AUTOMATIC SPEECH PROCESSING METHODS FOR BIOACOUSTIC SIGNAL ANALYSIS: A CASE STUDY OF CROSS-DISCIPLINARY ACOUSTIC RESEARCH

John G. Harris and Mark D. Skowronski

Computational Neuro-Engineering Lab
University of Florida
Gainesville, FL, USA

ABSTRACT

Automatic speech processing research has produced many advances in the analysis of time series. Knowledge of the production and perception of speech has guided the design of many useful algorithms, and automatic speech recognition has been at the forefront of the machine learning paradigm. In contrast to the advances made in automatic speech processing, analysis of other bioacoustic signals, such as those from dolphins and bats, has lagged behind. In this paper, we demonstrate how techniques from automatic speech processing can significantly impact bioacoustic analysis, using echolocating bats as our model animal. Compared to conventional techniques, machine learning methods reduced detection and species classification error rates by an order of magnitude. Furthermore, the signal-to-noise ratio of an audible monitoring signal was improved by 12 dB using techniques from noise-robust feature extraction and speech synthesis. The work demonstrates the impact that speech research can have across disciplines.

1. INTRODUCTION

Advances in automatic processing of speech has produced many useful algorithms. For example, knowledge about the production and perception of speech has led to several celebrated algorithms like linear prediction [1], overlap-and-add synthesis [2], mel frequency cepstral coefficients [3], and audio codecs like code-excited linear prediction [4] and the popular MP3. These algorithms have significantly impacted society and the way we communicate. Furthermore, while the goal of a general-purpose automatic speech recognition (ASR) system has not yet been fully realized, ASR research has led the advances in the machine learning paradigm (i.e., data-driven methods) through the use of hidden Markov models (HMMs) and Gaussian mixture models (GMMs) [5]. ASR research has also produced several robust algorithms that elegantly improve recognition performance such as embedded temporal derivatives (Δ and $\Delta\Delta$ coefficients) [6], cepstral mean subtraction [7], perceptual linear prediction [8], and, more recently, human factor cepstral coefficients [9].

While the techniques for analyzing human speech have steadily advanced over the past few decades, analysis of other bioacoustic signals has remained static. In this paper, we demonstrate several examples which showcase the potential that automatic speech processing methods have on bioacoustic analysis. For our work, we have chosen the echolocating bat as the model animal for several reasons. Bats represent a quarter of all mammal species and are found in every ecosystem on the planet except for the extreme latitudes [10], and bats play a significant environmental role in helping to control insect populations, pollinate flowers, and disperse seeds. Furthermore, the vast majority of bats belong to the suborder Microchiroptera, all of which are known to employ echolocation [11]. Echolocation is the use of acoustic chirps for the purposes of hunting and navigation [12] and have been recorded and analyzed for the detection and classification of bats [13, 14, 15] as well as for real-time audio monitoring of the predominantly ultrasonic echolocation calls [16]. In the following section we describe the use of machine learning methods, specifically the HMM and GMM, and robust feature extraction techniques for acoustic detection and species classification which reduce error rates by an order of magnitude compared to conventional methods. Next, the robust features extracted for detection and classification are used to synthesize echolocation calls, increasing the signal-to-noise ratio (SNR) by 12 dB compared to heterodyning and frequency division techniques.

Email: harris@cnel.ufl.edu, markskow@cnel.ufl.edu

2. EXPERIMENTS AND RESULTS

For all experiments, about 3000 echolocation calls were hand-labeled from field recordings of 5 species: *Pipistrellus bodenheimeri, Molossus molossus, Lasiurus borealis, L. cinereus semotus,* and *Tadarida brasiliensis*. Echolocation calls are characterized by a modulated fundamental frequency with duration on the order of 10 ms and constantly emitted during flight at a rate of about 10 calls per second (i.e., 10% duty cycle).

