AUTOMATIC SONG-TYPE CLASSIFICATION AND SPEAKER IDENTIFICATION OF NORWEGIAN ORTOLAN BUNTING (*EMBERIZA HORTULANA*) VOCALIZATIONS

Marek B. Trawicki & Michael T. Johnson

Marquette University
Department of Electrical and Computer Engineering
Speech and Signal Processing Lab, 518
P.O. Box 1881
Milwaukee, WI USA 53201-1881
{marek.trawicki, mike.johnson}@marquette.edu

Tomasz S. Osiejuk

Adam Mickiewicz University
Department Behavioural Ecology
Umultowska 89, 61-614
Poznań, Poland
http://www.behaecol.amu.edu.pl
osiejuk@amu.edu.pl

ABSTRACT

This paper presents an approach to song-type classification and speaker identification of Norwegian Ortolan Bunting (*Emberiza Hortulana*) vocalizations using traditional human speech processing methods. Hidden Markov Models (HMMs) are used for both tasks, with features including Mel-Frequency Cepstral Coefficients (MFCCs), log energy, and delta (velocity) and delta-delta (acceleration) coefficients. Vocalizations were tested using leave-one-out cross-validation. Classification accuracy for 5 song-types is 92.4%, dropping to 63.6% as the number and similarity of the songs increases. Song-type dependent speaker identification rates peak at 98.7%, with typical accuracies of 80-95% and a low end at 76.2% as the number of speakers increases. These experiments fit into a larger framework of research working towards methods for acoustic censusing of endangered species populations and more automated bioacoustic analysis methods.

1. INTRODUCTION

There is currently great interest in the study of Norwegian Ortolan Bunting (*Emberiza Hortulana*) vocalizations [1]. Even though Ortolan Buntings are not currently on any endangered species' lists, there has been a steady decline in their population over the past fifty-years, most recently at a rate of about 8% per year during 1996 – 2000 [2]. Traditional reasons for such population loss - variation in habitat, mortality during migration, nest egg loss, and human and animal predators - have not been the primary causes for this species. Instead, research has indicated that there is a surplus of isolated male buntings that attracts only about two-thirds of female buntings for breeding purposes [2]. Through better understanding of the connection between male bunting vocalizations and their behavior patterns, steps may be possible to correct the population decline and reduce future risk of extinction. The ability to analyze and classify Ortolan Bunting vocalizations may contribute to the creation of better habitat management and species survival plans. Long term, such methods may lead to creation of acoustic technologies for remote monitoring and automatic censusing methods for many different species. In a recent report by the Oceanic Research committee of the International Council for Science, Christine Erbe of the Institute of Ocean Sciences states "For example, in my opinion, only acoustics has the potential for long-term automatic and objective censusing without recurring and ongoing costs of personnel and ship or plane time," [3] referring to marine mammal censusing.

There has historically been very little research in the automatic analysis of bird vocalizations. Anderson, Davis, and Margoliash employed the Dynamic Time Warping (DTW) algorithm to continuous recordings of Indigo Bunting (*Passerina cyanea*) and Zebra Finch (*Taeniopygia guttata*) to recognize birdsong syllables [4]. In their work, they discovered that HMMs usually outperformed DTW for relatively noisy recordings and highly confusable song-types. Kogan and Margoliash compared DTW and Hidden Markov Model (HMM) approaches on these same species [5] and found upwards of 97% syllable and song-type accuracy on low-noise signals. Härmä used sinusoidal syllable models across several species to perform recognition [6] and noted the benefits of using syllables rather than song-types. In each case, the researchers automated the entire process of vocalization analysis. This is a contrast to the dominant approach in the field, with many features commonly labeled by hand through interactive spectrogram analysis, a time-intensive and sometimes subjective process. A better strategy would be to have "automatic recognition followed by limited inspection by human experts" [5].

This paper is organized into the following sections: Data (Section 2), Feature Extraction and Models (Section 3), Experiments and Results (Section 4), Conclusions

(Section 5), Acknowledgements (Section 6), and References (Section 7).

