Acoustic model adaptation for ortolan bunting (*Emberiza hortulana* L.) song-type classification

Jidong Tao[a] and Michael T. Johnson
Speech and Signal Processing Laboratory, Marquette University, P.O. Box 1881, Milwaukee, Wisconsin 53233-1881

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

(Received 11 June 2007; revised 3 January 2008; accepted 4 January 2008)

Automatic systems for vocalization classification often require fairly large amounts of data on which to train models. However, animal vocalization data collection and transcription is a difficult and time-consuming task, so that it is expensive to create large data sets. One natural solution to this problem is the use of acoustic adaptation methods. Such methods, common in human speech recognition systems, create initial models trained on speaker independent data, then use small amounts of adaptation data to build individual-specific models. Since, as in human speech, individual vocal variability is a significant source of variation in bioacoustic data, acoustic model adaptation is naturally suited to classification in this domain as well. To demonstrate and evaluate the effectiveness of this approach, this paper presents the application of maximum likelihood linear regression adaptation to ortolan bunting (*Emberiza hortulana* L.) song-type classification. Classification accuracies for the adapted system are computed as a function of the amount of adaptation data and compared to caller-independent and caller-dependent systems. The experimental results indicate that given the same amount of data, supervised adaptation significantly outperforms both caller-independent and caller-dependent systems.
 [DOI: 10.1121/1.2837487]

PACS number(s): 43.66.Gf, 43.80.Ka, 43.72.Fx, 43.60.Uv [DOS] Pages: 1582–1590

I. INTRODUCTION

Hidden Markov models (HMMs) have been successfully applied to animal vocalization classification and detection in a number of species. Kogan and Margoliash (1998) and Anderson (1999) have shown that HMM-based classification is more robust to noise and more effective for highly confusable vocalizations than a dynamic time warping approach applied to the indigo bunting (*Passerina cyanea*) and zebra finch (*Taeniopygia guttata*). Other species in which HMM-based classification has been investigated include African elephants (*Loxodonta africana*) (Clemins *et al.*, 2005), beluga whale (*Delphinapterus leucas*) (Clemins and Johnson, 2005), ortolan bunting (*Emberiza hortulana* L.) (Trawicki *et al.*, 2005), red deer (*Cervus elaphus*) (Reby *et al.*, 2006), and rhesus macaques (*Macaca mulatto*) (Li *et al.*, 2007). HMM systems have been widely used to examine vocal repertoire, identify individuals, and classify vocalizations according to social context or behavior.

Typically, such classification systems are caller-independent (CI), meaning that the examples used for training the classifier come from a different set of individuals than those used for testing. In contrast to this, systems for human speech recognition are often speaker-dependent (SD), i.e., trained on the same individual who will be using the system, since given sufficient individual-specific training data SD systems have better performance than speaker-independent (SI) systems. When individual-specific training data are limited, an alternative is to use a speaker-adapted (SA) system. In this case a SI system is trained first and then the classification models are adapted with some individual-specific data, called adaptation data, to better account for individual variability in speech and pronunciation patterns. SA systems will typically have better overall accuracy than either SI or SD systems for small or moderate amounts of adaptation data. The error rate of a SD system may be as low as one-third that of a comparable SI speech recognition system tested on the same data (Hazen, 1998; Lee *et al.*, 1991), because individual speech differences are minimized in the SD system. The goal of using adaptation is to achieve performance approaching that of an ideal SD system using only limited amounts of speaker-specific data (Kuhn *et al.*, 2000).

Similarly, it is possible to develop analogous classification systems for animal vocalizations that are caller-dependent (CD) or caller-adapted (CA). The goal of this approach is to maximize the accuracy of the classifier while minimizing the amount of labor required to analyze and transcribe the collected data. Previous studies in animal vocalization analysis have found that individual vocal variability is one of the most important cues impacting vocalization related behavior study in bioacoustics (Reby *et al.*, 2006). Individual variability in acoustic structure has been described in many species such as bottlenose dolphins (*Tursiops truncates*) (Parijs *et al.*, 2002; Janik *et al.*, 2006), zebra finches (*Taeniopygia guttata*) (Vignal *et al.*, 2004), and Belding's ground squirrels (*Spermophilus beldingi*) (McCowan and

[a]Electronic mail: vjdtao@hotmail.com

Hooper, 2002). In ortolan buntings, song vocalization has been found to differ significantly between individuals in terms of repertoire content (Osiejuk *et al.*, 2003) and tonality (Osiejuk *et al.*, 2005). These differences have strong influence on species biology as ortolan bunting males were recently shown to discriminate between neighbors and strangers by song (Skierczyński *et al.*, 2007) and to differentiate response to songs composed of syllables originating from local or foreign population (Osiejuk *et al.*, 2007). This would imply that a CA system for animal vocalization analysis and classification should yield measurable improvements in overall accuracy and performance. Because both the data collection and analysis/transcription processes are much more difficult and time-consuming for most animal species than for human speech, utilizing a CA system to reduce the overall data requirements for developing automated classification systems may result in significant cost-savings. Additionally, cross comparisons of CD, CI, and CA recognition models have the potential to yield significant insight into the source of individual vocal variability.

The aim of this study is to demonstrate the use of adaptation for animal vocalization classification and examine the data requirements and degree of improvement provided by a CA system over comparable CI and CD systems. The CA system implemented for this task is based on the maximum likelihood linear regression (MLLR) technique (Leggetter and Woodland, 1995). The MLLR method works by clustering the states in an HMM into groups using a regression tree, then learning a maximum likelihood (ML) linear transformation for each group. The regression-based transformations tune the HMM mean and covariance parameters to each new individual represented by the adaptation data. To ensure all parameters can be adapted, a global transformation can be used for all HMMs in the system if only a small amount of adaptation data is presented, so that MLLR adaptation can improve recognition performance even with very limited adaptation data (Leggetter and Woodland, 1995). Results indicate that CA does in fact provide substantial performance improvement over both CI and limited-data CD systems.

