METHODS FOR CLASSIFICATION OF NOCTURNAL MIGRATORY BIRD VOCALIZATIONS USING PSEUDO WIGNER-VILLE TRANSFORM

Anand Patti, Geoffrey A. Williamson

Illinois Institute of Technology
Electrical and Computer Engineering
Chicago, Illinois

ABSTRACT

Many species of birds in Americas vocalize during nocturnal migration flights. Acoustic detection and classification of the calls show potential for study of the natural history of these migrant birds. In particular, information about the species' composition and number of birds involved in migration movements may be obtainable through acoustic techniques. Other methods such as radar monitoring may have capability only to assess the number, but not the composition. Mel Frequency Cepstral Coefficients-Gaussian Mixture Model-based methods (MFCC-GMM), Mel Frequency Cepstral Coefficients-Hidden Markov Model-based methods (MFCC-HMM) and spectrogram correlation-based methods have been proposed to automate the recognition/classification of the nocturnal flight calls. Here we investigate the choice of Pseudo Wigner-Ville Transform (PWVT) on MFCC-HMM-based classifier and correlation-based classifier performance. We use a collection of recordings of nocturnal flight calls of several species of thrushes and other bird species with similar calls to evaluate and compare classifiers.

Index Terms— Acoustic signal detection, spectrogram, classification algorithm

1. INTRODUCTION

Automated recognition of animal sounds from continuous field recordings assume significance in a variety of situations, including population monitoring, the study of animal behavior, prevention of harmful human/animal interactions, in biological studies, and ornithology [1], [2]. These recordings are often noisy or clipped, calling for the use of reliable automatic techniques rather than conventional manual techniques. Longer battery life and cheap memory have dramatically increased the amount of audio to analyze and hence manual inspection of spectrographs is often error-prone and involves multiple human experts which makes the identification unreliable and expensive. Thus there is a need for automated analysis techniques to generate reliable constituent labels for each sound [3].

Statistical classifiers applied to various characteristics of marine mammal sounds have been proposed [4]. The cross-correlation of spectrograms has been used to recognize marine mammal vocalizations [5], [6]. Machine identification of bird sounds has also been suggested to help prevent bird/aircraft collisions at airfields [7], with the classification based on speech analysis techniques. Another application of automated acoustic monitoring is in regards to nocturnal flight calls of migrant birds [8], [9], [10], [11], [12] and the cross-correlation of spectrograms has been used [13] for the purpose of studying the animals' range and distribution. Automated acoustic detection of nocturnal bird calls in conjunction with other methods such as radar monitoring shows potential in providing information about the number of birds involved in migration movements. Acoustic monitoring can also provide information about relative species composition in such movements while other techniques may not [14].

In this study we used recordings of six bird species, namely, the Gray-cheeked Thrush (GCTH) - *Catharus minimus*, the Hermit Thrush (HETH) - *Catharus guttatus*, the Scarlet Tanager (SCTA) - *Piranga olivacea*, the Swainson's Thrush (SWTH) - *Catharus ustulatus*, the Veery (VEER) - *Catharus fuscescens* and the Wood Thrush (WOTH) - *Hylocichla mustelina*. These six species have calls of similar duration and spectral content. We assume that calls from this set of species can be separated from the audio events captured at a recording station, and that the problem becomes one of determining which of the six species was the source of a particular audio event. Using audio recordings of species mentioned above, we investigate the effect of choice of time-frequency representations (TFRs) on classifier performance. We describe the PWVT-MFCC-HMM-based and the PWVT correlation based classifier and compare their performance to the Short Term Fourier Transform (STFT)-MFCC-HMM-based [15], [16] and the STFT correlation based classifiers.

2. PREPROCESSING

The audio signals recorded as described in section 4 below were first preprocessed in the following steps.

2.1. Band pass filtering

The bird species of interest vocalize in the frequency range between 2 kHz and 5 kHz. Therefore the recorded signal was band-pass filtered with cutoff frequencies at 2 kHz and 5 kHz. This filter was designed using the $fir1$ command in MATLAB ©, which implements the classical method of windowed linear-phase FIR digital filter design. The filter was designed using a Kaiser window and was of order 112.

2.2. Signal Normalization

Due to variations in the recording environment, flight patterns, and distance of the bird from the recording equipment, the calls were recorded with varying signal strengths, and therefore were normalized to prevent scaling errors.