2.1. Detection

Detection is typically achieved by comparing only short-term energy estimates to a threshold [14], completely neglecting frequency information. To improve detection performance, we increased the amount of information available to the detector by using frame-based frequency and log-energy estimates along with temporal derivatives [17]. Fundamental frequency was estimated using a zero-padded FFT following spectral mean subtraction (similar to cepstral mean subtraction), which reduced noise power and equalized the noise floor of recordings from different measurement platforms. The threshold of energy was replaced with a threshold of log likelihood difference using a pair of GMMs to model the distribution of features from both the calls *and* the background [18]. The GMMs were trained using 25 Gaussian kernels and standard maximum likelihood training with full covariance matrices [19]. The detection experiment results for the GMM and baseline energy detectors are shown in Figure 1. The accuracy of the GMM detector at equal sensitivity and specificity was 96%, compared to 68% for the energy detector. Therefore,

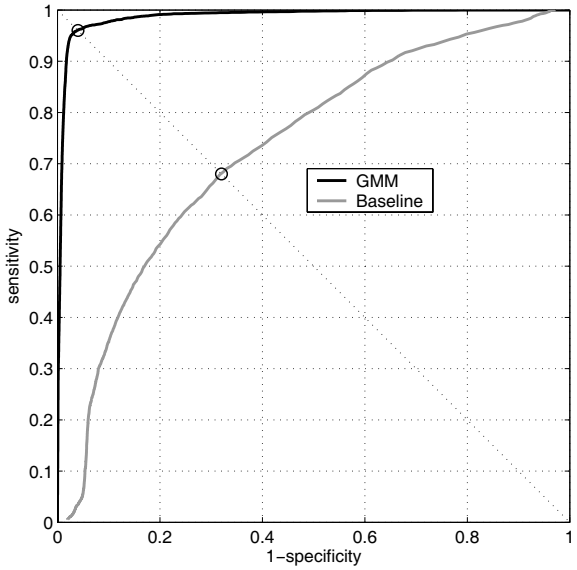


Fig. 1. Receiver operator characteristic curves for the GMM and baseline energy detectors. The circles on the ROC curves indicate the operating points with equal sensitivity and specificity.

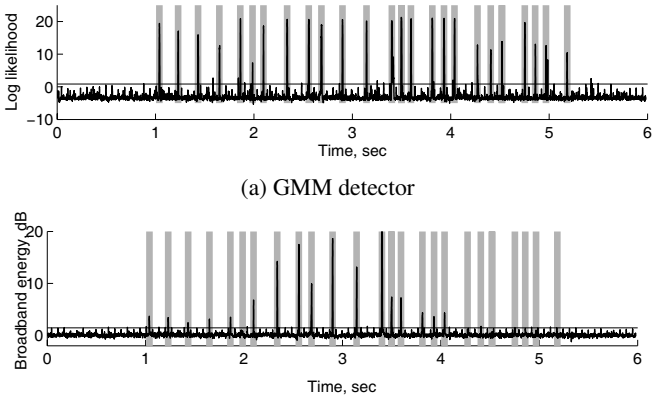
the GMM detector error was 8 times lower than the baseline energy detector at equal sensitivity and specificity.

An example of detector performance for a single pass of calls is shown in Figure 2. The pass contains 25 hand-labeled calls, denoted by gray bars in the figure. Figure 2 demonstrates how the broadband energy detector distinguishes only the most prominent calls, near the middle of the pass, from the background noise, while the GMM detector outputs peaks significantly above the background output level at the locations of all of the hand-labeled calls.

2.2. Classification

A classification experiment was performed on hand-labeled calls to determine the species of the bat that produced each call. Three classifiers were tested in the experiment: a baseline discriminant function analysis (DFA) classifier commonly used in the bat literature [20, 21], a GMM classifier, and an HMM classifier. The DFA classifier was trained with features most commonly used in the literature: minimum frequency, maximum frequency, frequency of peak energy, and duration [13, 14]. All baseline features were determined from the noise-robust features used by the GMM detector. In addition, both machine learning classifiers employed the same feature vectors used with the GMM detector. The classifiers were tested in a cross-validation experiment. For each of 20 trials, 50% of all calls were uniformly randomly selected to train the three classifiers while the remaining 50% of the calls were used to test the classifiers.

The cross-validation classification results are reported as the confusion matrices in Tables 1 and 2 for the GMM and DFA classifiers, respectively. A t-test between the GMM and HMM overall average scores across all trials showed that the scores were not significantly different ($p > 0.9$). For the GMM and HMM classifiers, the accuracy was $99.4 \pm 0.2\%$ correct, while the accuracy for the baseline DFA classifier was $83.1 \pm 1.1\%$ correct. Thus, the DFA classification error was 28 times larger than the error of the machine learning classifiers.



