THESIS

# BALANCING SPEED AND PRECISION: A COMPARATIVE STUDY OF ASR SYSTEMS IN MULTIMODAL COLLABORATIVE ENVIRONMENTS

Submitted by

Corbyn Terpstra

Department of Computer Science

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Summer 2025

Master's Committee:

Advisor: Nathaniel Blanchard

Sudipto Ghosh
Anne Cleary

# ABSTRACT

BALANCING SPEED AND PRECISION: A COMPARATIVE STUDY OF ASR SYSTEMS IN MULTIMODAL COLLABORATIVE ENVIRONMENTS

Automatic Speech Recognition (ASR) systems are increasingly critical for analyzing collaborative problem-solving (CPS) tasks, yet their segmentation and transcription accuracy in dynamic, multimodal environments remain underexplored. This study evaluates the performance of OpenAI's Whisper (Large, Medium, Turbo) and Vosk ASR systems in segmenting and transcribing collaborative dialogue, with a focus on implications for CPS annotation workflows. Leveraging a dataset of triads solving a multimodal task—comprising oracle (human-segmented), Google-segmented, and Whisper-segmented audio—we measure transcription accuracy via Word Error Rate (WER) and assess segmentation alignment through start time deviations, segment length ratios, and pause dynamics. Results reveal that while Whisper Turbo achieves the lowest overall WER (52.5%), its semantic segmentation strategy fragments coherent CPS moves, complicating annotation. Conversely, Vosk's pause-based approach under-segments rapid exchanges, obscuring interruptions and cross-talk. The study highlights a fundamental tension: Whisper prioritizes intent preservation at the cost of over-segmentation, while Vosk and Google ASR sacrifice nuance for efficiency. Annotation fidelity is further eroded by ASR-induced errors, including insertions (e.g., hallucinated phrases during silence) and temporal misalignments. These findings underscore the need for hybrid segmentation strategies and adaptive annotation frameworks that explicitly account for ASR limitations. Practical recommendations are proposed, including model-specific post-processing and context-aware annotation tools. By bridging technical evaluation with real-world application, this work advances the design of ASR systems tailored for collaborative environments, ensuring their outputs align with the complexities of human interaction.

ACKNOWLEDGEMENTS

I would like to thank Dr. Nathaniel Blanchard for being such an amazing friend and advisor during my graduate studies and final years of my undergraduate degree. Without him, I wouldn't have started my graduate degree in the first place. I would also like to thank Dr. Sudipto Ghosh and Dr. Anne Cleary for agreeing to be part of my committee, without them this would not be possible. Finally I would like to thank my friends and family for believing in me during this past year and specifically month, and pushing me to keep going when I was struggling to keep my head up.

# DEDICATION

*I would like to dedicate this thesis to my partner Ainsley, who has always placed unbounded faith in me.*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# Chapter 1

# Introduction

Automatic Speech Recognition (ASR) systems have become indispensable in modern technology, enabling seamless interactions with virtual assistants like Siri, real-time transcription services, and accessibility tools for the hearing impaired. These systems excel in controlled, one-on-one settings where clear audio inputs and turn-based dialogues simplify segmentation and transcription. However, their performance in collaborative group environments—marked by overlapping speech, variable microphone setups, and ambient noise—remains a critical unsolved challenge. Collaborative problem-solving (CPS) tasks, which integrate speech with gestures, gaze, and object manipulation, demand ASR systems capable of parsing dynamic, multimodal communication. Yet, the very features that make human collaboration rich and adaptive—interruptions, rapid turn-taking, and disfluencies—pose significant hurdles for automated speech processing.

Prior research has illuminated the limitations of ASR in group settings. For instance, Google ASR segments speech at pauses, often truncating mid-sentence or merging adjacent utterances, while Whisper prioritizes semantic coherence, over-segmenting continuous thoughts into disjointed fragments. These discrepancies degrade annotation fidelity in CPS workflows, where precise utterance boundaries are vital for labeling behaviors like interruptions or shared reasoning. Foundational work by [1] compared oracle (human-segmented) transcripts with automated outputs, revealing systemic mismatches: Whisper's over-segmentation fractured coherent CPS moves, while Google ASR's under-segmentation obscured cross-talk. However, this work focused narrowly on broad comparisons between systems, leaving unexplored performance variations across ASR architectures (e.g., model size, speed optimizations) and their implications for real-world deployment.

This study bridges these gaps by conducting a granular evaluation of three Whisper variants (Large, Medium, Turbo) and Vosk, an open-source alternative, in a multimodal CPS dataset. We analyze transcription accuracy through Word Error Rate (WER) and segmentation alignment via

start time deviations, segment length ratios, and pause dynamics. Our findings reveal a fundamental tension: Whisper Turbo achieves the lowest WER (52.5%) but fragments discourse through over-segmentation, while Vosk's pause-based approach under-segments rapid exchanges, erasing critical CPS indicators. For example, Whisper's semantic strategy split hypotheses at conjunctions (e.g., "But if we try this…"), disrupting logical flow, whereas Vosk merged interruptions into single segments, conflating labels like "interrupts" and "confirms understanding." These challenges are compounded by cross-talk, where all models prioritized dominant speakers, omitting 41% of non-dominant participant words.

By quantifying these trade-offs, we advance the design of ASR systems tailored for collaborative environments. Our results underscore the need for hybrid segmentation strategies that harmonize semantic coherence with pause dynamics, as well as adaptive annotation frameworks resilient to ASR-induced errors. This work not only refines technical evaluations of ASR performance but also provides actionable insights for developers and researchers aiming to deploy speech technologies in classrooms, workplaces, and other dynamic group settings—ensuring that the voices of collaboration are not just heard, but understood.

## 1.1 The Importance of Automatic Speech Recognition

Automatic Speech Recognition (ASR) is a transformative technology that converts spoken language into written text, enabling machines to interpret and respond to human speech. Its applications span virtual assistants (e.g., Siri, Alexa), real-time transcription services, accessibility tools for the hearing impaired, and educational technologies. ASR systems are particularly critical in collaborative environments, where understanding multimodal communication—such as dialogue intertwined with physical actions—is essential for modeling group dynamics. However, the accuracy and usability of ASR systems depend on their ability to segment speech into meaningful units and transcribe them correctly, tasks complicated by natural speech phenomena like disfluencies, overlapping dialogue, and pauses.

