Herpetological Conservation and Biology 4(3):384-388.
Submitted: 21 January 2009; Accepted: 24 April 2009.

EFFICACY OF AUTOMATIC VOCALIZATION RECOGNITION SOFTWARE FOR ANURAN MONITORING

J. HARDIN WADDLE[1], TYLER F. THIGPEN[2], AND BRAD M. GLORIOSO[2]

[1] *U.S. Geological Survey, National Wetlands Research Center, Lafayette, Louisiana 70506, USA, e-mail: waddleh@usgs.gov*
[2] *IAP Worldwide Services Inc., National Wetlands Research Center, Lafayette, Louisiana 70506, USA*

***Abstract*.—Surveys of vocalizations are a widely used method for monitoring anurans, but it can be difficult to coordinate standardized data collection across a large geographic area. Digital automated recording systems (ARS) offer a low-cost method for obtaining samples of anuran vocalizations, but the number of recordings can easily overwhelm human listeners. We tested Song Scope, an automatic vocalization recognition software program for personal computers to determine if this type of machine learning approach is currently a viable solution for anuran monitoring. For three species, Song Scope scanned more than 200 h of recordings in 3-20 h at the settings we chose. The software misidentified true calls (false positive) at rates of 2.7%-15.8% per species and failed to detect calls (false negative) in 45%-51% of recordings. There exists a tradeoff between false positive and false negative errors, which can be adjusted by setting the minimum criteria for the recognition software. Users of this approach should carefully consider their reasons for monitoring and how they intend to use the data before creating a large monitoring network.**

Key Words.—ARS; automated recording system; frog call; machine learning; monitoring program; study design

INTRODUCTION

Anuran vocalization surveys are widely used as a technique for monitoring the occurrence of populations of amphibians (Weir and Mossman 2005; Weir et al. 2005) or for estimating the relative abundance of calling male anurans (Zimmerman 1994; Nelson and Graves 2004). Data collectors must be trained to distinguish species by sound and be familiar with survey protocol (Genet and Sargent 2003), and a large number of data collectors are required to simultaneously survey many sites. Anuran activity patterns are dynamic and affected by changes in climatic conditions (Ossen and Wassersug 2002) and by the presence of other anuran species (Given 1987, 1990). Automated Recording Systems (ARS) programmed to record audio at predetermined times have been used to remotely monitor amphibians (Bridges and Dorcas 2000). The advantage of ARS is that they can record at any programmed time at any number of sites, including times and places where it may be difficult for trained observers to be (Hsu et al. 2005). Recent advances in ARS technology include the ability to record high-quality digital audio files directly to removable digital media (Acevedo and Villanueva-Rivera 2006). This technology has become more affordable, as has the storage of digital data, spawning increased interest in networks of ARS monitoring stations.

The availability of efficient, cost-effective ARS means that a single amphibian researcher can easily obtain hundreds of hours of recordings of frog calls. The expense in person-hours of listening to these recordings is not trivial, whether one is using trained volunteers or paid technicians. Although machine-learning methods for identifying recordings of frog calls have been developed (Taylor et al. 1996; Yen and Fu 2001; Brandes et al. 2006), only very recently has an automated bioacoustic identification software package for a personal computer platform become commercially available. A reliable automated vocalization recognition system would be very useful for networks of many ARS set up to collect anuran vocalization monitoring data.

We evaluated the use of Song Scope™ Bioacoustics Monitoring Software (Ver. 2.1A; Wildlife Acoustics Inc., Concord, Massachusetts, USA; US $499.95) as a tool for automatic scanning of large numbers of digital audio field recordings for vocalizations of anuran species. Our objective was to assess the efficacy of an automated recording and scanning system for anuran vocalization monitoring. Because we are interested in the potential of this system for automated monitoring, our chief consideration was to achieve the highest accuracy possible in identification of calls. Therefore we chose to minimize false positive identifications at the cost of increased chances for false negative identifications. We highlight key considerations when designing a network of autonomous digital recorders for computer-based methods of anuran vocalization identification.

MATERIALS AND METHODS

We deployed five commercially available ARS units (Song Meter™ Model SM1; Wildlife Acoustics Inc., Concord, Massachusetts, USA; US $599.95), one at each of five sites in the Atchafalaya River Basin in south-

central Louisiana, USA. We programmed these units to make eight digital recordings of 5 min length, spaced 1 h apart beginning 30 min past sunset each night. This sampling was conducted from June through September 2008.