2.3. Noise Suppression using Spectral Subtraction

Background noise acoustically added to the call can degrade identification. Because all of the acoustic bird call data were recorded in

non-lab conditions, there was background noise needed to be suppressed to increase the signal-to-noise ratio and improve detectability. A method of noise suppression using spectral subtraction was proposed by Steven Boll [17]. In our study a simplified implementation of Boll's noise suppression method was applied to the bird call data assuming the noise to be stationary.

2.4. Activity Detection and Clipping

Since the recordings were of varying durations, the signal was clipped to eliminate sections of the recording with no bird call activity and to standardize the duration of all calls. This was done by using a moving average energy detector, which detects the presence of a bird call when the energy is above a certain threshold. The duration of the call is set to 5400 samples (0.245 sec sampled at 22050 Hz).

3. CLASSIFIERS

This section describes the Psuedo Wigner-Ville Transform (PWVT) used in the two classifiers suggested in this paper: the MFCC-HMM-based and the saturated correlation based classifiers for acoustic data recognition.

3.1. Pseudo Wigner Ville Transform

The Wigner distribution was originally defined by Ville in 1948 [18] using the analytic signal. Extending the definition to the Pseudo (time localized window h) Wigner distribution [19], [20], [21] of a real signal $s(t)$, we have

$$W(t,\omega) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} h(\tau)z^*(t-\frac{1}{2}\tau)z(t+\frac{1}{2}\tau)e^{-j\tau\omega}d\tau \quad (1)$$

where, $z(t)$ is the analytic signal associated with the real signal $s(t)$. It is calculated in the time domain as:

$$z(t) = s(t) + jH[s(t)] \quad (2)$$

where, $H[\cdot]$ stands for the Hilbert transform.

3.2. PWVT-MFCC-HMM-based method

The STFT-MFCC-HMM approach has been successfully applied to the classification of Mexican antbirds [22], [7]. Here the steps for the method that uses the PWVT as the TFR have been given.

3.2.1. PWVT Computation

We computed the PWVT of the pre-processed audio signal as given by Equation 1, varying window sizes (WD) from 10 to 45 ms in intervals of 5 ms. The step size (ST) parameter is varied between 2 ms and the window sizes, since $WD - 5$ must be greater than ST. Figure 1b shows the PWVT of the Gray-cheeked thrush call.

3.2.2. MFCC Computation

MFCCs for each of the time windowed PWVT segments were computed. MFCC computation includes: Non-uniformly spaced (Mel-scaled) Filterbank processing, Log Energy Computation, and Inverse Discrete Fourier Transform (IDFT) as described by Davis in [22]. Since the log power spectrum is real and symmetric, the IDFT reduces to a DCT.

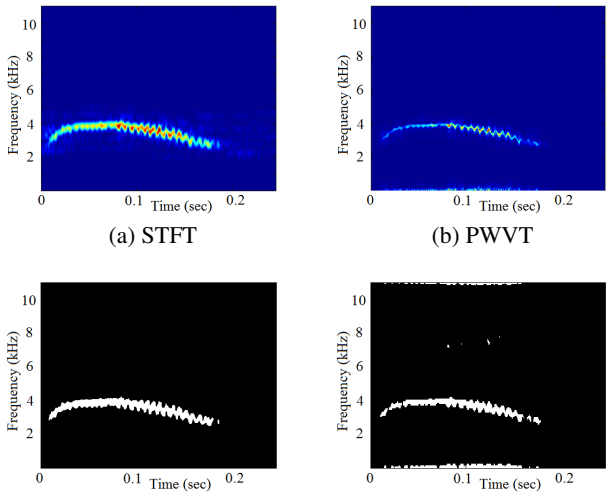

Fig. 1. Gray-cheeked thrush call.

The number of MFCCs (CP) that are computed and used as feature vectors for every time segment (window) were varied between discrete values of 7, 10, 13, 15, 17, 21, 25, 27, 31 and 35 and the Number of Mel filter bins (BD) were varied between discrete values of 7, 10, 13, 15, 17, 21, 27, and 31.