## 1.2 Evaluating ASR Performance: Metrics and Challenges

The performance of ASR systems is typically measured using Word Error Rate (WER), a benchmark metric calculated as:

$$\text{WER} = \frac{\text{Substitutions } (S) + \text{Deletions } (D) + \text{Insertions } (I)}{\text{Number of Reference Words } (N)} \times 100\% \tag{1.1}$$

Substitutions (S) occur when the ASR system replaces a correct word in the reference transcript with an incorrect word (e.g., transcribing "blue" as "green" in "Place the blue block"). Deletions (D) refer to the omission of words present in the reference transcript (e.g., skipping "30" in "The weight is 30 grams"). Insertions (I) involve adding words absent in the reference (e.g., inserting "definitely" into "This block feels heavier"). N denotes the total number of words in the reference transcript, serving as the denominator for WER calculations.

WER quantifies discrepancies between ASR outputs and human-annotated "oracle" transcripts. While WER provides a broad measure of transcription accuracy, it does not fully capture nuances in segmentation, such as how systems handle pauses, interruptions, or overlapping speech.

Segmentation parameters—including pause times between segments, segment lengths, and pause thresholds before segmentation—are equally critical. For instance, systems may split speech at pauses exceeding a specific duration (e.g., 300ms), but inconsistent thresholds can lead to over-segmentation (breaking coherent thoughts into fragments) or under-segmentation (lumping multiple utterances together). These errors propagate into downstream tasks, such as annotating collaborative problem-solving (CPS) behaviors, where precise utterance boundaries are vital for labeling interruptions or confirmations.

Pause times in this study refer to the duration of silence (measured in milliseconds or seconds) that ASR systems use as thresholds to segment continuous speech into discrete utterances. For example, Google ASR employs a default pause threshold of 300 ms to split audio at silences, while Vosk relies on conservative voice activity detection (VAD) to identify pauses. These pauses are critical for segmentation alignment metrics, including pre-segmentation pause thresholds (delays

before transcription starts), inter-utterance pauses (silence between segments), and intra-utterance pauses (silence within segments that may trigger over-segmentation).

## 1.3 Prior Research and Gaps

Recent studies [1] highlight the challenges of relying on automatic segmentation for annotation tasks. Their work compared oracle transcripts with outputs from Google ASR and OpenAI's Whisper, revealing significant inconsistencies:

Whisper often over-segmented speech, creating more utterances than oracle transcripts.

Google ASR under-segmented speech, merging utterances and omitting segments mistaken for noise.

Annotations derived from automatic segments frequently misaligned with oracle labels, obscuring critical CPS indicators like interruptions.

These findings underscore the need to evaluate how segmentation strategies and transcription errors impact both WER and semantic coherence. However, prior work focused on broad comparisons between systems, leaving unexplored the performance variations across different versions of the same ASR architecture or the role of open-source alternatives.

## 1.4 Advancing the Research: A Deeper Dive into ASR Systems

This thesis expands on existing research by conducting a granular analysis of three Whisper models (Large, Medium, Turbo) and Vosk, an open-source ASR toolkit. Each Whisper variant offers trade-offs. Whisper's Large model offers high accuracy, however it is computationally intensive. Whisper's Medium model offers balanced performance for real-time applications. Whisper's Turbo model is optimized for speed, which is attained through focusing solely on the English language and sacrificing accuracy. Finally, the Vosk toolkit offers lightweight, offline modes, and customization to its models that are unavailable with Whisper.

By measuring WER, pause times, segment lengths, and pre-segmentation pause thresholds, this study intends to investigate the following questions. How do segmentation strategies differ

across Whisper's model sizes? Does Vosk's offline processing offer advantages in noisy or low-resource environments? How do segmentation errors correlate with WER and annotation fidelity in collaborative tasks?

## 1.5 Related Work

The performance of Automatic Speech Recognition (ASR) systems in collaborative problem-solving (CPS) environments intersects with research on multimodal interaction, CPS competency frameworks, and advancements in ASR technology. Below, we synthesize scholarly works that contextualize and inform this study.

The gap between oracle data and real-world data has been identified previously [2]. Other works have pointed out the need to move away from oracle transcriptions in pursuit of AI applications for real-world use cases [3, 4]. The use of automatic segmentation of speech for modeling tasks is becoming increasingly widespread [5, 6, 7].

What is distinct with this work is that here we focus our analysis on the fine details of ASR systems and their efficacy in CPS tasks. For example, [4] refused to use human transcriptions in a multimodal sentiment challenge because such transcripts were not true to real-world contexts; however, they did not comment on how the labeling of sentiment might change were those annotations done on automatically extracted data.

Here, we explicitly focus on that challenge. We explore the implications of segmentation and transcription methods when annotating CPS for groups. CPS is a critical skill used in many areas of life [8], and AI agents for group settings will need some way of representing group state. Work has been done to model CPS at the utterance level [9, 10]. The framework defined by [11] captures CPS at three levels and identifies specific actions that indicate different types of collaborative actions and their impact on group state, as fragmented or merged utterances can disrupt these processes—a challenge highlighted in the present study. Similarly, PISA 2015 findings revealed global deficiencies in CPS skills among students, emphasizing the need for tools that accurately capture collaborative dialogue dynamics [12].

Recent ASR advancements, such as Transformer-based models (e.g., Whisper, Conformer), have improved robustness in noisy settings but struggle with speaker overlap and domain adaptation [13]. For instance, Whisper's semantic segmentation prioritizes intent preservation but fragments rapid exchanges, while Vosk's pause-based method under-segments cross-talk—trade-offs quantified in this study [14][13]. Federated learning and transfer learning techniques show promise for customizing ASR to collaborative tasks but require extensive labeled data, a limitation echoed in the current dataset's demographic bias (80% male, 60% Caucasian) [13][12].

While prior work establishes ASR's potential in CPS contexts, gaps remain in evaluating model-specific trade-offs (e.g., Whisper variants, Vosk) and linking segmentation errors to annotation workflows. This study bridges these gaps by quantifying the impact of ASR architectures on CPS label fidelity and proposing adaptive frameworks for real-world deployment.

# Chapter 2

# Methodology

This chapter details the experimental design, dataset, and analytical procedures used to evaluate the performance of OpenAI's Whisper (Large, Medium, Turbo) and Vosk ASR systems. The study builds on the framework established by [1] to measure transcription accuracy, segmentation alignment, and pause dynamics in three segmentation groups: Oracle, Google-segmented, and Whisper-segmented.