II. DATA

A. Species under study

Ortolan buntings (*Emberiza hortulana* L.) are the focus of the current study. The species has declined steadily the last 50 years in Western Europe, and is currently listed in Norway as critically endangered on the Norwegian red-list. The population size is now only about 100 singing males and declines an average of 8% annually (Dale, 2001; Steifetten and Dale, 2006). The initial decline of the Norwegian population was probably due to the habitat loss related to changes in agriculture practices (Dale, 2001). However, 10 years of intensive study revealed that the main reason for the continuous decrease is female-biased dispersal pattern, which in isolated and patchy population seriously affects sex ratio, behavior of males, and breeding success measured at the population level (Dale *et al.*, 2005, 2006; Steifetten and Dale, 2006). It is hoped that increasing our understanding of male ortolan bunting vocalizations will enable us to better understand breeding behavior and reduce the risk of extinction.

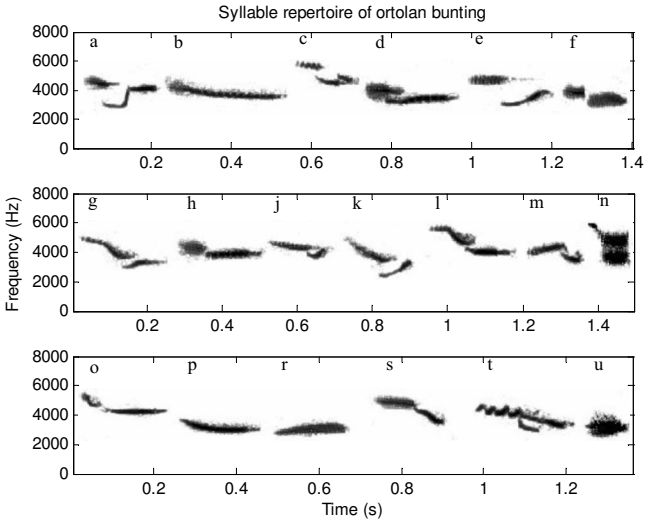

FIG. 1. Complete set of the 19-syllable repertoire of ortolan bunting.

Norwegian ortolan bunting vocalizations were collected from County Hedmark, Norway in May of 2001 and 2002 (Osiejuk *et al.*, 2003). The birds covered an area of approximately 500 km^2 on 25 sites, and males were recorded on 11 of those sites. A team of one to three research members who recognized and labeled the individual male buntings visited the sites. Overall, the entire sample population in 2001 and 2002 contains 150 males, 115 of which were color-ringed for individual identification. Because there are no known acoustic differences between the ringed and nonringed males, all data were grouped together for experimental use.

Ortolan buntings communicate through fundamental acoustical units called syllables (Osiejuk *et al.*, 2003). Figure 1 depicts the 19-syllable vocal repertoire used in this data set. Individual songs are grouped into song-type categories, e.g., *ab*, *cb*, that indicate the sequence of syllable types present. Each song type has many specific song variants, e.g., *aaaab*, *aaabb*, which indicate the exact repetition pattern. Figure 2 shows spectrograms of three specific type *ab* songs, song variants *aaaab*, *aaabb*, and *aaaabb*. The waveforms in Figs. 1 and 2 are low background noise exemplars, taken from different individuals to illustrate the repertoire.

B. Data collection

Vocalizations were recorded in the morning hours between 04:00 and 11:00 in each site, using a HHB PDR 1000 Professional DAT recorder with a Telinga V Pro Science parabola, a Sony TCD-D8 DAT recorder with a Sennheiser ME 67 shotgun microphone or an Aiwa HS-200 DAT recorder with a Sennheiser ME 67 shotgun microphone. All recordings were digitally transferred from Technics SV-DA 10 recorder via a SPDIF cable to a PC workstation with SoundBlaster Live! 5.1 at a sampling rate of 48 kHz with 16-bit quantization. For a more detailed description of the methods used to record the vocalizations, see Osiejuk *et al.* (2003, 2005).

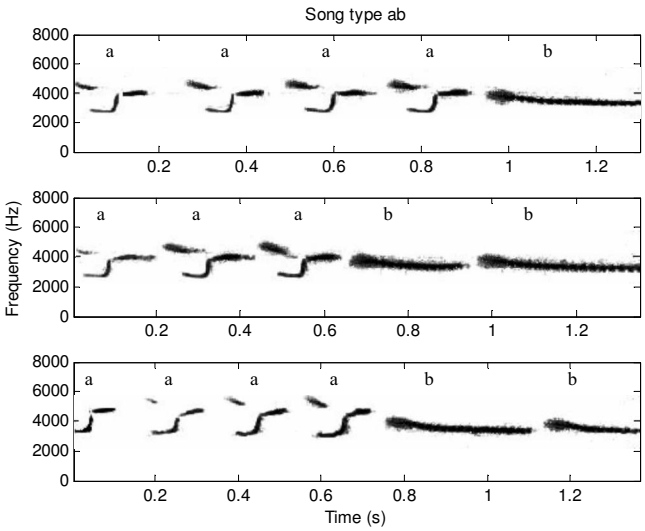

FIG. 2. *ab*-type song variation in ortolan bunting.

