A deep dive into microphone hardware for recording collaborative group work

Mariah Bradford, Paige Hansen, Ross Beveridge, Nikhil Krishnaswamy, Nathaniel Blanchard
Computer Science Department
Colorado State University
mbrad@colostate.edu

ABSTRACT

Classroom environments are challenging for artificially intelligent agents primarily because classroom noise dilutes the interpretability and usefulness of gathered data. This problem is exacerbated when groups of students participate in collaborative problem solving (CPS). Here, we examine how well six popular microphones capture audio from individual groups. A primary usage of audio data is automatic speech recognition (ASR), therefore we evaluate our recordings by examining the accuracy of downstream ASR using the Google Cloud Platform. We simultaneously captured the audio of all microphones for 11 unique groups of three participants first reading a prepared script, and then participating in a collaborative problem solving exercise. We vary participants, noise conditions, and speech contexts. Transcribed speech was evaluated using word error rate (WER). We find that scripted speech is transcribed with a surprisingly high degree of accuracy across groups (average WER = 0.114, SD = 0.044). However, the CPS task was much more difficult (average WER = 0.570, SD = 0.143). We found most microphones were robust to background noise below a certain threshold, but the AT-Cardioid and ProCon microphones were more robust to higher noise levels. Finally, an analysis of errors revealed that most errors were due to the ASR missing words/phrases, rather than mistranscribing them. We conclude with recommendations based on our observations.

Keywords

Group work, speech recognition, collaborative, spontaneous speech, microphones

1. INTRODUCTION

The ubiquity of technology and computers in the classroom has provided unparalleled opportunities to uncover new insights about how people learn in an increasingly digital environment. Successful analysis of learning settings and outcomes — whether through traditional data mining and information extraction, general machine learning, or targeted applications of artificial intelligence (such as intelligent agents in the classroom) — must be able to handle multiple human inputs like structured questions and answers, free-form text, and, importantly, naturalistic human speech that grounds so much of education.

Over the years, various works have assessed the feasibility of utilizing automatic speech recognition (ASR) systems in various environments [1, 4, 6, 7, 8] — of particular note, in 2015, Blanchard et al. [4] evaluated state-of-the-art speech recognition technology on teachers wearing high-quality headset microphones in classrooms, reporting an average word error rate (WER) of 0.44. While [4] found a wide range of ASR performance across major platforms, modern ASR performance is now largely consistent [6]; thus, the primary focus of our work is on which hardware should be used to collect the data to be processed, from real classroom environments. This work highlights budget and performance trade-offs for researchers and practitioners to consider when deploying hardware for studying group work or collaborative problem solving (CPS) [10].

For our evaluation, we simultaneously recorded group audio from multiple microphones. Groups participated in discussion under various conditions that simulate a classroom environment. We then used automatic speech recognition (ASR) to generate separate transcripts from the audio streams of each microphone and used word error rate as our primary metric for evaluating hardware performance. Finally, we took a deep dive into specific errors in the automatic transcription under various conditions to better anticipate how a downstream system might be influenced by said errors. The result is, to our knowledge, a novel assessment of audio sensor hardware for recording collaborative problem solving in a classroom-based data mining system.

2. METHODOLOGY

2.1 Recording setup

We evaluated six microphones: an Audio-Technica ATR2100x-USB (AT-ATR), an Audio-Technica U891Rb Cardioid Condenser Boundary Microphone (AT-Cardioid), an Audio-Technica U891RbO Omnidirectional Condenser Boundary Microphone (AT-Omni), a Blue-Yeti, an MXL AC-404 ProCon (ProCon), and a Saramonic SmartMic. Microphones were selected to represent a range of technologies and price points. We assumed that, in a real-world context, one microphone would be used for a single group.

The Blue-Yeti was set to medium-low gain and omnidirectional recording, which are empirically-derived settings motivated by the manufacturer guidelines to maximize the volume of the recording while minimizing clipping. The AT-Cardioid, AT-Omni, and AT-ATR were plugged into a mixer, with the gain set following the same procedure as the Blue-Yeti. The Saramonic was plugged into an auxiliary port (AUX). All other microphones were plugged in with USB. Microphones were all placed on one table five feet from participants. Participant and microphone locations were marked to ensure consistency across groups. Using Adobe Audition, all microphones were synchronously recorded on separate tracks.

Table 1 breaks down our participant demographics. Participants were all students in the Computer Science Department at Colorado State University, all over the age of 18. 26 participants considered English as their first language. Other first languages in the participant pool included Gujarati, Korean, Spanish, Turkish, and Urdu. Providing demographic information was optional and recorded anonymously. Participant groups were recruited by request and all personal information was de-identified.

