Classroom sound can be used to classify teaching practices in college science courses

Melinda T. Owens[a,1], Shannon B. Seidel[b,1], Mike Wong[c,1], Travis E. Bejines[b], Susanne Lietz[a], Joseph R. Perez[b], Shangheng Sit[a], Zahur-Saleh Subedar[a], Gigi N. Acker[d,e], Susan F. Akana[f], Brad Balukjian[g], Hilary P. Benton[a,h], J. R. Blair[a], Segal M. Boaz[i], Katharyn E. Boyer[a,j], Jason B. Bram[d], Laura W. Burrus[a], Dana T. Byrd[a], Natalia Caporale[k], Edward J. Carpenter[a,j], Yee-Hung Mark Chan[a], Lily Chen[a], Amy Chovnick[i], Diana S. Chu[a], Bryan K. Clarkson[l], Sara E. Cooper[h], Catherine Creech[m], Karen D. Crow[a], José R. de la Torre[a], Wilfred F. Denetclaw[a], Kathleen E. Duncan[h], Amy S. Edwards[h], Karen L. Erickson[h], Megumi Fuse[a], Joseph J. Gorga[n], Brinda Govindan[a], L. Jeanette Green[o], Paul Z. Hankamp[p], Holly E. Harris[a], Zheng-Hui He[a], Stephen Ingalls[a], Peter D. Ingmire[a,q], J. Rebecca Jacobs[h], Mark Kamakea[r], Rhea R. Kimpo[a,s], Jonathan D. Knight[a], Sara K. Krause[t], Lori E. Krueger[u,v], Terrye L. Light[a], Lance Lund[a], Leticia M. Márquez-Magaña[a], Briana K. McCarthy[w], Linda J. McPheron[x], Vanessa C. Miller-Sims[a], Christopher A. Moffatt[a], Pamela C. Muick[u,y], Paul H. Nagami[a,g,z], Gloria L. Nusse[a], Kristine M. Okimura[aa], Sally G. Pasion[a], Robert Patterson[a], Pleuni S. Pennings[a], Blake Riggs[a], Joseph Romeo[a], Scott W. Roy[a], Tatiane Russo-Tait[bb], Lisa M. Schultheis[h], Lakshmikanta Sengupta[p], Rachel Small[cc], Greg S. Spicer[a], Jonathon H. Stillman[a,j], Andrea Swei[a], Jennifer M. Wade[dd], Steven B. Waters[w], Steven L. Weinstein[a], Julia K. Willsie[l], Diana W. Wright[e,ee], Colin D. Harrison[ff], Loretta A. Kelley[gg], Gloriana Trujillo[hh], Carmen R. Domingo[a], Jeffrey N. Schinske[d,h], and Kimberly D. Tanner[a,2]

[a]Department of Biology, San Francisco State University, San Francisco, CA 94132; [b]Department of Biology, Pacific Lutheran University, Tacoma, WA 98447; [c]Center for Computing for Life Sciences, San Francisco State University, San Francisco, CA 94132; [d]Department of Biology, De Anza College, Cupertino, CA 95014; [e]Nutrition, Food Science, and Packaging Department, San Jose State University, San Jose, CA 95192; [f]Biology Department, City College of San Francisco, San Francisco, CA 94112; [g]Biology Department, Laney College, Oakland, CA 94607; [h]Department of Biology, Foothill College, Los Altos Hills, CA 94022; [i]Biology Department, Las Positas College, Livermore, CA 94551; [j]Romberg Tiburon Center for Environmental Studies, San Francisco State University, Tiburon, CA 94920; [k]Department of Neurobiology, Physiology, and Behavior, University of California, Davis, CA 95616; [l]Department of Biological Science, Diablo Valley College, Pleasant Hill, CA 94523; [m]Department of Biology, Portland Community College, Portland, OR 97219; [n]Math and Sciences Department, Diablo Valley College, San Ramon, CA 94582; [o]Science and Technology Division, Cañada College, Redwood City, CA 94061; [p]Biology Department, College of San Mateo, San Mateo, CA 94402; [q]Division of Undergraduate Education and Academic Planning, San Francisco State University, San Francisco, CA 94132; [r]Life Science Department, Chabot College, Hayward, CA 94545; [s]Science/Mathematics/Technology Division, Skyline College, San Bruno, CA 94066; [t]Life Sciences Department, Palomar College, San Marcos, CA 92069; [u]Biology Department, Solano Community College, Fairfield, CA 94534; [v]Department of Biological Sciences, California State University, Sacramento, CA 95819; [w]Biology Department, Los Medanos College, Pittsburg, CA 94565; [x]Science Department, Berkeley City College, Berkeley, CA 94704; [y]Biological Sciences Department, Contra Costa College, San Pablo, CA 94806; [z]Department of Biological Science, Holy Names University, Oakland, CA 94619; [aa]Department of Earth and Climate Sciences, San Francisco State University, San Francisco, CA 94132; [bb]Department of Curriculum and Instruction, STEM Education, University of Texas at Austin, Austin, TX 78712; [cc]Department of Chemistry and Biochemistry, San Francisco State University, San Francisco, CA 94132; [dd]Department of Biology, University of San Francisco, San Francisco, CA 94117; [ee]Biological, Health & Environmental Sciences Division, DeAnza College, Cupertino, CA 95014; [ff]School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA 30332; [gg]Kelley, Petterson, and Associates, Inc., San Francisco, CA 94127; and [hh]Office of the Vice Provost for Teaching and Learning, Stanford University, Stanford, CA 94305

Edited by Bruce Alberts, University of California, San Francisco, CA, and approved January 31, 2017 (received for review November 20, 2016)

Active-learning pedagogies have been repeatedly demonstrated to produce superior learning gains with large effect sizes compared with lecture-based pedagogies. Shifting large numbers of college science, technology, engineering, and mathematics (STEM) faculty to include any active learning in their teaching may retain and more effectively educate far more students than having a few faculty completely transform their teaching, but the extent to which STEM faculty are changing their teaching methods is unclear. Here, we describe the development and application of the machine-learning–derived algorithm Decibel Analysis for Research in Teaching (DART), which can analyze thousands of hours of STEM course audio recordings quickly, with minimal costs, and without need for human observers. DART analyzes the volume and variance of classroom recordings to predict the quantity of time spent on single voice (e.g., lecture), multiple voice (e.g., pair discussion), and no voice (e.g., clicker question thinking) activities. Applying DART to 1,486 recordings of class sessions from 67 courses, a total of 1,720 h of audio, revealed varied patterns of lecture (single voice) and nonlecture activity (multiple and no voice). We also found that there was significantly more use of multiple and no voice strategies in courses for STEM majors compared with courses for non-STEM majors, indicating that DART can be used to compare teaching strategies in different types of courses. Therefore, DART has the potential to systematically inventory the presence of active learning with ∼90% accuracy across thousands of courses in diverse settings with minimal effort.

active learning | evidence-based teaching | science education | lecture | assessment

Current college STEM (science, technology, engineering, and mathematics) teaching in the United States continues to be lecture-based and is relatively ineffective in promoting learning (1, 2). Undergraduate instructors continue to struggle to engage, effectively teach, and retain postsecondary students, both generally and particularly among women and students of color (3, 4). Federal analyses suggest that a 10% increase in retention of undergraduate STEM students could address anticipated STEM workforce shortfalls (5). Replacing the standard lecture format with more active teaching strategies has been shown to increase

