering IV Building, Box 951594, Los Angeles, California 90095

George Kossan, Martin L. Cody, and Charles E. Taylor
Department of Ecology and Evolutionary Biology, University of California, Los Angeles 621 Charles Young Drive South, Los Angeles, California 90095

(Received 11 July 2014; revised 26 November 2014; accepted 17 December 2014)

Annotation of phrases in birdsongs can be helpful to behavioral and population studies. To reduce the need for manual annotation, an automated birdsong phrase classification algorithm for limited data is developed. Limited data occur because of limited recordings or the existence of rare phrases. In this paper, classification of up to 81 phrase classes of Cassin's Vireo is performed using one to five training samples per class. The algorithm involves dynamic time warping (DTW) and two passes of sparse representation (SR) classification. DTW improves the similarity between training and test phrases from the same class in the presence of individual bird differences and phrase segmentation inconsistencies. The SR classifier works by finding a sparse linear combination of training feature vectors from all classes that best approximates the test feature vector. When the class decisions from DTW and the first pass SR classification are different, SR classification is repeated using training samples from these two conflicting classes. Compared to DTW, support vector machines, and an SR classifier without DTW, the proposed classifier achieves the highest classification accuracies of 94% and 89% on manually segmented and automatically segmented phrases, respectively, from unseen Cassin's Vireo individuals, using five training samples per class.
 [http://dx.doi.org/10.1121/1.4906168]

[MRB] Pages: 1069–1080

I. INTRODUCTION

Bird vocalizations are compositions of short individual units or syllables, each generally lasting less than a second. Sound recordings of bird vocalizations are helpful in behavioral and population studies (Brandes, 2008; Mennill, 2011), especially in dense vegetation environment where visual identification is difficult. Typically, longer and more elaborate songs are used for mate attraction and territory declaration; shorter calls are often used for family member contacts and identification, predator announcement, or food information communication (Catchpole and Slater, 2008). Automated bird song classification has already proved useful for species identification (Härmä, 2003; Frommolt *et al.,* 2008; Trifa *et al.,* 2008; Agranat, 2009; Graciarena *et al.,* 2011), individual recognition (Kirschel *et al.,* 2011), and classification of particular syllables or phrases expressed by birds with complex song structures (Kogan and Margoliash, 1998; Ranjard and Ross, 2008; Hansson-Sandsten *et al.,* 2011; Sasahara *et al.,* 2012; Tachibana *et al.,* 2014). Much of this research has been reviewed by Brandes (2008) and Mennill (2011). Such applications will gain importance with an increasing general interest in "soundscape ecology" (Pijanowski *et al.,* 2011; Stowell *et al.,* 2013; Glotin *et al.,* 2013). In this paper, the task of interest is automated classification of birdsong phrases of the Cassin's Vireo (*Vireo cassinii*) species. Syllable/phrase level birdsong annotations can facilitate studies to better understand bird communication, including song syntax (Berwick *et al.,* 2011; Briefer *et al.,* 2013) and common phrases used for conveying certain information (e.g., mate attraction and territory declaration).

Bird call or song element classification becomes especially challenging when the song repertoire is diverse; some species have thousands of distinct phrases in their lexicons (Catch-pole and Slater, 2008). In our experience, the frequencies at which individual bird song elements are observed often resembles a Zipf–Mandelbrot distribution (Silagadze, 1997; McCowan *et al.,* 2005), where some phrases are heard many times, while others are rare. Communication in other species, including humans, show similar disparity in word or other language unit usage; thus a premium is placed on the ability of automated classifiers to correctly classify bird song phrases using only a few training samples per phrase. Further, the amount of training data available may be limited by the logistics of bird song recordings in certain geographical locations of interest and by the labor intensity of human experts to identify phrase types in complex songs. The ability of an automated classifier to perform accurate phrase classification with limited data will reduce the need for human experts to perform manual annotation on new recordings.

[a)]Portions of this work were presented in "Evaluation of a sparse representation-based classifier for bird phrase classification under limited data conditions," Proceedings of Interspeech, Portland, OR, Sept. 2012.

[b)]Author to whom correspondence should be addressed. Electronic mail: tleengee.ee.ucla@gmail.com

Since bird phrase classification is similar to automatic speech recognition (ASR), techniques that were proposed for ASR have been applied to bird songs. One example is dynamic time warping (DTW) template-based techniques. DTW has also been used for classifying birdsong elements and achieves good classification performance in birdsong elements (Anderson *et al.*, 1996; Kogan and Margoliash, 1998; Kaewtip *et al.*, 2013; Meliza *et al.*, 2013) with a few templates per class. Since a DTW classifier derives its class decision from the template that is most similar to the test sample after time warping, its performance degrades when the individual templates are not able to cover within-class variations found in the test set. Hidden Markov models (HMMs) (Kogan and Margoliash, 1998; Chu and Blumstein, 2011) and neural networks (Ranjard and Ross, 2008) have also been successfully applied to birdsong element recognition. The techniques can accumulate more information and generalize better than template-based techniques, but they generally require a large amount of data to appropriately train model parameters. Support vector machines (SVM) have also demonstrated good classification performance on birdsongs (Fagerlund, 2007; Sasahara *et al.*, 2012; Tachibana *et al.*, 2014) as well as in other domains with limited training data (Zhang *et al.*, 2001). However, under very limited training data conditions (with three or less training samples per class), over-fitting tends to occur because almost all the training samples are used to define the SVM's maximum margin hyperplane. Besides the various classification techniques, different features have also been used to represent birdsong segments (containing multiple syllables) or individual birdsong elements. These include using the time-frequency spectrogram explicitly (Anderson *et al.*, 1996; Kogan and Margoliash, 1998; Neal *et al.*, 2011; Tan *et al.*, 2012; Kaewtip *et al.*, 2013), frequency and energy trajectories (Härmä, 2003; Chen and Maher, 2006), Mel-frequency cepstral coefficients (MFCCs) (Kogan and Margoliash, 1998; Trifa *et al.*, 2008; Graciarena *et al.*, 2011), and other spectrographic image-based features (Lee *et al.*, 2013).

In our previous work (Tan *et al.*, 2012), we introduced an exemplar-based sparse representation (SR) classification technique for birdsong phrase classification. This relatively new SR classification technique was first proposed for human face recognition, where it achieves a high accuracy with just seven facial images per individual (Yang *et al.*, 2007). Each exemplar is a feature vector extracted from a training sample. The SR classifier seeks to represent the test feature vector by a sparse linear combination of exemplars, which is found by solving a l_1 minimization problem. In Tan *et al.* (2012), the features are explicitly derived from the spectrogram of each phrase segment. A spectrogram with a fixed number of time frames is generated for each phrase segment by varying the inter-frame overlap (the shorter the segment, the larger the amount of overlap). Since manual phrase identification is performed via spectrogram inspection, we used features that are explicitly derived from spectrograms to avoid hand-crafting features that can discriminate phrase classes with very similar dominant frequency trajectories. Sparsity-based techniques that rely on time-frequency acoustic energies in spectrograms as features have reported good performance in speech applications (Gemmeke *et al.*, 2011; Yilmaz *et al.*, 2013). The spectrogram-based feature extraction framework in Tan *et al.* (2012) also yields good classification performance on manually segmented phrases from the same bird individuals.

In this paper, an improved classification algorithm that combines DTW with a two-pass SR classification (abbreviated as DTW-SR-2pass) is proposed. Instead of varying the amount of inter-frame overlap in spectrograms, the inter-frame overlap is kept constant, and DTW is used to produce time-warped training phrase spectrograms that have the same number of frames as the test phrase spectrogram. Spectrogram normalization and dimension reduction are performed to obtain the feature vectors for SR classification. This paper is organized as follows. The Cassin's Vireo (CAVI) database is described in Sec. II. Separate test sets are formed to facilitate performance evaluation on manually or automatically segmented phrases from bird individuals present or absent in the training set. Section III presents technical details of the proposed DTW-SR-2pass classifier, while Sec. IV describes the algorithms used for comparative purposes. Finally, classification results, discussion, and conclusion can be found in Secs. V–VII, respectively.

II. THE CAVI DATABASE

The Cassin's Vireo (CAVI) species is found commonly in many coniferous and mixed-forest bird communities in far western North America. Birdsong phrases for classification were obtained from song recordings of male CAVIs because only the males of this species give full songs. Their songs have 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…" (Goguen and Curson, 2002). Figure 1 shows the spectrogram of a CAVI song segment containing two different phrases, each consisting of two syllables.