2. DATA

Norwegian Ortolan Bunting vocalization data was collected from County Hedmark, Norway in May of 2001 and 2002 [1]. Although the birds covered an area of approximately 500 km^2 on twenty-five sites, males (color-ringed and non-color-ringed) were only recorded on eleven of those sites. The sites were visited by a team of one to three research members who recognized and labeled the individual male buntings. Overall, the entire sample population in 2001 and 2002 contains 150 males, where 115 of them are of the color-ringed variety. Because there are no known acoustic differences between the ringed and non-ringed males, all data was grouped together for these experiments.

Ortolan Buntings communicate with each other through fundamental acoustical units called syllables, analogous to phonetic units in human speech. Figure 1 depicts the entire 20-syllable vocal repertoire. To produce a song, the syllables are joined in sequence, creating multiple song-types, e.g. *ab*, *cb*, *huf*, and many specific song variants, e.g. *aaaabb*, *ccccbbb*, *hhhuff*. Table 1 lists the 10 most common song-types, along with their relative frequencies and durations.

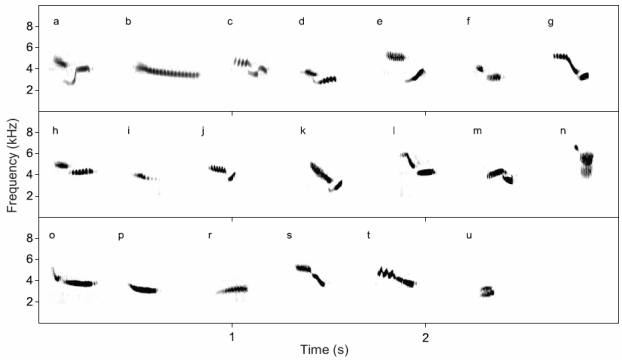

Figure 1: Syllable Repertoire of Ortolan Bunting

In this data set, there are a 63 song-types with 234 distinct variants [1].

Song-Type	Average Time (s)	Average Length (syllables)	Frequency Count
ab	1.74	7.12	3393
cb	1.63	6.63	1575
cd	1.66	7.54	897
eb	1.56	6.08	895
ef	1.67	7.97	617
gb	1.39	5.58	448
guf	1.71	8.12	390
h	1.16	5.00	268
huf	1.62	8.24	250
jufb	1.84	9.35	238

Table 1: Song-Types with Associated Frequencies and Durations

3. FEATURE EXTRACTION AND CLASSIFICATION MODELS

3.1. Feature Extraction

Mel-Frequency Cepstral Coefficients [7] are the most common feature representation in modern speech processing systems. Based on the Mel frequency scale for human perception, MFCCs have demonstrated superior performance and are well suited to capture significant acoustic and perceptual information in the signal [8]. Although avian species do not have this same perceptual scale, their auditory systems are similar in structure to humans, and the physical characteristics of the basilar membrane produce a logarithmic frequency characteristic similar to the Mel scale.

Extracting MFCC features from the vocalizations begins by segmenting the waveform into frames. The frames are then parameterized into speech vectors containing 12 Mel-Frequency Cepstral Coefficients (MFCCs), plus log energy, and delta (velocity) and delta-delta (acceleration) coefficients. Figure 2 shows the algorithm for computing the MFCCs from each frame. The Fast Fourier Transform (FFT) is calculated from the pre-emphasized and Hamming windowed signal frame, then the FFT magnitude is input to a series of Mel-Scale spaced filterbanks. Finally, the log filterbank amplitudes are used to calculate the MFCCs using the Discrete Cosine Transform (DCT). Log energy and time derivatives (delta and delta-delta) are then appended to the feature vector [9].

For these experiments, a frame size of 25 ms and step size of 10 ms has been used, based on results from prior work with the Zebra Finch and Indigo Bunting [5].