### 2.0.1 Dataset and Task Design

The data set comprises 10 triads (3 participants per group) collaborating on a multimodal problem solving task involving dialogue and manipulation of physical objects. Participants weighed five blocks that had weights following the Fibonacci sequence, and were asked to place them on a worksheet and decide as a group the weight of each block. A sixth "mystery" block was then given, and without the use of the scale, the group had to conclude the weight of the block. Finally, the researcher asked what the weight of the next block would be, if there were to be another block.

Key features include 170 minutes of audiovisual recordings, manually segmented and transcribed utterances with precise timestamps, and automatically segmented audio using Google ASR and Whisper (from the original study), subsequently human-transcribed for ground truth comparison. The demographic of the dataset contains participants from a pool of those aged 19-35 and following an 80% male, 60% Caucasian split. The single microphone setup simulates an environment where individual microphones are impractical.

### 2.0.2 Preprocessing and Segmentation

Human annotators manually segmented audio into utterances using temporal transcription methods. This means that each utterance was manually segmented to the millisecond and then transcribed. These manual, human completed segmentation and transcriptions are labeled as Oracle

segmentation and Oracle transcription. Further segmentation methods were applied to the full raw audio, including Google ASR segmentation and Whisper ASR segmentation using the Whisper Large-v2 model. Google ASR uses a pause based approach, with a default of 300 ms, whereas Whisper ASR uses a semantic segmentation technique which prioritizes sentence boundaries. The three segmentation groups were generated and had a total number of segments reaching 2,873 Oracle segments, 1,822 Google segments, and 3,013 Whisper segments.

## 2.1 ASR Systems and Configurations

Four different systems were evaluated to capture diverse design philosophies. Three different Whisper ASR models were used alongside an offline capable model known as Vosk. The Large-v3, Medium, and Turbo models were used for Whisper with their default settings, and Vosk had the 0.3.45 model running with default configurations as well. Each audio file, Oracle segments, Google segments, Whisper segments, and full audio were processed independently by all aforementioned ASR systems.

### 2.1.1 Annotation and Label Transfer

In the initial research, expert annotators labeled utterances using the CPS framework from [11], coding behaviors such as interrupts, confirms understanding, and proposes solution. Interrupts are defined as Overlapping speech disrupting another participant, confirms understanding is when there are verbal affirmations such as "yeah, that makes sense", and proposes solution is when a participant offers a hypothesis or tests a hypothesis, such as "maybe this block is 20 grams". Labels from the oracle segments were then mapped to the automated segments using temporal overlap and majority voting measures. Temporal overlap assigns labels if more than 50% of the automated segment overlapped with the Oracle segment. Majority voting takes merged segments and inherits the labels if at least one overlapping Oracle segment had the label present.

## 2.2 Evaluation Metrics

### 2.2.1 Transcription Accuracy

Word Error Rate (WER) was computed using Python's jiwer library to compare ASR outputs against human-transcribed segments. The following formula was used for the computation of WER.

$$\text{WER} = \frac{\text{Substitutions } (S) + \text{Deletions } (D) + \text{Insertions } (I)}{\text{Number of Reference Words } (N)} \times 100\% \tag{2.1}$$

Substitutions (S) occur when the ASR system replaces a word in the reference transcript with an incorrect word. For example, if the reference transcript said "Place the blue block on the scale.", and the ASR Output were to be "Place the green block on the scale.". There would be 1 substitution, swapping "blue" for "green". Deletions (D) occur when the ASR system omits a word present in the reference transcript. For example if the reference transcript was "The weight is *30* grams.", and the ASR output for the same transcript were "The weight is grams.", there would be 1 deletion. Insertions (I) occur when the ASR system adds a word not present in the reference transcript. For example, the reference transcript of "This block feels heavier.", and the ASR output of "This block definitely feels heavier." has 1 insertion, being the word "definitely".

Using this definition, and Python's jiwer package, WER, substitutions, deletions, and insertions were gathered at the segment level. The rate of substitutions, insertions, and deletions were also then found on the group and model level. This is to keep in line with the initial research methodology and data collection.

### 2.2.2 Segmentation Alignment

Segmentation alignment was assessed through start time deviations, segment length, and pause dynamics. Start time deviations are the millisecond difference between the beginning of the audio segment and the ASR transcription. Segment length is the discrepancies between utterance duration, measured at the millisecond level. Pause dynamics were analyzed by measuring pause

durations between sub-segments generated when ASR systems split oracle segments, as well as during full audio breakdown with the models in the current research.

## 2.3 Experimental Workflow

Raw audio files were segmented by Oracle, Google ASR and Whisper ASR, and then transcribed by all systems. WER was computed using Python's jiwer library and custom scripts were written to analyze start time deviations, segment lengths, and pauses between segments. over 50% of transcripts were manually reviewed to identify error patterns, and group-wise comparisons assessed WER differences across models and segmentation methods.

### 2.3.1 Integration of Prior Work

This methodology extends the foundational work of [1], which first compared oracle and automated segmentation. Key advancements include an expanded ASR evaluation through use of Vosk, and Whisper models Turbo and Medium, multimodal annotation by explicitly linking segmentation errors to CPS label inaccuracies, and error typology by classifying ASR errors and their impact on annotation. By unifying preprocessing, segmentation, and evaluation, this framework bridges the gap between theoretical ASR research and practical CPS annotation workflows.

# Chapter 3

# Results

## 3.1 Initial Results

The foundational study [1] compared Oracle transcripts with automated outputs from Google ASR and Whisper, revealing critical insights into segmentation biases and their impact on CPS annotation.

### 3.1.1 Segmentation Discrepancies

Whisper produced significantly more segments (3,013 utterances) when compared to Google ASR (1,822) or Oracle annotators (2,873) (Table 3.1). For example, Group 5 had 406 Whisper segments vs. 237 Oracle segments, reflecting Whisper's over segmentation of continuous speech. Annotations derived from automated segments frequently misaligned with oracle labels. For instance, Google ASR merged three distinct utterances ("Weren't those both thirty?", "No, this is twenty," and "Twenty and then. . . ") into one segment, conflating "interrupts" and "confirms understanding" labels (Figure 3.1).

| Group | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Whisper | 297 | 201 | 391 | 293 | 406 | 278 | 311 | 354 | 136 | 346 |
| Google | 139 | 151 | 254 | 128 | 146 | 153 | 380 | 235 | 90 | 146 |
| Oracle | 229 | 207 | 337 | 195 | 237 | 227 | 590 | 338 | 134 | 379 |

**Table 3.1:** # of utterances per group determined by each segmentation method. Totals: Whisper - 3,013 utterances; Google - 1,822 utterances; Oracle - 2,873.