Author contributions: M.T.O., S.B.S., M.W., J.N.S., and K.D.T. designed research; M.T.O., S.B.S., M.W., T.E.B., S.L., J.R.P., S.S., Z.-S.S., G.N.A., S.F.A., B.B., H.P.B., J.R.B., S.M.B., K.E.B., J.B.B., L.W.B., D.T.B., N.C., E.J.C., Y.-H.M.C., L.C., A.C., D.S.C., B.K.C., S.E.C., C.C., K.D.C., J.R.d.I.T., W.F.D., K.E.D., A.S.E., K.L.E., M.F., J.J.G., B.G., L.J.G., P.Z.H., H.E.H., Z.-H.H., S.I., P.D.I., J.R.J., M.K., R.R.K., J.D.K., S.K.K., L.E.K., T.L.L., L.L., L.M.M.-M., B.K.M., L.J.M., V.C.M.-S., C.A.M., P.C.M., P.H.N., G.L.N., K.M.O., S.G.P., R.P., P.S.P., B.R., J.R., S.W.R., T.R.-T., L.M.S., L.S., R.S., G.S.S., J.H.S., A.S., J.M.W., S.B.W., S.L.W., J.K.W., D.W.W., C.D.H., L.A.K., G.T., C.R.D., J.N.S., and K.D.T. performed research; M.T.O., S.B.S., M.W., J.N.S., and K.D.T. contributed new reagents/analytic tools; M.T.O., S.B.S., M.W., T.E.B., S.L., J.R.P., S.S., Z.-S.S., J.N.S., and K.D.T. analyzed data; and M.T.O., S.B.S., M.W., T.E.B., J.R.P., J.N.S., and K.D.T. wrote the paper.

Conflict of interest statement: K.D.T., J.N.S., M.W., S.B.S., and M.T.O. have filed a provisional patent on the subject of this report, DART (US Provisional Patent Application No. 62/398,888).

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]M.T.O., S.B.S., and M.W. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: kdtanner@sfsu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1618693114/-/DCSupplemental.

PSYCHOLOGICAL AND COGNITIVE SCIENCES

Significance

Although the United States needs to expand its STEM (science, technology, engineering, mathematics) workforce, United States postsecondary institutions struggle to retain and effectively teach students in STEM disciplines. Using teaching techniques beyond lecture, such as pair discussions and reflective writing, has been shown to boost student learning, but it is unknown what proportion of STEM faculty use these active-learning pedagogies. Here we describe DART: Decibel Analysis for Research in Teaching, a machine-learning–derived algorithm that analyzes classroom sound to predict with high accuracy the learning activities used in classrooms, and its application to thousands of class session recordings. DART can be used for large-scale examinations of STEM teaching practices, evaluating the extent to which educators maximize opportunities for effective STEM learning.

retention, and hundreds of millions of dollars have been invested by national and federal agencies to this end (2). Even for those students retained in STEM, active-learning pedagogies have been repeatedly demonstrated to produce superior learning gains with large effect sizes compared with lecture-based pedagogies (6–9). All of the evidence suggests that shifting large numbers of STEM faculty to include even small amounts of active learning in their teaching may retain and more effectively educate far more students than having a few faculty completely transform their teaching (10).

The extent to which large numbers of STEM faculty are changing their teaching methods to include active learning is unclear. What proportion of United States STEM faculty use anything but lecture with question/answer (Q/A) of individual students? What is the probability that a student would encounter any active learning across all STEM courses in a single department or institution? To address these questions, one would need a measurement tool that could systematically inventory the presence and frequency of active learning not only in one course but also across dozens of departmental courses, multiple STEM departments, and thousands of colleges and universities. Currently available classroom observation tools [e.g., Teaching Dimensions Observation Protocol (TDOP), Reformed Teaching Observation Protocol (RTOP), Classroom Observation Protocol for Undergraduate STEM (COPUS), Practical Observation Rubric To Assess Active Learning (PORTAAL)] (11–14) require trained human observers and are not feasible for addressing questions at this scale. Previous research into using automatic classification of classroom activities largely focuses on K–12 education and has either required special recording equipment (15, 16), analyzed small numbers of teachers (17–19), or did not focus on active-learning pedagogies (17), making these methods insufficient for large-scale analysis of the presence of active learning in college classrooms.

To meet this need, we developed DART: Decibel Analysis for Research in Teaching. DART is a machine-learning–based algorithm that can rapidly analyze thousands of audio-recorded class sessions per day, with minimal costs and without need for human observers, to measure the use of teaching strategies beyond traditional lecture in undergraduate STEM courses. Below we describe the development and validation of DART and report results from over 60 STEM courses drawn from community colleges and a 4-y university.

Results

Our key insight from observations of classroom environments was that nonlecture activities are typically associated with either unusually high noise levels (e.g., pair discussions, small group discussions) or unusually low noise levels (e.g., individual clicker question response, minute paper writing). This suggests that variation in the sound level of a classroom may indicate variation in teaching strategies. To test this hypothesis, an initial 45 audio recordings from 8 instructors teaching different courses (Table 1, pilot group) were analyzed by extracting audio decibel levels at a 2-Hz sampling rate (every 0.5 s) and graphing sound waveforms. To analyze DART's performance in diverse teaching settings, these instructors were purposefully drawn from an atypical pool consisting of people from many different institutions who had undergone over 40 h of professional development in scientific teaching. To determine if patterns of variation in waveforms correlated with activity types, a three-person team listened to all recorded class sessions and individually annotated them using six emergent annotation codes (lecture with Q/A, discussion, silent, video, transition, and other) (Table S1). Sound-level patterns in class sessions primarily using lecture with Q/A were visibly different from the patterns in class sessions with varied learning activities (Fig. 1 *A* and *C*).

Developing an Algorithm to Automate the Classification of Classroom Noise. To develop DART, human annotations were used to design and optimize a machine-learning–based algorithm that reports what types of activities are going on in a classroom based on sound waveforms. To do this task, we applied methods from the field of audio segmentation, which applies machine learning to classify sound into different categories based on statistical characterizations (20). Because some of the human annotation categories yielded waveforms that were statistically similar to each other, we collapsed the six human annotation categories into four activity prediction modes with distinct waveform profiles: single voice, multiple voice, no voice, and other. Lecture with Q/A and video were aggregated into the mode "single voice"; discussion and transition were aggregated into the mode "multiple voice"; silent was assigned to the mode "no voice"; and other was assigned to the mode "other" (Table S1).

To prepare the classroom audio-recording waveforms for the optimization procedure, we tagged each 0.5-s sample of sound from each recording from the pilot group (640,152 samples in total) with three pieces of data: its label from human annotation (S for single voice, M for multiple voice, or N for no voice), the normalized mean volume of the 15-s window of audio around it, and the normalized SD in that window's volume (Fig. S1*A*). Both the mean volume and the SD of the volume of each sample were normalized with respect to their class session.

Then, to sort the samples into the four prediction modes (single voice, multiple voice, no voice, and other), we used an ensemble of binary decision trees comprised of four nodes connected serially. A binary decision tree is a series of decisions to either sort or not sort a given input into a certain category based on the values of the input. Here, the inputs were the 0.5-s samples of classroom audio, and the sorting decisions were based on each sample's normalized mean volume and SD of the volume. In our tree, each node represented one activity prediction mode, and the nodes for each mode were connected in order of decreasing frequency from the pilot data, so

Table 1. Overview of DART study participants

Group	Instructors	Courses	Class sessions	Recorded hours (h)
Pilot group				
Community college	8	8	45	65
Large-scale analysis				
Community college	27	35	712	970
Four-year university	22	32	774	750
All	49	67	1,486	1,720

Total number of instructors, courses, class sessions, and hours recorded in each group.