(a) GMM detector

(b) Baseline energy detector

Fig. 2. Detector outputs for a single pass of 25 hand-labeled calls from *Lasiurus borealis* using (a) a Gaussian mixture model detector, and (b) a baseline detector. The gray bars denote the locations of hand-labeled calls, and the horizontal black lines denote the thresholds for equal sensitivity and specificity, denoted by the circles in Figure 1.

Table 1. Confusion matrix for **GMM** classifier from cross-validation experiment. Values in each cell are the average percentage of calls from the hand-labeled species for each row that were classified as the species for each column over 20 trials. The species are *Pb: Pipistrellus bodenheimeri, Mm: Molossus molossus, Lb: Lasiurus borealis, Lc: Lasiurus cinereus semotus,* and *Tb: Tadarida brasiliensis*. The overall percent correct over 20 trials was $99.4 \pm 0.2\%$.

	Pb	*Mm*	*Lb*	*Lc*	*Tb*
Pb	99.6	0	0	0	0.4
Mm	0.03	96.2	0	0	3.7
Lb	0	0	99.8	0.2	0
Lc	0	0	0.2	99.8	0
Tb	0	0.2	0	0	99.8

2.3. Audio monitoring

Echolocation calls from nearly all species of bats are above the frequency range of human hearing, varying from 20 kHz to 250 kHz. To monitor bats in the field, bat detectors use any of three strategies to shift the echolocation calls into the audible range: 1) time expansion, 2) frequency division, and 3) heterodyning [16]. Time expansion simply reduces the sampling rate by a factor of 10 or 20 during playback, perfectly preserving frequency and temporal structure at the cost of missing large temporal blocks of data during playback. Frequency division detectors typically employ a zero-crossing counter: when the counter is a multiple of the division factor, the binary output of the counter is flipped. Frequency division distorts the amplitude of the original signal, so a short-term energy estimate may be used to provide a temporal envelope for the frequency-divided signal. Heterodyning modulates the input by a carrier frequency, shifting the high-frequency echolocation call into the audible range.

Table 2. Confusion matrix for **DFA** classifier from cross-validation experiment, similar to Table 1. The overall percent correct over 20 trials was $83.1 \pm 1.1\%$.

	Pb	*Mm*	*Lb*	*Lc*	*Tb*
Pb	97.1	0.2	2.7	0	0
Mm	0.6	76.7	4.1	17.3	1.3
Lb	1.2	16.9	79.6	0.3	2.1
Lc	0	1.1	0.3	89.7	8.8
Tb	0	6.6	5.4	16.5	71.4

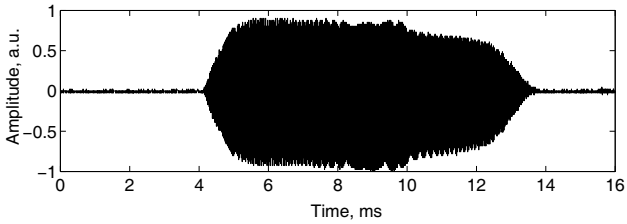
The carrier frequency may be fixed at a preset frequency (e.g., 25 or 40 kHz) or tunable in a bat detector. Heterodyning perfectly preserves amplitude information but does not compress the bandwidth of calls, which may span 50 kHz, into the audible range. Neither frequency division nor heterodyning increase the duration of calls, so shifted chirps sound like periodic clicks and pops.

The effective monitoring range of bat detectors is limited by noise, either from the recording environment, electronic noise, or noise from the frequency division or heterodyning mechanisms. The intensity of echolocation calls decreases as a function of distance to the bat due to spherical spreading and atmospheric attenuation. For example, a 60 kHz tone propagating as a spherical wavefront in air at 25° Celsius and 50% relative humidity is attenuated by 50 dB at a distance of 7 m and 70 dB at a distance of 14 m from the source [22]. To improve the SNR of a recorded signal, the noise-reduction techniques for feature extraction used in the detection and classification experiments described above were employed to find short-term estimates for the fundamental frequency and peak energy. The frequency and energy estimates from non-overlapping frames were then used to synthesize the echolocation and background signals. For a set of frequencies $\omega(k)$ and amplitudes $a(k)$ for each frame $k = [1 \dots K]$, the synthetic output $x(n)$ using frames of length L was constructed according to the following expressions:

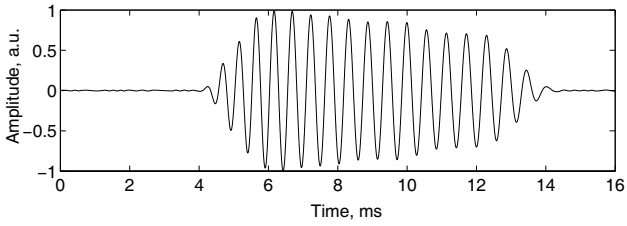
$$\begin{aligned} n_k &= [1 \dots L] + (k-1)L \\ A_k &= a(k-1) + (a(k) - a(k-1))[1 \dots L] \\ x(n_k) &= A_k \sin\left(2\pi \frac{\omega(k)}{\beta} \frac{n_k}{f_s} + \theta\right) \\ \theta &\leftarrow \theta + 2\pi \frac{\omega(k)}{\beta} \frac{L}{f_s} \end{aligned} \quad (1)$$

where f_s is the sampling rate, β is a frequency division factor, n_k is the index of length L into x for frame k, $a(0) = a(1)$ by default, and $\theta = 0$ initially. The term A_k was used to linearly interpolate between $a(k)$ and $a(k-1)$ for frame k and produce smooth amplitude transitions between frames.

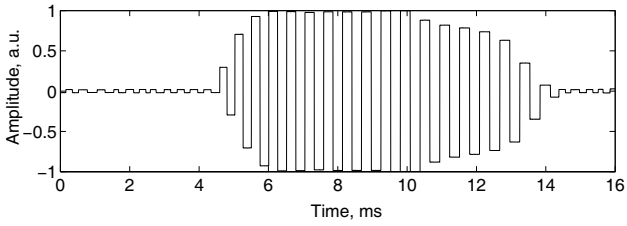
Figures 3(a) through (c) show the original time series of a call from *Lasiurus borealis*, a synthesized output of the original signal with a frequency division factor $\beta = 20$, and a frequency division output with temporal envelope and $\beta = 20$, respectively. Peak energy for the frequency division output in Figure 3(c) was 34.5 dB above the noise floor, while peak energy for the synthetic output in Figure 3(b) was 46.5 dB above the noise floor, a difference in SNR of 12 dB. According to Lawrence and Simmons [22], a decrease in the noise floor by 12 dB would extend the range of a 60 kHz signal that was originally 50 dB above the noise floor at 25° C and 50% relative humidity from 7 m to 11 m. A detection sphere with radius 11 m has almost 4 times the volume as a detection sphere of radius 7 m, significantly increasing the coverage of the detector without changing the measurement equipment.



(a) Raw signal



(b) Synthetic output



(c) Frequency division output

Fig. 3. Example of audio monitoring signals from a single call from *Lasiurus borealis*. The raw signal in (a) was used to create the synthetic output in (b) according to Eq. 1 and the frequency division output with amplitude envelope in (c) using a frequency division factor $\beta = 20$. The peak SNR in (b) was 46.5 dB, compared to 34.5 dB in (c).

3. DISCUSSION

The above examples of automatic detection, classification, and audio monitoring of echolocating bats demonstrate the significant impact that methods commonly used in automatic speech processing can have on the analysis of non-human bioacoustic signals. Much of the conventional analysis methods for echolocation signals heavily rely on expert intervention, which contributes to the subjectivity of results among researchers and stifles progress. Robust automated analysis methods for echolocating bats would also significantly reduce the tedium of analyzing by hand the vast data typically generated from acoustic experiments due to the high sampling rate and long recording sessions. Twelve hours of data recorded at 200 kHz, 16 bits per sample, from 4 recorders would generate about 69 GB of data and require about 60 minutes for an expert to thoroughly hand label calls from each minute of data [18]. Robust automated methods would allow bat researchers to spend less time managing data and more time collecting data.

4. CONCLUSIONS

Methods from automatic speech processing can significantly impact bioacoustic research across disciplines. In this paper, we demonstrated the impact of methods developed for human speech have on the detection, species classification, and audio monitoring of echolocating bats. Using spectral mean subtraction to reduce the noise around the fundamental frequencies of echolocation calls, frequency and peak energy estimates were extracted according to the frame-based machine learning paradigm and used to train GMMs for detection and a GMM and HMM for species classification. The GMM detector produced 8 times fewer errors compared to a conventional energy-based detector, and the machine learning classifiers produced 28 times fewer errors compared to a conventional DFA classifier. As ASR research discovered 2 decades ago for human speech, machine learning algorithms use more information and account for variations in the data better than human-expert based methods. Furthermore, a synthetic signal, constructed from frequency and energy features used for detection and classification, was constructed which reduced the noise floor by 12 dB compared to the raw signal and significantly increased the volume of the detection sphere around a detector given a particular example recording environment.