### 3.1.2 Error Patterns

Google ASR and Whisper exhibited similar overall WER but had divergent error profiles. Where Google had an average WER of 0.573, with an elevated substitution rate of 0.259 and deletion rate of 0.132, Whisper had an average WER of 0.542 and an elevated insertion rate of

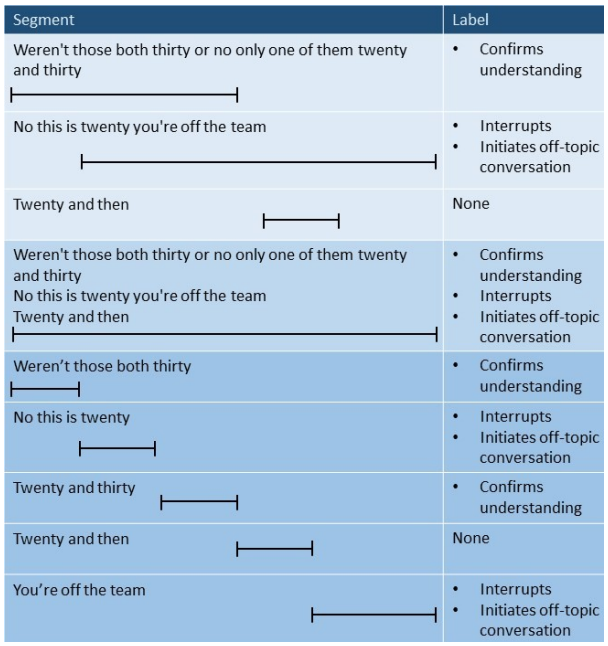


**Figure 3.1:** Overlap between oracle (top), Google (middle), and Whisper (bottom) segments. Right column shows the CPS indicator annotated for each utterance.

0.294. This is due to Whisper often generating and inserting phrases such as "Thank you" during silence (Table 3.2).

### 3.1.3 Qualitative Observances

Whisper's semantic strategy fragmented CPS moves by splitting at conjunctions, while Google ASR's pause-based method merged interruptions. Annotators reported difficulties coding automated segments due to missing visual context.

## 3.2 Current Results: Expanded Analysis of Whisper and Vosk

Building on prior findings, this study evaluated three Whisper variants, being Large-v3, Medium, and Turbo, as well as Vosk across segmentation groups, yielding granular insights into speed-precision trade-offs.

| | Google | | | | Whisper | | | |
|---|---|---|---|---|---|---|---|---|
| Group | WER | Sub. rate | Del. rate | Ins. rate | WER | Sub. rate | Del. rate | Ins. rate |
| 1 | 0.571 | 0.252 | 0.113 | 0.206 | 0.534 | 0.193 | 0.045 | 0.296 |
| 2 | 0.459 | 0.211 | 0.128 | 0.120 | 0.416 | 0.177 | 0.040 | 0.200 |
| 3 | 0.539 | 0.236 | 0.117 | 0.186 | 0.527 | 0.177 | 0.047 | 0.303 |
| 4 | 0.529 | 0.267 | 0.154 | 0.170 | 0.572 | 0.201 | 0.040 | 0.332 |
| 5 | 0.631 | 0.262 | 0.173 | 0.195 | 0.581 | 0.175 | 0.060 | 0.346 |
| 6 | 0.581 | 0.252 | 0.077 | 0.252 | 0.525 | 0.191 | 0.041 | 0.293 |
| 7 | 0.610 | 0.260 | 0.155 | 0.196 | 0.650 | 0.209 | 0.064 | 0.377 |
| 8 | 0.532 | 0.259 | 0.137 | 0.137 | 0.486 | 0.200 | 0.048 | 0.238 |
| 9 | 0.571 | 0.274 | 0.180 | 0.118 | 0.514 | 0.229 | 0.084 | 0.202 |
| 10 | 0.645 | 0.306 | 0.087 | 0.252 | 0.612 | 0.202 | 0.054 | 0.356 |
| Average | 0.573 | 0.259 | 0.132 | 0.183 | 0.542 | 0.195 | 0.052 | 0.294 |
| SD | 0.063 | 0.018 | 0.038 | 0.079 | 0.086 | 0.018 | 0.033 | 0.105 |

**Table 3.2:** WER, substitution rate, deletion rate, and insertion rate by group during initial research.

| Model | Google | Whisper | Oracle | Overall |
|---|---|---|---|---|
| Large | 0.392 | 0.744 | 0.456 | 0.531 |
| Medium | 0.405 | 0.767 | 0.460 | 0.544 |
| Turbo | 0.385 | 0.742 | 0.447 | 0.525 |
| Vosk | 0.665 | 0.865 | 0.694 | 0.741 |

**Table 3.3:** Average WER by Model and Segmentation Group Based on Segmented Audio

### 3.2.1 Transcription Accuracy

#### Pre-segmented Audio

Whisper Turbo achieved the lowest WER of 0.525, outperforming Vosk and all other Whisper models (Table 3.3). Performance of each model varied by segmentation method. On Google segments, Turbo reached an average WER of 0.385, comparatively Turbo reached a WER of 0.742 on Whisper segmented audio.

For substitutions and deletions, Vosk has the highest reported rate, being at or above 0.800 on average. Once again, Whisper showed a tendency to insert phrases into silence, raising the Turbo model to have the highest insertion rate of 0.294 (Table 3.4).

| | Whisper: Large | | | | Whisper: Medium | | | |
|---|---|---|---|---|---|---|---|---|
| | WER | Sub. rate | Del. rate | Ins. rate | WER | Sub. rate | Del. rate | Ins. rate |
| Google | 0.392 | 0.678 | 0.574 | 0.292 | 0.405 | 0.677 | 0.559 | 0.292 |
| Whisper | 0.744 | 0.689 | 0.540 | 0.243 | 0.767 | 0.675 | 0.547 | 0.262 |
| Oracle | 0.456 | 0.639 | 0.593 | 0.183 | 0.460 | 0.648 | 0.585 | 0.184 |
| | **Whisper: Turbo** | | | | **Vosk** | | | |
| | WER | Sub. rate | Del. rate | Ins. rate | WER | Sub. rate | Del. rate | Ins. rate |
| Google | 0.385 | 0.686 | 0.553 | 0.294 | 0.665 | 0.886 | 0.840 | 0.200 |
| Whisper | 0.742 | 0.715 | 0.525 | 0.258 | 0.865 | 0.843 | 0.750 | 0.147 |
| Oracle | 0.447 | 0.668 | 0.571 | 0.196 | 0.694 | 0.865 | 0.876 | 0.128 |

