

# Voice recognition technology as a tool for behavioral research

DAVID J. WHITE, ANDREW P. KING, and SHAN D. DUNCAN  
*Indiana University, Bloomington, Indiana*

Behavioral research often requires the acquisition and processing of large volumes of data. Most current techniques for recording behavior constrain the amount and type of data that can be measured. We developed and tested a system that uses voice recognition technology to collect data on the social interactions and singing patterns of cowbirds (*Molothrus ater*) living outdoors in a semi-natural environment. We spoke observation data into a wireless microphone that transmitted the data to a computer in the laboratory. After collection, the data were automatically checked for errors and then were entered into a database. Overall, the system performed at extremely high levels of accuracy. Furthermore, owing to the removal of constraints on observers such as breaking visual contact with subjects and manual data entry into a database, we were able to increase the amount of data collected and to collect new measures of social interactions that have not been available to us in the past. We tested the system under the challenging circumstances of field observation, and it performed above our expectations. In a laboratory setting, if transmission difficulties are removed, voice recognition could be even more accurate. We recommend voice recognition as a powerful new tool for the variety of research fields in which measuring behavior is involved.

Observational research often requires the unobtrusive acquisition of large volumes of data. Innovations in computer power and the increasing ease of use of modern software now provide the potential to analyze behavior on a scale of complexity that has been impossible in the recent past. This potential, however, has rarely been realized due to the inability to collect large enough data sets. The removal of limitations of data collection and the exploitation of modern computer resources could result in dramatic new advances in the study of behavior.

Voice recognition technology (VRT) provides a way to alleviate the longstanding problem of capturing complex and rapid sequences of behavior. VRT not only allows more data to be collected, but also provides the opportunity to take fundamentally new measures of behavior, both of intricate detail and of larger context. Voice recognition offers advantages for researchers in the lab and especially in the field. VRT (1) allows researchers to maintain visual contact with mobile subjects, spending less time looking at a data sheet or keypad and then trying to relocate individ-

ual subjects, (2) removes encumbrance of data-recording paraphernalia, providing freedom of movement, (3) frees researchers from having to enter data manually into a computer, and (4) requires relatively little setup time and can be modified (e.g., new measures can be added) without significant training time. These advantages also reduce the opportunity for human error at each stage of data collection.

Of course, other technologies exist for automated data collection (Table 1). Technologies such as voice recorders or keypad systems that interface with specialized software are available and some are affordable (e.g., Barrett & Barrett, 2001; Noldus, Trienes, Hendriksen, Jansen, & Jansen, 2000). For our purposes, however, we have never been able to find a system that provides all four of the advantages of voice recognition listed above. For example, keypad systems can be used to enter data directly into a computer but they add encumbrance to observers and also add an extra level of complexity to data collection, since observers must learn and key in the code for the observed behavior. Adding new measures to a keypad data-collection protocol requires significant system reprogramming and observer retraining. Alternatively, tape recorders can be used to record speech but require significant time to transcribe the dictation into data. We have, therefore, to this point been reliant on traditional paper-and-pencil methods for collection of developmental measures in brown-headed cowbirds (*Molothrus ater*).

Despite the potential of voice recognition as a research tool in alleviating the constraints of data collection, VRT has rarely been used as a data collection technique in the lab (but see Grasso & Grasso, 1994) and never, to our knowledge, in the field. The goal of this study was to test the feasibility of using VRT as a tool for recording behav-

---

S.D.D. is at the Center for the Integrative Study of Animal Behavior at Indiana University. We wish to acknowledge the following sources of funding: NSF, Howard Hughes Biomedical Institute, and the Department of Psychology at Indiana University. We would also like to acknowledge Anne Smith for conducting early work in our lab with a voice recognition system. We would also like to thank Lee Drickamer, Vicenc Quera, Jonathan Vaughan, and an anonymous reviewer for making valuable comments on an earlier draft of this article. Research was approved by the Animal Care and Use Committee of Indiana University (Study 99-108). Correspondence should be addressed to D. J. White, Indiana University, Department of Psychology, 1101 E. 10th St., Bloomington, IN 47405 (e-mail: dajwhite@indiana.edu).