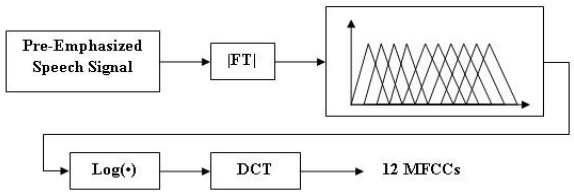

Figure 2: MFCC Feature Extraction Algorithm

3.2. Acoustic Models

Hidden Markov Models (HMMs) are the most common approach for applications such as speech recognition and speaker identification [10, 11]. Compared with the statistical analysis, spectrogram correlation, matched filter, and Artificial Neural Network (ANN) methods commonly used in bioacoustics, HMMs have the advantage of being able to incorporate a non-linear time alignment between models and waveforms. In addition, the structure of HMMs allows for more complex recognition models and language constraints.

As displayed in Figure 3, HMMs are finite-state machines, with transition between states representing changes in time and probability distributions within each state representing feature patterns within spectrally stationary portions of the waveform. Specifically, the models comprise states $S = \{S_1, S_2, \ldots, S_T\}$, transitional probabilities a_{ij}, and output probabilities $b_j(\cdot)$. Output probabilities are typically Gaussian Mixture Models (GMMs) whose distributions are a weighted sum of multi-variant Gaussian densities. For speech applications, HMMs are constrained as being left-to-right to model time progression.

In both the song-type and speaker identification experiments done here, each HMM is a 3-state single-mixture Gaussian model representing one syllable.

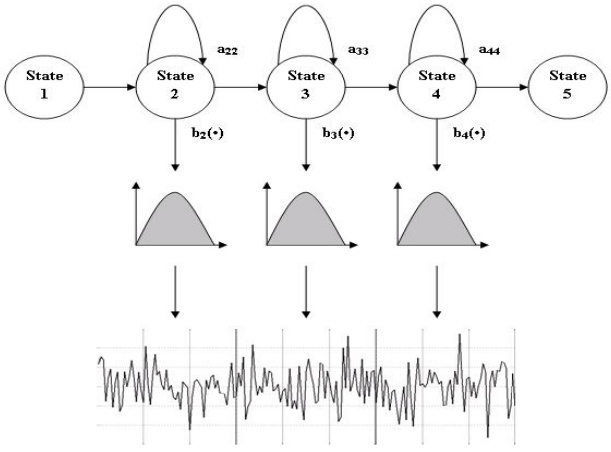

Figure 3: Example of a 3-State HMM

3.3. Language Models

In order to model the variants of each song-type, a grammar model was constructed. Figure 4 shows the topology of these grammar constraints, which allow for recognition of only valid song variants. During classification, each syllable in this topology is represented by its associated 3-state HMM. The ability of the HMM approach to incorporate constraints such as these is a significant benefit for acoustic analysis within structured domains, such as the song variants of the Ortolan Buntings.

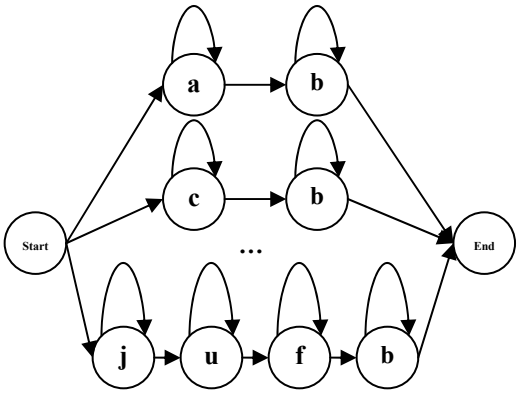

Figure 4: Language Model

4. EXPERIMENTS & RESULTS

Experiments include both song-type classification and speaker identification. There is existing prior work on the use of MFCC-based features to classify both Indigo Bunting (*Passerina Cyanea*) and Zebra Finch (*Taeniopygia Guttata*) song-types with accuracies of 92.3% (Indigo Bunting) and 83.9% (Zebra Finch) [5]. There is not as of yet existing work in speaker-identification to indicate baseline results.

Figure 5 illustrates the entire classification and identification process. In both experiments, the HMMs are 3-state, left-to-right models with a single Gaussian Mixture Model (GMM) per state. The Baum-Welch Expectation Maximization (EM) algorithm is employed to re-estimate the model parameters. For classification, the Viterbi algorithm is used to find the maximum likelihood state sequence of the test waveform given the model. Experiments are implemented using leave-one-out cross-validation. The programming toolkit HTK 3.1.1 from Cambridge University is used to implement the various HMMs [12].