Manual phrase annotation was done by two human expert annotators. Phrase identity and time boundaries of each phrase in the song were annotated based on visual spectrogram inspection using the PRAAT software (Boersma and Weenink, 2011). Phrase types were categorized based on their frequency trajectories on spectrograms, and the label of

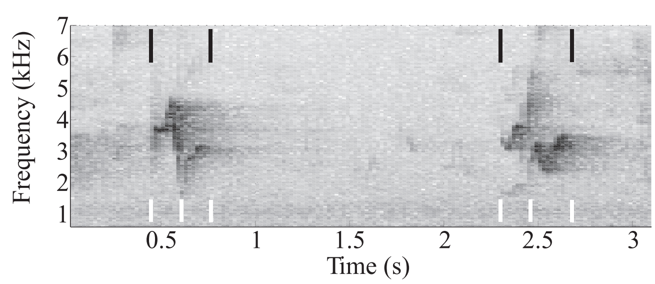

FIG. 1. A spectrogram of a Cassin's Vireo song segment. The phrase boundaries are marked by black lines, while the syllable boundaries are marked by white lines.

the phrase was assigned to the subjectively matching spectrogram in the CAVI phrase catalogue. This phrase catalogue was built up from scratch. Whenever a phase segment with a subjectively different frequency trajectory from the existing spectrograms was found, its spectrogram was added to the catalogue, and a new phrase label was created.

The song recordings were conducted in a mixed conifer-oak forest at approximately 800 m elevation (38°29′04″N, 120°38′04″W), near the city of Volcano in California. Songs were recorded using a Marantz PMD 670 portable compact flash audio recorder and a Sennheiser omni-directional microphone with a Telinga parabolic reflector. Each sound file (WAV-format, 16-bit, mono, 44.1 kHz sampling rate, with an average file duration of 3.66 min) contained songs from a single CAVI, with occasional songs/calls of other species in the background. One or more files were recorded per CAVI individual. Two separate data collections were performed. The first collection was done between April and June 2010 when two different males, denoted by CAVI1 and CAVI2, on adjacent territories were recorded. The second data collection was done between April and August 2012, and song recordings from four territorial males, denoted by CAVI3–CAVI6, were used in this study. All recordings with some metadata, and the phrase catalogue are available online at http://taylor0.biology.ucla.edu/birdDBQuery/. Figures 2(a)–2(f) show the distribution of the phrases for each CAVI (the total recording duration is noted in each subfigure header). Each phrase spans between 0.12 and 1.25 s, with a mean duration of 0.36 s. Each individual CAVI has 10–55 unique phrase types (which is dependent on the number of songs recorded from the individual CAVI), and 101 unique phrase classes are observed in the combined CAVI dataset.

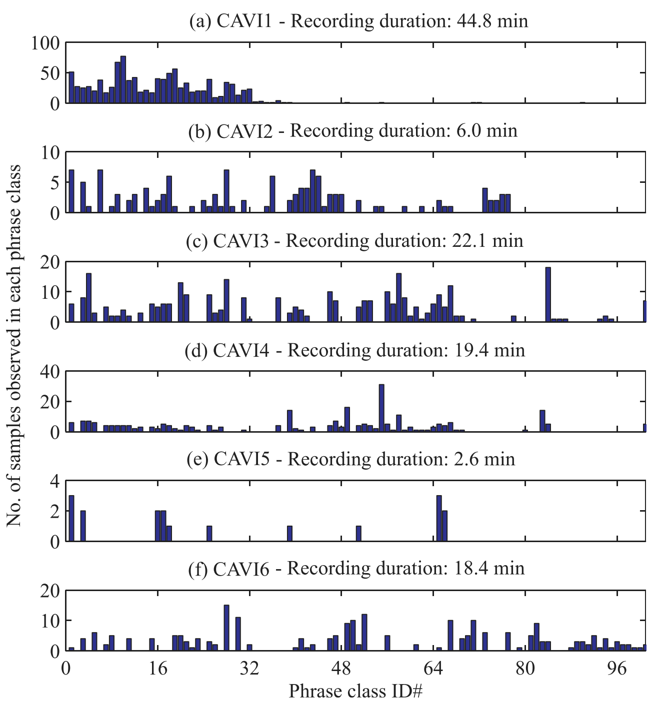

FIG. 2. (Color online) Number of samples observed in each phrase class from each of the six CAVI individuals.

A. Training and test sets

CAVI3, CAVI4, and CAVI5 in the 2012 collection are selected as the "training CAVIs" for their variety of phrase classes, so that a larger number of phrase classes is available to train the classifier. The remaining CAVIs—CAVI1, CAVI2, and CAVI6—are the "test CAVIs." The training phrase segments are extracted based on the manual annotations. Phrase classes with at least n samples or tokens from the training CAVIs are used for training. Let K be the total number of classes that satisfies this condition. To evaluate classification performance on manually/human-segmented and automatically/machine-segmented phrases from the training and test CAVIs, four separate test sets are formed. Test sets A1 and A2 contain human-segmented and machine-segmented phrases (in the K classes) from the training CAVIs, respectively, that are not used for training. Test sets B1 and B2 contain human- and machine-segmented phrases (in the K classes) from the test CAVIs, respectively. Table I shows the value of K and the number of phrase samples in the training or test sets for each value of $n = 1, 2, \ldots, 5$. Note that the number of training phrase samples is equal to $n \times K$.

The automatic phrase segmentation algorithm described in Härmä (2003) is used to obtain the machine-segmented test phrases in test sets A2 and B2. This time-frequency energy-based bird syllable detection and segmentation algorithm is also used in Lee *et al.* (2006) and Chen and Li (2013). An online MATLAB code (Lindermuth, 2010) of this algorithm is modified so that more than 90% of human-segmented phrase segments are detected, and most of the detected time boundaries fall within 100 ms of the human-segmented ones. The global and local energy input parameters to the MATLAB code are both set to 15 dB. This detects every segment with a local-maximum spectrographic amplitude L_m that is greater than $G_m - 15$ dB. $G_m = \max_{f,t} 10 \log_{10} \|S(f,t)\|^2$ is the global-maximum time-frequency value of input power spectrogram $\|S(f,t)\|^2$, where f and t denote the frequency and frame indices, respectively. Each detected segment's frame boundaries are the last consecutive frames (prior and beyond the frame containing L_m) with spectrographic amplitudes that are at least $L_m - 15$ dB.

TABLE I. The total number of phrase samples in the training and test sets as the training samples per class n varies from 1 to 5. These samples belong to one of the K phrase classes that has at least n samples from the training CAVIs. "Training" refers to the training set. Tests A1 and B1 contain human-segmented phrases from the training and test CAVIs, respectively. Tests A2 and B2 contain machine-segmented phrases from the training and test CAVIs, respectively.

		No. of phrase samples in				
n	K	Training	Test A1	Test B1	Test A2	Test B2
1	81	81	419	1269	396	1218
2	67	134	352	1208	336	1159
3	61	183	291	1138	277	1049
4	50	200	241	1061	228	1021
5	43	215	198	821	190	788

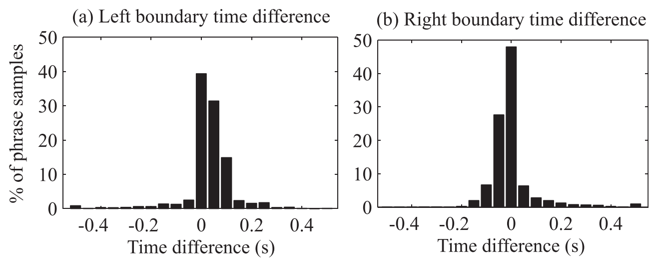

FIG. 3. Histograms of time boundary differences between machine-segmented and human-segmented phrases. (a) and (b) show the histograms of left and right boundary time differences, respectively.

The algorithm is run separately for every 10-s segment of each sound file to improve detection and segmentation accuracy of this algorithm in the presence of energy variations across the sound file. Two successive detected syllables are merged into a single detected segment if the in-between pause is less than 0.5 s and also less than the sum of these two syllables' durations. Only segments of duration between 0.1 and 1.5 s that have time overlap with a human-segmented phrase are used to form test sets A2 and B2. Histograms of the time boundary difference between the machine- and human-segmented phrases (machine-segmented time boundary minus human-segmented time boundary) are plotted in Fig. 3. Figures 3(a) and 3(b) show the histograms of time differences observed at the left/starting and right/ending time boundaries, respectively. The histograms are roughly Gaussian in shape, with 35%–50% of the time differences falling within ± 0.025 s (see the bars centered at the time difference of 0 s), and about 5% of the phrase segments has a time boundary difference exceeding 0.2 s.

B. Spectrographic representations of some phrase classes

Figure 4 shows the linear frequency spectrograms of 9 CAVI phrase classes. It can be observed that the acoustic signatures of some classes are very similar to each other. For example, the first two classes in the first row resemble each other except at the beginning and end of the phrases. Phrase classes in the last row also have very similar ascending and descending frequency signatures. Some phrases have a long pause between syllables, which can be observed in the spectrograms plotted in the second row of Fig. 4. Figure 5 shows the Mel-frequency (Mel) spectrogram representation of the same phrases plotted in Fig. 4. Mel-spectrograms are used in the DTW-SR-2pass classification algorithm described in Sec. III.