**Table 3.4:** Average WER, substitution rate, deletion rate, and insertion rate based on segmented audio.

### Full Audio

Full audio segmentation is the result of running the model on the original audio video file as was used in the initial research for segmentation purposes. This was done to see the difference between hybrid approaches like using Google segmentation with Whisper transcription. During current research, all three whisper models created more segments than previously, with Turbo almost doubling the utterance count (Table 3.5. Meanwhile, Vosk showed an utterance number close to the original Google segmentation, showing the similarities in the pause based-approach when segmenting the files. For WER, all Whisper models performed at a lower caliber when transcribing the full audio file on Oracle segments. Vosk achieved a lower WER compared to the pre-segmented audio files in all segmentation types with Oracle transcriptions. All models showed a 0.2 or greater decrease in WER when moving to full audio and using Whisper segmentation and Oracle transcription. Whisper Medium is the only Whisper model that improved on full audio when compared to the Google segmented Oracle transcripts (Table 3.6).

| Group | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Large | 324 | 269 | 561 | 380 | 465 | 344 | 894 | 548 | 183 | 498 |
| Medium | 332 | 192 | 363 | 276 | 318 | 289 | 811 | 436 | 172 | 384 |
| Turbo | 399 | 326 | 576 | 361 | 529 | 451 | 1002 | 690 | 262 | 494 |
| Vosk | 162 | 132 | 237 | 139 | 150 | 145 | 425 | 226 | 95 | 150 |

**Table 3.5:** # of utterances per group determined by each segmentation method. Totals: Large: 4,466 utterances; Medium: 3,573 utterances; Turbo: 5,090 utterances; and Vosk: 1,861 utterances.

| Model | Google | Whisper | Oracle | Overall |
|---|---|---|---|---|
| Large | 0.465 | 0.467 | 0.511 | 0.481 |
| Medium | 0.373 | 0.378 | 0.472 | 0.408 |
| Turbo | 0.493 | 0.488 | 0.522 | 0.501 |
| Vosk | 0.624 | 0.624 | 0.681 | 0.643 |

**Table 3.6:** Average WER by Model and Segmentation Group Based on Full Audio

| | Group | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Google | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Avg $\pm$ SD |
| Large | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | $0.000 \pm 0.000$ |
| Medium | 0.000 | 0.000 | 0.003 | 0.000 | 0.000 | 0.000 | 0.004 | 0.000 | 0.000 | 0.000 | $0.000 \pm 0.001$ |
| Turbo | 0.000 | 0.000 | 0.007 | 0.000 | 0.006 | 0.005 | 0.008 | 0.000 | 0.003 | 0.000 | $0.002 \pm 0.003$ |
| Vosk | 0.447 | 0.779 | 0.702 | 0.863 | 0.562 | 0.486 | 0.912 | 0.649 | 0.796 | 0.469 | $0.667 \pm 0.161$ |
| Whisper | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Avg $\pm$ SD |
| Large | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | $0.000 \pm 0.000$ |
| Medium | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | $0.000 \pm 0.000$ |
| Turbo | 0.002 | 0.000 | 0.005 | 0.001 | 0.000 | 0.004 | 0.001 | 0.001 | 0.000 | 0.000 | $0.001 \pm 0.002$ |
| Vosk | 0.486 | 0.635 | 0.577 | 0.485 | 0.256 | 0.632 | 0.946 | 0.669 | 0.908 | 0.384 | $0.598 \pm 0.203$ |
| Oracle | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Avg $\pm$ SD |
| Large | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | $0.000 \pm 0.000$ |
| Medium | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | $0.000 \pm 0.000$ |
| Turbo | 0.000 | 0.000 | 0.003 | 0.000 | 0.004 | 0.007 | 0.000 | 0.000 | 0.003 | 0.000 | $0.002 \pm 0.002$ |
| Vosk | 0.382 | 0.656 | 0.545 | 0.600 | 0.383 | 0.409 | 0.484 | 0.550 | 0.520 | 0.236 | $0.476 \pm 0.118$ |

**Table 3.7:** Average Pause Before Transcription in Seconds

### 3.2.2 Segmentation Alignment

Whisper showed nearly perfect alignment with the highest average start time deviation of 8 ms, while Vosk introduced delays from 0.236 seconds in Oracle group 10, to 0.946 seconds in Whisper group 7, due to Vosk's conservative Voice Activity Detection (VAD) (Table 3.7).

Vosk's pause-based segmentation strategy produced shorter segments than the Whisper models with the exception of the Medium Whisper model on Google segments. Vosk had average lengths from 0.898 seconds to 2.735 seconds, while no Whisper model ever had an average segment length below 1 second (Table 3.8).

| | Group | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Google | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Avg ± SD |
| Large | 2.545 | 4.363 | 4.871 | 3.253 | 3.722 | 3.398 | 3.395 | 3.139 | 4.872 | 2.117 | 3.568 ± 0.869 |
| Medium | 2.216 | 2.394 | 2.317 | 2.121 | 1.869 | 2.297 | 2.417 | 2.266 | 1.915 | 1.878 | 2.169 ± 0.201 |
| Turbo | 2.055 | 2.412 | 3.196 | 2.508 | 2.098 | 2.207 | 2.958 | 2.702 | 2.953 | 1.947 | 2.504 ± 0.412 |
| Vosk | 2.320 | 1.775 | 2.312 | 2.238 | 2.659 | 2.407 | 1.692 | 2.028 | 1.731 | 2.735 | 2.190 ± 0.354 |
| Whisper | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Avg ± SD |
| Large | 3.344 | 2.538 | 3.726 | 1.969 | 2.757 | 3.941 | 3.801 | 3.278 | 2.164 | 2.696 | 3.021 ± 0.661 |
| Medium | 2.168 | 2.320 | 2.445 | 1.867 | 1.681 | 2.241 | 2.498 | 2.141 | 1.945 | 1.900 | 2.121 ± 0.253 |
| Turbo | 2.711 | 2.743 | 3.018 | 1.939 | 2.371 | 2.789 | 3.525 | 2.661 | 1.886 | 2.461 | 2.610 ± 0.460 |
| Vosk | 1.240 | 1.596 | 1.514 | 0.898 | 0.910 | 1.176 | 1.502 | 1.338 | 1.099 | 1.180 | 1.245 ± 0.231 |
| Oracle | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Avg ± SD |
| Large | 2.410 | 4.350 | 4.287 | 2.986 | 2.865 | 3.017 | 2.115 | 2.884 | 3.619 | 1.526 | 3.006 ± 0.848 |
| Medium | 1.969 | 2.136 | 2.374 | 1.812 | 1.869 | 2.061 | 2.085 | 2.171 | 1.940 | 2.020 | 2.044 ± 0.154 |
| Turbo | 1.853 | 2.988 | 3.110 | 2.024 | 2.012 | 2.086 | 3.196 | 2.525 | 2.568 | 2.441 | 2.480 ± 0.464 |
| Vosk | 1.983 | 1.590 | 2.060 | 1.921 | 2.247 | 2.038 | 1.381 | 1.875 | 1.466 | 1.381 | 1.794 ± 0.297 |