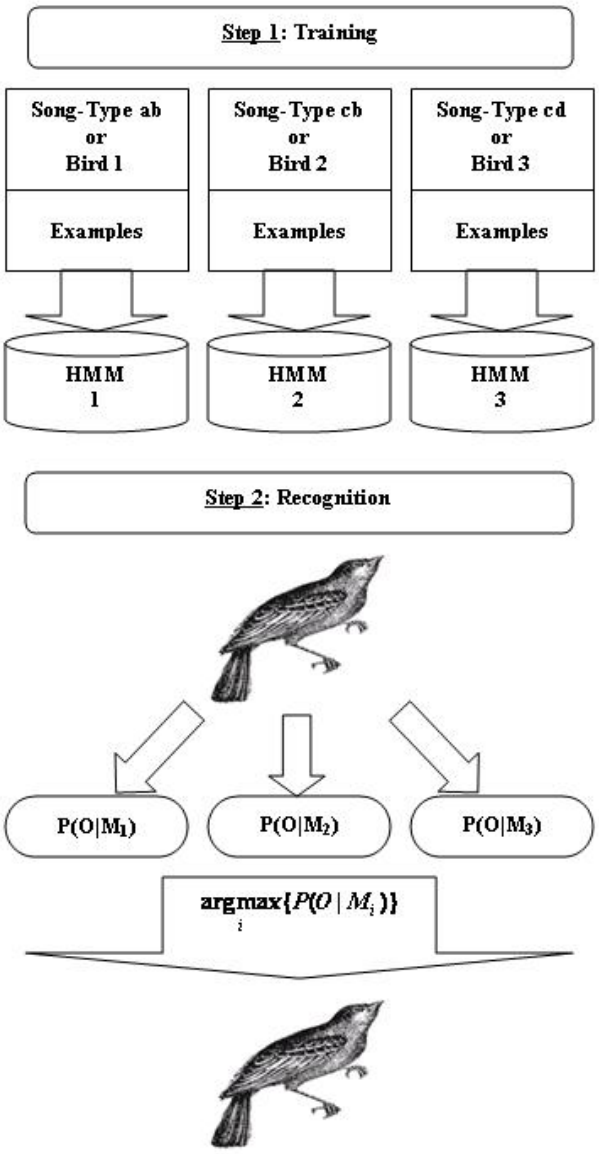

Figure 5: Training and Recognition Algorithm

4.1. Song-Type Classification

Speaker independent song-type classification experiments were performed using a subset of the Ortolan Bunting data described in Section 2. 100 exemplars of each song-type, including multiple song-variants from different speakers, were used to train each HMM, giving one model per song type. Experiments were run with an increasing number of song types, starting with the 5 most frequent and adding one type at a time until reaching 10.

Overall accuracies as a function of the number of song-types are shown in Figure 6. Results range from 92.4% for the 5 song-type experiment down to 63.6% for the 10 song-type experiment. In particular, calls *h*, *huf* and *jufb* created substantial confusion with each other and with other song-types with common syllables. Given the similarity of the components of these song-types and the shortness of some of the examples, particularly for song-type *h*, the HMM was not able to correctly separate them in all cases. In addition, there were substantial confusions between *ef* and *eb* and between *cb* and *gb* for some experiments. In the 10 song-type experiment, shown in Figure 7, song-type *ef* was classified as *eb* for all 100 exemplars, and song-type *huf* was classified as *h* 97 times out of 100. The 63.6% accuracy number thus reflects an average of many cases with fairly high (>90%) accuracy coupled with a few cases with extremely poor accuracy.

The overall average accuracy for song-type classification across these experiments was 83.3%.