5. REFERENCES

[1] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, Upper Saddle River, NJ, 1978.

[2] J. B. Allen and L. R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proc. IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[3] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Sign. Process.*, vol. 28, no. 4, pp. 357–366, 1980.

[4] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): high-quality speech at very low bit rates," in *Int. Conf. Acoust., Speech, and Sign. Process.*, Tampa, FL, Apr. 1985, pp. 937–940, IEEE.

[5] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in *Readings in Speech Recognition*, A. Waibel and K.-F. Lee, Eds., pp. 267–296. Kaufmann, San Mateo, CA, 1990.

[6] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, and Sign. Process.*, vol. 29, no. 2, pp. 254–272, Apr 1981.

[7] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Am.*, vol. 55, no. 6, pp. 1304–1312, June 1974.

[8] H. Hermansky, "Perceptual linear prediction (PLP) analysis for speech," *J. Acoust. Soc. Am.*, vol. 87, pp. 1738–1752, 1990.

[9] M. D. Skowronski and J. G. Harris, "Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 116, no. 3, pp. 1774–1780, September 2004.

[10] M. Brock Fenton, *Bats*, Facts on File, New York, NY, 1992, ISBN: 0-816-02679-3.

[11] M. Brock Fenton, *Communication in the Chiroptera*, Indiana University Press, Bloomington, IN, 1985, ISBN: 0-253-31381-3.

[12] D. R. Griffin, *Listening in the dark: the acoustic orientation of bats and men*, Comstock Pub. Associates, Ithaca, NY, 1958, ISBN: 0801493676.

[13] M. B. Fenton and G. P. Bell, "Recognition of species of insectivorous bats by their echolocation calls," *J. Mammal.*, vol. 62, no. 2, pp. 233–243, May 1981.

[14] M. K. Obrist, "Flexible bat echolocation: the influence of individual, habitat and conspecifics on sonar signal design," *Behav. Ecol. Sociobiol.*, vol. 36, pp. 207–219, 1995.

[15] W. L. Gannon, R. E. Sherwin, and S. Haymond, "On the importance of articulating assumptions when conducting acoustic studies of habitat use by bats," *Wild. Soc. Bull.*, vol. 31, no. 1, pp. 45–61, 2003.

[16] S. Parsons, A. M. Boonman, and M. K. Obrist, "Advantages and disadvantages of techniques for transforming and analyzing chiropteran echolocation calls," *J. Mammal.*, vol. 81, no. 4, pp. 927–938, Nov. 2000.

[17] M. D. Skowronski and J. G. Harris, "Automatic detection of microchiroptera echolocation calls from field recordings using machine learning algorithms," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2552, Apr. 2005, 149th meeting of the Acoustical Society of America, Vancouver, Canada, May 16-20.

[18] M. D. Skowronski and J. G. Harris, "Acoustic detection and classification of microchiroptera using machine learning: lessons learned from automatic speech recognition," *J. Acoust. Soc. Am.*, 2005, submitted.

[19] K. Murphy, "Bayes network toolbox for matlab," 2005, URL: http://www.cs.ubc.ca/ ∼murphyk/ Software/BNT/bnt.html [Oct. 21, 2005].

[20] R. F. Lance, B. Bollich, C. L. Callahan, and P. L. Leberg, "Surveying forest-bat communities with Anabat detectors," in *Bats and Forests Symposium*, R. M. R. Barclay and R. M. Brigham, Eds., Victoria, B.C., CA, 1996, pp. 175–184, Res. Br., B.C. Min. For.

[21] M. K. Obrist, R. Boesch, and P. F. Fluckiger, "Variability in echolocation call design of 26 Swiss bat species: consequences, limits and options for automated field identification with a synergetic pattern recognition approach," *Mammalia*, vol. 68, no. 4, pp. 307–322, Dec. 2004.

[22] B. D. Lawrence and J. A. Simmons, "Measurements of atmospheric attenuation at ultrasonic frequencies and the significance for echolocation by bats," *J. Acoust. Soc. Am.*, vol. 71, no. 3, pp. 585–590, Mar. 1982.