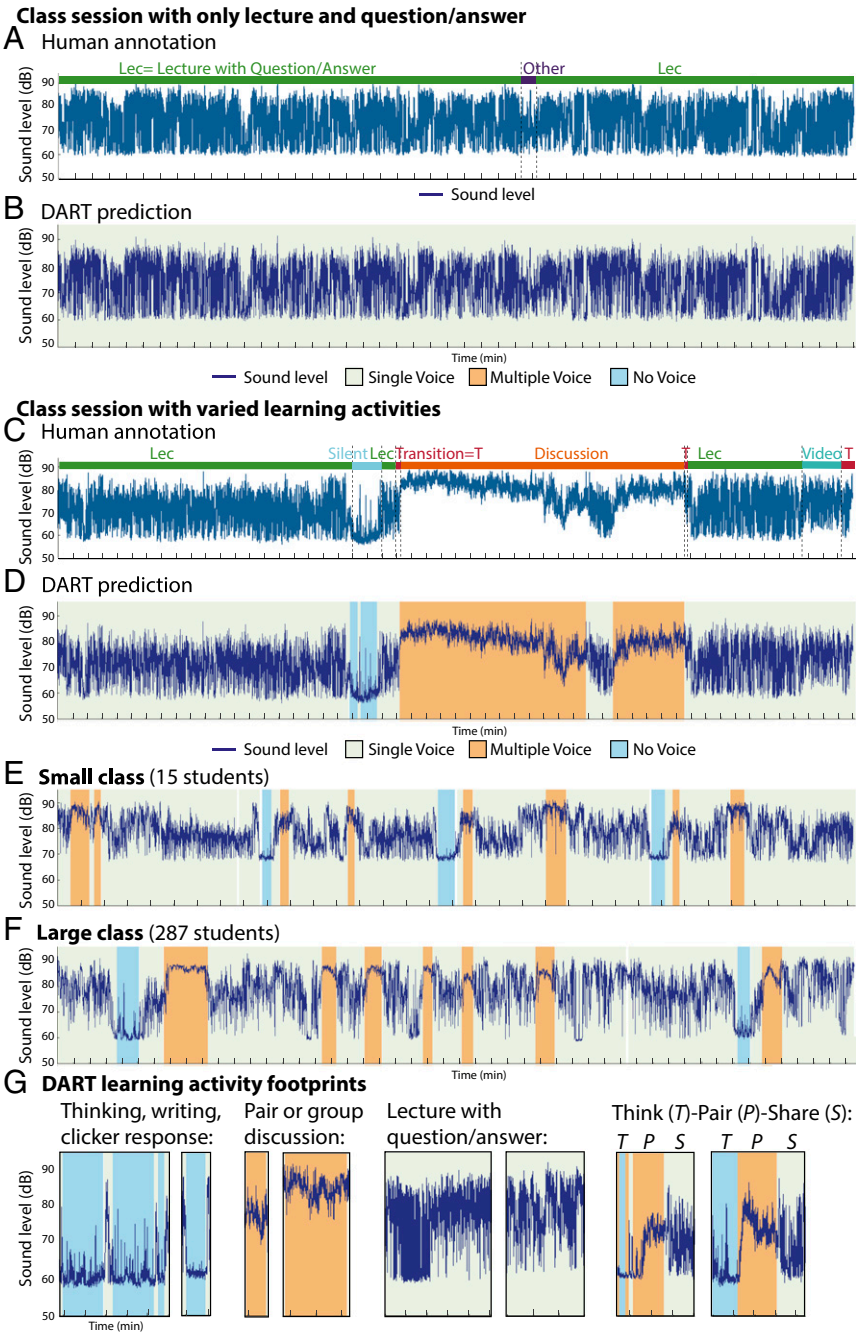

Fig. 1. Sound analysis can differentiate lecture and nonlecture classroom activities. All: Sound levels over time sampled at 2 Hz, with each tickmark indicating 2 min. Typical results are shown. (*A*) Class session with mostly lecture (94 min) with human annotation codes indicated above the waveform. (*B*) Background color indicates DART prediction for the recording shown in *A*. (*C*) Class session with varied learning activities (108 min) with human annotation codes indicated. (*D*) Background colors indicate DART predictions for recording in *C*. (*E*) DART prediction, small class (n = 15 students; 98 min). (*F*) DART prediction, large class (n = 287 students; 49 min). (*G*) Examples of DART learning activity footprints from different class sessions: thinking, writing, or clicker response; pair or group discussion; lecture; think-pair-share.

that the dominant class activity (single voice) was detected first, and less-frequent class activities follow (multiple voice, no voice, and other, in that order) (Fig. S1*B*). This ordering emphasized the importance of predicting the common activities correctly while allowing some prediction flexibility for the less-frequent activities.

Next, we optimized the selection parameters that would determine which audio samples were sorted into which activity modes. To accomplish this, we used machine learning, specifically grid search (Fig. S1 *C* and *D*). Grid search is a brute-force method to select the optimal selection parameters for each mode by first evaluating each possible combination of the two selection parameters, the normalized average volume and the normalized average SD, and then choosing the pair of parameter values that yielded the model with the best match to human annotation, defined as the fewest number of errors. This grid search process was conducted three times—once each for single voice, multiple voice, and no voice—to find the optimal parameters for each activity prediction mode. For more details of the development of the DART algorithm, refer to *SI Methods, Development of DART Algorithm with Machine Learning*.

We found that the resulting algorithm, DART, is able to classify each 0.5-s sample of a recording into one of three DART prediction modes: single voice, multiple voice, or no voice. (The final algorithm never categorizes samples as other, probably because the human annotation "other" was assigned only 0.9% of the time to a variety of instances that were difficult to categorize in the pilot data.) Single-voice samples, characterized by one person speaking at a time (e.g., lecture, question/answer, and so forth), were of average volume but high variance. Single voice typically indicated nonactive teaching strategies given that only a single active voice was heard, with all other individuals passively listening. In contrast,

multiple-voice samples, characterized by many people speaking simultaneously (e.g., pair discussions), were of high mean volume and low variance. No-voice samples, characterized by quiet throughout the classroom (e.g., silent writing), were of low mean volume and low variance. As verified by human annotations, multiple and no voice generally indicated active learning because many or all students actively were engaged in a task.

DART Classifies Classroom Noise with High Accuracy. To assess the accuracy of DART, we compared DART's classifications of classroom noise to the human annotations in various ways, both in the original dataset of 45 class sessions collected from 8 instructors and a new, larger dataset comprised of 1,486 class sessions collected from 49 instructors, representing 67 courses taught across 15 community colleges and a 4-y university, a total of 1,720 h of recordings (Table 1). Qualitatively, we saw that DART was able to differentiate between lecture and nonlecture classroom activities. For example, DART predicted a class session that was annotated as 98% lecture with Q/A to be solely single voice (Fig. 1 *A* and *B*) and a class session with varied activities, like silent writing and discussion, to have a variety of modes (Fig. 1 *C* and *D*). DART identification of varied learning activities was robust in both small and large classes (Fig. 1 *E* and *F*). Its predictions reveal that waveform "footprints" are indicative of specific teaching techniques (Fig. 1*G*). For example, the common active learning technique "think-pair-share" actually consists of three distinct activities in response to an instructor's question to the class: first students silently think or write about the answer, then they discuss it in pairs or small groups, and finally some students share their responses individually with the class. A human would annotate these three phases, in order, as silent, discussion, and lecture with Q/A. Similarly, DART assigns no voice (think), multiple voice (pair), and single voice (share) (Fig. 1*G*).