Table 1  
Comparison of Data Collection Methods

Variables	Traditional Methods		Computer-Assisted Methods	
	Pencil/Paper	Audio Recordings	Keyboard Assisted <sup>a</sup>	Voice Recognition <sup>a</sup>
Setup time <sup>b</sup>	Low	Low	Medium to high	Low to medium
Training time <sup>b</sup>	Low to medium	Low to medium	Medium to high <sup>c</sup>	Low
Number of subjects	Limited to form/paper size	Unlimited	Limited to keyboard mappings	Limited to screen size
Observer's focus <sup>d</sup>	Split—subjects and data entry	Solely on subjects	Split—subjects and data entry	Split—subjects and data entry
Encumbrance	Hands required	Hands free	Hands required	Hands required
Resulting data computer ready <sup>e</sup>	No	No	Yes	Yes
Automatic error checking <sup>f</sup>	No	No	Yes	Yes
Cost of system <sup>g</sup>	Low	Low	Medium to high	Low to medium

Note—Data was recorded on data sheets with the paper-and-pencil method and on a tape recorder with the audio-recording method. The keyboard-assisted method used Observer, ObsWin, and ProCoder, the voice recognition method used IBM ViaVoice, and the PDA-based method used Newton and PalmPilot. <sup>a</sup>Observer software can be found at <http://www.noldus.com>, ObsWin at <http://pssgumi.bham.ac.uk/obswin/obswin2.htm>, ProCoder at <http://kc.vanderbilt.edu/~jont/procoder.html>, ViaVoice at [http://www.ibm.com/software/speech/For software for the Apple Newton, see Ethoscribe \(http://www.tumasci.com\), for the Palm Pilot see Outdoor Explorer \(http://www.phenotyping.de/Outdoorexplorer.html\)](http://www.ibm.com/software/speech/For software for the Apple Newton, see Ethoscribe (http://www.tumasci.com), for the Palm Pilot see Outdoor Explorer (http://www.phenotyping.de/Outdoorexplorer.html)). <sup>b</sup>Low  $\leq$  2 h, Medium = 2–10 h, High  $\geq$  10 h. <sup>c</sup>Training time to map behavior codes to specific keys or keystroke combinations can take as much as two-thirds the total training time (James Ha, personal communication, August 2000). <sup>d</sup>Focus of observer's attention, for example, to glance down at the data sheet or computer screen. <sup>e</sup>Data is in computer ready format for further analysis, saving time and avoiding transcription/translation related errors. <sup>f</sup>Error checking is carried out by the computer on entered codes; the observer does not need to manually verify each observation. <sup>g</sup>Depends on whether base computer system for recording the observations is available. Low cost = \$100–1,000, Medium = \$1,000–\$5,000, High = \$5,000+.

ior in an outdoor, semi-natural environment. For the system to succeed, we required it to have (1) high levels of accuracy, (2) robustness of computer performance, (3) usability outside in many environmental conditions, and (4) functionality at a distance from the receiving computer. Below, we describe how we developed and tested a method of recording behavior with the use of voice to measure social interactions and vocal behavior of cowbirds in large, outdoor, wire-mesh aviaries. For reference, we compared the VRT method's performance with that of a paper-and-pen method of recording behavior that we have established in the past to be a reliable method of measuring patterns of behavior predictive of important developmental outcomes (White, King, & West, in press).

## METHOD

We used IBM ViaVoice Millennium Pro Edition voice recognition software. We tested a number of other software packages (including Lernout & Hauspie's Voice Xpress, CSLU's Speech Toolkit, Dragon Systems' Naturally Speaking, and Phillips's Freespeech). At the time of evaluation (summer 1999), ViaVoice proved to be the most mature technology, provided the best support, and offered the best cross-platform compatibility. ViaVoice also provided options for manually disabling features of the software, such as commands and wizards that, if opened accidentally, can stop the program (a serious problem for observers not able to see the monitor). Finally, ViaVoice provided the most options for programming customizable commands (see below).