C. Data organization

The data set used here is a subset of the Osiejuk *et al.* data (Osiejuk *et al.*, 2003, 2005) including 60 song types and 19 syllables from 105 individuals. In selecting data for this study, calls containing syllables which were identified in only a single individual or a single song type were not included. Different individuals were selected for the training and testing/adaptation sets, balanced to get full coverage of all syllables in each set.

The protocol used to separate the data into training, test, and adaptation sets is as follows:

(1) Remove calls containing syllables identified in only a single individual or a single song type. This gives a resulting data set of 105 individuals, 60 call types, and 19 syllables.
(2) Select individuals for testing/adaptation.
  (a) Sort song types in increasing order according to number of examples.
  (b) Starting with the least common song type, select the individual with the highest number of examples in that song type (minimum two examples).
  (c) Repeat this process for each song type until the individuals selected for testing cover all 60 types.
  This results in a set of 30 individuals for testing/adaptation.
(3) Create explicit test and adaptation data sets by randomly dividing the data into test and adaptation sets for each selected individual, subject to a maximum of 30 vocalizations in each set for any one individual and song type.
(4) Group the remaining individuals into a training data set, again reducing the number of examples to a maximum of 30 for any one individual and song type.

Descriptive statistics of the resulting training, test, and adaptation sets are shown in Table I. From the above-detailed process it is clear that the 75 individuals in the training set are disjoint from the 30 individuals in the test/adaptation data, while the test and adaptation sets share the same group of individuals. All three sets have a full representation of syllables. Note that the training set does not cover the full range of 60 song types, but is still sufficient for training syllable-level HMMs for classification, as discussed in Sec. III. The size of the adaptation set is the same as that of the test set to allow the data to be used for training caller-dependent models as well as to allow a large range of variation for examining the impact of adaptation data quantity on performance.

TABLE I. Distribution of the number of individuals, song types, and vocalizations, and vocalizations with associated frequencies on individual, song type and syllable for training, test, and adaptation sets.

	Training set	Test set	Adaptation set
Number of individuals	75	30	30
Number of song types	53	60	60
Number of syllables	19	19	19
Number of vocalizations	2039	864	886
Mean vocalizations/caller	27.2	28.8	29.5
Mean vocalizations/type	38.5	14.4	14.8
Mean vocalizations/syllable	107.3	45.5	46.6

III. METHODS

A. Feature extraction

The primary features used in this HMM classification system are Greenwood function cepstral coefficients (GFCCs) (Clemins *et al.*, 2006; Clemins and Johnson, 2006). GFCCs are a species-specific generalization of mel frequency cepstral coefficients (MFCCs) (Huang *et al.*, 2001), one of the most common feature sets used in human speech recognition. The process for computing cepstral coefficients begins with segmenting vocalizations into evenly spaced appropriately sized windows (based on the frequency range and vocalization patterns of the species). For each window, a log magnitude fast Fourier transform (FFT) is computed and grouped into frequency bins. A discrete cosine transform is then taken to transform the log magnitude spectrum into cepstral values. For GFCCs, the frequency scale of the FFT is warped according to the Greenwood function (Greenwood, 1961) to provide a perceptually scaled axis. To do this, the parameters of the Greenwood function are estimated from the upper and lower bounds of the species' hearing range along with a warping constant of $k=0.88$ (LePage, 2003). Details of the warping equations and GFCC feature extraction process can be found in Clemins *et al.* (2006) and Clemins and Johnson (2006). Given basic information about a species frequency range, GFCCs provide an accurate and robust set of features to describe spectral characteristics over time.

In addition to the base set of GFCC features, energy is computed on the original time-domain data, and velocity and acceleration coefficients representing the first- and second-order rates of change are added. For the experiments described here, the vocalizations are segmented using 5 ms Hamming windows, with a 2.5 ms overlap. Twelve GFCCs plus normalized log energy along with velocity and acceleration coefficients are calculated, for a total of 39 features.

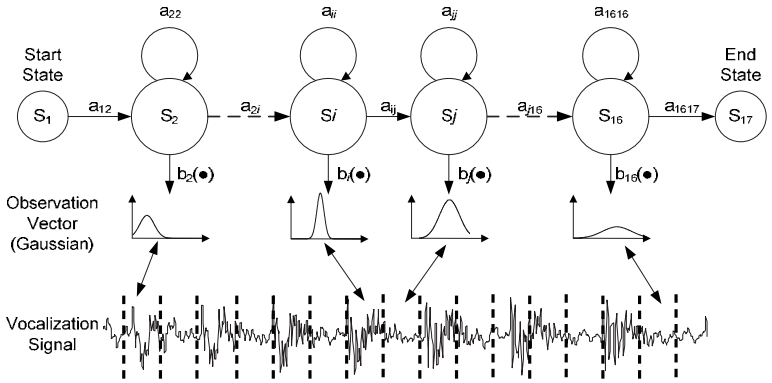

FIG. 3. A 15-state left-to-right hidden Markov model (HMM). Each state emits an observable vector of 39 GFCCs that are characterized by a single Gaussian model.

Frequency warping is done using a given hearing range from 400 to 7200 Hz, with 26 triangular frequency bins spaced across that range. Velocity and acceleration coefficients are computed using a five-window linear regression.

B. Acoustic models

HMMs (Rabiner and Juang, 1993) are the state-of-the-art approach for continuous speech recognition tasks. HMMs are statistical finite-state machines, where states represent spectrally stationary portions of the vocalization and transitions between states represent spectral transitions. This results in the ability to model spectral and temporal differences between an example vocalization and a trained HMM, with an implicit nonlinear time alignment.