**Table 3.8:** Average Length of Utterances in Seconds

## 3.3 Pause Dynamics

Whisper's pre-segmented audio had on average the longest between utterance times, with Turbo exhibiting the longest average pause across Whisper models at 0.929 seconds. Vosk also showed the longest average pause of 2.082 seconds on the Whisper segments, with an average of around 1.2 second for both Oracle and Google segments. Vosk also had the highest average pause time across all models and segmentation types (Table 3.9). Within the utterances, any pause at a conjunction triggered Whisper's segmentation process which fractured hypothesis testing utterances. All models prioritized the loudest speaker, transcribing mainly only the person who was loudest in an utterance. If there were multiple people speaking at the same time, there was very low chance that any model would pick up both speakers and transcribe what they said individually.

## 3.4 Synthesis of Initial and Current results

Both studies confirm Whisper's tendency to over-segment and Vosk/Google's under-segmentation. Current results quantify speed-precision trade-offs with Turbo as a model running 8 times as fast as the Large model, and introduce Vosk's noise vulnerability.

| | Group | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Google | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Avg $\pm$ SD |
| Large | 0.292 | 0.250 | 0.291 | 0.243 | 0.234 | 0.293 | 0.200 | 0.190 | 0.246 | 0.184 | $0.242 \pm 0.039$ |
| Medium | 0.105 | 0.057 | 0.064 | 0.087 | 0.107 | 0.037 | 0.131 | 0.048 | 0.084 | 0.055 | $0.078 \pm 0.029$ |
| Turbo | 0.361 | 0.262 | 0.581 | 0.327 | 0.219 | 0.480 | 0.152 | 0.216 | 0.411 | 0.191 | $0.320 \pm 0.132$ |
| Vosk | 1.195 | 1.002 | 1.492 | 1.105 | 1.297 | 1.192 | 1.282 | 1.173 | 1.372 | 0.960 | $1.207 \pm 0.154$ |
| Whisper | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Avg $\pm$ SD |
| Large | 0.568 | 1.303 | 1.509 | 1.029 | 0.586 | 0.798 | 0.371 | 0.745 | 1.584 | 0.452 | $0.895 \pm 0.417$ |
| Medium | 0.624 | 0.606 | 2.124 | 0.576 | 0.318 | 0.517 | 0.372 | 0.648 | 1.260 | 0.175 | $0.722 \pm 0.541$ |
| Turbo | 0.551 | 1.420 | 1.487 | 0.811 | 0.386 | 0.893 | 0.948 | 0.346 | 1.690 | 0.755 | $0.929 \pm 0.442$ |
| Vosk | 1.733 | 1.794 | 2.656 | 1.437 | 2.021 | 2.425 | 1.913 | 2.140 | 3.456 | 1.240 | $2.082 \pm 0.608$ |
| Oracle | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Avg $\pm$ SD |
| Large | 0.324 | 0.529 | 0.293 | 0.239 | 0.212 | 0.297 | 0.224 | 0.223 | 0.220 | 0.178 | $0.274 \pm 0.095$ |
| Medium | 0.145 | 0.293 | 0.175 | 0.046 | 0.086 | 0.045 | 0.115 | 0.060 | 0.090 | 0.209 | $0.126 \pm 0.076$ |
| Turbo | 0.290 | 0.570 | 0.564 | 0.243 | 0.323 | 0.353 | 0.130 | 0.211 | 0.121 | 0.191 | $0.300 \pm 0.152$ |
| Vosk | 1.243 | 1.554 | 1.465 | 1.325 | 1.154 | 1.113 | 1.235 | 1.188 | 1.380 | 0.802 | $1.246 \pm 0.198$ |

**Table 3.9:** Average Time Between Utterances in Seconds

# Chapter 4

# Discussion

## 4.1 Initial Insights: Lessons from Prior Work

The foundational study [1] revealed systemic challenges in deploying ASR systems for collaborative annotation tasks, particularly the tension between segmentation strategies and CPS label fidelity. Tensions such as fragmenting continuous speech into disjointed utterances disrupted CPS coherence. For example, splitting "But if we try this…" at the conjunction “but” severed hypothesis-testing logic, complicating annotation of "proposes solution." Merging rapid exchanges into single segments also conflated labels like "interrupts" and "confirms understanding" reducing annotation specificity.

Whisper’s tendency to generate phrases like "Thank you" during silence introduced semantic noise, while Google ASR’s substitutions altered the task-critical details, changing the semantic meaning. Annotators also reported frustration when coding automated segments, as the missing visual context obscured the speaker’s intent.

The study underscored the need for hybrid approaches that balance semantic and pause-based segmentation while accounting for real-world acoustic noise.

## 4.2 Current Insights: Expanding the Framework

Building on the initial findings, the current study dissects the interplay between model architecture, segmentation strategy, and annotation workflows, offering nuanced recommendations for ASR deployment in collaborative settings.

### 4.2.1 Model-Specific Trade-Offs

Whisper Turbo achieved the lowest WER on the segmented audio with a 0.525, as well as having the fastest processing speed of all Whisper models. However, Turbo did fragment many seg-

ments on short pauses and conjunctions. Vosk, while incredibly fast and offline capable, deleted far too many words for it to be viable in any context when using audio segments. If we look specifically at group 7 Vosk had an average deletion rate of over 0.8 across all segmentation methods, showing just how Vosk's model is unable to handle fragmented noise and multiple speakers. Whisper Medium balanced speed and accuracy, being twice as fast compared to Whisper large and having the lowest average WER across all models when looking at full length audio.