Table 1: Demographics

Gender	Male	20
	Female	12
	Nonbinary	1
Native Language	English	26
	Non-English	5
	Bilingual	2
Age	18-24	28
	25-31	5

Data was gathered in two types of tests: 1) a prespecified, scripted recording, and 2) a collaborative problem solving task, which are described in Section 2.2 and Section 2.3. In each session, participants sat facing toward the microphone array, and performed both tasks (script first, followed by collaborative task). Each test contained variant noise conditions and was run with 11 ($N = 33$) collaborative triads. Groups were split into two conditions: a "noise" condition where generic "classroom noise" was played at 50% volume from a speaker 10 feet away from the microphone array, and a condition without background noise. The "classroom noise" was pre-recorded classroom sounds including indeterminate chatter. The speaker used was a JBL Bluetooth Flip4 with two 8W amplifiers, corresponding to a maximum 12 dB amplification of the source audio, which was intended to simulate ambient background noise of an average classroom.

Across all microphones, 7:50:34 total hours of audio were recorded. Specifically, the noise condition constituted 3:40:05 hours of audio, with the remaining 4:10:29 comprising the non-noise condition. Additionally, the scripted task comprised 2:22:38 hours of recording, while the collaborative task constituted 5:27:56 hours of recording.

Each recording from each microphone was separately processed through Google Cloud Automatic Speech Recognition (ASR). Word error rate (WER) was calculated for the full session. The prewritten script served as the ground truth transcript for the scripted portion of the experiment. For the collaborative tasks, transcripts were manually transcribed by researchers, with a lead researcher subsequently verifying the correctness of all the transcripts.

2.2 Group Script Reading Task

In this condition, the participants read a specified script where each participant played a distinct "role" (teacher, student 1, or student 2). The script was a transcription of a real classroom interaction involving two students and a teacher with no overlapping speech.

2.3 Fibonacci Weights Collaborative Problem Solving Task

The Fibonacci Weights exercise is a collaborative problem solving (CPS) group activity where the participants work together to determine the relative weights of five differently colored cubes. The masses of each cube correspond to the Fibonacci sequence. The participants are given a scale, a 10g calibration weight, the cubes, and a worksheet on which to log the weight of each cube when it is determined. The task invites CPS, leading to explicit and implicit coordination, free-form utterances, and overlapping speech.

2.4 Evaluation Metric

We evaluate microphone performance based on ASR performance for word error rate (WER), given by:

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{H + S + D}$$

where S = the number of substitutions, D = the number of deletions, I = the number of insertions, N = the number of words in the ground truth transcript, and H = the number of successes.

3. RESULTS

3.1 Cross-Task Performance

Figure 1 shows the word error rate (WER) distribution for each microphone across both tasks. A high-level analysis shows that results were relatively consistent across all microphones, with the average WER for every microphone all within 1 standard deviation. The AT-Cardioid showed the best performance, with an average WER of 0.319 ($\sigma = 0.235$). The worst-performing microphone was the Blue-Yeti, with an average WER of 0.365 ($\sigma = 0.278$).

Table 2 provides statistics describing the performance of each microphone across all groups in the script reading task.

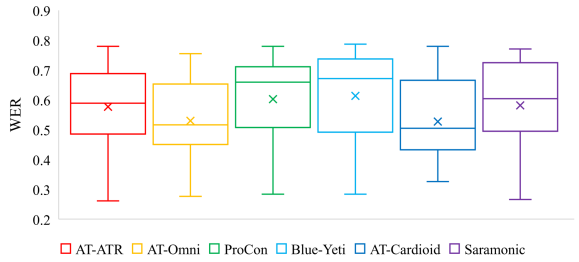

Figure 1: Box-and-whisker plot showing word error rate for each microphone across all groups in the collaborative problem solving task.

Table 2: Descriptive statistics of WER across microphones in the script reading task.

Microphone	μ WER	σ
AT-ATR	0.115	0.051
AT-Cardioid	0.112	0.050
AT-Omni	0.112	0.042
Blue-Yeti	0.117	0.040
ProCon	0.120	0.051
Saramonic	**0.107**	0.035
All	0.114	0.044

Table 3 provides the same in the collaborative problem solving (Fibonacci Weights) exercise.

In the script reading task, all recordings resulted in a low WER within a very small numerical window of each other, with a low standard deviation across groups. The scripted nature of the task allow for clean turn taking, and a review of the recordings confirm there is little to no overlapping speech during the reading of the script. However, this is not how speech in group work typically manifests. Rather, there are overlaps, incomplete sentences, interruptions, etc. Speakers mumble, speak fast, and rely on implicit communication strategies that are not captured in audio recordings, and therefore on information that is not captured in an automatic speech recognition (ASR) platform's underlying language model. Put simply, if a typical speech recognition algorithm encounters actual audio as captured in group work, we can expect the WER to be much higher.