In this work, each of the 19 ortolan bunting syllables is modeled with a 15-state left-to-right HMM, as illustrated in Fig. 3. Each state S_j is entered according to a transition probability a_{ij} from the previous state S_i. An observation feature vector o_t at time t is generated from the current state S_j based on a probability distribution $b_j(o_t)$, which in this work is a diagonal covariance Gaussian model.

During the training process, the Baum–Welch algorithm for expectation maximization (EM) (Baum *et al.*, 1970; Moon, 1996) is used to estimate the HMM parameters that maximize the joint likelihood of all training observation sequences. For classification, the Viterbi algorithm (Forney, 1973) is used to find the model sequence having the highest likelihood match to the sequence of test features.

C. Maximum likelihood linear regression adaptation

Once an HMM has been trained, the model parameters can be adapted to tailor the model to more domain-specific data. The key parameters for adaptation are the means and variances corresponding to each state distribution $b_j(o_t)$. In the MLLR adaptation approach, two linear transformation matrices are estimated for each state, one for the mean vector and one for the covariance matrix, under a maximum likelihood criteria function (Leggetter and Woodland, 1995). The underlying principle is to provide a reestimation approach that is consistent with maximizing the HMM likelihood while keeping the number of free parameters under control, thus requiring a smaller amount of adaptation data and allowing for rapid adaptation. MLLR has been widely used to obtain adapted models for both new speakers and new environmental conditions (Huang *et al.*, 2001).

In order to maximize the use of adaptation data, the required linear transformation matrices for each state are grouped into broad acoustic/syllable classes so that the overall number of free parameters is significantly less than the number of mean vectors. This is accomplished by building a regression class tree to cluster states with similar distributions into regression classes, the members of which share the same linear transformation.

The regression class tree is constructed so as to cluster together components that are close acoustically, using the original CI model set (independent of any new data). A centroid splitting algorithm using a Euclidean distance measure is applied to construct the tree (Young *et al.*, 2002). The terminal nodes or leaves of the tree specify the finest possible resolution groupings for transformation, and are termed the base (regression) classes. Each Gaussian component from the CI model set belongs to one specific base class.

The amount and type of adaptation data that is available determines exactly which transformations are applied to the original model. This makes it possible to adapt all models, even those for which there were no observations in the adaptation data, because the regression tree representation allows for adaptation to be done based on similar models that are present in the data. When more adaptation data are available, a larger number of unique transformations are applied, in accordance with the structure of the regression tree.

Specifically, the mean vector μ_i for each state can be transformed using

$$\hat{\mu}_i = A_c m_i + b_c = W_c \mu_i, \tag{1}$$

where m_i is the original mean vector for state i, $\mu_i = [1 \quad m_i^T]^T$ is the extended mean vector incorporating a bias vector b_c, A_c is the transformation matrix for regression class C, and W_c is the corresponding extended transformation matrix $[b_c \quad A_c]$.

While the regression tree itself is built from the caller-independent models, the number of regression classes C actually implemented for a particular set of adaptation data is variable, depending on the data's coverage of the classes. A tiny amount of adaptation data would result in only a single transformation matrix being used across all classes, or even no adaptation at all.

The required transformation matrix W_c for adapting the mean vector μ_i as indicated in Eq. (1) is obtained using the EM technique. The resulting reestimation formula W_c is given by

$$w_q = \left[\sum_t \sum_{i\in C} \xi_t(i)\Sigma_i^{-1} x_t \mu_i^T\right]_q \left[\sum_{i\in C}\left[\sum_t \xi_t(i)\Sigma_i^{-1}\right]_{qq} (\mu_i \mu_i^T)_q\right]_q^{-1}, \tag{2}$$

where w_q is the qth row vector of W_c being estimated, $\xi_t(i)$ is the occupancy likelihood of state i, Σ_i and μ_i are the corresponding diagonal covariance matrix and the extended mean vectors, and x_t is the adaptation data feature vector at time t. The subscripts q and qq in this equation are used to indicate the corresponding row and diagonal element of a matrix, respectively, for compactness of representation.

The Gaussian covariance Σ_i for state i is transformed using

$$\hat{\Sigma}_i = (\Sigma_i^{1/2})^T H_i \Sigma_i^{1/2}, \tag{3}$$

where the diagonal linear transformation matrix H_i is estimated via

$$H_i = \frac{\sum_{j\in C} (\Sigma_j^{-1/2})^T \left[\sum_t \xi_t(j)(x_t-\mu_j)(x_t-\mu_j)^T\right]\Sigma_j^{-1/2}}{\sum_{j\in C}\sum_t \xi_t(j)}. \tag{4}$$

Typically, transformation matrices converge in just a few iterations. At each iteration, all matrices are initialized to the identity transformation, and recognition likelihood statistics are accumulated over the data using the current model. Means alone or both means and variances are then updated using Eqs. (2) and (4). Typically the impact of variance adaptation is much less significant than that of mean adaptation. Transforming the variances can still be significant, however, because by nature variances in a CI system, which come from many individuals, are higher than those of the corresponding CD systems.

To implement the adaptation process, transformation matrices are initialized to the identity matrix. Using the original CI model and the prebuilt regression tree, state occupancies are calculated for all possible states, and the occupation counts are grouped for each class in the tree and compared to a threshold to determine exactly which transformations are to be applied. Following this, several iterations of Eqs. (2) and (4) are run to estimate and apply the mean and variance transformation matrices and create a new adapted model.