Birdsong phrase classification is challenging because acoustic signatures of samples from the same class are not identical. This intra-class differences can be due to individual bird differences (Mennill, 2011), different sound propagation properties in the forest (Richards and Wiley, 1980; Nemeth *et al.*, 2006), and phrase segmentation errors. Each column in Fig. 6 shows the spectrographic differences that can be present between two human-segmented phrase samples from the same class. Differences in segmentation boundaries and phrase energy variations across time are evident. Small segmentation errors or differences can exist in human-segmented phrases due to subjectiveness, especially in determining the end points. For example, the human-annotated time boundary at the end of the phrase in Fig. 6(b-i) is marked earlier than that in Fig. 6(b-ii) because the phrase energy reaches a human-determined intensity threshold at that point. Segmentation errors tend to be larger with an

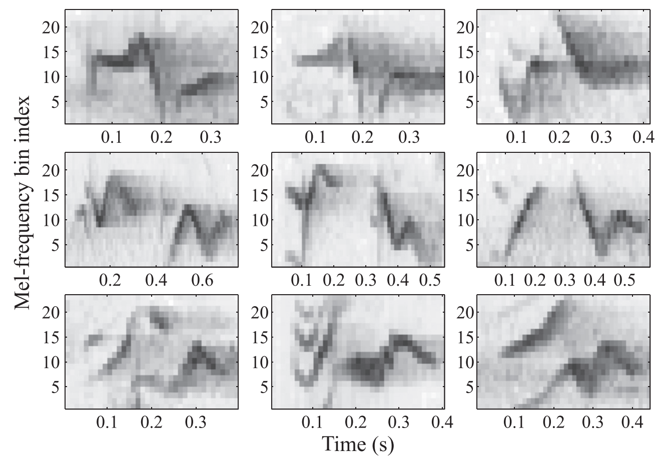

FIG. 5. Mel-spectrograms of the 9 phrase classes of Fig. 4.

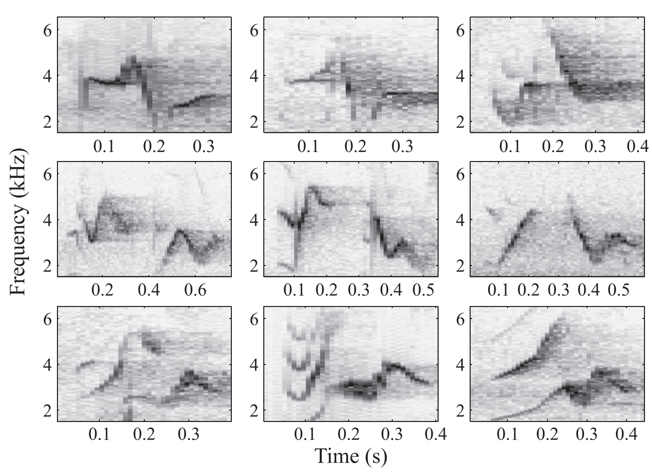

FIG. 4. Linear frequency spectrograms of 9 different phrase classes.

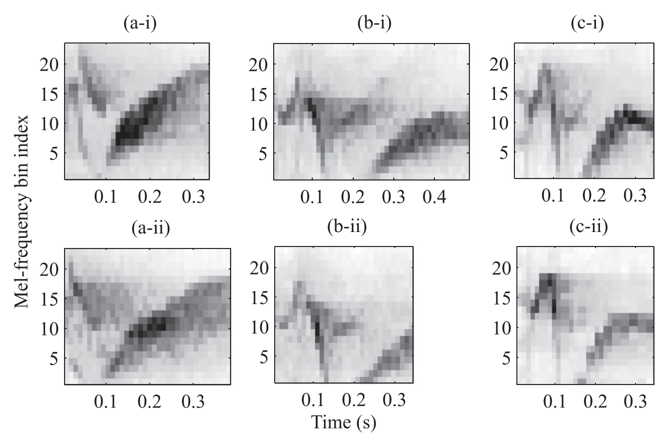

FIG. 6. Mel-spectrograms to show differences between phrase samples from the same class. Each of the three columns contains two samples from one particular phrase class.

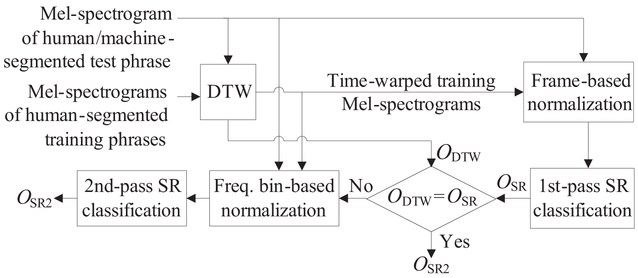

FIG. 7. Flow chart of the proposed two-pass sparse representation classification framework. O_X is the class decision of classifier X, where X is either DTW, SR (from 1st-pass SR) or SR2 (from 2nd-pass SR).

automated segmentation algorithm because energies can vary greatly for different birds and recording environments.

III. PROPOSED DTW-SR-2PASS CLASSIFICATION ALGORITHM

The flow chart in Fig. 7 summarizes the proposed bird phrase classification algorithm, abbreviated DTW-SR-2pass. The algorithm applies DTW on the training Mel-spectrograms, followed by a two-pass SR classification. Mel-spectrogram generation and the signal processing performed in each stage are described in the following subsections.

A. Mel-spectrogram generation

A Mel-spectrogram is computed for each phrase segment in both the training and testing sets. Figure 8(a) shows the various steps involved in generating each Mel-spectrogram. The segment is first downsampled from 44.1 to 20 kHz because little phrase energy is found above 10 kHz. It is then split into multiple 20 ms frames with 50% overlap (10 ms shift). A Hamming window is applied to each frame, followed by a 512-point FFT to obtain the power spectrum of each frame. The Mel-spectrogram is generated by applying a Mel-filter bank [shown in the Fig. 8(b)] that covers the 1.5–6.5 kHz frequency range in which most of the CAVI phrase energy falls within. A different frequency range can be used when classifying phrases from other bird species. When energies in multiple linear frequency bins are combined to form each Mel component of coarser frequency resolution, energy transitions across frequency are smoother (compare the spectrograms in Figs. 4 and 5). This leads to smaller intra-class component-wise distances, thus the Mel-spectrogram is chosen over its linear counterpart.

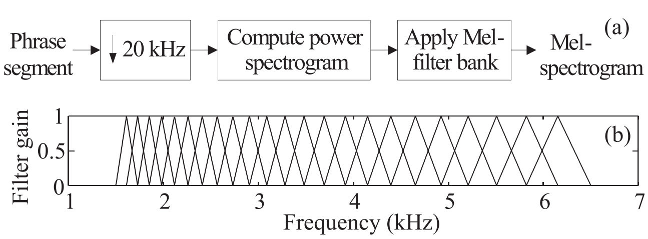

FIG. 8. (a) Block diagram on Mel-spectrogram generation. (b) Mel-filter distribution in the frequency range of interest (1.5–6.5 kHz).

B. DTW of the training phrases using Mel-spectrograms

DTW is applied on each training Mel-spectrogram to yield a time-warped version that is more similar to and has the same number of frames as that in the test Mel-spectrogram (the phrase class of which is to be determined). This is done by computing the optimal path through a similarity matrix Θ, where each $\Theta(i,j)$ is a cosine similarity measure [see Eq. (1)] that emphasizes the similarity in spectral shape, instead of energy values, between vectors s_i and r_j in the ith frame of the test Mel-spectrogram and the jth frame of the training Mel-spectrogram, respectively. $\|\bullet\|_2$ in Eq. (1) represents the l_2 vector norm. Cosine similarity is computed on the linear power scale Mel-spectrogram, which is also used in Kaewtip *et al.* (2013) and Meliza *et al.* (2013),

$$\Theta(i,j) = \frac{s_i^T r_j}{\|s_i\|_2 \|r_j\|_2} = \left(\frac{s_i}{\|s_i\|_2}\right)^T \left(\frac{r_j}{\|r_i\|_2}\right). \tag{1}$$