### 4.2.2 Segmentation Strategy Impacts

Whisper's semantic based approach preserved the intent in monologues but fractured much of the collaborative dialogue. As it was difficult for the models to differentiate when someone new was speaking, all Whisper models began to segment during short pauses after conjunctions believing that a new speaker was interrupting. Google ASR's pause-based method, while efficient, merged many of the interruptions and cross talk, over-complicating many of the utterances and confusing the ASR systems.

Some recommendations that could prove to aid in the automatic annotation of group work, as well as lower the WER are to have hybrid segmentation, increasing error mitigation, and have cost aware deployment of models. Hybrid segmentation involves merging Whisper's semantic output with adaptive pause thresholds, causing the conjunction based segment splits to fuse back together. This could also integrate visual cues from the original audio visual data to resolve ambiguities in the multimodal pipelines. Error mitigation would deploy context-aware tools to flag ASR hallucinations, such as the phantom "Thank you" insertions that are prevalent in Whisper models, or pairing Vosk with a noise-robust VAD model such as WebRTC's VAD or Silero VAD to reduce the amount of deletions. Finally, cost aware development would prioritize which model to use based on the situation present. For example, in educational assessment, the Medium model from whisper performs both quickly and accurately, and if you are in a resource constrained setting, Vosk can be used to leverage WER for offline capability.

To mitigate Whisper's tendency to insert hallucinated phrases (e.g., "Thank you") during silence and reduce over-segmentation, Google ASR's pause-based segmentation could be leveraged as a preprocessing filter. By first applying Google ASR to identify and remove non-speech sections (e.g., prolonged silences or background noise), Whisper could then process only the filtered, speech-dense segments. This hybrid approach would capitalize on Google's efficiency in isolating viable speech via pause thresholds (300 ms default), while preserving Whisper's semantic coherence for transcription. For instance, Google's segmentation could prune silent intervals where Whisper erroneously inserts text, thereby lowering insertion errors and improving annotation fidelity in collaborative dialogues.

## 4.3 Synthesis of Insights

Both studies confirm that ASR systems designed for one-on-one interactions falter in group settings due to acoustic complexity and segmentation biases. However, with the current results we have quantified the trade-off between speed and accuracy is not necessarily there, with Vosk being the fastest but having a vulnerability to high noise environments.

## 4.4 Limitations

In typical one-on-one ASR applications, each speaker is recorded on a dedicated microphone channel, ensuring clean audio separation and minimizing crosstalk. However, in this study, all three participants shared a single microphone, resulting in a mixed audio signal with overlapping speech, background noise, and inconsistent vocal clarity. This setup inherently increases the base noise floor and complicates speaker identity. Consequently, ASR systems faced challenges in accurately segmenting and transcribing utterances, particularly during interruptions or rapid turn-taking. The muddied audio likely inflated substitution and deletion errors across all models, as systems struggled to distinguish speech from noise or separate simultaneous voices. This also affects Whisper's semantic segmentation and Vosk's pause-based approach, which could have underperformed due to the lack of clear acoustic boundaries. Had each participant worn a lapel microphone (or used

a multi-channel recorder), post-processing could isolate individual speakers, reducing noise and improving WER.

The demographic diversity within the dataset used was 80% male and 60% Caucasian, potentially causing model performance bias. The diversity of vocal pitch, accent, and speech pattern are difficult to increase when the dataset itself is limited in who has participated. Models may have performed better or worse when participants had stronger variance in vocal pitch or accent, but it is difficult to tell with the current results.

While ASR systems automate segmentation and transcription, their evaluation depends on high-quality oracle transcripts, which are labor-intensive to produce. Training annotators to transcribe consistently, such as when handling filler words like "um" or partial words, requires months of practice, and even expert transcribers may disagree on ambiguous segments. When transcribers aren't properly trained, inaccuracies within oracle transcripts artificially inflate ASR WER, as the "ground truth" itself is flawed. There is also the fact of human fatigue to look at. Over time, transcribers may inconsistently apply rules, introducing noise into the reference data due to fatigue. Since every audio segment that was created as the result of Google ASR or Whisper ASR was also transcribed manually to prevent downstream segmentation errors or inconsistent results, there is a chance that fatigue muddied some of the Oracle transcriptions.

After manually going through over 50% of the transcriptions for each model and segmentation group, it was clear to tell that the WER that was being received from each of the utterances was flawed. There is no inherent better transcription method, however it should be noted that temporal transcriptions are not the way that Whisper transcribes audio. Thus, even when the same information is captured, because the annotators in the original paper transcribed temporally, there is a higher WER. Whisper mainly transcribes by speaker, where if there are two people speaking at the same time, it will separate the two speakers into their own segments, whereas the transcription method that was used has words appear in the transcript as they are said, no matter the context or flow of sentence.

## 4.5 Future Directions

For future directions, I believe that a new dataset could be used to verify the results gotten from this paper. The current dataset shows inherent bias with the participating members of the experiment, and could be expanded to fit a wider, more generalizable audience. This same experiment could also be run with members who all speak another language, so we can test if different language model ASRs are producing the same relative results as what we see here.

Further analysis could also make use of other ASR models, as the tip of the iceberg has only been scratched with Google, Whisper, and Vosk. While Google and Whisper have extremely prominent ASR models, there are many more that have yet to be tested exhaustively on group dynamics. This could even expand to using Whisper's smaller models, such as tiny or small.

Another experimentation route could combine visual input with the audio segmentation to more accurately segment data by combining segments that are part of one sentence. This could be done with the WTD, as there is visual footage already incorporated within the dataset.

A direction that could potentially develop the understanding of the ASR models further is changing the oracle transcription methodology. Instead of having temporal transcripts, having speaker based transcription or dominant speaker transcription and comparing the word error rates. It would also be interesting to see if giving the utterances more of a buffer in the Oracle segmentation would increase or decrease the WER. This avenue could also lead toward changing the default settings in the ASR systems away from their default to see if there are better features for group work compared to solo work.