MLLR adaptation can be used in various different modes. Supervised adaptation refers to adaptation done using data accompanied by expert transcriptions, so that the process is applied to known model components. It is also possible to implement unsupervised adaptation, where before each adaptation iteration a recognition pass is performed to determine which models to adapt. Clearly, if the initial CI models are not a good match to the new domain, unsupervised adaptation could potentially fail to improve or even degrade the overall system by adapting an incorrect selection of models using the new data. It is also possible to apply

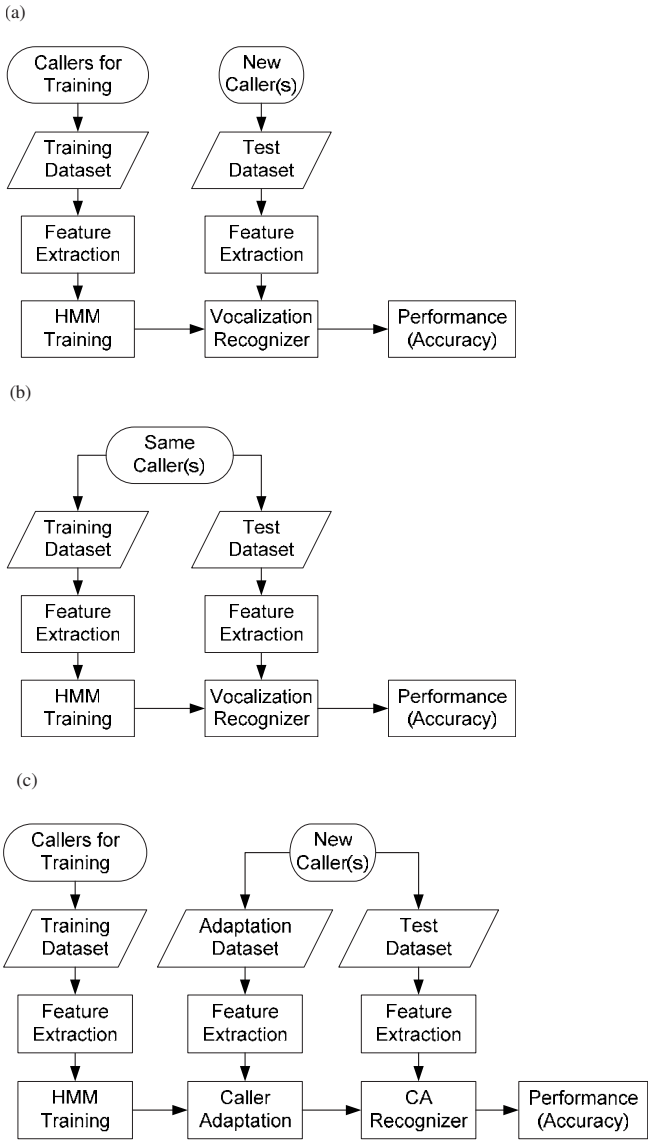

FIG. 4. Vocalization recognition systems. (a) Caller-independent, with separate individuals for the training and testing data. (b) Caller-dependent, with training and testing data coming from the same group of individuals. (c) Caller-adapted, with separate training and testing data, but with a portion of the testing data pulled out and used for adaptation.

adaptation methods either statically, where the entire amount of adaptation data is used together, or incrementally, where adaptation is done repeatedly as the amount of adaptation data increases.

D. Song-type recognition experiments

Song-type recognition experiments were implemented on the ortolan bunting data set as previously described. The goal of these experiments is to compare how well a CA HMM system performs compared to a baseline CI system. For reference, a fully CD system was also implemented.

The recognition models used for all experiments were 15 state single Gaussian HMMs with diagonal covariance matrices. The feature vector used for classification, as described previously, was a 39-element vector that included 12 GFCCs plus normalized log energy, accompanied by delta and delta–delta coefficients. The software toolkit HTK ver-

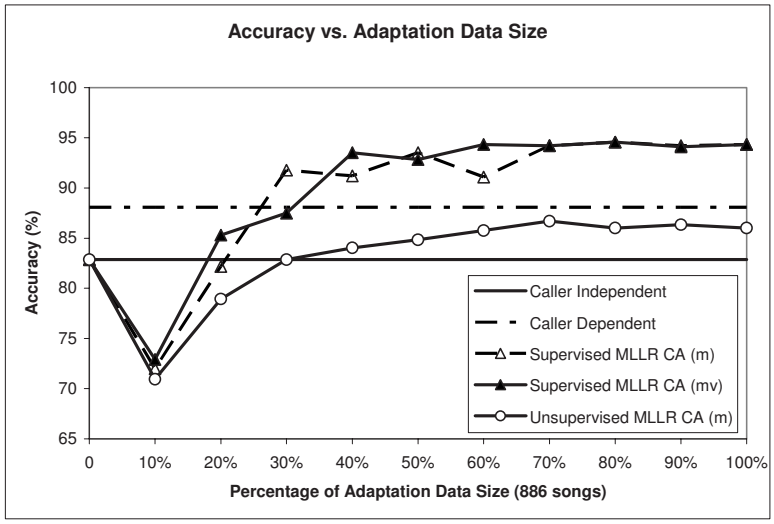

FIG. 5. The CI, CD, and CA system performance with varying amounts of adaptation data. MLLR adaptation was run in both supervised and unsupervised modes on the same data. Two types of adaptation in supervised mode are used: mean (m) only, and both mean and variance (mv).

sion 3.2 (Young *et al.*, 2002) was used to implement the HMMs, perform adaptation, and analyze classification performance. There were 19 different HMMs trained for each system, one for each syllable.

The following song-type recognition systems were implemented for comparison:

CI: the baseline caller-independent models. The system diagram for the CI system is shown in Fig. 4(a). There was no overlap between the training individuals and test individuals, with 75 and 30 individuals in the two data sets, respectively.