3.2.3. Vector Quantization

Vector Quantization (VQ) is a process of mapping vectors of a large vector space to a finite number of regions in that space. Each region is called a cluster and is represented by its center (called a centroid). A collection of all the centroids makes up a codebook. Although the codebook is smaller than the original sample, it still accurately represents bird call characteristics. The only difference is that there will be some spectral distortion due to quantization effects.

The VQ method implemented in this study is the K-means clustering algorithm described by Lloyd, S. P. in 1957 [23].

3.2.4. Hidden Markov Models

A HMM is a statistical tool that can model a discrete time dynamical system described by a Markov process with unknown parameters [24]. In our research, a bird call can be considered as a sequence of observations produced by such a dynamical system.

An HMM for each class (each class refers to a bird species) is used to model the temporal evolution of the vector of features (Codebook vectors corresponding to MFCCs) extracted from a call signal at discrete time step, and recognition is done by looking at which HMM is most likely to produce a given sequence of observations.

An HMM is a quintuple model (Ω_X, Ω_O, A, B, Π), where Ω_X = [S_1, ..., S_N] is a finite set of N distinct states, while Ω_O = [o_1, ... ,o_K] is the set of possible observation symbols (codebook vectors). λ = (A, B, Π) denotes the parameters of the hidden Markov chain, with A_{NxN} as the transition probabilities matrix, B_{NxK} the probabilities of observing each symbol for each state, and Π_{1xN} the distribution of the initial state (see illustration in Figure 2 for the implemented model).

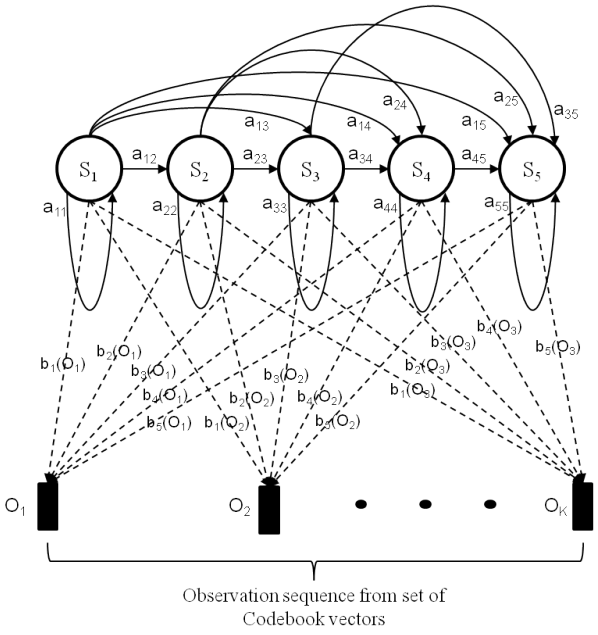

Fig. 2. Hidden Markov Model.

The number of Codebook Vectors (CB) are varied between discrete values of 10, 13, 15, 17, 21, 27, 31, 35, 40, 45, 50.

3.3. PWVT Correlation-based Method

A method of classification of nocturnal migratory bird calls using saturated spectrogram correlation was introduced in [13]. Here we describe the use of the Pseudo Wigner-Ville transform (PWVT) as the TFR. The method applied to each bird call proceeds as follows:

3.3.1. PWVT Computation

After pre-processing the audio files, we compute Pseudo-Wigner Ville transform with window sizes (WD) varying between 40 and 200 samples (1.8ms - 9ms) and step sizes (ST) varying between 20 and 150 samples (0.9ms - 6.65ms) as will be seen in the section describing the choice of parameters.

3.3.2. Scaling

The PWVTs are scaled so that the maximum value achieved equals 100. The PWVT values then lie in the interval [0, 100].

3.3.3. PWVT Saturation

The raw PWVT is saturated to binary values (+1: signal present, 0: signal absent) using an iterative thresholding technique. The iterative thresholding proceeds as follows.

1. Step 1: Select an arbitrary threshold value.
2. Step 2: Separate the PWVT in two groups: one group with energy lower than the threshold and the other group with energy higher than the threshold.
3. Step 3: Calculate the means of the energies in the first group, then in the second group. Average the means to produce a new threshold. Return to Step 2 until the threshold stabilizes.

3.3.4. Correlation detection

The saturated PWVT is correlated with a template signal. The highest correlation value among the six species connotes a positive recognition.