We operated the software on a Pentium III, 500-MHz IBM-compatible computer (Compaq Deskpro EP), with a 128-bit PCI Sound Blaster sound card (Creative Technology), running Microsoft Windows 98. The Pentium III processor provided voice transcription fast enough to be impossible for us to overload the system by speaking quickly. We used a solid-state, wireless, omni-directional lapel microphone (Telex WT 150; Telex Communications) and receiver system (Telex FMR 150) that could be used in a variety of weather conditions. The microphone was small and lightweight, allowing us to move around when taking data and to be unencumbered, freeing our hands to use binoculars. Speech was transcribed into text in Microsoft Word 2000 word processing software. Although ViaVoice does allow dictation into other applications, the word processor application allows speech to be recorded in real time with the text and to be saved, which was beneficial for locating and correcting any errors made by the observer. After transcription, we exported Microsoft Word files into a database (4th Dimension v. 6.5.1; ACI). We programmed the database to match the incoming text to a list of possible codes, to detect and automatically correct consistently made errors.

## Training

We trained the system to recognize voice patterns of data collectors by using the enrollment scripts and instructions provided by the ViaVoice software. We then developed and trained in ViaVoice a custom vocabulary of approximately 200 words that included acronyms for the colors of the leg bands of each of 85 birds and codes for automated navigation (see below). We trained words in our custom vocabulary by dictating and correcting errors repeatedly until recognition accuracy surpassed approximately 95%.

Upon reaching 95% accuracy in the lab, we began using the system in aviaries. The four aviaries were constructed of wood and were completely enclosed in wire mesh. They were located outdoors, a distance of 30–100 m from the receiving computer. Aviaries housed groups of 14–25 cowbirds. Groups differed in age and sex composition (for details of experimental procedures, see White et al., in press).

ViaVoice provides the capability to program macro codes for navigation and dictation. Several options for macros are available as built-

in functions, such as adding the current date or time. Other macros, such as behavioral code, can be created manually. A macro is made by creating a command word and then either by selecting a built-in function or by creating one in the provided text box. For example, we used the voice command copulation to enter a combination of customized dictation and navigation functions. It inputs the behavioral code for a copulation followed by the current time, several commas (for database use), and then returns the cursor to the next line of text. This (albeit simple) macro is processed extremely rapidly, never causing any lag between the time the command is spoken and the time that is reported (the reported time, however, is not displayed in units more precise than seconds). Although macros in ViaVoice are limited to simple commands, creating macros is fast, easy, and requires no programming experience. ViaVoice also offers developer software that can be used to program new speech applications and to increase the potential for customizing the system.

### Procedure

We used the system to collect data on two types of social behaviors of cowbirds. We measured social assortment by recording near-neighbor associations, and we measured singing interactions by recording males' production and use of song.

**Near neighbors.** We recorded near neighbors by sampling all birds in each aviary in 7-min blocks and noting any bird within 30 cm of any other bird. We collected 151.43 h of data from January through June 2000. Although 2 observers could not take data in the same aviary at the same time (since they would interfere with each other's voice recognition), each observer did collect data in each aviary each morning. For the voice method of data collection, we spoke two leg band acronyms (focal bird band then near-neighbor band) into the microphone. For the paper method of data collection, we wrote out near-neighbor band acronyms on data sheets that listed all focal bands in the aviary.

**Singing.** To measure the birds' singing patterns, we conducted song censuses for 219.75 observation hours from May through July 2000. In song censuses, we scanned aviaries in 15-min blocks, recording any male that vocalized and noting whether the vocalization was directed to any other bird, and if so, whether there was any response from the recipient of the vocalization. We recorded similar data with a paper method, recording vocalizations on data sheets. However, since the singing interactions were so frequent and rapid, it was impossible with the paper method to note responses by recipients to vocalizations.

## RESULTS

Initial enrollment took on average 2 h. Although it was possible to complete a partial enrollment in less than 15 min, accuracy improved noticeably with increased enrollment. Training words from our custom vocabulary took 20 min. Since voice recognition accuracy improved after we repeatedly corrected misrecognized words, getting the system to perform at 95% accuracy required approximately 10 h of correction.