Figure 9 illustrates the type I local path constraints imposed (Myers *et al.*, 1980) that limit the minimum and maximum scale of time-warping to 0.5 and 2, respectively. The intermediate accumulative score $D(i,j)$ is computed recursively in Eq. (2), for $i = 0, 1, \ldots, M-1$, and for $j = -T, -T+1, \ldots, -1, 0, 1, \ldots, N+T-1$, where M and N denote the number of frames in the test and training segments, respectively. T is the additional number of frames extended beyond the training segment boundaries that are used to generate each training Mel-spectrogram. In Eq. (2), the Θ values are weighted by 0.5 for path 1 so as not to double count the similarity score with the same frame of the test spectrogram, i.e., s_i. This makes d, the final accumulative score of the optimal path (between the test spectrogram and the current training spectrogram) computed in Eq. (3), comparable across all training spectrograms or samples,

$$D(i,j) = \max\begin{cases} D(i-1,j-2) + \frac{1}{2}\Theta(i,j-1) + \frac{1}{2}\Theta(i,j), & \text{path 1} \\ D(i-1,j-1) + \Theta(i,j), & \text{path 2} \\ D\left(i-2,j-1\right) + \Theta(i-1,j) + \Theta(i,j), & \text{path 3,} \end{cases} \tag{2}$$

$$d = \max_{j\in[N-T-1,\ldots,N+T-1]} D(i = M-1, j). \tag{3}$$

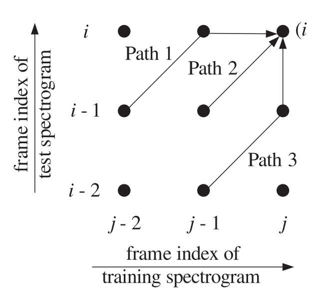

FIG. 9. The possible paths allowed to reach an arbitrary point (i,j) in the DTW grid when searching for the optimal path. The frame indices of the test and training Mel-spectrograms are denoted by i and j, respectively.

For each training sample, d is computed such that the optimal path is selected from all paths that stop within the last $2T+1$ frames of the extended training Mel-spectrogram while covering all the test frames. The paths are allowed to start at any frame within the first $2T+1$ frames of the extended training Mel-spectrogram. This flexibility (that resulted from setting $T > 0$) improves classification accuracies when there is phrase position variation within the segment due to inconsistently determined time boundaries. Large phrase position variations can exist when an automatic segmentation algorithm is used. The phrase class of the training sample that yields the highest d, over all training samples, is the classification decision of this DTW classifier, O_{DTW}. The DTW decision is used in a subsequent step to decide if a second-pass of SR classification is necessary.

To obtain the time-warped training spectrogram, $\tilde{R} = [\tilde{r}_0, \ldots, \tilde{r}_{M-1}]$, which has the same number of frames as the test spectrogram for SR classification, the optimal path is backtracked as shown in Eq. (4). Figure 10 shows an example of a training Mel-spectrogram before and after DTW to match the test Mel-spectrogram.

$$\begin{aligned}
&\text{If path 1,} \quad \tilde{r}_i = \frac{1}{2}(r_j + r_{j-1}),\\
&\text{If path 2,} \quad \tilde{r}_i = r_j,\\
&\text{If path 3,} \quad \tilde{r}_i = \tilde{r}_{i-1} = r_j.
\end{aligned} \tag{4}$$

C. Feature normalization prior first-pass SR classification

To emphasize spectral shape similarity over spectral amplitude similarity, all time-warped training Mel-spectrograms and the test Mel-spectrogram are normalized to have an Euclidean norm of 1 for every frame prior to amplitude log-compression. The Euclidean norm of 1 is chosen for simplicity. This normalization is implicitly done in computing DTW's cosine similarity measure [see Eq. (1)]. This frame-based energy normalization reduces within-class dissimilarity between phrases that have different energy variations across time as observed in the phrases shown in Figs. 6(c-i) and 6(c-ii).

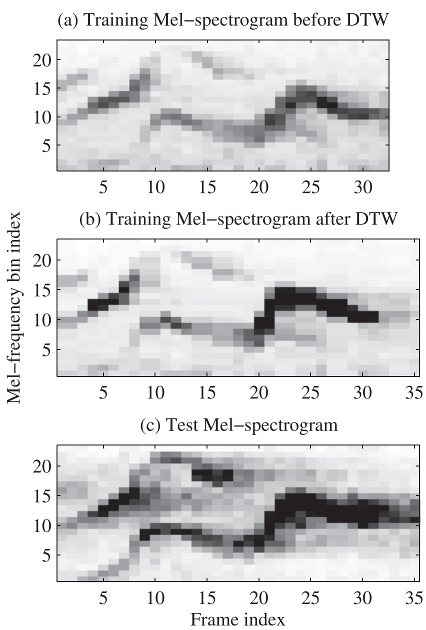

FIG. 10. An example to illustrate the effect of DTW on the training Mel-spectrogram. (a) and (b) show the training Mel-spectrograms before and after DTW, respectively, to match the test Mel-spectrogram shown in (c). Amplitude log-compression has been applied to give a clearer visual display of the spectrograms in this figure.

D. Dimension reduction and SR classification

After frame normalization and log-compression, the frames of each Mel-spectrogram are concatenated to form a single feature vector per sample. The dimension of the feature vector is then reduced to $p = \min(m, P)$ by performing a principal component analysis (PCA) on the feature vectors from all $m = Kn$ training Mel-spectrograms matrix, where P is a user-specified dimension. K is the total number of phrase classes in the training set, and n is the number of training samples per class, as defined in Sec. II A. The final p-by-1 feature vector is obtained by normalizing the dimensionally reduced vector to unit length (this is done in both training and test sets). The rationale for performing dimension reduction and unit length normalization is related to the l_1 minimization expression used in the SR classification algorithm, which is explained in the following paragraph.

The SR classification algorithm summarized in Eqs. (5)–(8), follows "Algorithm 1" described in Yang *et al.* (2007) for a 100-subject face recognition application. It is also known as exemplar-based sparse classification because each representation is an exemplar—a feature vector extracted from each individual training sample rather than learned dictionary representations derived from a large pool of training samples. The SR classifier finds a sparse linear combination of training feature vectors that best represents the test feature vector, b. This linear combination is found by solving for a sparse vector $x = [x[1], x[2], \ldots, x[m]]^T \in \mathbf{R}^{m\times 1}$ via the l_1 minimization convex optimization problem defined in Eq. (5), where $\|\bullet\|_1$ represents the l_1 vector norm

$$\min \|x\|_1 \text{ subject to } \|\mathbf{A}x - b\|_2 \le \varepsilon. \tag{5}$$

Each column, $a_i \in \mathbf{R}^{p\times 1}$, in dictionary matrix $\mathbf{A} = [a_1, a_2, \ldots, a_m] \in \mathbf{R}^{p\times m}$ contains one exemplar or feature vector from the training set. Reducing the feature dimension, p, to the smaller value of two m and P values ensures that Eq. (5) is not an over-determined linear system of equations such that a solution x can always be found. The exemplars in the dictionary matrix are generally normalized to unit length (Wright *et al.*, 2009; Gemmeke *et al.*, 2011) so as not to bias toward the selection of exemplars with larger feature values when the l_1 solver tries to minimize $\|x\|_1$. The difference tolerance, ε is usually set to a small value to allow small differences between the test feature vector and the one reconstructed using a sparse linear combination of the training feature vectors. Normalizing the test feature vector to unit length enables ε to be fixed to a constant value instead of varying it proportionally to the Euclidean

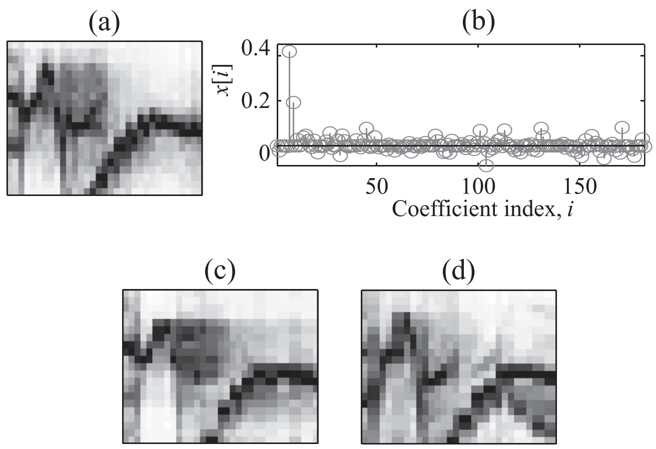

FIG. 11. (a) Example of a test Mel-spectrogram that is correctly classified in the first pass of SR classification. (b) The sparse solution x computed with the test feature vector extracted from the Mel-spectrogram in (a). The training Mel-spectrograms corresponding to the two largest coefficients in x, are shown in (c) and (d).

norm of each test vector. The spg_bpdn function in the SPGL1 MATLAB toolbox (van den Berg and Friedlander, 2007, 2008) is used to solve Eq. (5). After the solution, x, is found, the residual vector r_k between b and $\mathbf{A}\delta_k(x)$ is computed in Eq. (6). The $\delta_k(x)$ function outputs a vector, y, the coefficients of which $y[i] = x[i]$, if a_i is a training feature vector from class k. All other coefficients in y are set to zero, as shown in Eq. (7),

$$r_k = b - \mathbf{A}\delta_k(x), \quad \text{for } k = 1, 2, \ldots, K, \tag{6}$$

$$\delta_k(x) = y, \text{ where } y[i] = \begin{cases} x[i], & \text{if } a_i \in \text{class } k \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

The class that yields the minimum residual norm $\|r_k\|_2$ is the output decision of the SR classifier O_{SR} as shown in Eq. (8). This decision of the SR classifier is compared with O_{DTW} that is computed during DTW. If they are the same, it is the final class decision of the proposed algorithm,

$$O_{\mathrm{SR}} = \arg\min_k \|r_k\|_2. \tag{8}$$

Figure 11 shows an example of a sparse solution x computed using Eq. (5), for a test phrase that is correctly classified. Figure 11(a) shows the Mel-spectrogram of this test phrase segment, and Fig. 11(b) plots the sparse x computed with the test feature vector extracted from this Mel-spectrogram. Figures 11(c) and 11(d) plot the training Mel-spectrograms that correspond to the two largest x coefficients in Fig. 11(b). These spectrograms are of the same phrase class as the test Mel-spectrogram.