Three anuran species known to be calling during that period were selected for automatic detection using the Song Scope software: Green Treefrog, *Hyla cinerea*, American Bullfrog, *Lithobates* (aka *Rana*) *catesbeianus*, and Bronze Frog, *Lithobates* (aka *Rana*) *clamitans*. Clear examples of vocalizations of each target species from the ARS recordings we obtained were isolated temporally and spectrally, and labeled to species in Song Scope using the spectrogram visualization tools of the software. Song Scope creates a composite call from the characteristics of the example calls and uses it as a recognizer file for comparison when determining if sounds match the target sound. To create a viable recognizer file for each species, we adjusted a variety of settings in Song Scope following the guidelines in the software manual. We adjusted sample rate, frequency range, and minimum frequency to help isolate the target call, and the maximum durations for syllable, syllable gap, and song were set to capture best an individual frog call.

We used 230 individual vocalizations from 17 audio files to create a recognizer file for the Green Treefrog, which has a highly variable call. We used 64 vocalizations from eight audio files to generate the American Bullfrog recognizer file, and for the Bronze Frog recognizer file, we used 49 vocalizations from six audio files. All of the files recorded on the five ARS units were batch scanned for each of the three species using algorithm 2.0 in Song Scope. Results are filtered in Song Scope by setting the minimum values for "quality" and "score", proprietary measures of the similarity of a vocalization to the target vocalization from the recognizer. The quality (scale of 0.00 to 9.99) represents the statistical distribution of the model parameters from the recognizer algorithm and the score (scale of 0 to 100%) represents the statistical fit of the vocalization to the recognizer model. We varied the minimum settings of quality and score by species based on the judgment of the software operator to achieve the best possible accuracy, while excluding the fewest actual calls of the species possible. Upon batch scanning of sound files, Song Scope creates a results window that reports the file name, time offset from the beginning of the recording, quality, and score of each vocalization identified as matching the recognizer.

To ascertain the accuracy of the software, we randomly selected 100 files for each species from those in which Song Scope indicated the species was present. An expert human listener determined whether each of the instances where Song Scope identified the vocalization of the target species was correct. A false positive occurred when the software indicated that a species was present at a given time during the file, but the human listener could not detect it. In these cases, we proposed an explanation, when possible, as to why the software was in error. We also randomly selected 100 files from those in which the software did not find the species and used those files to search for false negatives. If the human listener heard the given species at any instance during the file, this constituted a false negative. We used the same human listener in all analyses to reduce observer variability.

RESULTS

Moisture affected two of our ARS units and they failed at 36 and 42 days, but the other three units provided more than 70 days of recordings. We retrieved 2,432 audio files (202.7 h) from the five ARS units. It took approximately 40 person-hours per species to locate and isolate example calls and fine-tune software settings for the development of a recognizer file. Song Scope took approximately 3 h to scan all of the files for Green Treefrog vocalizations, 15 h to scan for American Bullfrog vocalizations, and 20 h to scan for Bronze Frog vocalizations.

The Song Scope software detected 1,755 Green Treefrog vocalizations in 773 audio files. The true positive rate for Green Treefrog calls in the 100 files that were manually listened to was 84.2% (Table 1). Most of the sounds misidentified by the software as Green Treefrog calls were actually bird calls or calls of other frog species (Table 2). The human listener heard Green

TABLE 1. Number of detections by Song Scope of each species in the 100 randomly-chosen audio files in which the software detected the species, and the number of those detections that were confirmed by the human listener as true or false.

Species	Software Detections	True Positive	True Positive Rate	False Positive	False Positive Rate
Green Treefrog (*Hyla cinerea*)	221	186	84.2%	35	15.8%
American Bullfrog (*Lithobates* [=*Rana*] *catesbeianus*)	482	469	97.3%	13	2.7%
Bronze Frog (*Lithobates* [=*Rana*] *clamitans*)	1749	1525	87.2%	224	12.8%

TABLE 2. Source of sounds misidentified by Song Scope as a vocalization of each anuran species (i.e., false positive identification; Table 1).

Species	Birds	Other Frogs	Insects	Rain or Noise	Unknown
Green Treefrog (*Hyla cinerea*)	16	15	0	2	2
American Bullfrog (*Lithobates* [=*Rana*] *catesbeianus*)	2	3	4	4	0
Bronze Frog (*Lithobates* [=*Rana*] *clamitans*)	46	149	0	17	12

Treefrog vocalizations in 45 of the 100 randomly selected files with no detections by the software (i.e., rate of false negatives was 45%).

Song Scope detected 6,346 vocalizations of Bullfrogs in 1,336 audio files, with a true positive rate of 97.3%. Most of the misidentifications of Bullfrogs by the software were due to insect sounds and noise. We heard Bullfrog vocalizations on 48 of the 100 randomly selected files that were not selected by Song Scope as having Bullfrog calls.