4. EXPERIMENTAL SETUP

The dataset on which the experiments have been performed consists of audio recordings captured using a hypercardioid dynamic handheld microphone (Samson R31S, Samson Technologies, Hauppauge, NY) mounted in a parabolic dish connected to a desktop PC running Syrinx-PC software (John Burt, Seattle WA, http://www.syrinxpc.com). The recording equipment was assembled and operated by Mr. Paul Sweet in Zion, Illinois. The audio signals were sampled at 22.05 kHz with 16-bit resolution. Each has a duration of about 1 second. The dataset consisted of a total of 21,652 recorded calls between the six species.

The experiments were first conducted using 100 high quality bird call samples ('clean' subset) from each bird species and then the entire 'noisier' dataset. We used two different experimental setups not only because TFR MFCC-HMM-based methods require training (i.e. estimation of a model), but also because the TFR correlation methods need only an appropriate choice of a template. The results of both setups are comparable. These are described in the following subsections.

4.1. TFR-MFCC-HMM-based Method

As mentioned above, the TFR-MFCC-HMM-based method requires a large training dataset for the estimation of the HMM model for each species. Since the number of calls available for testing and training is limited, we use the k-fold cross-validation method to train-test the data with a 100 calls from each species. In our study, we partitioned the data into 10 complementary sets of 10 calls each, where 90 calls are used for training and 10 for testing per round of cross-validation and the cross-validation process was repeated 10 times and the results averaged.

In order to assess the performance of classifiers on the entire dataset, we train (estimate a model) using 70 percent randomly selected calls from the entire dataset and test using the remaining 30 percent.

4.2. TFR-Correlation-based Method

We tested the TFR-correlation by choosing a template call for each bird species that gave us the highest detection rates (R). Each template call was chosen from the set of 100 calls available from each bird species. These templates were then used for the computation of other classification measures. A similar approach was used when experimenting with the entire dataset.

5. PERFORMANCE METRICS & NOTATIONS

5.1. Detection Rate (R)

The detection rate is a measure that is simply the fraction of calls that are identified correctly. Let us define T_{ij} as the number of files of type j, identified by the classifier as type i. Then, R is defined by

$$R = 100 \frac{\sum_i T_{ii}}{\sum_{ij} T_{ij}} \ \%, \tag{3}$$

where T_{ii} is the number of bird calls of type i correctly detected as bird i.

5.2. Root Mean Square Error in Count (E_c)

The E_c is a measure of the percentage root mean square error between the actual number of bird calls of a species i (M_i) and the number of calls identified as species i ($\widehat{M_i}$). This metric is helpful from an application perspective as it indicates how well the classifier does in accurately counting the number of bird calls of a particular species in the vicinity of the recording station. It gives us a measure of the deviation of the number of identifications from the actual number. The derivation and description of how E_c actually captures this is provided in detail in [25].

R, $E_{c(equal)}$ (where the $M_i s$ are equal)and $E_{c(prior)}$ (where the $M_i s$ are proportional to the actual probability of occurence) are the three performance metrics used to evaluate the classifiers.

We define the following six metrics that will be used throughout our analysis and discussion:

1. R_{ns} - Detection Rate with noise suppression.
2. $E_{c(equal_{ns})}$ - Mean Square Error of count with equal probability of occurrence of each species with noise suppression.
3. $E_{c(prior_{ns})}$ - Mean Square Error of count with prior probabilities of occurrence of each species with noise suppression. Prior probabilities are estimated as the fraction of each species among the 21,652 available calls.

6. RESULTS

6.1. Classification results by bird species - 'clean' subset

6.1.1. Classifier as a detector

Table 1 below shows how detection rates (average of the 'best' five) compared for each of the four classifiers as detectors of the six species individually.

Table 1. Classification results based on R_{ns} - 'clean' subset

Methods	Detection Rates R_{ns}					
	GCTH	HETH	SCTA	SWTH	VEER	WOTH
STFT-MFCC-HMM	96%	87%	64%	68.6%	49.6%	55.6%
PWVT-MFCC-HMM	96%	85.6%	48.6%	57.6%	55.6%	60.8%
STFT Correlation	80.8%	95%	82.8%	84.6%	78%	7%
PWVT Correlation	93.2%	4.6%	42.2%	68%	62.2%	26.6%

6.1.2. Classifier as a counter with equal probability of occurrence

Table 2 below shows how $E_{c(equal)}$ (average of the 'best' five) compared for each of the four classifiers as counters (with equal probability of occurrence of bird species) of the six species individually.