We also assessed DART's accuracy quantitatively by measuring how often DART predictions matched the human annotations. In the original dataset used for optimizing the algorithm, DART classification matched the human annotations 90% of the time across all modes. In comparison, human annotators agreed with each other only 93% of the time, showing that DART was almost as accurate at identifying classroom activities as human annotators were. To see if this high rate of accuracy was retained in a new context, we randomly chose one class session from each of the 67 courses recorded as part of the new, larger dataset, performed human annotation, and compared DART's classifications to the human annotation. We again obtained a very high accuracy of 87%, suggesting that DART can accurately applied to many different classroom contexts.

To further assess DART's ability to discern the presence of activities that may indicate active learning or traditional lecture, we used signal-detection theory to analyze DART's accuracy by mode. In the original dataset, we used signal-detection theory to discriminate for each mode (single voice, multiple voice, and no voice) between correct inclusions (hits) and incorrect exclusions (misses) (21). We also used this method to determine the rates of correct exclusions (correct rejections) and incorrect inclusions (false alarms) for each of the three modes (21). The results are given in Fig. 2. DART correctly identifies nearly all instances of lecture and Q/A as single voice (hit rate = 98.0%) (Fig. 2*A*). In addition, the false-alarm rates for multiple voice and no voice are low (2.3% and <0.1%, respectively) (Fig. 2 *B* and *C*). Combined, these rates mean that most errors over- rather than underestimate lecture, minimizing the potential for falsely indicating the presence of active learning in class sessions.

DART Can Be Used to Perform Large-Scale Analysis of Classrooms. We sought to explore how DART could be used to analyze classroom audio recordings on a larger scale, so we performed DART analysis on the larger dataset consisting of 1,720 h of recordings of 67 courses. DART analysis revealed that in these courses, a range of instructional strategies were represented. Although all courses (n = 67) used single voice a majority of the time, ranging from 69 to 100%, among individual class sessions (n = 1,486), time spent in single voice ranged from 15 to 100% (Fig. 3 *A* and *B*). Within a course, we observed that the time spent in single voice could vary from 15% in one class session to 90% in another class session (Fig. 3*C*). In addition, some instructors that had no multiple or no voice in some class sessions nevertheless spent up to 37% of the time in these categories in another class session within the same course (Fig. 3*D*). This within-course variability highlights the need for a tool that can efficiently analyze every class session of a course.

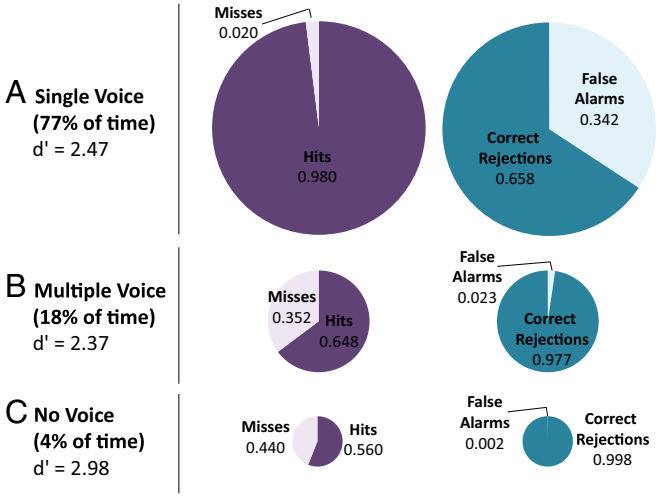

Fig. 2. DART accurately identifies single voice and conservatively estimates multiple and no voice. Recordings from eight instructors from two colleges teaching one course each were used to produce this data. Pie charts on the *Left* show rates for hits (dark purple) and misses (light purple) and on the *Right* show rates for correct rejections (dark teal) and false alarms (light teal) for each DART mode. Both the number in parentheses and the area of the pie chart represent the proportion of each mode present in human annotations. d′, the sensitivity index, is a measurement of the difference between the signal and noise distributions. (*A*) Single voice, (*B*) multiple voice, (*C*) no voice.

To determine the likelihood a student experienced active learning in any one of these courses, we calculated the percentage of class sessions within each course that included any multiple or no voice (<100% single voice). Whereas only 31% of the courses had multiple or no-voice activities in all class sessions, 88% of courses had multiple or no-voice activities in at least half of their class sessions (Fig. 3*D*), indicating that many of these instructors are using active-learning strategies, which is likely unusual among undergraduate STEM instructors.

DART also has the potential to reveal differences in how courses are taught across instructors and courses in particular departments or institutions. In this course sample, we found that the percentage of time spent in multiple or no voice did not vary by instructor gender (n = 36 female, n = 26 male; P = 0.10) but was significantly higher in courses for biology majors (n = 32) than nonbiology majors (n = 35; P = 0.01) (Fig. 3 *D* and *E*).

Discussion

In summary, we have described the development and validation of DART, an analytical tool that uses sound levels to predict classroom activities, as well as results from applying DART to 67 STEM courses. We show that DART is robust to varying class sizes and can determine the presence and quantity of single-voice (e.g., lecture), multiple-voice (e.g., pair or group discussion), or no-voice (e.g., clicker question, thinking, or quiet writing) learning activities with ∼90% accuracy. At this level of accuracy, ease, and time efficiency (∼5 min per 2-h class session), one could analyze and draw broad conclusions about millions of hours of class sessions at periodic intervals over time. Because DART only analyzes sound