E. Second-pass of SR classification

If $O_{\mathrm{DTW}} \neq O_{\mathrm{SR}}$, a second (2nd)-pass of SR classification is activated. This SR classification is performed using only the training samples from the two conflicting phrase classes and with a modification to the feature normalization step. Instead of frame normalization, frequency bin normalization is performed on the time-warped training Mel-spectrograms and the original test Mel-spectrogram, such that the Euclidean norm of the values in each frequency bin is equal to one. This provides a different perspective of the features to the SR classifier in the 2nd-pass. Subsequent steps after feature normalization, follow the same procedure as that used in the 1st-pass, which include log-compression, dimension reduction, and sparse vector computation. The class with the smaller residual norm of the two classes in the 2nd-pass, is the final class decision of the proposed algorithm. Note that the computation time required to perform the l_1-minimization optimization in the 2nd-pass is much less than that used in the 1st-pass because the dictionary matrix involved is significantly smaller.

The parameters of the proposed DTW-SR-2pass algorithm are tuned using the set of values listed in Table II. Classification results of the optimal parameters are shown in Sec. V, while performance variations with different parameter values are presented in Sec. VI B.

IV. COMPARATIVE ALGORITHMS

A. Sparse representation (SR) classifier

This SR classifier is similar to our previous work in Tan *et al.* (2012), in which no DTW is involved. A 64-frame Mel-spectrogram is computed using a phrase-duration-dependent frame shift on each phrase segment, followed by the same frame normalization, log-compression, dimension reduction, and 1st-pass SR classification framework described in Secs. III C and III D. The output of the 1st-pass SR classification is the decision of this SR classifier.

B. Support vector machine (SVM) classifier

The multi-class SVM classifier is implemented using a popular software library known as LIBSVM (Chang and Lin, 2011). Like the SR classifier described in Sec. IV A, this SVM classifier does not use DTW, and a 64-frame Mel-spectrogram is computed. The input feature vector to the SVM is the frame-concatenated vector of this Mel-spectrogram. Dimension reduction is omitted, and the linear kernel is used because experiments show that this configuration yields better classification performance under the limited training data condition compared to using dimension reduction or the Gaussian radial basis function (RBF) kernel. The innate multi-class SVM classifier in LIBSVM uses an one-against-one decomposition scheme. However, this one-against-one decomposition scheme breaks down when there are less than three samples per class to train the SVM. Hence an one-against-all decomposition strategy is implemented

TABLE II. The various values used for parameter optimization of the proposed algorithm.

Parameter	Values
PCA dimension, P	32, 64, 128
Difference tolerance, ε	0.025, 0.05, 0.1, 0.2
DTW frame extension, T	0, 5, 10, 20, 30, 40

via a code modification. The optimal regularization factor (also known as soft margin parameter), C, that gives the highest classification accuracy on the training samples is selected from the set $\{2^{-1}, 2^{0}, \ldots, 2^{3}\}$.

C. Dynamic time warping (DTW) classifier

The DTW-based classifier used for performance comparison is the same as the one described in Sec. III B. The classification accuracy is calculated based on the decision of the DTW classifier, O_{DTW}.

D. DTW-SVM-2pass classifier

The DTW-SVM-2pass classifier is similar to the proposed DTW-SR-2pass classifier. The SR classification stage (including dimension reduction) in each pass is replaced by the linear SVM classifier described in Sec. IV B. It enables assessment of the performance difference between the two classification techniques (SR and SVM) with the proposed DTW framework.

E. DTW-SR-1pass classifier

This DTW-SR-1pass classifier is similar to our proposed DTW-SR-2pass algorithm, but without the 2nd-pass of SR classification. The decision, O_{SR} (see Sec. III D), is used to compute the classification accuracy of this DTW-SR-1pass classifier. It enables assessment of the performance gain obtained with the additional 2nd-pass of SR classification.

V. RESULTS

In this section, the classification accuracies (Acc.) of each classifier are presented. In each experiment, n training samples per phrase class from the training CAVIs are randomly selected, and the average results of five such experiments are computed for each parameter setting (using the values in Table II) with each classification algorithm. Section V A contains the results for human-segmented phrases in Tests A1 and B1, while Sec. V B contains the results for machine-segmented phrases in Tests A2 and B2.

A. Classification accuracy on human-segmented test phrases

Figure 12 shows the accuracies on Tests A1 and B1 obtained by the comparative algorithms and the proposed DTW-SR-2pass classifier at their optimal parameter settings. Test A1 contains human-segmented test phrases from the training CAVIs, while Test B1 contains human-segmented test phrases from the test CAVIs. The values of P, ε, and T for the DTW-SR-based classifiers that give the best performance on this dataset are 128, 0.05, and 5, respectively. The results of the DTW and DTW-SVM-2pass classifiers shown are obtained with T set to 5 (which is also the best value for each classifier). For the SR classifier, the best results shown is obtained with $P = 128$ and $\varepsilon = 0.1$.

The results in Fig. 12 show that the proposed DTW-SR-2pass classifier outperforms other classification algorithms in all cases. With $n = 5$ training samples per phrase class, it achieves the highest classification accuracies of 98.6% and 94.3% on human-segmented phrases in Tests A1 and B1, respectively. The proposed classifier also has the highest averaged classification accuracies of 92.4% and 90.6% (averaged Acc. over different n values) on Tests A1 and B1, respectively. Its performance gain over those algorithms increases as n decreases. Both SVM and DTW-SVM-2pass classifiers perform worse than their SR-based counterparts. The averaged accuracies of the DTW-SR-1pass and DTW-SR-2pass are also better than that of the DTW classifier. It is also observed that the SR and SVM classifiers without DTW perform worse than classifiers with DTW. The performance differences in averaged Acc. between classifiers with and without DTW are approximately 3% for Test A1 and 13% for Test B1. This shows that DTW helps to reduce intra-class variations in human-segmented phrases, especially when the phrases are recorded from bird individuals not found in the training set.

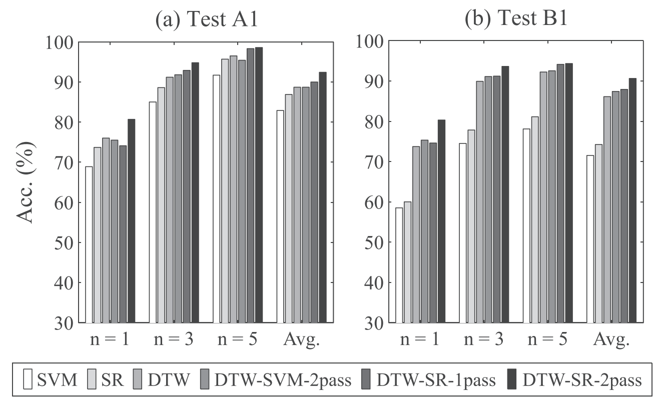

FIG. 12. Classification accuracies (%) of test sets A1 and B1 (n is the number of training samples per phrase class). (a) Test A1 contains human-segmented test phrases from the training CAVIs, while (b) Test B1 contains human-segmented test phrases from the test CAVIs. n refers to the number of training samples per phrase class.

Error analysis of Test B1 results for the case of $n = 5$ reveals that about 50% of the errors are misclassifications to classes that have very similar dominant frequency trajectories, and about 35% of the misclassified phrases are affected by noise interferences from other birds and segmentation errors. For example, the test phrase in Fig. 13(a) is misclassified to another phrase class shown in Fig. 13(b) of similar acoustic signature, while the noise-corrupted test phrase in Fig. 13(c) is misclassified to the phrase class shown in Fig. 13(d). Another reason for the misclassifications is the presence of intra-class variations that cannot be well-represented by a linear combination of dynamically time-warped training samples of the same class.