Table 2. Classification results based on $E_{c(equal_{ns})}$ - 'clean' subset

Methods	Counting Error $E_{c(equal_{ns})}$					
	GCTH	HETH	SCTA	SWTH	VEER	WOTH
STFT-MFCC-HMM	1.74	5.01	5.17	8.54	5.16	7.84
PWVT-MFCC-HMM	2.04	12.45	6.66	5.07	5.42	5.18
STFT Correlation	8.75	16.97	13.68	15.69	24.43	93
PWVT Correlation	6.21	95	57	29.60	28.96	53.2

6.1.3. Classifier as a counter with prior probability of occurrence

Table 3 below shows how $E_{c(prior)}$ (average of the 'best' five) compared for each of the four classifiers as counters (with equal probability of occurrence of bird species) of the six species individually.

Table 3. Classification results based on $E_{c(prior_{ns})}$ - 'clean' subset

Methods	Counting Error $E_{c(prior_{ns})}$					
	GCTH	HETH	SCTA	SWTH	VEER	WOTH
STFT-MFCC-HMM	7.9	125.52	152.06	29.10	106.8	330.2
PWVT-MFCC-HMM	3.5	183.37	201.7	40.32	84.6	343.0
STFT Correlation	12.53	145.65	210.5	14	224.0	93
PWVT Correlation	29.01	92.67	944.74	28.32	240.84	79.59

6.2. Classification results by bird species - entire dataset

Table 4 below shows how detection rates compared for each of the four classifiers as detectors of the six species individually (trained/tested on entire dataset).

Table 4. Classification results based on R_{ns} - 'noisier' dataset

Methods	Detection Rates R_{ns}					
	GCTH	HETH	SCTA	SWTH	VEER	WOTH
STFT-MFCC-HMM	79.61%	60%	45.80%	37.60%	18.57%	36.36%
PWVT-MFCC-HMM	80.58%	54.37%	36.77%	35.63%	10.52%	32.95%
STFT Correlation	62.22%	61.84%	52.61%	48.37%	49.39%	2.04%
PWVT Correlation	86.47%	2.819%	22.82%	43.91%	41.13%	14.67%

7. RELATION TO PRIOR WORK AND DISCUSSION

As seen in Table 1, the PWVT-MFCC-HMM-based classifier identified the Gray-cheeked thrush correctly on average 96% of the time and was comparable to the existing STFT-MFCC-HMM based method described in the classification of Mexican antbirds. [22], [7]. A reasonable margin of error for classifiers as counters (with equal probability of occurrence of bird species) of individual bird species is 2%. We see that Gray-cheeked thrush $GCTH$ was counted correctly by the PWVT-MFCC-HMM-based classifiers within the $E_{c(equal)} = 2\%$ margin and was comparable to the existing STFT-MFCC-HMM-based classifier. GCTH was identified/counted more accurately than the other species because of its unique spectral 'hook' at the beginning of the call. Another reason for the success of the GCTH in this study is the consistency of the spectral features within the species. A good counter that accounts for prior knowledge of the number of birds of each species must not have an error greater than 5%. The GCTH again was counted correctly by the PWVT-MFCC-HMM classifier with an error $< 4\%$.

As a detector, the PWVT correlation classifer and the PWVT-MFCC-HMM classifier detected the GCTH correctly 86% and 80% of the time in comparison to the STFT correlation [13] and the STFT-MFCC-HMM methods that detected the Gray-cheeked thrush correctly only 62% and 79% of the time. The PWVT due to its inherent property of concentrating energies about the instantaneous frequency aids classification in a correlation type of classifier applied to 'noisier' data.

We also observe a marginal improvement in the performance of classifiers when the noise suppression as described by Steven Boll [17] was applied to the raw audio file.

The Pseudo Wigner-Ville improves classifier performance when applied to 'noisier' low quality recordings, but doesn't make much of a difference when applied to 'clean' high quality recordings. The performance of PWVT-based classifiers is within acceptable levels for Gray-cheeked thrush calls, but needs improvement for others.

8. REFERENCES

[1] 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 Acoustical Society of America*, vol. 103(4), pp. 2185 – 2196, April 1998.