CD: the caller-dependent models. The system diagram for the CD system is shown in Fig. 4(b). The training and testing data were separate but came from the same individuals. The training data used for the CD experiments was the same as the adaptation data used for the CA experiments.

CA: the caller-adapted models. The system diagram for the MLLR adaptation systems is shown in Fig. 4(c). The training and testing data were from separate individuals, and the test data were further split into adaptation data and final test data. Three different CA experiments were implemented: supervised mean adaptation, supervised mean and variance adaptation, and unsupervised mean adaptation.

In order to see how the amount of adaptation data affected the results, each adaptation method was implemented multiple times, using increasing amounts of adaptation data. This was done in 10% increments, starting with 0% (no adaptation, equivalent to the initial CI system), then 10%, 20%, and so on up to 100% (full adaptation set in use).

IV. RESULTS

Overall results of the adaptation process can be seen in Fig. 5. The baseline CI system has an 82.9% accuracy, while the CD system has an 88.1% accuracy. Unsupervised adaptation of the means has a peak accuracy of 86.7% and a final accuracy of 86.0% using the full data set. Supervised adaptation yields the highest accuracy, 94.3% overall, representing a net gain of 11.4 percentage points (66% reduction in error) over CI and 6.2 percentage points (52% reduction in error) over CD.

The supervised adaptation using means (m) and that using both means and variances (mv) show a different pattern for lower amounts of data, but reach exactly the same accuracy, 94.3%, as the adaptation data increases. The supervised methods significantly outperform both the CI and the CD systems, reaching the performance level of the CI system at about 20% data and that of the CD at about 30% data. The unsupervised adaptation results, as expected, trail those of the supervised system, but are still able to significantly outperform the baseline CI system.

A. Detailed recognition results across specific individuals

Table II displays the comparison of CI, CD, and CA (full adaptation set) for each individual in the test set, along with the distributions of song types and syllables for each.

Note that in a few cases, 9 out of 30 individuals, the CD system actually gives a lower accuracy than the original CI system. In two of these cases, even the CA system still has a lower accuracy than the CI system. Comparing the CD to the CA systems, only one individual has an accuracy that is lower in the adapted system.

B. Detailed recognition results as a function of song-type frequency

In order to examine the recognition accuracy as a function of how often each song occurs (i.e., the amount of data in the training and test sets for each song), an additional analysis is done by rank-ordering the songs according to frequency of occurrence and plotting the accuracy.

The overall system recognition cumulative accuracies by classified song types are shown in Fig. 6. The CI results drop from 95.5% for the most common *ab* song down to 82.9%

TABLE II. Vocalization recognition comparison among CI, CD, and CA for each new individual, with the distributions of adapted songs, song types, and syllables of each individual bird. Overall accuracies with variances are CI 82.9 ± 16.4%, CD 88.1 ± 10.1%, and CA 94.3 ± 7.3%.

Caller ID	Adapt songs	song types	Syllables	CI (%)	CD (%)	CA (%)
2044	30	ab, cb, c, a	a, b, c	53.3	83.3	100
2049	30	huf, h, jufb, juf, hu	b, f, h, j, u	60	90	100
347	30	ab, gb, hufb, ghuf	a, b, f, g, h, u	63.3	93.3	96.7
2004	30	cb, eb, huf, h, jufb, hufb, c, cufb	b, c, e, f, h, j, u	56.7	90	90
2046	8	cb, er	c, b, e, r	66.7	83.3	100
2026	22	h, jufb, juf, hu, ju	b, f, h, j, u	68.2	72.7	100
2029	30	cb, gb, guf, gufb, gcb, gluf	b, c, f, g, l, u	46.7	70	76.7
385	30	h, ef, e	e, f, h	83.3	96.7	100
502	30	ab, cb, cufb, cf, cfb, tb, sfb	a, b, c, f, s, t, u	83.3	93.3	100
2022	30	huf, jufb, juf, j, ju	b, f, h, j, u	80	90	96.7
2010	30	ab, eb, ef	a, b, e, f	86.2	96.6	100
1303	40	cb, gb, guf, c, g, gh, gu, ch	b, c, f, g, h, u	70	86.7	83.3
205	30	ab, p, pb	a, b, p	73.3	60	83.3
165	30	cd, eb, cdb, suf, tb, sb, tuf	c, d, e, f, s, t, u	90	83.3	96.7
384	30	eb, cufb, cuf	b, c, e, f, u	93.3	93.3	100
430	30	ab, huf, h, hd, a	a, b, d, f, h, u	93.3	93.3	100
176	60	eb, huf, guf, hufb, luf, gufb, lufb	c, d, e, f, s, t, u	69	72.4	74.1
1201	30	gb, h, hb, gh, ghb, hgb	b, g, h	83.3	80	86.7
2038	30	ab, cb, cd, cdb	a, b, c, d	93.3	96.7	96.7
39	30	hd, gd	h, g, d	100	100	100
106	30	ab, kb, a, k	a, b, k	100	100	100
413	25	jufb, jb	b, f, j, u	100	100	100
1030	30	cd, od	c, d, o	93.3	83.3	93.3
1903	30	gb, nu, nuf, n	b, f, g, n, u	100	90	100
2011	30	cb, ghuf, gh, ghu	b, c, g, h, u	100	100	100
2021	30	gb, huf, h, hufb, ghuf, gh, ghufb, hu	b, f, g, h, u	96.7	80	96.7
2025	13	h, gh, hr	g, h, r	91.7	83.3	91.7
2030	30	ab, cb, kb, kab	a, b, c, k	100	100	100
314	30	gb, h, jd, hr	b, d, g, h, j, r	100	93.3	96.7
239	28	ab, kb, luf, mluf, mufl	a, b, f, k, l, m, n	100	95.7	95.7
Total	886	**60 song types**	**19 syllables**	**82.9**	**88.1**	**94.3**

for the least frequent song type, *sfb*. It is interesting to note that the impact of song-type frequency is less pronounced for the CA systems as compared to the CI and CD systems, with smoother accuracy curves across song type.