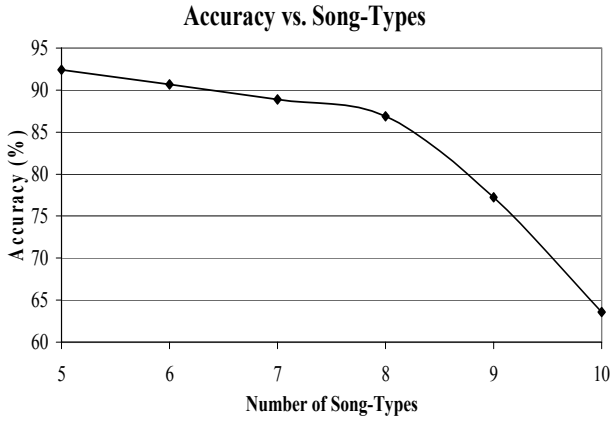

Figure 6: Classification Accuracy vs. Number of Song-Types

4.2. Speaker Identification

Song-type dependent speaker identification experiments were also performed using a subset of the Ortolan Bunting data described in Section 2. The number of exemplars used for each speaker was based on the number of recordings available for the different song-types. Beginning with a group of 5 speakers, chosen arbitrarily by order of occurrence within the data, additional speakers were added one by one, and the speaker identification experiment was repeated each time. This process was implemented for each of the 10 most frequent call types, as listed in Table 1 in Section 2.

Figure 8 illustrates the classification accuracies as a function of both song-type and number of speakers. The

		Classification									
		ab	cb	cd	eb	ef	gb	guf	h	huf	jufb
Labels	ab	98	0	0	1	0	0	0	0	0	1
	cb	3	74	0	10	0	0	2	4	0	7
	cd	0	6	92	1	0	0	1	0	0	0
	eb	1	0	0	98	0	0	0	1	0	0
	ef	0	0	0	100	0	0	0	0	0	0
	gb	0	33	1	46	0	0	4	12	0	4
	guf	0	0	5	0	0	0	94	1	0	0
	h	0	11	0	1	0	0	0	88	0	0
	huf	0	0	2	0	0	0	1	97	0	0
	jufb	0	3	1	2	0	0	0	2	0	92