B. Classification accuracy on machine-segmented test phrases

Classification accuracies (Acc.) of the comparative algorithms and the proposed DTW-SR-2pass on Tests A2 and B2 are tabulated in Fig. 14. Test A2 contains machine-segmented test phrases from the training CAVIs, while Test B2 contains machine-segmented test phrases from the test CAVIs. The optimal values of P and ε for the SR and DTW-SR-based classifiers are the same as those for the human-

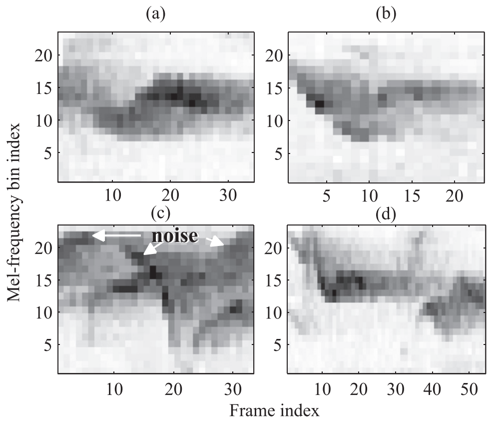

FIG. 13. Examples of human-segmented phrases that are misclassified. Test phrase in (a) is misclassified to the phrase class in (b) of similar acoustic signature. Test phrase in (c) is misclassified to the phrase class in (d) due to noise interference.

segmented test phrases (in Sec. V A). For the DTW frame extension parameter T it is found that increasing it from 5 to 30 frames improves the classification accuracy of the machine-segmented bird phrases, which have larger segmentation inconsistencies than human-segmented phrases. The results of the DTW and DTW-SVM-2pass classifiers shown are also obtained with T set to 30.

Classification accuracies on machine-segmented test phrases in Fig. 14 show similar performance trends as those observed in Fig. 12. The proposed DTW-SR-2pass classifier outperforms other algorithms in all cases, achieving accuracies of 93.7% and 89.4% with $n=5$ training samples per phrase class and averaged accuracies of 84.2% and 82.4% for Tests A2 and B2, respectively. DTW-SR-2pass classifier's performance gain over other algorithms also increases as n decreases; this shows its robustness to very limited training data. Performance differences in averaged accuracies between classifiers with and without DTW are nearly 40% in both Tests A2 and B2. This shows that DTW is essential for good classification performance on machine-segmented phrases.

It is also observed that the classification accuracies on machine-segmented test phrases are lower than those obtained for human-segmented test phrases. This decrease is mainly due to segmentation errors present in the machine-segmented phrase segments. A shorter segment with missing portions of the test phrase results in misclassification to classes that are more similar to the detected sub-segment. For example, the phrase shown in the top left corner of Fig. 5 may be classified to the phrase on its right if some of its beginning and end portions are missing. Test phrase segments with additional portions that contain substantial background noise also tend to be misclassified.

Note that human- instead of machine-segmented training phrases are used for the classification of machine-segmented test phrases because the latter yields about 2% lower classification accuracy (absolute difference) on average, for all classification algorithms evaluated, compared to the case when human-segmented training phrases are used.

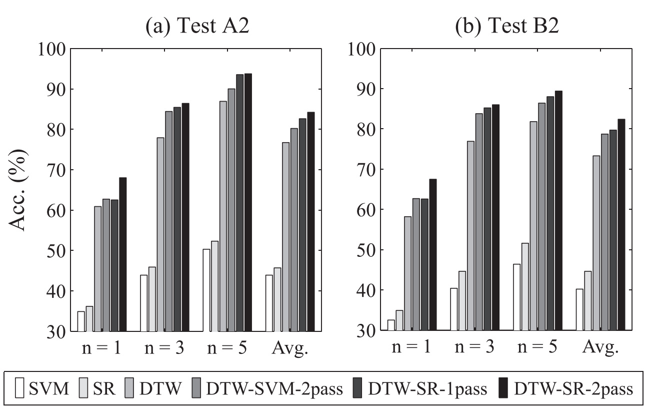

FIG. 14. Classification accuracies (%) of test sets A2 and B2. (a) Test A2 contains machine-segmented test phrases from the training CAVIs, while (b) Test B2 contains machine-segmented test phrases from the test CAVIs.

VI. DISCUSSION

A. On the importance of DTW

On the human-segmented test phrases, the SR and SVM classifiers without DTW perform worse than a DTW classifier. A larger performance degradation is observed on phrases from test CAVIs in Test B1 than phrases from training CAVIs in Test A1. This shows the benefit of using DTW to reduce the time mismatch between the phrases recorded from different bird individuals. Error analysis reveals that within-class time variations exists due to individual bird differences in producing phrases from the same class, and the subjective differences between human annotators in determining the phrase segment boundaries. Figure 15 shows an example of a spectrographic feature mismatch between human-segmented phrases of the same class when a phrase-duration-dependent frame shift is used (to compute Mel-spectrograms with a fixed number of frames), as done in classifiers without DTW. The Mel-spectrograms in Figs. 15(a) and 15(b) are computed from the phrase segments previously shown in Figs. 6(b-i) and 6(b-ii), respectively.

The importance of DTW in reducing time boundary mismatch is further exemplified on machine-segmented bird phrases. The SR and SVM classifiers without DTW

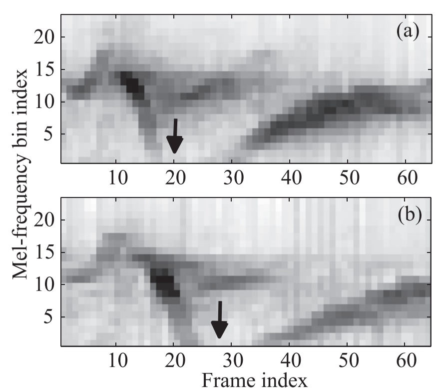

FIG. 15. Time mismatch observed between Mel-spectrograms computed using a phrase-duration-dependent frame shift. These 64-frame Mel-spectrograms in (a) and (b) are computed from the two human-segmented phrases shown in Figs. 6(b-i) and 6(b-ii), respectively. The arrows on the 64-frame Mel-spectrograms in (a) and (b) note the time instances of the low frequency regions in the phrase segments.

suffer a large performance degradation on phrases from both the training and test CAVIs (30%– 45% absolute decrement in classification accuracy in contrast to the 5%–15% decrement observed in classifiers with DTW) in comparison to their performance on human-segmented bird phrases. This is because only those machine-segmented phrase segments with time boundaries that are very close to their human-segmented counterparts are correctly classified by classifiers without DTW.

B. On the performance gain of DTW-SR-2pass over DTW-SR-1pass

When the results of DTW-SR-1pass classifier are compared with those of the DTW-SR-2pass classifier, it is evident that the largest performance gain is observed in the case when there is only $n=1$ training sample per phrase class. The 2nd-pass of the SR classification aims to resolve the conflicting class decisions of DTW-SR-1pass and DTW. The DTW classifier compares the test sample to every training sample separately. It obtains the correct decision if there is a training sample in the correct class that has the highest cosine similarity with the test sample. In contrast, the SR classifier generally classifies correctly if the sparse solution in Eq. (5) is concentrated at the coefficients corresponding to the correct phrase class. When there is more than one training sample per class, the SR classification framework has the advantage of using a linear combination of (all or a few) training samples from the correct phrase class (together with some training samples from other classes) to reconstruct a vector that resembles the test feature vector. When there is only one training sample per phrase class, the advantage of DTW-SR-1pass over DTW with a linear combination framework is reduced. Even if the single training sample from the correct class is closest to the test sample, but not by a large margin compared to other training samples, there is a larger possibility of finding a sparse linear combination of training samples from multiple classes that resembles the test sample in which the coefficient corresponding to the correct class is not the largest. Analysis reveals that among the conflicting class decisions in which either DTW or DTW-SR-1pass is correct, DTW-SR-1pass is correct for approximately 77% of them in the case when $n=5$, and this percentage decreases to 53% when $n=1$. When only the two conflicting classes' training feature vectors (obtained from frequency bin-normalized Mel-spectrograms) are used in the 2nd-pass of SR classification, the SR classifier is able to derive a correct class decision for 70%–90% of these cases. Note that normalizing by frequency bin in the 2nd-pass improves the classification accuracy of the 1st-pass by more than 2% (absolute difference when averaged over Acc. obtained for all values of n), while using frame normalization in the 2nd-pass gives an improvement of less than 1%. The difference in classification performance between frequency bin-based versus frame-based normalized features in the 2nd-pass is found to be statistically significant with a p-value less than 0.01 via the McNemar test (McNemar, 1947). On the other hand, if frequency bin-based normalized features were to be used in the 1st-pass, the classification accuracy of DTW-SR-1pass would be less than that obtained with just the DTW classifier.