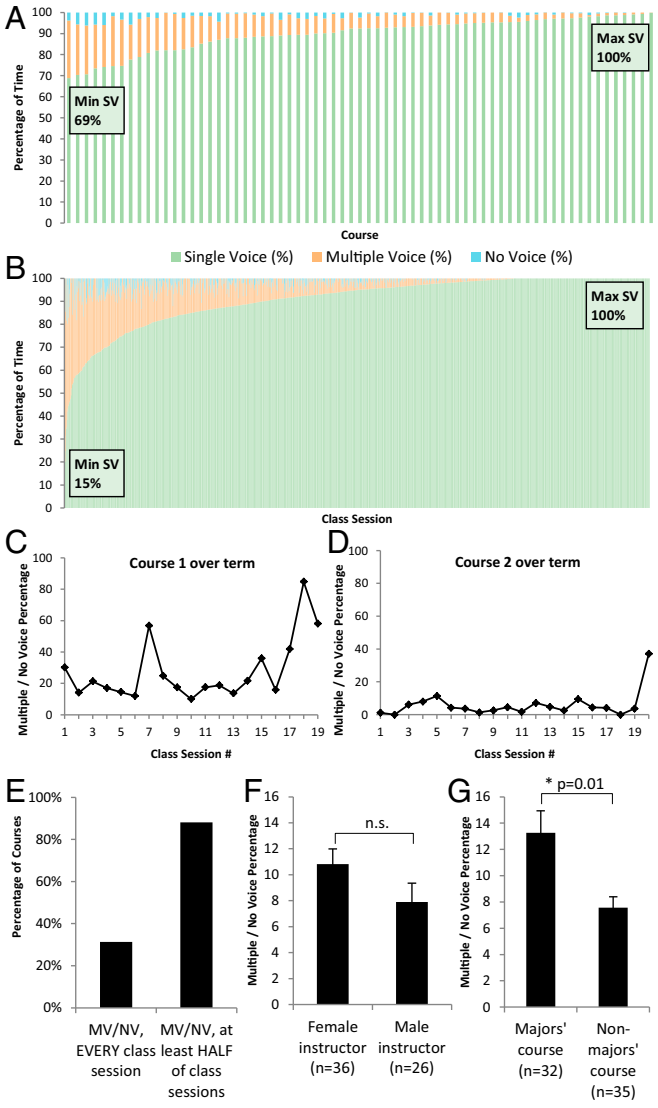

Fig. 3. DART can be used to analyze large numbers of courses. (*A*) Percentage of absolute time spent in single voice (SV), multiple voice (MV), and no voice (NV) for all eligible courses (n = 67). Courses ordered in increasing order of single voice percentage. Boxes indicate minimum and maximum percentages spent in single voice. (*B*) Percentage of absolute time spent in various modes for all class sessions from eligible courses (n = 1,486). Class sessions ordered in increasing order of single voice. Boxes indicate minimum and maximum percentages spent in single voice. (*C* and *D*) Percentage of time spent in multiple or no voice in each class session in time order for two representative courses, course 1 and course 2. (*E*) Proportion of courses where all class sessions have some multiple or no voice (<100% single voice) (*Left*) and where at least half of all class sessions have some multiple or no voice (*Right*). (*F*) Average time spent in multiple or no voice for courses with one female (n = 36) or male (n = 26) instructor (cotaught courses excluded). Error bars represent SE. n.s.: P = 0.10. (*G*) Average time spent in multiple or no voice for biology majors' (n = 32) and nonbiology majors' (n = 35) courses. Error bars represent SE. *P = 0.01.

levels, it protects the anonymity of instructors and students. Furthermore, because DART detected differences in the extent of nonlecture in courses for nonbiology majors versus biology majors, DART additionally promises to reveal differences among other types of courses, instructors, disciplines, and institutions that were previously not feasible for study.

DART is relevant to many educational stakeholders, from individual instructors to institutions and researchers. For individual instructors, DART holds additional promise as a tool for individual instructor professional development. Although previous studies have shown that many STEM faculty aspire to change their teaching (22), detailed observations of classroom videos suggest that instructors overestimate the extent to which they have integrated reformed teaching practices in their classrooms (23). DART could provide instructors with quick and quantitative evidence for instructor self-study. DART can easily identify those class sessions with minimal to no learning activities for students and enable faculty to specifically target how they spend their limited time for pedagogical innovation. For disciplinary programs or departments and the faculty developers that support their teaching efforts, DART could supplement ongoing program assessment, providing insight into the nature of the learning activities happening in different courses with varying student outcomes. It could quickly reveal differences in the teaching strategies used across a department, allowing faculty to have discussions of teaching goals across the curriculum. For institutions, DART may provide a means for describing to prospective students and skeptical parents the added value of a STEM education at their particular campus. Increasingly, parents and students seek information about the added value of an education at particular institution, going beyond academic reputation and research profile, and DART could help institutions make transparent the extent to which their students experience active engagement and their faculty use pedagogically effective teaching methods in their courses. Finally, for federal and private agencies attempting to foster change in STEM faculty teaching practices, DART has the potential to dramatically increase systematic measurement of classroom practices and expand insights being gained from current evaluation approaches through self-report, occasional classroom observations, and time-consuming videotape analyses. In addition, although DART emerged from studies of STEM classrooms, DART also has the potential to address similar inquiries about university classrooms in other subjects or about precollege settings. DART's efficiency could allow for studying correlations between DART's quantitative metrics and a variety of variables associated with STEM courses, including positive outcomes, such as overall learning gains, pass rates, and success in follow-on courses, as well as negative outcomes, such as persistent achievement gaps correlated with student gender or cultural background. It is important to note that DART is not suitable for ranking or evaluating individual instructors, both because of the possibility of errors and because DART is not intended to measure the quality of teaching. Although much research has established that any form of active learning appears to produce higher learning gains than lecture alone (9), it is not known how much or what patterns of active learning may be adequate or optimal for learning.

So, what proportion of STEM instructors in the United States and internationally regularly use teaching strategies beyond lecture? What is the probability that an undergraduate STEM student would have the opportunity to speak, write, or discuss their ideas with peers in every class session? Analyzing classroom noise can quickly and anonymously reveal what is happening in classrooms, making DART a measurement tool with the potential to systematically inventory the presence of active learning across all types of higher education institutions. Given pressing needs to expand and diversify STEM workforces in the United States and beyond, DART can also be used to characterize the extent to which educators are maximizing opportunities for effective STEM learning. Because DART will be available online at dart.sfsu.edu, thousands of instructors, students, or other stakeholders could soon perform DART analyses, opening a variety of new lines of research and inquiry.

Methods

Audio Recording. Each audio recording analyzed as part of this paper was obtained from Sony audio recorder model ICD-PX333. Decibel analysis has also been completed using recordings made on the iPhone Voice Memo App,

as well as live-recording the sound levels in the classroom using the iPhone Decibel 10th App. Instructors were given audio recorders and asked to record every class session of at least one of the courses they were teaching. They were instructed to place the audio recorders at the front of the classroom (e.g., on a lectern) with the microphone pointing in the general direction of students. Before analysis, recordings were trimmed by hand at the beginning and end to exclude noise associated with student arrival and departure.

Instructor Population. Courses analyzed in this study were taught by collaborators on the Talk Matters Project, an advanced collaborative scientific teaching research project. Participating instructors in this project were drawn from two faculty development programs focusing on scientific teaching: Community College Biology Faculty Enhancement through Scientific Teaching (CCB FEST), for community college biology faculty; and Biology Faculty Explorations in Scientific Teaching (Biology FEST), for biology faculty in a single 4-y university. They included part-time, full-time, and tenured/tenure-track faculty teaching a variety of biology courses, including lower- and upper-division courses and courses for biology majors and nonbiology majors. Course enrollments ranged from 4 to 287 students with a median course size of 48 students.

Faculty were recruited in two phases, a pilot phase in Spring 2014 and a large-scale analysis phase in Spring 2015. The research done was a collaboration between dozens of faculty instructors, and as a result there were no human subjects and no need for informed consent. Each instructor who contributed recordings has a letter of collaboration on file with San Francisco State University's Institutional Review Board, which approved the research described in this report in exempt protocols #E14-141a-d. For more information about faculty recruitment and participation rates, see *SI Methods, Participant Recruitment* and Table S2.

Human Annotation of Pilot Data. The development of annotation codes was an iterative process. A team of three people annotated a total of 45 class session recordings split between the 8 instructors in the pilot group. Initially, human annotation was unstructured, and coders were charged to individually listen to audio files, observe audio waveforms, and develop codes that correlated with the types of activities occurring in class sessions. For each new activity lasting more than 15 s, annotators indicated a start time (minutes and seconds) and a code. Emergent codes from all three annotators were compared and collapsed into six categories (lecture with Q/A, discussion, silent, transition, video, and other) (Table S1). The predominant annotation code of this set was lecture with Q/A, which took up 73.5% of the time, followed by discussion at 13.8%. Silent, transition, video, and other each took up less than 5% of the time (Table S1).

One class session from each of the instructors (17% of total annotation) was used to test interrater reliability; all other class sessions were annotated by only one person. The mean Fleiss' κ, a metric appropriate for measuring agreement between multiple annotators for categorical ratings, was $\kappa = 0.567$, indicating moderate to substantial agreement (24). Fleiss' κ was calculated by hand and in Excel. In addition, annotators agreed with each other 93.2% of the time, also showing good interrater reliability.