Table 3: Descriptive statistics of WER across microphones in the Fibonacci Weights Task.

Microphone	μ WER	σ
AT-ATR	0.577	0.151
AT-Cardioid	**0.527**	0.137
AT-Omni	0.530	0.128
Blue-Yeti	0.613	0.157
ProCon	0.602	0.150
Saramonic	0.581	0.151
All	0.570	0.143

In the collaborative problem solving task, which contains free-form speech with many overlapping utterances and some disfluencies, we also find that all microphones performed comparably. The clear difference in performance between the scripted condition and collaborative problem solving condition implies free-form speech (e.g., contains natural interruptions, overlaps, sentence fragments, disfluencies, etc.) was the primary difference in ASR performance. In Section 3.2 we quantify these effects. While the Saramonic was (marginally) the best-performing microphone in the scripted condition, but the AT-Cardioid and AT-Omni produced the lowest mean WER for the Fibonacci Weights recordings.

3.2 Word-Level Analysis

Increased word error rates can be due to three primary factors: *deletions*, where a word in the ground truth transcript is omitted; *insertions*, where an additional word not in the ground truth transcript is added; and *substitutions*, where one word in the ground truth transcript is swapped for a different word.

The majority of errors from the ASR were deletions, while substitution and insertion rates stay relatively stable across both tasks. The most commonly deleted words were relatively constant across microphones. This likely means that the deletions were not due to the quality of the recording, but rather the ASR model itself.

A qualitative analysis of the most frequently deleted words exposes the commonality of such terms in the collaborative problem solving task. For instance, the most commonly deleted word weighted by frequency is the demonstrative pronoun "this". "This" was dropped 41% of the time and correctly transcribed 53% of the time. In the collaborative problem solving activity we can expect this word to be used frequently, which contributes to the high weighted deletion rate. Likewise, other most commonly deleted words, weighted by frequency, include common particles like "so" or "oh,", and acknowledgments like "yeah" or "okay." However, some words that in context are important and contentful, such as "10" or "20" (referring to the masses of the weights), are also frequently deleted. See supplemental material for the complete analysis of these commonly missed words.

3.3 Ablation of noise

Classroom environments are by nature noisy [5, 9]. Thus, we ran additional experiments where we increase the noise testing, we had four groups of two participants read the preprepared script from Section 2.2 multiple times. With each trial we increased the level of the background noise played by the speaker (a 10% increase was equivalent to a 1.6W increase in speaker power). The net effect of this trial was to control for participants (vocal profile) and ground truth transcript (speech content), while varying background noise level.

Figure 2 shows the results of the noise ablation test. The effect of background noise up to 50% (9 dB) was negligible across all microphones, confirming the results in Section 3.1. When the background noise level is greater than 50% (9 dB), we see the word error rate start to increase significantly with each 10% increase in background noise. With the background noise played at maximum speaker volume, with certain microphones (e.g., Blue-Yeti, Saramonic), we see word error rates start to approach the word error rates demonstrated in the free-from Fibonacci Weight collaborative task, even though the participants in this experiment

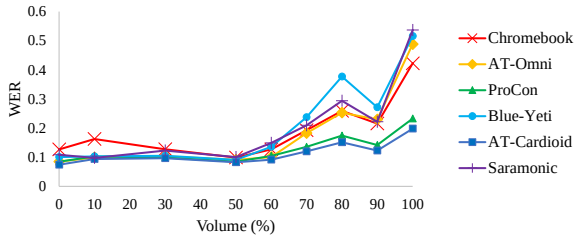

Figure 2: WER vs. background noise volume.

Table 4: Cost of microphones

Microphone	Cost
Chromebook*	$0.00
Saramonic	$25.00
AT-ATR	$99.00
ProCon	$99.95
Blue-Yeti	$129.99
AT-Cardioid	$319.00
AT-Omni	$319.00

*Baseline microphone

were reading the "well-behaved" prepared script. When the speaker is at maximum volume, this condition should better approximate the noise conditions of a real classroom [5, 9].