C. Parameter sensitivity analysis of DTW-SR-2pass's classification performance

The proposed DTW-SR-2pass classification performance is relatively consistent compared to DTW-SR-1pass across the different parameter values of P (PCA dimension) and ε (difference tolerance) investigated. Figures 16(a) and 16(b) plot the averaged accuracies (Acc. averaged across n) achieved with different P values, while ε and T (DTW frame extension) are fixed at their optimal value, for human-segmented phrases (in Test B1), and machine-segmented phrases (in Test B2) from the test CAVIs, respectively. Similarly, Figs. 16(c) and 16(d) plot the accuracy variations with ε (while P and T are fixed at their optimal value), for human- and machine-segmented test CAVIs' phrases, respectively. The DTW classifier has a constant Acc. in Figs. 16(a)–16(d) because its algorithm is independent of P and ε. The Acc. range of DTW-SR-2pass is less than 2% in the range of P and ε evaluated, while the Acc. range of DTW-SR-1pass is larger. There are also some values of P and ε for which DTW-SR-1pass performs worse than the DTW classifier. The reduced Acc. variation observed for the proposed DTW-SR-2pass classifier is mainly because the 2nd-pass

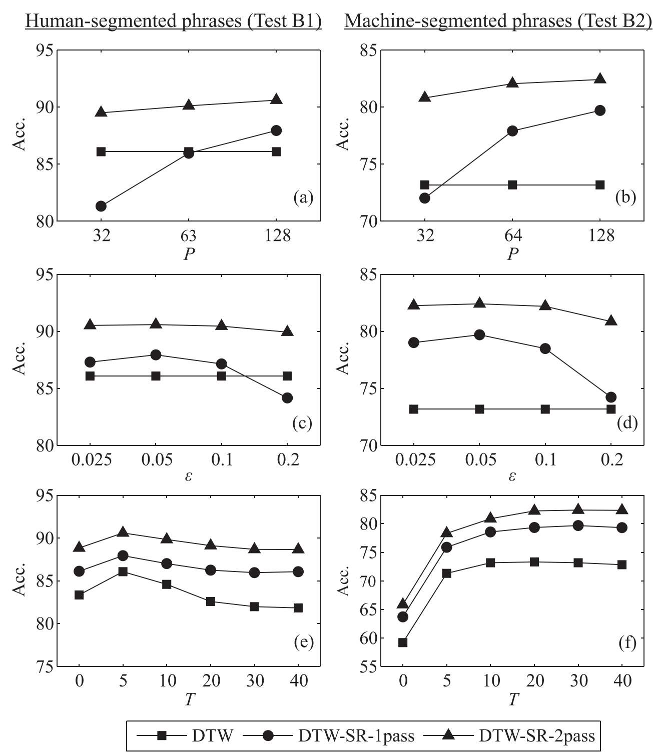

FIG. 16. Sensitivity analysis of the classification accuracy (Acc.) of DTW-SR-2pass to different values of PCA dimension (P), difference tolerance (ε), and DTW frame extension (T). The averaged Acc. (averaged across n) obtained by the DTW, DTW-SR-1pass, and DTW-SR-2pass classifiers on human-segmented test CAVIs' phrases in Test B1 with different values of P, ε, and T, are plotted in (a), (c), and (e), respectively. The classifiers' averaged Acc. on machine-segmented test CAVIs' phrases in Test B2 as P, ε, and T vary, are plotted in (b), (d), and (f), respectively.

TABLE III. Real-time (RT) performance in $\times$RT of the DTW-SR-2pass classifier with various values of ε, P, T, and n.

$n=3$ and $T=5$				$\varepsilon=0.05$ and $P=128$			
$\varepsilon\backslash P$	32	64	128	$T\backslash n$	1	3	5
0.025	1.09	1.15	1.19	0	0.53	1.09	1.15
0.05	0.97	1.03	1.04	10	0.56	1.09	1.26
0.1	0.86	0.91	0.92	20	0.59	1.19	1.34
0.2	0.84	0.86	0.87	30	0.65	1.29	1.47

takes into account of the DTW classifier's decision, which is independent of P and ε.

Figures 16(e) and 16(f) plot the averaged accuracies achieved by DTW, DTW-SR-1pass, and DTW-SR-2pass with different T values (while P and ε are fixed at their optimal value) for human- and machine-segmented test CAVIs' phrases, respectively. As T varies, the Acc. variations of DTW, DTW-SR-1pass, and DTW-SR-2pass are similar to one another. Comparing these figures, we conclude that increasing T can result in performance degradation (about 2%) for human-segmented phrases. For machine-segmented phrases, on the other hand, increasing T beyond 0 is beneficial.

D. Computation time of DTW-SR-2pass

Table III shows the computation time of the proposed DTW-SR-2pass classification algorithm (written in MATLAB) on an Intel i7 2.67 GHz processor with 4 GB RAM (without parallel-computing). It varies from 0.5 to 1.5 $\times$ real-time (RT), depending on the parameter settings and the number of training samples used. A computation time of 0.5$\times$RT means that the algorithm takes an average of 0.15 s to derive the class decision of a bird phrase segment of duration 0.3 s. In general, the computation time increases with T because DTW is done on a larger number of frames. The computation time also increases as n and P increase because there are more elements in the dictionary matrix A on which the l_1-minimization in Eq. (5) is performed. On the other hand, the l_1-minimization requires less computation time when a larger tolerance ε is used.

VII. SUMMARY AND CONCLUSION

A DTW-SR-2pass classification framework that combines DTW with a two-pass SR classification is proposed for limited data birdsong phrase classification. Mel-spectrograms of the training samples are dynamically time-warped to the Mel-spectrogram of each test sample to obtain training Mel-spectrograms with the same number of time frames as the test Mel-spectrogram. This is followed by feature normalization, which involves frame-based energy normalization and amplitude log-compression. PCA dimension-reduced feature vectors are used in the 1st-pass SR classification. If the class decision of the 1st-pass SR classification is different from the DTW classifier's decision (computed as a by product at the DTW stage), a 2nd-pass SR classification is performed using the training samples solely from the two conflicting classes. Frequency bin-based (instead of frame-based) energy normalization is performed in the 2nd-pass.

The phrases used in this study are extracted from songs of the Cassin's Vireo. The training set contains a few training samples ($n=1$–5) per phrase class from a few bird individuals, and the test data are divided into four sets to evaluate classification performances on phrases recorded from bird individuals present or absent in the training set, and on human- or machine-segmented phrases separately. Compared to the DTW, SVM-based, and SR (without DTW) classifiers evaluated, the proposed DTW-SR2pass classifier achieves the highest classification accuracies on all test conditions. DTW is essential for achieving good classification performances on machine-segmented phrases, and phrases recorded from bird individuals that are not found in the training set. Performing the 2nd-pass SR classification (on phrases with conflicting DTW and 1st-pass SR decisions) leads to increasing performance gains over the comparative algorithms as n decreases, and reduces the parameter sensitivity of the algorithm's performance.

Future work will include development of discriminative and noise-robust feature extraction schemes to improve the algorithm's classification performance on phrase classes that are very similar to each other in the presence of noise interferences (such as overlapping birdsongs). We plan to add the in-set/out-of-set verification module developed in our previous paper (Tan *et al.*, 2013) to the DTW-SR-2pass algorithm to enable automated detection of new phrase classes that are not found in the training set. We also intend to extend the proposed framework to phrase classification of other bird species, and species identification.

ACKNOWLEDGMENTS

This work was supported in part by National Science Foundation Award Nos. 0410438 and IIS-1125423.

Agranat, I. (**2009**). "Automatically identifying animal species from their vocalizations," in *Proceedings of the IOA International Conference Bioacoustics*, pp. 80–98.

Anderson, S. E., Dave, A. S., and Margoliash, D. (**1996**). "Template-based automated recognition of birdsong syllables from continuous recordings," J. Acoust. Soc. Am. **100**, 1209–1219.

Berwick, R. C., Okanoya, K., Beckers, G. J., and Bolhuis, J. J. (**2011**). "Songs to syntax: The linguistics of birdsong," Trends Cogn. Sci. **15**, 113–121.

Boersma, P., and Weenink, D. (**2011**). "PRAAT: Doing phonetics by computer (version 5.2.22) [computer program]," http://www.praat.org (Last viewed 4/15/2011).

Brandes, T. S. (**2008**). "Automated sound recording and analysis techniques for bird surveys and conservation," Bird Conserv. Int. **18**, S163–S173.

Briefer, E. F., Rybak, F., and Aubin, T. (**2013**). "Does true syntax or simple auditory object support the role of skylark song dialect?," Anim. Behav. **86**, 1131–1137.

Catchpole, C. K., and Slater, P. J. B. (**2008**). *Bird Song: Biological Themes and Variations* (Cambridge University Press, Cambridge, UK), pp. 1–274.