Measurement of DART's Accuracy.

Pilot data. In the final model used for DART, model prediction accuracy was found to be 89.5% accurate overall on the pilot data. The accuracy was found by calculating the percentage of time the prediction mode matched the human annotation for all 66 annotations (of 45 class sessions; some class sessions were annotated by multiple people). As noted above, by the same metric, the human annotators achieve an accuracy of 93.2%, because human annotators did not always agree. We also analyzed the accuracy of DART with signal-detection theory (Fig. 2). Signal-detection theory calculations of hit, miss, false positive, and correct rejection rates used equations outlined in Stanislaw and Todorov (21) and were calculated in Excel.

For further analyses of DART's accuracy on the pilot group data, see *SI Materials and Methods, Further DART Accuracy Measures* and Fig. S2. Common DART errors are described in Table S3.

Large-scale analysis data. To calculate DART's accuracy on the large-scale analysis data set, one class session from each of this dataset's 67 courses was randomly chosen and annotated by a new two-person team trained in annotation using the previous annotation team's codes and files. We compared how often the human annotations matched DART's predictions, obtaining an accuracy of 87%.

DART Analysis of a Large Set of Courses. Fifty-seven instructors recorded at least one class session in 78 distinct courses. Of these 78 courses, we only included nonlaboratory biology courses where at least 30% of class sessions were recorded. Therefore, we excluded three courses for being laboratories and eight courses for having low numbers of recordings, giving an inclusion rate of 67 of 78 = 85.9%.

DART was used to calculate the time spent in single voice, multiple voice, and no voice for each class session. To compare DART data between different groups of courses, we used *t* tests in Excel on logit-transformed DART data, to correct for using percentage data.

ACKNOWLEDGMENTS. We thank John Coley, Heidi McLaughlin, Sarah Bissonnette, Kristin de Nesnera, the National Science Foundation-funded Community College Biology Faculty Enhancement through Scientific Teaching community, and the Howard Hughes Medical Institute-funded Biology Faculty Enhancement through Scientific Teaching community for insightful discussion and support. We also thank the Center for Computing for Life Sciences at San Francisco State University for extensive support. This work was funded by Howard Hughes Medical Institute Undergraduate Science Education Award 52007556 and National Science Foundation Transforming Undergraduate Education in Science, Technolgy, Engineering, and Mathematics Award DUE-1226361.

1. Arum R, Roksa J (2010) *Academically Adrift: Limited Learning on College Campuses* (Univ of Chicago Press, Chicago).
2. Singer SR, Nielsen NR, Schweingruber HA, eds (2012) *Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering* (National Academies, Washington, DC).
3. Seymour E, Hewitt NM (1997) *Talking About Leaving: Why Undergraduates Leave The Sciences* (Westview Press, Boulder, CO).
4. Graham MJ, Frederick J, Byars-Winston A, Hunter A-B, Handelsman J (2013) Increasing persistence of college students in STEM. *Science* 341(6153):1455–1456.
5. President's Council of Advisors on Science and Technology (2012) *Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics* (Executive Office of the President, Washington, DC).
6. Eddy SL, Hogan KA (2014) Getting under the hood: How and for whom does increasing course structure work? *CBE Life Sci Educ* 13(3):453–468.
7. Hake RR (1998) Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *Am J Phys* 66(1):64–74.
8. Halloun IA, Hestenes D (1985) The initial knowledge state of college physics students. *Am J Phys* 53(11):1043–1055.
9. Freeman S, et al. (2014) Active learning increases student performance in science, engineering, and mathematics. *Proc Natl Acad Sci USA* 111(23):8410–8415.
10. Fairweather J (2008) Linking evidence and promising practices in STEM undergraduate education. *NRC workshop on Evidence on Selected Promising Practices in Undergraduate Science, Technology, Engineering, and Mathematics (STEM) Education* (Board of Science Education, National Research Council, The National Academies, Washington, DC). Available at https://nsf.gov/attachments/117803/public/Xc–Linking_Evidence–Fairweather.pdf. Accessed September 9, 2016.
11. Hora MT, Oleson A, Ferrare JJ (2008) *Teaching Dimensions Observation Protocol (TDOP)* (Wisconsin Center for Education Research, Madison, WI).
12. Sawada D, et al. (2002) Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol. *Sch Sci Math* 102(6):245–253.
13. Smith MK, Jones FHM, Gilbert SL, Wieman CE (2013) The Classroom Observation Protocol for Undergraduate STEM (COPUS): A new instrument to characterize university STEM classroom practices. *CBE Life Sci Educ* 12(4):618–627.
14. Eddy SL, Converse M, Wenderoth MP (2015) PORTAAL: A classroom observation tool assessing evidence-based teaching practices for active learning in large science, technology, engineering, and mathematics classes. *CBE Life Sci Educ* 14(2):14:ar23.
15. Donnelly PJ, et al. (2016) Multi-sensor modeling of teacher instructional segments in live classrooms. *Proceedings of the 18th ACM International Conference on Multimodal Interaction - ICMI 2016* (ACM, New York), pp 177–184.
16. Wang Z, Pan X, Miller KF, Cortina KS (2014) Automatic classification of activities in classroom discourse. *Comput Educ* 78:115–123.
17. Li Y, Dorai C (2006) Instructional video content analysis using audio information. *IEEE Trans Audio Speech Lang Process* 14(6):2264–2274.
18. Donnelly PJ, et al. (2016) Automatic teacher modeling from live classroom audio. *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization - UMAP '16* (ACM, New York), pp 45–53.
19. Brdiczka O, Maisonnasse J, Reignier P (2005) Automatic detection of interaction groups. *Proceedings of the 7th International Conference on Multimodal Interfaces - ICMI '05* (ACM, New York), p 32.
20. Lu L, Zhang H-J, Li SZ (2003) Content-based audio classification and segmentation by using support vector machines. *Multimedia Syst* 8(6):482–492.
21. Stanislaw H, Todorov N (1999) Calculation of signal detection theory measures. *Behav Res Methods Instrum Comput* 31(1):137–149.
22. Savkar V, Lokere J (2010) *Time to Decide: The Ambivalence of the World of Science Toward Education.* (Nature Education, Cambridge, MA).
23. Ebert-May D, et al. (2011) What we say is not what we do: Effective evaluation of faculty professional development programs. *Bioscience* 61(7):550–558.
24. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* 33(1):159–174.

Supporting Information

Owens et al. 10.1073/pnas.1618693114

SI Methods

Development of DART Algorithm with Machine Learning. To develop DART, the human annotations were used to design and optimize a machine-learning algorithm that reports what types of activities are going on in a classroom based on sound waveforms. As stated in the main text, to do this task we applied methods from the field of audio segmentation, which applies machine learning to classify sound into different categories based on statistical characterizations (15). Because some of the human annotation categories yielded waveforms that were statistically similar to each other, we collapsed the eight human annotation categories into four activity prediction modes with distinct waveform profiles: single voice, multiple voice, no voice, and other. Lecture with Q/A and video were aggregated into the mode single voice, discussion and transition were aggregated into the mode multiple voice, silent was assigned to the mode no voice, and other was assigned to the mode other (Table S1).