3.3.1 *Correlation with price*

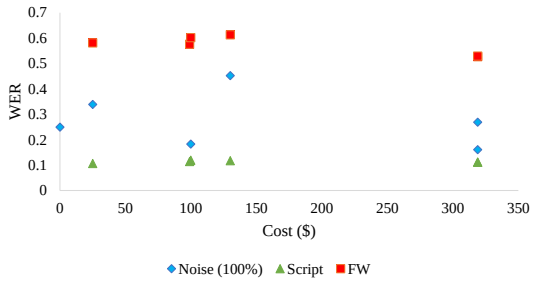

Figure 3: Microphone cost plotted against group mean WER in scripted, Fibonacci weights, and 100% noise ablation conditions.

Finally, Figure 3 shows the mean word error rates across groups in three of the conditions: the script reading, the collaborative task and the noise ablation test session where noise was set to 100%. Together, these three plots show that there is no clear relationship between the cost of the microphone hardware and the ASR performance over associated recordings. Note, the Chromebook served as a "zero dollar" baseline because all other microphones required a laptop to connect to; the microphone integrated with a laptop that students already have access to would entail zero dollars in additional purchases for an in-class recording system.

4. LIMITATIONS AND DISCUSSION

We assessed the quality of a number of microphones by considering how downstream tasks like automatic speech recognition (ASR) performed. At a high level, we found microphones were largely comparable; however, diving deeper, even when the speech is canonically well-behaved the influence of background noise causes dramatic differences in word error rate across microphones at different rates. We found that in the most extreme noise cases, the AT-Cardioid and the ProCon were the most robust to noise, with word error rates (WER) below 0.2 (the no-noise baseline performance was about 0.1). Our overall pick is the ProCon, because it too is highly robust to noise, the price is a third of the AT-Cardioid, and WER is arguably comparable.

Independent of the microphone hardware, off-the-shelf automatic speech recognition performs well (WER less than 0.2) when the recorded speech is canonically well-behaved (e.g., few overlaps, disfluencies, simultaneous speech). However, when participating in group work, there *are* explicit overlaps, disfluencies, and simultaneous speech. Indeed, the presence of these may indicate healthy, productive, educational group dynamics. Still, when recorded speech contains these artifacts, word error rate soars to surprising levels, not dissimilar to the word error rates reported by Blanchard et al. [4] in their study of live teacher speech. Interestingly, they also evaluated scripted and unscripted speech, and found scripted speech error rates were similar to unscripted. Our results showcase that the technology driving automatic speech recognition has improved substantially, since scripted speech is now transcribed with a far lower WER — just not for the kind of naturalistic speech used in collaborative group-work.

There are several limitations of our study. First, we did not verify our conclusions with alternative ASRs (e.g., solutions from Microsoft or Amazon). Several microphones have settings like gain which we hand-tuned, but could be further experimented with. There are, of course, a plethora of microphones we did not include in our test. Finally, we have done our best to replicate classroom environments in a lab setting, but, for now, our evaluation is limited to the lab.

Since most of the word error rate is due to deleted words, a future analysis should consider if it matters if those words are dropped. Taking the transcribed text and evaluating the performance of a further downstream task such as abstract meaning representation (AMR) parsing [2, 3, 11] on it, and on the ground truth transcript, would demonstrate if the dropped words are resulting in significant lost information. However, downstream tasks are likely much more context/project dependent, and being able to predict possible errors in the transcription by understanding limitations of the data gathering process can aid in appropriately designing the downstream tasks by, for example, accounting for the likelihood of missing stop words.

5. ACKNOWLEDGEMENTS

We thank Rosy Southwell, Isaac Courchesne-Owades, and Yongxin Liu for their contributions to data collection and the processing pipeline. This work was supported in part by the National Science Foundation (NSF) under grant number DRL 1559731 to Colorado State University.

6. REFERENCES

[1] M. H. Asyrofi, F. Thung, D. Lo, and L. Jiang. Crossasr: Efficient differential testing of automatic speech recognition via text-to-speech. In *2020 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 640–650. IEEE, 2020.

[2] L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider. Abstract meaning representation (amr) 1.0 specification. In *Parsing on Freebase from Question-Answer Pairs. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle: ACL*, pages 1533–1544, 2012.

[3] L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186, 2013.

[4] N. Blanchard, M. Brady, A. M. Olney, M. Glaus, X. Sun, M. Nystrand, B. Samei, S. Kelly, and S. D'Mello. A study of automatic speech recognition in noisy classroom environments for automated dialog analysis. In *International conference on artificial intelligence in education*, pages 23–33. Springer, 2015.

[5] J. E. Dockrell and B. M. Shield. Acoustical barriers in classrooms: The impact of noise on performance in the classroom. *British Educational Research Journal*, 32(3):509–525, 2006.

[6] F. Filippidou and L. Moussiades. Benchmarking of IBM, Google and Wit Automatic Speech Recognition Systems. *Artificial Intelligence Applications and Innovations*, 583:73–82, May 2020.