# Chapter 5

# Conclusion

This thesis explored the performance of automatic speech recognition (ASR) systems in segmenting and transcribing collaborative problem-solving (CPS) interactions, with a focus on their implications for group dynamic ASR. By evaluating OpenAI's Whisper (Large, Medium, Turbo) and Vosk models across Oracle, Google-segmented, and Whisper-segmented data, this work illuminates critical trade-offs between transcription spped, accuracy, segmentation strategies, and real-world usability. Below, we summarize key contributions, reflect on broader implications, acknowledge limitations, and outline future directions.

## 5.1 Summary of Findings

Whisper Medium achieved the lowest word error rate during full length audio transcription on Google segmentation, excelling in noisy environments, while also having decent processing speed. Whisper Medium was also had the highest WER when transcribing pre-segmented audio using the Whisper segmentation. Whisper Large and Turbo models performed better, with the exception of Whisper segmentation, on the pre-segmented audio when compared to the full audio tracks. Vosk performed at a lower standard than all Whisper models regardless of full length audio or pre-segmented audio.

Whisper exclusively had a buffer of less that 10 ms before it began listening for audio during transcription. Vosk made use of a buffer of, on average, at least 200 ms at the start of utterances where it would delete any audio during that time, increasing WER. During full audio transcription, Vosk's buffer grew to be at least 2 full seconds per each group, with the longest buffer being over 10 seconds.

Whisper's segmentation method proved to be the least effective when it came to measuring WER, with Google segmentation providing the lowest WER across all models. The only point

that outperformed Google segmented audio was full audio transcription using Google segmented oracle transcripts using the Turbo model.

## 5.2 Final Remarks

The quest to harmonize ASR capabilities with the complexities of human collaboration is both a technical and conceptual challenge. While Whisper and Vosk offer powerful tools for automating speech preprocessing, their segmentation strategies and error profiles necessitate careful alignment with annotation goals. This thesis underscores that there is no universal solution—instead, practitioners must strategically balance model size, segmentation logic, and task requirements. By embracing hybrid approaches and evolving annotation frameworks to address ASR realities, we can pave the way for AI systems that genuinely enhance collaborative learning and problem-solving.

In the dynamic interplay of speech, silence, and collaboration, ASR systems are both bridge and barrier. This work charts a path toward turning these barriers into bridges—ensuring that the voices of collaboration are not just heard, but understood.

# References


[1] Corbyn Terpstra, Ibrahim Khebour, Mariah Bradford, Brett Wisniewski, Nikhil Krishnaswamy, and Nathaniel Blanchard. How good is automatic segmentation as a multimodal discourse annotation aid? *arXiv preprint arXiv:2305.17350*, 2023.

[2] Nathaniel Blanchard, Patrick J. Donnelly, Andrew M. Olney, Borhan Samei, Brooke Ward, Xiaoyi Sun, Sean Kelly, Martin Nystrand, and Sidney K. D'Mello. Semi-automatic detection of teacher questions from human-transcripts of audio in live classrooms. *International Educational Data Mining Society*, 2016.

[3] Fabrizio Morbini, Kartik Audhkhasi, Kenji Sagae, Ron Artstein, Doğan Can, Panayiotis Georgiou, Shrikanth Narayanan, Anton Leuski, and David Traum. Which asr should i choose for my dialogue system? In *Proceedings of the SIGDIAL 2013 Conference*, pages 394–403, 2013.

[4] Nathaniel Blanchard, Daniel Moreira, Aparna Bharati, and Walter Scheirer. Getting the subtext without the text: Scalable multimodal sentiment classification from visual and acoustic modalities. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 1–10, 2018.

[5] Mariah Bradford, Paige Hansen, J. Ross Beveridge, Nikhil Krishnaswamy, and Nathaniel Blanchard. A deep dive into microphone hardware for recording collaborative group work. In *Proceedings of the 15th International Conference on Educational Data Mining*, page 588, 2022.

[6] Mariah Bradford, Paige Hansen, Kenneth Lai, Richard Brutti, Rachel Dickler, Leanne Hirshfield, James Pustejovsky, Nathaniel Blanchard, and Nikhil Krishnaswamy. Challenges and opportunities in annotating a multimodal collaborative problem-solving task. In *Interdisciplinary Approaches to Getting AI Experts and Education Stakeholders Talking Workshop, AIEd*, 2022.

[7] Iliana Castillon, Videep Venkatesha, Hannah VanderHoeven, Mariah Bradford, Nikhil Krishnaswamy, and Nathaniel Blanchard. Multimodal features for group dynamic-aware agents. In *Interdisciplinary Approaches to Getting AI Experts and Education Stakeholders Talking Workshop, AIEd*, 2022.

[8] Arthur C. Graesser, Stephen M. Fiore, Samuel Greiff, Jessica Andrews-Todd, Peter W. Foltz, and Friedrich W. Hesse. Advancing the Science of Collaborative Problem Solving. *Psychological Science in the Public Interest*, 19(2):59–92, nov 2018. Publisher: SAGE Publications Inc.

[9] Angela E. B. Stewart, Zachary Keirn, and Sidney K. D'Mello. Multimodal modeling of collaborative problem-solving facets in triads. *User Modeling and User-Adapted Interaction*, 31(4):713–751, September 2021.

[10] Mariah Bradford, Ibrahim Khebour, Nathaniel Blanchard, and Nikhil Krishnaswamy. Automatic detection of collaborative states in small groups using multimodal features. In *Proceedings of the 124th International Conference on Artificial Intelligence in Education*, 2023.

[11] Chen Sun, Valerie J. Shute, Angela Stewart, Jade Yonehiro, Nicholas Duran, and Sidney D'Mello. Towards a generalized competency model of collaborative problem solving. *Computers & Education*, 143:103672, 2020. Publisher: Elsevier.

[12] Xuyan Tang, Yan Liu, and Marina Milner-Bolotin. Investigating student collaborative problem-solving competency and science achievement with multilevel modeling: Findings from pisa 2015. *PLOS ONE*, 18(12):1–17, 12 2023.

[13] Hamza Kheddar, Mustapha Hemis, and Yassine Himeur. Automatic speech recognition using advanced deep learning approaches: A survey. *Information Fusion*, 109:102422, September 2024.

[14] Hannah VanderHoeven, Brady Bhalla, Ibrahim Khebour, Austin Youngren, Videep Venkatesha, Mariah Bradford, Jack Fitzgerald, Carlos Mabrey, Jingxuan Tu, Yifan Zhu, Kenneth Lai,

Changsoo Jung, James Pustejovsky, and Nikhil Krishnaswamy. Trace: Real-time multimodal common ground tracking in situated collaborative dialogues, 2025.