To prepare the classroom audio recording waveforms for the optimization procedure, we tagged each 0.5-s sample of sound from each recording from the pilot group (640,152 samples in total) with three pieces of data: its label from human annotation (S for single voice, M for multiple voice, or N for no voice), the normalized mean volume of the 15-s window of audio around it, and the normalized SD in that window's volume (Fig. S1*A*). Both the mean volume and the SD of the volume of each sample were normalized with respect to their class session.

Then, to sort the samples into the four prediction modes (single voice, multiple voice, no voice, and other), we used an ensemble of binary decision trees comprised of four nodes connected serially. A binary decision tree is a series of decisions to either sort or not sort a given input into a certain category based on the values of the input. Here, the inputs were the 0.5-s samples of classroom audio, and the sorting decisions were based on each sample's normalized mean volume and SD of the volume. In our tree, each node represented one activity prediction mode, and the nodes for each mode were connected in order of decreasing frequency from the pilot data, so that the dominant class activity (single voice) was detected first, and less-frequent class activities follow (multiple voice, no voice, and other, in that order) (Fig. S1*B*). This ordering emphasized the importance of predicting the common activities correctly while allowing some prediction flexibility for the less-frequent activities.

To optimize the cut-off parameters that would sort the audio samples at each node, we used machine learning, specifically 10-fold stratified cross validation with grid search (Fig. S1 *C* and *D*). This process was repeated three times to find the optimal cut-off parameters for each node.

We first created stratified folds for cross-validation. To create the folds for, for example, the single-voice/nonsingle-voice node, all samples annotated by humans as single voice were equally and randomly divided into 10 single voice groups (S_1–S_{10}), whereas all other samples were equally and randomly divided into 10 nonsingle voice groups (NS_1–NS_{10}) (Fig. S1*C*). A fold consisted of a set of all 20 of these groups (i.e., all of the pilot data) with one pair of groups, for example S_1 and NS_1, designated as the "test set," whereas the remaining 18 groups were designated the "validation set" (Fig. S1*C*). All 10 such folds were created, each with a different pair of groups being designated the test set (Fig. S1*C*).

We then performed a grid search to look for the optimal cut-offs for each mode. Different combinations of mean volume in window and SD of the window volume were tried as cut-off parameters on each of the 10 folds (Fig. S1*D*). For each fold, the error rates (percentage of samples where the computer and human annotations did not match) for the validation set and the test set were calculated. The parameters were first tested at a low resolution (0.5 SD intervals), and the parameters that yielded the lowest validation error were then explored at a higher resolution (0.01 SD intervals). The combination of cut-offs with the lowest average validation error over all 10 folds was selected for DART (Fig. S1*D*). The test error was used as an estimate of generalized model performance. This approach avoided selecting a model that overfit the data and overestimated prediction performance.

As a side note, the final algorithm, DART, never predicts "other." As this mode was marked by humans only 0.9% of the time, the fact that this mode is never used does not greatly affect model accuracy.

Participant Recruitment. Participants were recruited in two phases, a pilot phase in spring 2014 and a large-scale analysis phase in spring 2015.

Pilot group. Data from this group were used to train the DART algorithm. We invited participants to be in the pilot group if they: (*i*) were a community college biology instructor teaching at a quarter-system institution, (*ii*) had attended a weeklong intensive scientific teaching institute, (*iii*) were teaching one or more nonlaboratory courses in spring 2014. Thirteen participants were invited and nine accepted. Data from one participant was excluded because of leaving the course midquarter. Therefore, eight instructors participated for an overall participation rate of 61.5%.

Large-scale analysis group. Data from this group were used to test the effectiveness of the DART algorithm for large-scale analyses. Therefore, we invited a much larger set of participants, all community college or comprehensive university instructors who had previously attended a scientific teaching institute. Seventy-five community college instructors at either semester or quarter institutions were invited; at the time of invitation, it was unknown whether they were teaching in spring 2015. Twenty-eight agreed, for a participation rate of 36.8% (Table S2). These instructors were given a $500 stipend for their collaboration. Forty-two comprehensive university instructors were invited, all of whom were teaching a course in spring 2015. Thirty-one instructors agreed, for a participation rate of 73.8% (Table S2). These instructors were given summer salary or course release in a future term. Although the participation rates for these sets of instructors varied greatly, these differences are anticipated given differences in the method of recruitment and teaching context.

Further DART Accuracy Measures. We analyzed DART's accuracy by prediction mode for the pilot data. First, we quantified how often each human annotation code was classified into each DART prediction mode and vice versa (Fig. S2). Quantification of how often each human annotation code was classified into each DART prediction mode shows that nearly all of the times annotated as lecture with Q/A were assigned correctly to single voice (98.5% of the time) (Fig. S2*A*). For comparison, DART was modestly accurate at identifying discussion as multiple voice (73.9%) and less accurate at identifying silent as no voice (56.0%) (Fig. S2*A*); the latter result is not surprising considering the presence of extraneous classroom sounds or instructor comments or human speech may cause DART to incorrectly classify silent activities as single voice (34.8%). Common DART errors are described in Table S3. In addition, quantification of which human annotations were classified as single voice shows that single voice was primarily composed of lecture with Q/A (86.4%) (Fig. S2*B*). For comparison, multiple voice was mostly composed of discussion (75.2%), and no voice was overwhelmingly composed of silent (91.6%) (Fig. S2*B*).