Chang, C. C., and Lin C. J. (**2011**). "LIBSVM: A library for support vector machines," ACM Trans. Intell. Syst. Technol. **2**, 27:1–27:27, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm (Last viewed 3/18/2012).

Chen, S. S., and Li, Y. (**2013**). "Automatic recognition of bird songs using time-frequency texture," in *Proceedings of IEEE Computer Intelligence Communication Network*, pp. 262–266.

Chen, Z., and Maher, R. C. (**2006**). "Semi-automatic classification of bird vocalization using spectral peak tracks," J. Acoust. Soc. Am. **120**, 2974–2984.

Chu, W., and Blumstein, D. T. (**2011**). "Noise robust bird song detection using syllable pattern-based hidden Markov models," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 345–348.

Fagerlund, S. (**2007**). "Bird species recognition using support vector machines," EURASIP J. Adv. Signal Process. **2007**, 1–8.

Frommolt, K.-H., Bardeli, R., and Clausen, M., eds. (**2008**). "Computational bioacoustics for assessing biodiversity," BfN Scripten-234 (Bundesamt für Naturschutz, Bonn, Germany), pp. 1–156.

Gemmeke, J. F., Virtanen, T., and Hurmalainen, A. (**2011**). "Exemplar-based sparse representations for noise robust automatic speech recognition," IEEE Trans. Audio Speech Lang. Process. **19**, 2067–2080.

Glotin, H., Clark, C., LeCun, Y., Dugan, P., Halkias, X., and Sueur, J., eds. (**2013**), *Proceedings of the 1st Workshop on Machine Learning for Bioacoustics*, 1, joint to International Conference on Machine Learning (ICML), pp. 1–104.

Goguen, C. B., and Curson, D. R. (**2002**). "Cassin's Vireo (*Vireo cassinii*), the birds of North America online (A. Poole, Ed.)," Ithaca: Cornell Lab of Ornithology, http://bna.birds.cornell.edu/bna/species/615 (Last viewed 7/18/2012).

Graciarena, M., Delplanche, M., Shriberg, E., and Stolcke, A. (**2011**). "Bird species recognition combining acoustic and sequence modeling," in *Proceedings of IEEE ICASSP*, pp. 341–344.

Hansson-Sandsten, M., Tarka, M., Caissy-Martineau, J., Hansson, B., and Hasselquist, D. (**2011**). "A SVD-based classification of bird singing in different time-frequency domains using multitapers," in *Proceedings of EURASIP European Signal Processing Conference (EUSIPCO)*, pp. 966–970.

Härmä, A. (**2003**). "Automatic recognition of bird species based on sinusoidal modeling of syllables," in *Proceedings of IEEE ICASSP*, pp. 545–548.

Kaewtip, K., Tan, L. N., Alwan, A., and Taylor, C. E. (**2013**). "A robust automatic bird phrase classifier using dynamic time-warping with prominent region identification," in *Proceedings of IEEE ICASSP*, pp. 768–772.

Kirschel, A. N. G., Cody, M. L., Harlow, Z. T., Promponas, V. J., Vallejo, E. E., and Taylor, C. E. (**2011**). "Territorial dynamics of Mexican Antthrushes *Formicarius moniliger* revealed by individual recognition of their songs," Ibis **153**, 255–268.

Kogan, J. A., and Margoliash, D. (**1998**). "Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study," J. Acoust. Soc. Am. **103**, 2185–2196.

Lee, C.-H., Hsu, S.-B., Shih, J.-L., and Chou, C.-H. (**2013**). "Continuous birdsong recognition using Gaussian mixture modeling of image shape features," IEEE Trans. Multimedia **15**, 454–464.

Lee, C. H., Lee, Y. K., and Huang, R. Z. (**2006**). "Automatic recognition of bird songs using cepstral coefficients," J. Info. Tech. App. **1**, 17–23.

Lindermuth, M. (**2010**). "Harma syllable segmentation in MATLAB," http://www.mathworks.com/matlabcentral/fileexchange/29261-harma-syllable-segmentation (Last viewed 5/1/2013).

McCowan, B., Doyle, L. R., Jenkins, J., and Hanser, S. F. (**2005**). "The appropriate use of Zipf's law in animal communication studies," Anim. Behav. **69**, F1–F7.

McNemar, Q. (**1947**). "Note on the sampling error of the difference between correlated proportions or percentages," Psychometrika **12**, 153–157.

Meliza, C. D., Keen, S. C., and Rubenstein, D. R. (**2013**). "Pitch- and spectral-based dynamic time warping methods for comparing field recordings of harmonic avian vocalizations," J. Acoust. Soc. Am. **134**, 1407–1415.

Mennill, D. J. (**2011**). "Individual distinctiveness in avian vocalizations and the spatial monitoring of behavior," Ibis **153**, 235–238.

Myers, C., Rabiner, L., and Rosenberg, A. E. (**1980**). "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," IEEE Trans. Acoust. Speech Signal Process. **28**, 623–635.

Neal, L., Briggs, F., Raich, R., and Fern, X. Z. (**2011**). "Time-frequency segmentation of bird song in noisy acoustic environments," in *Proceedings of IEEE ICASSP*, pp. 2012–2015.

Nemeth, E., Dabelsteen, T., Pedersen, S. B., and Winkler, H. (**2006**). "Rainforests as concert halls for birds: Are reverberations improving sound transmission of long song elements?," J. Acoust. Soc. Am. **119**, 620–626.

Pijanowski, B. C., Villanueva-Rivera, L. J., Dumyahn, S. L., Farina, A., Krause, B. L., Napoletano, B. M., Gage, S. H., and Pieretti, N. (**2011**). "Soundscape ecology: The science of sound in the landscape," Bioscience **61**, 203–216.

Ranjard, L., and Ross, H. (**2008**). "Unsupervised bird song syllable classification using evolving neural networks," J. Acoust. Soc. Am. **123**, 4358–4368.

Richards, D. G., and Wiley, R. H. (**1980**). "Reverberations and amplitude fluctuations in the propagation of sound in a forest: Implications for animal communication," Am. Nat. **115**(3), 381–399.

Sasahara, K., Cody, M. L., Cohen, D., and Taylor, C. E. (**2012**). "Structural design principles of complex bird songs: A network-based approach," PLoS One **7**, e44436.

Silagadze, Z. (**1997**). "Citations and the Zipf–Mandelbrot's law," Complex Syst. **11**, 487–499.

Stowell, D., Musevic, S., Bonada, J., and Plumbley, M. D. (**2013**). "Improved multiple birdsong tracking with distribution derivative method and Markov renewal process clustering," in *Proceedings of IEEE ICASSP*, pp. 468–472.

Tachibana, R. O., Oosugi, N., and Okanoya, K. (**2014**). "Semi-automatic classification of birdsong elements using a linear support vector machine," PLoS One **9**, e92584.

Tan, L. N., Kaewtip, K., Cody, M. L., Taylor, C. E., and Alwan, A. (**2012**). "Evaluation of a sparse representation-based classifier for bird phrase classification under limited data conditions," in *Proceedings of Interspeech*, pp. 2522–2525.

Tan, L. N., Kossan, G., Cody, M. L., Taylor, C. E., and Alwan, A. (**2013**). "A sparse representation-based classifier for in-set bird phrase verification and classification with limited training data," in *Proceedings of IEEE ICASSP*, pp. 763–767.

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

van den Berg, E., and Friedlander, M. P. (**2007**). "SPGL1: A solver for large-scale sparse reconstruction," http://www.cs.ubc.ca/labs/scl/spgl1 (Last viewed 5/1/2013).

van den Berg, E., and Friedlander, M. P. (**2008**). "Probing the pareto frontier for basis pursuit solutions," SIAM J. Sci. Comput. **31**, 890–912.

Wright, J., Yang, A., Ganesh, A., Sastry, S., and Ma, Y. (**2009**). "Robust face recognition via sparse representation," IEEE Trans. Pattern Anal. Mach. Intell. **31**, 210–227.

Yang, A., Wright, J., Ma, Y., and Sastry, S. (**2007**). "Feature selection in face recognition: A sparse representation perspective," UC Berkeley Tech. Report UCB/EECS-2007-99, pp. 1–17. http://www.eecs.berkeley.edu/Pubs/TechRpts/2007/EECS-2007-99.html (Last viewed 8/9/2013).

Yilmaz, E., Gemmeke, J. F., Van Compernolle, D., and Van Hamme, H. (**2013**). "Embedding time warping in exemplar-based sparse representations of speech," in *Proceedings of IEEE ICASSP*, pp. 8076–8080.

Zhang, L., Lin, F., and Zhang, B. (**2001**). "Support vector machine learning for image retrieval," in *Proceedings of IEEE International Conference on Image Processing*, pp. 721–724.