[7] V. Këpuska and G. Bohouta. Comparing speech recognition systems (microsoft api, google api and cmu sphinx). *Int. J. Eng. Res. Appl*, 7(03):20–24, 2017.

[8] F. Morbini, K. Audhkhasi, K. Sagae, R. Artstein, D. Can, P. Georgiou, S. Narayanan, A. Leuski, and D. Traum. Which asr should i choose for my dialogue system? In *Proceedings of the SIGDIAL 2013 Conference*, pages 394–403, 2013.

[9] B. M. Shield and J. E. Dockrell. The effects of environmental and classroom noise on the academic attainments of primary school children. *The Journal of the Acoustical Society of America*, 123(1):133–144, 2008.

[10] C. Sun, V. J. Shute, A. Stewart, J. Yonehiro, N. Duran, and S. D'Mello. Towards a generalized competency model of collaborative problem solving. *Computers & Education*, 143:103672, 2020. Publisher: Elsevier.

[11] C. Wang, N. Xue, and S. Pradhan. A transition-based algorithm for amr parsing. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 366–375, 2015.

APPENDIX

A. WORD-LEVEL ANALYSIS OF ERRORS

Table 5: Top 20 deletions weighted by rate of occurrence

AT-ATR	AT-C	AT-O	Blue-Yeti	ProCon	Saram.
this	this	this	this	this	this
so	so	so	so	the	so
the	the	the	the	so	the
is	is	is	is	is	is
yeah	yeah	yeah	yeah	yeah	yeah
we	i	we	we	we	we
i	we	i	i	i	10
10	one	one	one	10	i
one	10	10	10	one	one
that	that	that	that	that	that
it	it	it	it	it	it
to	okay	okay	to	okay	to
okay	to	to	okay	to	okay
20	a	20	and	and	and
it's	and	a	a	20	a
a	oh	and	20	a	20
and	20	think	oh	be	be
be	think	it's	think	think	like
oh	be	oh	be	it's	it's
think	it's	be	like	oh	oh

Since substitution and insertion rates stay relatively stable across both tasks, it seems clear that in the free-from collaborative problem solving task, the nature of the speech captured in the recordings leads to more deletions. Therefore it becomes more important to understand the nature of the high deletion rates in the collaborative problem solving task. Are there differences in the precise words that are deleted from the recordings made by each microphone? Are these words, when deleted, ones that are likely to have a negative effect on the performance of downstream analysis tasks using the transcription, e.g., parsing or classification?

Table 5 shows the top 20 most frequently deleted words for each microphone, weighted by the overall occurrence of that word in the ground truth, human-generated transcript. Data here was taken from recordings of the collaborative problem solving task only, to analyze the nature of the words being dropped in a group work environment.

Words w are ranked according to $\sum_{t \in T} D_t(w) \times \frac{\sum_{t \in T} C_t(w)}{\sum_{t \in T} N_t}$, where t is a transcript $\in T$ the set of all transcripts, $D_t(w)$ is the number of times w was deleted in transcript t, $C_t(w)$ is the total count of w in the ground truth transcript t, and N_t is the total number of words in the ground truth transcript t. Since this was the free-form activity and each ground truth transcript was different for each group performing the activity, we sum counts over all transcripts.

The most commonly deleted words were relatively constant across microphones. This likely means that the deletions were not due to the quality of the recording, but rather the ASR model itself. In fact, the only words that appear in the top 20 most deleted words that do *not* appear in the top-20 list for every microphone are "it's," which did not appear in the top 20 of the Blue-Yeti, "think," which did not appear in the top 20 of the Saramonic, and "like," which *only* appeared in the top 20 of the Saramonic and Blue-Yeti.

A qualitative analysis of the most frequently deleted words exposes the commonality of such terms in the collaborative problem solving task. For instance, the most commonly deleted word weighted by frequency is the demonstrative pronoun "this". "This" was dropped 41% of the time and correctly transcribed 53% of the time. In the collaborative problem solving activity we can expect this word to be used frequently, which contributes to the high weighted deletion rate. Another commonly deleted word is "one," a common continuation of "this" as in the bigram "this one," as might be used to refer to an object in a situated context. Likewise, other most commonly deleted words, weighted by frequency, include common particles like "so" or "oh,", and acknowledgments like "yeah" or "okay." However, some words that in context are important and contentful, such as "10" or "20" (referring to the masses of the weights), are also frequently deleted.

In reality, most of the deleted words are stop words, indicating that downstream processes should be robust. However, there are some task specific content words that may impede downstream tasks.