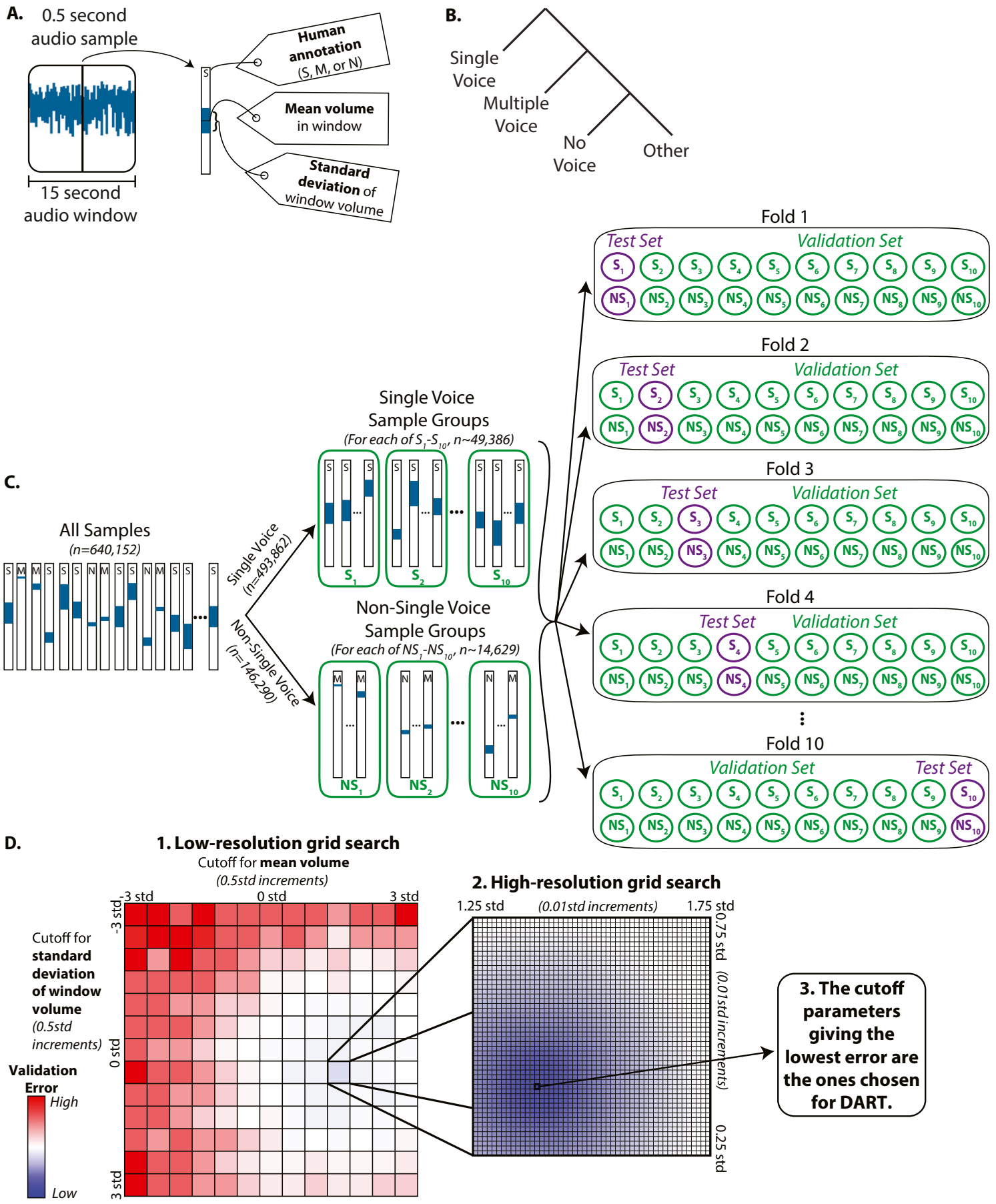

Fig. S1. Using machine learning to optimize the DART algorithm for classifying classroom noise as single voice, multiple voice, or no voice. (*A*) Each 0.5-s sample from each recording from the pilot group was tagged with its label from human annotation (S for single voice, M for multiple voice, or N for no voice), the mean volume of the 15-s window of audio around it, and the SD (std) in that window's volume. Mean volume and SD were normalized with respect to their class session. (*B*) Ensemble of binary decision trees used to classify classroom audio recordings. (*C* and *D*) Optimizing parameters for identifying nature of classroom noise samples using 10-fold stratified cross-validation with grid search. Example below shows the process of optimizing parameters for classifying samples as single voice. (*C*) Samples were sorted into single voice (n = 493,862) and nonsingle voice (n = 146,290) based on human annotation and further randomly and equally divided into 10 groups each (S_1–S_{10} and NS_1–NS_{10}). These groups were recombined 10 times to make 10 folds, each of which contained all of the data. Each fold had a different pair of groups (i.e., S_1/NS_1 or S_2/NS_2) designated as the test set, with all other groups forming the validation set. These folds were all tested using the grid search method that empirically tested all volume and SD parameters and measured error for each of these parameter sets. (*D*) Grid search for choosing cut-off parameters for classifying samples as either belonging to a given annotation category or not. Different combinations of mean volume in window and SD of the window volume were tried as cut-off parameters on each of the 10 folds. The error rates (percentage of samples where the computer and human annotations did not match) for the validation set and the test set were calculated and are represented as heat maps with red showing high-validation error and blue showing low-validation error for each fold. The parameters were first tested at a low resolution (0.5 SD intervals), and the parameters that yielded the lowest validation error were then explored at a higher resolution (0.01 SD intervals). The combination of cut-offs for mean volume and mean SD of volume with the lowest average validation error over all folds was selected for the final version of the DART algorithm. The test error was an estimate of generalized model performance.

A. Human Annotation	*Percentage of the time human annotation code was labeled by DART prediction as the following…*		
	Single Voice	**Multiple Voice**	**No Voice**
Lecture with Question/Answer	**98.5**	1.4	0.0
Video	**88.0**	12.0	0.0
Other	**73.0**	27.0	0.0
Transition	**65.0**	32.0	3.0
Discussion	25.5	**73.9**	0.6
Silent	34.8	9.2	**56.0**

B. DART Prediction	*Percentage of time DART prediction mode was labeled by human annotation as the following…*					
	Lecture with Question/Answer	**Discussion**	**Video**	**Transition**	**Silent**	**Other**
Single Voice	**86.4**	4.2	3.8	3.0	1.8	0.8
Multiple Voice	7.8	**75.2**	3.2	9.0	2.9	1.9
No Voice	0.7	3.4	0.0	4.3	**91.6**	0.0

Fig. S2. DART can accurately identify when lecture with Q/A occurs. (*A*) Percentage of the time each human annotation code was labeled by the DART prediction as single voice, multiple voice, or no voice. Shaded boxes represent the DART prediction mode that was most often assigned to that row's human annotation code. (*B*) Percentage of the time each DART prediction mode was labeled by each human annotation code. Shaded boxes represent the human annotation code that is most represented in that row's DART prediction mode.

Table S1. Description of human annotation codes

Human annotation code	Description of activity	Percentage of time (%)	DART mode
Lecture with Q/A	Instructor or another individual, including a student, speaking to class as a whole	73.5	Single voice
Video	Video recording played to the class as a whole	3.6	Single voice
Discussion	Multiple pairs or groups of students speaking with each other simultaneously	13.8	Multiple voice
Transition	Break during class or students switching between activities	3.8	Multiple voice
Silent	Student silently thinking and/or writing	4.3	No voice
Other	Activities that could not be coded into one of the above categories	0.9	Other

Activities represented by each human annotation code, the percentage of time each human annotation code was present in the pilot data, and the corresponding DART prediction mode for each human annotation code.

Table S2. Instructor participation rates for each phase of DART development

Group	Instructors invited	Instructors participated	Participation rate (%)
Pilot group	13	8	61.5
Large-scale analysis: Community college	76	28	36.8
Large-scale analysis: 4-y university	42	31	73.8

Table S3. Potential DART limitations and coding misclassifications

Classroom situation	Examples of diverse classroom situations	Human annotation code	Expected DART mode based on annotation	Actual DART prediction
(*A*) False-negative for single voice	Recorder is too far away from individual talking	Lecture with Q/A	Single voice	No voice
	Long pauses where instructor is silent (e.g., while writing on the board or working with equipment)	Lecture with Q/A	Single voice	No voice
	Student closest to audio recorder talking inappropriately while instructor is talking	Lecture with Q/A	Single voice	Multiple voice
	Instructor talking during a video with audio or video with loud music	Video	Single voice	Multiple voice
	Significant ambient or outside noise (e.g., loud fan, outside hall conversations, etc.) during single voice activity	Lecture with Q/A	Single voice	Multiple voice
(*B*) False-positive for single voice	Silent work in which an instructor or student speaks extraneously	Silent	No voice	Single voice
	Significant ambient or outside noise (e.g., loud fan, outside hall conversations, etc.) during silent activity	Silent	No voice	Single voice
	Small group or pair discussions in a very small (e.g., under four student) class	Discussion	Multiple voice	Single voice
	Small or pair discussions in which there is delayed/minimal student discussions	Discussion	Multiple voice	No/single voice
(*C*) Errors concerning other modes	Break during class: students remain in classroom	Transition	Multiple voice	Multiple voice
	Students left classroom for small group activity	Silent	No voice	No voice
	Choral response to instructor questions (lasting more than 15s)	Lecture with Q/A	Single voice	Single voice

(*A*) Classroom situations with a human annotation corresponding to single voice but DART prediction of multiple or no voice. (*B*) Classroom situations with a human annotation corresponding to multiple or no voice but DART prediction of single voice. (*C*) Classroom situations where predicted and actual modes match, but vary based on quality of activity.