V. CONCLUSIONS

This work demonstrates the advantages of using an acoustic model adaptation system in classifying ortolan bunting vocalizations. There are two key advantages illustrated by these experiments. The first is that a caller-adapted recognition system typically gives significant performance improvement over either caller-independent or caller-dependent systems. The second is that adaptation provides for extremely efficient data utilization, which is very important for bioacoustics tasks where data collection and labeling is difficult.

The results given here suggest that the classification accuracy of many systems could be improved using adaptation, since individual vocal variation is typically one of the most important factors affecting performance. In essence, caller adaptation works like a flexible interpolation between an independent and a fully dependent system. An adapted system starts from a baseline independent system, surpasses both independent and dependent systems, and approaches an ideal (well trained with unlimited size of individual identity labeled data) dependent system. As the amount of data increases, the specificity of the adaptation is improved through the creation of a larger regression tree.

Although a caller-dependent system with unlimited data is theoretically ideal, it is often impractical because of the large amount of data required for each individual to build well trained models. The performance of such a system will be low if the number of individual vocalizations used to train the HMMs is limited, and recognition accuracy for any new individuals who are not present in the training set will be especially low. In contrast, an adaptation-based system overcomes these limitations by taking advantage of existing well trained CI models. With a moderate amount of labeled adaptation data, an adapted system generally achieves better performance than a CI or even a CD system. Even in unsupervised mode where *no* transcriptions at all are used, an adapted system may still approach the accuracy of a caller-dependent system as the size of adaptation data increases, as illustrated in the experimental results in Fig. 5.

The underlying reason for the high accuracy of adaptation systems is data utilization efficiency. Vocalization data

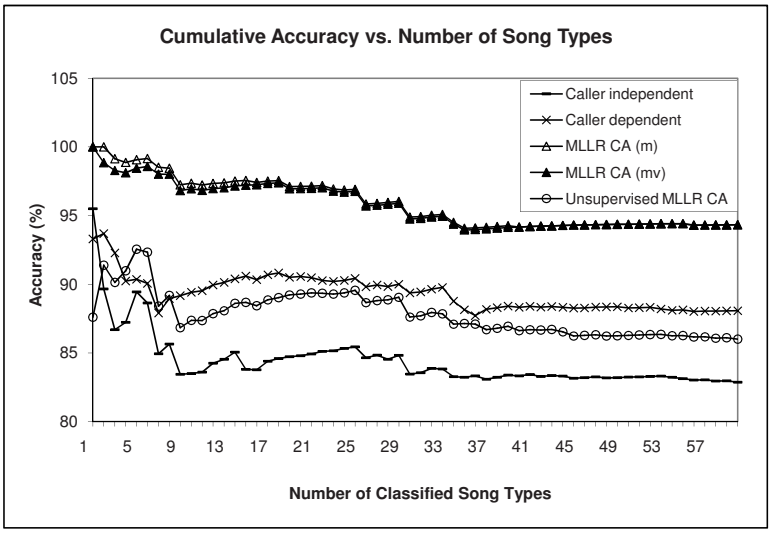

FIG. 6. Vocalization recognition cumulative accuracy vs number of classified song types for CI, CD, and CA systems.

collection is often difficult due to environmental constraints and inconsistent vocal repertoire across individuals, and data transcription or labeling is a time-consuming task requiring great expertise, so it is expensive to develop large data sets. Acoustic model adaptation is a natural solution to this problem, making the most effective use of the data regardless of how much is available. A CA system is initialized by starting from a CI system that is relative cheaper in effort required for data labeling because it does not need individual identity, then adapted using a much smaller amount of identity transcribed adaptation data to customize the system to the new individual vocal models. This enables a controlled trade-off between data labeling effort and system performance. In other words, maximum system performance is obtained with the minimum effort on data labeling.

The acoustic model adaptation methods presented here are applicable to a wide variety of species. Although each species has different vocal characteristics, individual vocal variability is nearly always present. Applying adaptation allows us to achieve high performance in classifying animal vocalizations with a small amount of available data.

ACKNOWLEDGMENTS

This material is based on work supported by National Science Foundation under Grant No. IIS-0326395. The study of ortolan buntings in Norway was supported by the Polish State Committee for Scientific Research, Grant Nos. 6 P04C 038 17 and 3 P04C 083 25 and Adam Mickiewicz University grant no. PBWB-301/2001.

Anderson, S. E. (**1999**). "Speech recognition meets bird song: A comparison of statistics-based and template-based techniques," J. Acoust. Soc. Am. **106**, 2130.

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

Clemins, P. J., and Johnson, M. T. (**2005**). "Unsupervised classification of beluga whale vocalizations," J. Acoust. Soc. Am. **117**, 2470(A).

Clemins, P. J., and Johnson, M. T. (**2006**). "Generalized perceptual linear prediction features for animal vocalization analysis," J. Acoust. Soc. Am. **120**, 527–534.

Clemins, P. J., Johnson, M. T., Leong, K. M., and Savage, A. (**2005**). "Automatic classification and speaker identification of African elephant (*Loxodonta africana*) vocalizations," J. Acoust. Soc. Am. **117**, 956–963.