Song Scope detected 32,242 vocalizations of Bronze Frogs in 1,984 audio files. The true positive rate for Bronze Frogs was 87.2%, and misidentifications were primarily from calls of other frog species and from birds. We detected Bronze Frog vocalizations in 51 of the 100 randomly chosen files in which Song Scope did not detect the species.

DISCUSSION

It took approximately one week of work to refine the settings in Song Scope for recognition of each of the three species studied, which is the same amount of training time Taylor et al. (1996) required for their computer-based recognition system. After the creation of a satisfactory recognizer is complete, it is a simple matter to scan additional digital files for the same species. Thus, one person could easily operate a network of many ARS field units and then easily scan the files for a particular species in little additional time. This is an advantage over manually listening to recordings, which requires at least as much time as the duration of the recordings. The disadvantage to automatic recognition of anuran vocalizations is the occurrence of false positive and false negative errors.

The rate of false positive errors observed for the three species in our test of Song Scope is a cause for concern. It is possible to reduce the error rate slightly with additional refinements to the settings in the software or with other training data, but we think it is unlikely to greatly reduce the false positive rate. There is a tradeoff between types of error in automatic recognition of vocalizations. The user must set the minimum values of quality and score very high to reduce false positive errors, but this will always increase the rate of false negative errors. Our settings for minimum quality and score helped minimize false positives but produced false negatives in 45%-51% of the files we examined.

It is difficult to infer patterns about the sources of misidentification of vocalizations for the three species we studied. Green Treefrogs had a false positive rate of 15.8% and birds and other frog species were the primary sources of misidentifications. American Bullfrogs had a very low rate of false positives, and a variety of sources accounted for the few misidentifications. Bronze Frog vocalizations and misidentifications were by far the most plentiful. The primary source of misidentifications for Bronze Frogs was other frogs. It is difficult to determine exactly what characteristics of these other sounds cause Song Scope to classify them as the target vocalization. The creation of the recognizer file is a subjective process that involves picking examples that are representative of the entire range of the species. It is possible to reduce the false positive rate with additional manipulation in the software, but the specific cause for the misidentifications is unknown. As we did in this study, users of automatic vocalization recognition software could estimate the overall occurrence of frog calls in the recorded files by determining the rates of false positive and false negative errors.

The tradeoff between error types underscores the importance of careful design of a monitoring program intended to rely on automatic recognition of anuran vocalizations. The specific goals of the project will help guide this implementation. For instance, if the goal were to survey for a rare species or to document accurately the beginning of the calling season for a species, then false negatives would need to be minimized. This would require scanning files with low minimums for quality and score; however, this would produce many more false positives. A human listener would be required to listen to the files selected by the software to verify actual vocalizations of the target species from among the identifications made by the software. However, if the goal of the monitoring is to implement a large network of monitoring stations and minimize the need for human listeners, the recognition software should be set to scan with the highest possible minimum values for quality and score to minimize false positives. This will certainly increase the number of false negative errors, but it will ensure that the majority of the identifications made by the software are accurate.

There is potential for the comparison of automatic vocalization recognition data with other sources of anuran vocalization data (e.g., North American Amphibian Monitoring Program, NAAMP; Weir and

Mossman 2005). The NAAMP monitoring protocol uses a code for calling intensity on a scale of 0–3, and Corn and Muths (2002) developed a call saturation index (CSI) by summing these scores and dividing by the number of samples per day. Researchers can easily derive surrogates for these measures from call recognition software, but a high degree of confidence in the accuracy of the software is required in order to compare these to manually gathered data. One must minimize both false positive and false negative errors in order to justify this type of comparison. Any comparison with manually collected vocalization data would require some estimate of the error rate for automatically identified calls.

The availability of low-cost, highly reliable digital field recorders makes the idea of a large network of autonomous recorders in the field very attractive. This equipment offers a method for highly reliable, standardized observations of anurans in remote locations at any date and time. It is now very easy to obtain more recordings than it is feasible to listen to manually. We caution researchers to keep in mind the cost of processing the data collected with such a network. There is a considerable amount of time needed to train vocalization recognition software and thoroughly test it, and storage of the digital data requires space and a plan for archiving back-up copies. Careful thought of how the data are to be used and how the monitoring program will handle errors made by the vocalization recognition software should be done before the monitoring is begun.

Acknowledgments.—Funding for this research was provided by the U.S. Geological Survey Amphibian Research and Monitoring Initiative. Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

LITERATURE CITED

Acevedo, M.A., and L.J. Villanueva-Rivera. 2006. Using automated digital recording systems as effective tools for the monitoring of birds and amphibians. Wildlife Society Bulletin 34:211–214.