Figure 7: Confusion Matrix of 10 Most Frequent Song-Types

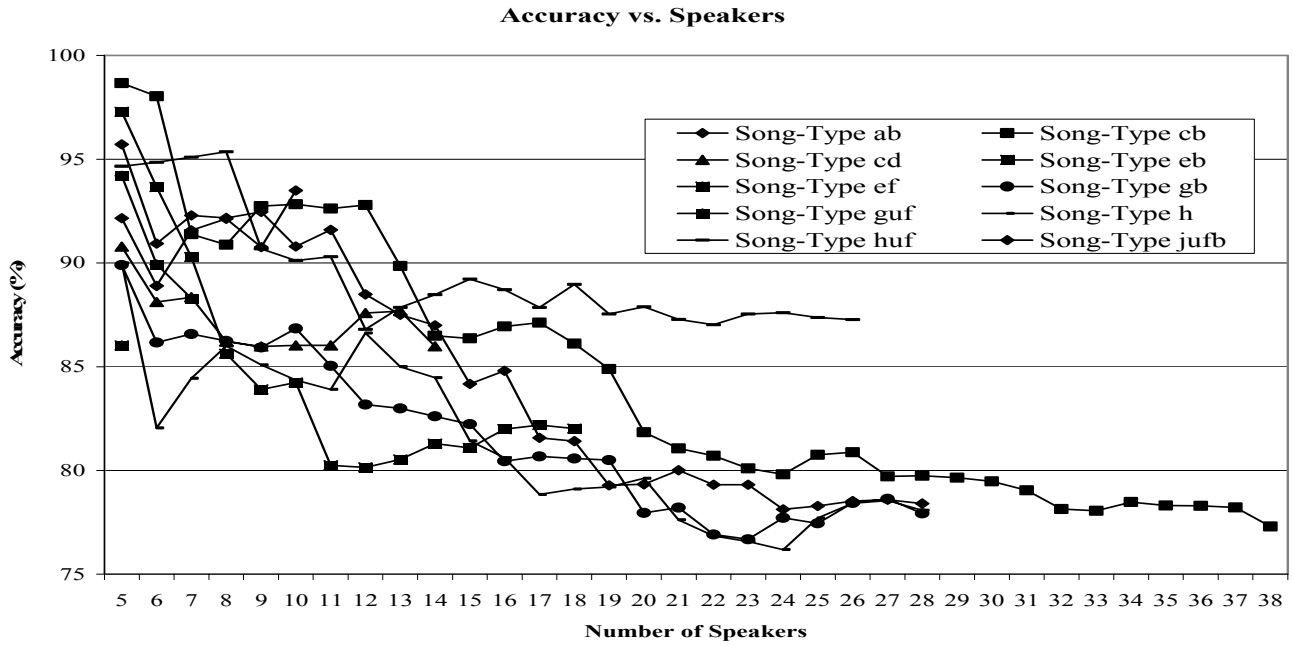

Figure 8: Classification Accuracy vs. Number of Speakers (Song-Type Dependent)

results varied from as high as 98.7% (song-type *cb*; 5 speakers) to as low as 76.2% (song-type *h*; 24 speakers). As would be expected, accuracies generally decreased as the number of speakers increased, typically starting above 90% and reaching asymptotes of 75-80% in most cases. The most successful song-type for speaker identification was perhaps *huf*, which varied from 95% down to about 87%. One interesting note about this is that the *huf* call type was one of the most inaccurate for the song-type classifications discussed in the previous section. The overall average accuracy for all speaker identification experiments was 84.8%.

5. CONCLUSION

Song-type classification and speaker identification experiments have been performed on a Norwegian Ortolan Bunting dataset. Accuracies averaged 83.3% for speaker independent song-type classification and 84.8% for song-type dependent speaker identification, with peak accuracies of 92-95%. These initial results using an HMM-based recognition system indicate good potential for automatic discrimination of calls and individuals from Ortolan Bunting recordings taken in their natural habitat.

Future work on this project will involve including a wider variety of song-types and speakers, modification of the feature extraction methods tailored toward the Ortolan Bunting vocal production mechanisms and frequency ranges [13], the use of additional features based on pitch and duration measures, and examination of model topology and GMM observation distributions to further improve performance.

6. ACKNOWLEDGEMENTS

This material is based on work supported by the National Science Foundation under Grant No. IIS-0326395.

7. REFERENCES

[1] T. S. Osiejuk, K. Ratynska, J. P. Cygan, and D. Svein, "Song structure and repertoire variation in ortolan bunting (Emberiza hortulana L.) from isolated Norwegian population," *Annales Zoologici Fennici*, vol. 40, pp. 3-16, 2003.

[2] S. Dale, "Causes of Population Decline in Ortolan Bunting in Norway," *Proceedings in 3rd International Ortolan Symposium*, pp. 33-41, 2001.

[3] C. Erbe, "Census of Marine Mammals," 2000.

[4] S. E. Anderson, D. S. Amish, and D. Margoliash, "Template-based automatic recognition of birdsong syllables from continuous recordings," *American Acoustical Society*, vol. 100, pp. 1209-1219, 1996.

[5] J. A. Kogan and D. Margoliash, "Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study," *Journal of the American Acoustical Society*, vol. 103, 1998.

[6] A. Härmä, "Automatic identification of bird species based on sinusoidal modeling of syllables," presented at International Conference on Acoustics, Speech, and Signal Processing, 2003.

[7] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. New York, NY: Macmillian Publishing Company, 1993.

[8] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, pp. 357-366, 1980.

[9] B. Gold and N. Morgan, *Speech and Audio Signal Processing*, 2000.

[10] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Recognition*, 2001.

[11] L. R. Rabiner and B. H. Juang, "An Introduction to Hidden Markov Models," *IEEE ASSP Magazine*, vol. 3, pp. 4-16, 1986.

[12] Cambridge University Engineering Department, *Hidden Markov Model Toolkit (HTK) Version 3.2.1 User's Guide*. Cambridge, MA, 2002.

[13] P. J. Clemins, "Automatic Classification of Animal Vocalizations," in *Electrical and Computer Engineering*. Milwaukee: Marquette University, 2005, pp. 137.