[2] A. Harma, "Automatic identification of bird species based on sinusoidal modeling of syllables," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 2003, vol. 5, pp. 545 – 548.

[3] J. A. Kogan and D. Margoliash, "Automated bird song recognition using dynamic warping and hidden markov models," *Journal of the Acoustical Society of America*, vol. 102(5), pp. 3175, November 1997.

[4] B. Pinkowski, "Robust fourier descriptors for characterizing amplitude-modulated waveform shapes," *Journal of the Acoustical Society of America*, vol. 95, pp. 3419 – 3423, 1994.

[5] D. K. Mellinger and C. W. Clark, "Recognizing transient low-frequency whale sounds by spectrogram correlation," *Journal of the Acoustical Society of America*, vol. 107, pp. 3518 – 3529, June 2000.

[6] D. Chabot, "A quantitative technique to compare and classify humpback whale (magapter novaeangliae) sounds," *Ethology*, vol. 77, pp. 89 – 102, 1988.

[7] C. Kwan, K. C. Ho, G. Mei, Y. Li, Z. Ren, R. Xu, Y. Zhang, D. Lao, M. Stevenson, V. Stanford, and C. Rochet, "An automated acoustic system to monitor and classify birds," *Journal of Applied Signal Processing*, pp. 1 – 19, 2006.

[8] R. R. Graber, "Nocturnal migration in illinois: different points of view," *Wilson Bull*, vol. 80, pp. 36 – 71, 1968.

[9] R. R. Graber and W. W. Cochran, "An audio technique for the study of the nocturnal migration of birds," *Wilson Bull*, vol. 71, pp. 220 – 236, 1959.

[10] W. R. Evans and D. K. Mellinger, "Monitoring grassland birds in nocturnal migration," *Studies in Avian Biology*, vol. 19, pp. 219 – 229, 1999.

[11] W. R. Evans, "Nocturnal flight call of bicknell's thrush," *Wilson Bull*, vol. 106(1), pp. 55 – 61, 1994.

[12] W. R. Evans and M. OBrien, "Flight calls of migratory birds: Eastern north american landbirds (cd-rom)," *Oldbird Inc. Ithaca, NY*, 2002.

[13] M. Marcarini, G. A. Williamson, and L. de Sisternes Garcia, "Comparison of methods for automated recognition of avian nocturnal flight calls," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2008*, April 2008, pp. 2029 – 2032.

[14] A. Farnsworth, "Flight calls and their value for future ornithological studies and conservation research," *The Auk*, vol. 122, no. 2, pp. 733 – 746, 2005.

[15] V. M. Trifa, A. N. Kirschel, C. E. Taylor, and E. E. Vallejo, "Automated species recognition of antbirds in a mexican rainforest using hidden markov models," *Journal of the Acoustical Society of America*, vol. 123(4), pp. 2424 – 2431, April 2008.

[16] A. Taylor, "Bird flight call discrimination using machine learning," *Journal of the Acoustical Society of America*, vol. 97, pp. 3370(A), May 1995.

[17] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," in *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1979, vol. 2, pp. 60 – 74.

[18] J. Ville, "Theorie et application de la notion de signal analytique," *Cables et Transmissions*, vol. A(1), pp. 61 – 74, 1948.

[19] B. Boashash, "Note on the use of the wigner distribution for time-frequency signal analysis," in *IEEE Transactions on Acoustics, Speech, Signal Processing*, September 1988, vol. 36(9), pp. 1518 – 1521.

[20] L. Cohen, "Introduction: a primer of time-frequency analysis," *Time Frequency Signal Analysis: Methods and Applications*, 1991.

[21] E. F. Velez and H. Garudadri, "Speech analysis based on smoothed wigner-ville distribution," *Time Frequency Signal Analysis: Methods and Applications*, 1991.

[22] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in *IEEE Transactions on Acoustic, Speech and Signal Processing*, April 1980, vol. 28(4), pp. 357 – 366.

[23] S. P. Lloyd, "Least square quantization in pcm," in *IEEE Transactions on Information Theory*, 1982, vol. 28(2), p. 129 137.

[24] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," in *Proceedings of the IEEE*, February 1989, vol. 77, pp. 257 – 286.

[25] A. V. Patti, "Methods for classification of nocturnal mirgratory bird vocalizations using time-frequency representations," M.S. thesis, Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, IL, 2012.