After the data were collected, we checked and corrected any errors in the data taken by the voice method and entered all data into the database. Correction time varied over time. For the first week, correcting and inputting data into the database for the voice method required approximately 10 min per block (40–60 min per day), but this declined to less than 1 min per block (2–5 min per day) within 2 weeks. The time required by the paper method to input data manually averaged 7 min per block (28–56 min per day) and remained constant.

VRT accuracy averaged 98.8% in the aviaries. This accuracy rate was based on a sample of 15,984 datapoints collected in May 2000 and included errors that occurred due to signal reception and ambient noise. When such errors occurred, they usually were displayed as added words or characters and rarely resulted in band-recognition errors. The database error-checking programs detected most band errors. Errors that occurred that were not detected in the database made up 0.06% of the sample. Upon completion of data collection, we retested accuracy in the laboratory. Based on a sample of 2,158 datapoints, accuracy in the lab averaged 99.3%.

A comparison of a random sample of 30 near-neighbor blocks taken by the voice and the paper methods revealed that the voice recognition method provided significantly more near-neighbor points per 7-min block (36.1 & 2.61 points) than did the paper method of data collection ( $17.4 \pm 0.88$ ; Mann-Whitney  $U$  test,  $U = 70$ ,  $p < .0001$ ). Furthermore, the variability in the number of near-neighbor points recorded with the voice method (range, 12–70 datapoints collected per block) was significantly greater than the variability in the number of near-neighbor points recorded with the paper method [range, 11–25 datapoints collected per block; Student's  $t$  test,  $t(58) = 5.49$ ,  $p < .0013$ . This difference resulted from an upper limit on collecting data by the paper method in two of the more active aviaries. In these aviaries, changes in Individual assortment were so frequent that many near-neighbor points could not be recorded with the use of the paper method. Because of the results from the voice method, we now have access to differences in the rates of social interactions across the different age and sex groups. This is a qualitatively new measure not available with the paper method.

A comparison of a random sample of 14 song census blocks taken with the voice and the paper methods revealed that significantly more songs were recorded with the voice method than with the paper method (71.6 vs. 50.8 songs per block;  $t(26) = 3.42$ ,  $p < .002$ ). In addition to allowing us to record more data, the voice method automatically entered the current time when we spoke, which provided us with detailed measures of the temporal sequence of behavior, measures not available with the paper method.

We obtained measures of interobserver reliability based the number of near-neighbor associations per bird recorded by the 2 observers over the entire study. Reliability was high for both the voice (Pearson's  $r = .91$ ,  $p < .001$ ) and the paper methods ( $r = .76$ ,  $p < .001$ ). Furthermore, there was a significant correlation in patterns of near-neighbor associations taken per bird between the two methods of data collection ( $r = .51$ ,  $p < .001$ ). We also obtained measures of interobserver reliability for singing patterns. We compared between observers the number of songs recorded for each male in the four aviaries. Interobserver reliability for the voice method was significantly higher than that for the paper method ( $r = .98$  and  $r = .36$ , respectively;  $z = 2.46$ ,  $p < .05$ ). There was also a significant correlation in the patterns of singing recorded per bird by the two methods ( $r = .90$ ,  $p < .001$ ). There was no

difference between observers in the amount of data collected per singing bird ( $t = -0.72$ , n.s.).

## DISCUSSION

Voice recognition, used to record social behavior in cowbirds in a semi-natural setting, proved to be a quantitative and qualitative advance over the paper method of recording behavior. VRT, limited only by speaking speed, allowed us to collect more data. In addition, VRT allowed us to collect qualitatively new measures of sequences of social behavior, detail on responses to song, and differences in rate of social interactions. VRT, working in concert with database error checking, performed at high levels of accuracy. Finally, VRT increased the efficiency of data input, essentially by removing all manual data entry (and with it, the potential for making errors in data entry).

We used commercially available and affordable equipment with little customization or programming. Although we did have the advantage of working in an area near a fully functioning laboratory, we tested voice recognition outside, under circumstances very challenging to the system, which included reception difficulties (transmission on wireless microphones from inside metal aviaries, on average 65 m away from the receiving computer), fluctuating ambient noise levels (including wind, rain, bird song, planes, helicopters, chainsaws, etc.), and the lack of any computer feedback. Despite these most difficult conditions, the system exceeded our expectations of accuracy, reliability, and efficiency.