Clemins, P. J., Trawicki, M. B., Adi, K., Tao, J., and Johnson, M. T. (**2006**). "Generalized perceptual features for vocalization analysis across multiple species," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Paris, France, Vol. **33**, pp. 1253–256.

Dale, S. (**2001**). "Causes of population decline in ortolan bunting in Norway," *Proceedings of the Third International Ortolan Symposium*, Poznan, Poland, pp. 33–41.

Dale, S., Lunde, A., and Steifetten, Ø. (**2005**). "Longer breeding dispersal than natal dispersal in the ortolan bunting," Behav. Ecol. Sociobiol. **16**, 20–24.

Dale, S., Steifetten, Ø., Osiejuk, T. S., Losak, K., and Cygan, J. P. (**2006**). "How do birds search for breeding areas at the landscape level? Interpatch movements of ortolan buntings," Ecography **29**, 886–898.

Forney, G. D. (**1973**). "The Viterbi Algorithm," Proc. IEEE **61**, 268–278.

Greenwood, D. D. (**1961**). "Critical bandwidth and the frequency coordinates of the basilar membrane," J. Acoust. Soc. Am. **33**, 1344–1356.

Hazen, T. J. (**1998**). "The use of speaker correlation information for automatic speech recognition," Ph.D. dissertation, MIT, Cambridge.

Huang, X., Acero, A., and Hon, H.-W. (**2001**). *Spoken Language Processing* (Prentice Hall, Upper Saddle River, NJ).

Janik, V. M., Sayigh, L. S., and Wells, R. S. (**2006**). "Signature whistle shape conveys identity information to bottlenose dolphins," Proc. Natl. Acad. Sci. U.S.A. **103**, 8293–8297.

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.

Kuhn, R., Junqua, J. C., Nguyen, P., and Niedzielski, N. (**2000**). "Rapid speaker adaptation in eigenvoice space," IEEE Trans. Speech Audio Process. **8**, 695–707.

Lee, C. H., Lin, C. H., and Juang, B. H. (**1991**). "A study on speaker adaptation of the parameters of continuous density hidden Markov models," IEEE Trans. Signal Process. **39**, 806–814.

Leggetter, C. J., and Woodland, P. C. (**1995**). "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Comput. Speech Lang. **9**, 171–185.

LePage, E. L. (**2003**). "The mammalian cochlear map is optimally warped," J. Acoust. Soc. Am. **114**, 896–906.

Li, X., Tao, J., Johnson, M. T., Soltis, J., Savage, A., Leong, K. M., and Newman, J. D. (**2007**). "Stress and emotion classification using jitter and shimmer features," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, HI, pp. IV1081–1084.

McCowan, B., and Hooper, S. L. (**2002**). "Individual acoustic variation in Belding's ground squirrel alarm chirps in the High Sierra Nevada," J. Acoust. Soc. Am. **111**, 1157–1160.

Moon, T. K. (**1996**). "The expectation-maximization algorithm," IEEE Signal Process. Mag. **13**, 47–60.

Osiejuk, T. S., Ratyńska, K., and Cygan, J. P. (**2007**). "What makes a 'local song' in a population of ortolan buntings without common dialect?," Anim. Behav. **74**, No. 1, pp. 121–130.
Osiejuk, T. S., Ratyńska, K., Cygan, J. P., and Dale, S. (**2003**). "Song structure and repertoire variation in ortolan bunting (*Emberiza hortulana* L.) from isolated Norweigian population," Ann. Zool. Fenn. **40**, 3–16.
Osiejuk, T. S., Ratyńska, K., Cygan, J. P., and Dale, S. (**2005**). "Frequency shift in homologue syllables of the Ortolan Bunting *Emberiza hortulana*," Behav. Processes **68**, 69–83.
Parijs, S. M. V., Smith, J., and Corkeron, P. J. (**2002**). "Using calls to estimate the abundance of inshore Dolphins: A case study with Pacific humpback dolphins (Sousa Chinensis)," J. Appl. Ecol. **39**, 853–864.
Rabiner, L. R., and Juang, B. H. (**1993**). *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliffs, NJ).
Reby, D., André-Obrecht, R., Galinier, A., Farinas, J., and Gargnelutti, B. (**2006**). "Cepstral coefficients and hidden Markov models reveal idiosyncratic voice characteristics in red deer (*Cervus elaphus*) stags," J. Acoust. Soc. Am. **120**, 4080–4089.
Skierczyński, M., Czarnecka, K. M., and Osiejuk, T. S. (**2007**). "Neighbour-stranger song discrimination in territorial ortolan bunting Emberiza hortulana males," J. Avian Biol. **38**, No. 4, pp. 415–420
Steifetten, Ø., and Dale, S. (**2006**). "Viability of an endangered population of ortolan buntings: The effect of a skewed operational sex ratio," Biol. Conserv. **132**, 88–97.
Trawicki, M. B., Johnson, M. T., and Osiejuk, T. S. (**2005**). "Automatic song-type classification and speaker identification of Norwegian ortolan bunting (*Emberiza hortulana*) vocalizations," IEEE Workshop on Machine Learning for Signal Processing, Mystic, CT, pp. 277–282.
Vignal, C., Mathevon, N., and Mottin, S. (**2004**). "Audience drives male songbird response to partner's voice," Nature (London) **430**, 448–451.
Young, S., Evermann, G., Hain, T., Kershaw, D., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (**2002**). *The HTK Book (for HTK Version 3.2.1)* (Cambridge University Engineering Department, Cambridge, UK).