Brandes, T.S., P. Naskrecki, and H.K. Figueroa. 2006. Using image processing to detect and classify narrow-band cricket and frog calls. Journal of the Acoustical Society of America 120:2950–2957.

Bridges, A.S., and M.E. Dorcas. 2000. Temporal variation in anuran calling behavior. Copeia 2000:587–592.

Corn, P.S., and E. Muths. 2002. Variable breeding phenology affects the exposure of amphibian embryos to ultraviolet radiation. Ecology 83:2958–2963.

Genet, K.S., and L.G. Sargent. 2003. Evaluation of methods and data quality from a volunteer-based amphibian call survey. Wildlife Society Bulletin 31:703–714.

Given, M.F. 1987. Vocalizations and acoustic interactions of the Carpenter Frog, *Rana virgatipes*. Herpetologica 43:467–481.

Given, M.F. 1990. Spatial distribution and vocal interaction in *Rana clamitans* and *R. virgatipes*. Journal of Herpetology 24:377–382.

Hsu, M.Y., Y.C. Kam, and G.M. Fellers. 2005. Effectiveness of amphibian monitoring techniques in a Tiawanese subtropical forest. Herpetological Journal 15:73–79.

Nelson, G.L., and B.M. Graves. 2004. Anuran population monitoring: comparison of the North American Amphibian Monitoring Progam's calling index with mark–recapture estimates for *Rana clamitans*. Journal of Herpetology 38:355–359.

Ossen, K.L., and R.J. Wassersug. 2002. Environmental factors influencing calling in sympatric anurans. Oecologia 133:616–625.

Taylor, A., G. Watson, G. Grigg, and H. McCallum. 1996. Monitoring frog communities: an application of machine learning. Pp. 1564–1569 *In* Anonymous (ed.). Proceedings of the 8th Innovative Applications of Artificial Intelligence Conference, Portland, Oregon, USA.

Weir, L.A., and M.J. Mossman. 2005. North American Amphibian Monitoring Program (NAAMP). Pp. 307–313 *In* Amphibian Declines: The Conservation Status of United States Species. Lannoo, M.J. (Ed.). University of California Press, Berkeley, California, USA.

Weir, L.A., J.A. Royle, P. Nanjappa, and R.E. Jung. 2005. Modeling anuran detection and site occupancy on North American Amphibian Monitoring Program (NAAMP) routes in Maryland. Journal of Herpetology 39:627–639.

Yen, G.G., and Q. Fu. 2001. Automatic frog calls monitoring system: a machine learning approach. International Journal of Computational Intelligence and Applications 1:165–186.

Zimmerman, B.L. 1994. Audio strip transects. Pp. 92–97 *In* Measuring and Monitoring Biological Diversity: Standard Methods for Amphibians. Heyer, W.R., M.A. Donnelly, R.W. McDiarmid, L.C. Hayek, and M.S. Foster (Eds.). Smithsonian Institution Press, Washington, D.C., USA.

HARDIN WADDLE is a research ecologist at the U.S. Geological Survey's National Wetlands Research Center in Lafayette, Louisiana, and a regional principal investigator for the Amphibian Research and Monitoring Initiative. He received his B.S. from Auburn University, his M.S. from Florida International University, and his Ph.D. from the University of Florida. Waddle's research focuses on applying quantitative ecological methods to address questions concerning the conservation and management of amphibians and reptiles. (Photographed by Tyler F. Thigpen).

TYLER F. THIGPEN is a student intern. She works with the South Central Amphibian Research Monitoring Initiative (SC ARMI) at the U.S. Geological Survey, National Wetlands Research Center in Lafayette, Louisiana. Tyler received a B.S. in wildlife biology and forest resources at the University of Georgia's Warnell School of Forestry and Natural Resources. She is currently working on an M.S. at the University of Louisiana at Lafayette. Her research focuses on the sublethal effects of contaminants on amphibians in the Atchafalaya River Basin in south-central Louisiana. (Photographed by Stephen Jones).

BRAD M. GLORIOSO is currently a general biologist for IAP Worldwide Services Inc. at the U.S. Geological Survey's National Wetlands Research Center in Lafayette, Louisiana. He received his B.S. from Southeastern Louisiana University, and his M.S. from Middle Tennessee State University, where his thesis focused on population ecology and feeding activity in stinkpots. While turtles remain his love, his current research focuses on long term amphibian monitoring in the Atchafalaya Basin as part of the Amphibian Research and Monitoring Initiative (ARMI). (Photographed by Matthew L. Niemiller).