Better performance than reported here would be attained were the system used in an actual field setting by removing transmission difficulties associated with conducting observations in metal enclosures. We used lapel microphones because they were less cumbersome, but in conditions where ambient noise is pronounced, unidirectional headset microphones would be more effective in dampening background sound. ViaVoice does not require specialized hardware and can operate from a notebook computer, provided that enough battery power or an electric power source exists. The use of a notebook would remove difficulties associated with wireless microphone transmission and the lack of feedback. Although we have concentrated here on performance in the field, voice recognition could be used in a laboratory setting at extremely high levels of accuracy, where reception problems, feedback issues, and environmental variables are removed.

The system would be less appropriate for research in which a vocalizing observer would be a distraction to the subjects or when multiple observers using separate systems would be working in close proximity. We have worked with 2 observers, using separate systems, as close as 8 m apart with no interference problems ever occurring. Observers who are required to be in close proximity would need to use directional microphones and speak quietly. ViaVoice has audio input controls that can be set to be sensitive to

quiet speaking. Our experience, however, has been that as voice amplitude decreases, errors increase. Thus, in such situations where sound is an issue or in areas where ambient noise is extreme, other observational systems may be more appropriate (see Table 1).

Our criticisms of the system are twofold. First, the reliability of the software could be improved. Although we never lost data to a computer crash, mysterious errors did occur in the daily operation of ViaVoice, including commands occasionally becoming inoperable. For example, the command *delete line* occasionally worked correctly in deleting the line of text on the current line. Other times, however, the command would be displayed, but no text would be deleted. Software upgrades have proven to reduce many of the bugs present in earlier versions, and ViaVoice was the most reliable package that we tested. Second, the investment in time to get the system trained and running at high levels of accuracy was significant and varied with individual speaking styles. The long-term return on this investment, however, in terms of efficiency, has been pronounced.

VRT has become the dominant method of data collection in our laboratory. We now use it to collect data on complex social interactions. These measures require the constant observation of subjects, frequent restructuring of measures, precise measures of the timing of sequences, and results in copious amounts of data that could not be managed without automatic transcription into a database. We have currently trained several other observers to use voice recognition, both female and male, and have seen no sex-related effects on the system's accuracy.

Overall, owing to the almost perfect accuracy of the system, the ability to increase the amount and type of data collected, the ability to maintain constant visual contact with subjects, the removal of almost all data-entry time (and associated errors), and the potential to add and change measures in the system rapidly and easily, we would recommend adding voice recognition to the ranks of automated data collection techniques (Table 1) as a serious tool for studying behavior.

The use of VRT, although providing the measurable advantages outlined above, has resulted in our becoming better observers. The removal of the demands of manual data recording has allowed us more time to observe and study the behavior unfolding in front of us. This advantage, although more difficult to quantify in terms of accuracy or efficiency, may be the most important benefit of voice recognition as a tool in behavior research.

## REFERENCES

- BARRETT, J. F., & BARRETT, D. J. (2001). An introduction to computerized experience sampling in psychology. *Social Science Computer Review*, *19*, 175-185.
- GRASSO, M. A., & GRASSO, C. T. (1994). Feasibility study of voice-driven data collection in animal drug toxicology studies. *Computers in Biology & Medicine*, *24*, 289-294.
- NOLDUS, L. P. J. J., TRIENES, R. J. H., HENDRIKSEN, A. H. M., JANSEN, H.,

& JANSEN, R. G. (2000). The Observer Video-Pro: New software for the collection, management, and presentation of time-structured data from videotapes and digital media files. *Behavior Research Methods, Instruments, & Computers*, 32, 197-206.

WHITE, D. J., KING, A. P., & WEST, M. J. (in press). Facultative devel-

opment in juvenile male cowbirds (*Molothrus ater*). *Behavioral Ecology*.

(Manuscript received December 29, 2000;  
revision accepted for publication August 23, 2001.)