How Good is Automatic Segmentation as a Multimodal Discourse Annotation Aid?

Corbyn Terpstra, Ibrahim Khebour, Mariah Bradford, Brett Wisniewski,
Nikhil Krishnaswamy and Nathaniel Blanchard
Department of Computer Science
Colorado State University
Fort Collins, CO, USA
`{cytoniv,nblancha}@colostate.edu`

Abstract

Collaborative problem solving (CPS) in teams is tightly coupled with the creation of shared meaning between participants in a situated, collaborative task. In this work, we assess the quality of different utterance segmentation techniques as an aid in annotating CPS. We (1) manually transcribe utterances in a dataset of triads collaboratively solving a problem involving dialogue and physical object manipulation, (2) annotate collaborative moves according to these gold-standard transcripts, and then (3) apply these annotations to utterances that have been automatically segmented using toolkits from Google and OpenAI's Whisper. We show that the oracle utterances have minimal correspondence to automatically segmented speech, and that automatically segmented speech using different segmentation methods is also inconsistent. We also show that annotating automatically segmented speech has distinct implications compared with annotating oracle utterances—since most annotation schemes are designed for oracle cases, when annotating automatically-segmented utterances, annotators must invoke other information to make arbitrary judgments which other annotators may not replicate. We conclude with a discussion of how future annotation specs can account for these needs.

1 Introduction

In order for Artificially Intelligent (AI) agents to interact with with an environment, they must first accurately perceive that environment. In real-world contexts, this necessitates automatically preprocessing various modalities for downstream procedures. For example, an AI agent to modulate classroom discourse needs to first identify distinct discourse components, but a single spoken utterance from a team member could contain multiple discourse components. The identification of each discourse component within the utterance could easily spiral into a doctoral thesis but overly fixating on this preprocessing step would make it extremely difficult to make substantive progress on AI agents themselves.

Typically, researchers default to "oracle" data, where one assumes the preprocessing step has been completed with human level accuracy (e.g., human transcriptions of speech, utterances segmented by dialogue move). However, in a real-world agent deployment, preprocessing of data that would be fed into the automated system will instead be handled by off-the-shelf software. Current practice in AI assumes the existence of suitable datasets that contain examples of the information an automated system is intended to extract. The task of developing the AI model entails solving for the function that best maps from the input samples to the desired outputs. If these datasets do not already exist, then the information that is to be learned must be annotated by humans.

Consider the scenario we focus on in this paper: a group collaborating to solve a problem involving the shared manipulation of physical objects. Multiple modalities are implicated in such a task—group members speak to each other, but also point or gesture, use body language, and manipulate objects to communicate meaning and intent. Specs intended for annotating *collaborative problem solving* (CPS) skills on display are intended to be used at the utterance level, and assume that the utterance has been segmented and transcribed by humans ("oracles"). There are many frameworks for modeling CPS that have been developed by researchers in the learning sciences (e.g., Roschelle and Teasley (1995); Cukurova et al. (2018); Andrews-Todd and Forsyth (2020); Sun et al. (2020)) and this literature stresses the multimodal nature of CPS (Dillenbourg and Traum, 2006). For example, the occurrence of an interruption or the content of cross-talk may not be immediately evident from the audio signal

alone, but watching the speakers interact may make it clear who is speaking when or what is said. High-quality annotation of oracle utterances of a multimodal task like CPS therefore relies on annotators attending to the multiple modalities implicated while making their decisions. If the annotations are performed without this information, or with this information scrambled somehow, we should expect this to affect the quality of the annotation. The question is, how much?

The development of such annotation schemes is typically conducted separately from the rapid preprocessing and scaling that AI practitioners are likely to encounter when they use such annotated data for model training. There may be little that AI practitioners can ask of spec developers given the risk involved with the development of an annotation spec (one can imagine the truly unpleasant experience of developing an annotation spec and finding, after innumerable modeling and annotation cycles, that meaning captured in the spec is not linked to the expected meaningful outcomes). Further, spec developers may (arguably rightly) say that they have no expectation that their spec will be used to train AI models, and that the problems that unfold should be solved by AI developers themselves. These are quite reasonable arguments—and unless the annotation spec is being explicitly developed for AI systems, annotators are unlikely to change (we strongly encourage annotations developed for AI to think deeply about these problems—but such thoughts are outside of the scope of this particular paper). Nonetheless, AI development relies on interoperable annotated data, and as AI practitioners ourselves, we conclude that AI practitioners must think deeply about traditional annotation schemes and how we can best accommodate them.

In this paper, the annotation scheme we use is the one developed by Sun et al. (2020) but the problem we address is independent of any particular spec. Namely, when an annotation spec designed for one utterance segmentation method is applied to utterances automatically segmented using a different method, the information retrieved is different from what the original spec intended to encode.

We annotate utterances for CPS using expert annotators, and we also have expert annotators label, as best they can, automatically-segmented utterances. We discuss common strategies to transfer oracle annotations to real-world annotations. We underscore, exactly, how disconnected the or-

acle utterance labels may be from the labels on automatically-segmented utterances. Finally, we discuss how, given even just two automatic utterance segmentation methods, achieving a gold-standard annotating become quickly intractable if the specification itself does not contain strategies for accommodating suboptimal preprocessing.

2 Related Work

The gap between oracle data and real-world data has been identified previously (Blanchard et al., 2016). Other works have pointed out the need to move away from oracle transcriptions in pursuit of AI applications for real-world use cases (Morbini et al., 2013; Blanchard et al., 2018). The use of automatic segmentation of speech for modeling tasks is becoming increasingly widespread (Bradford et al., 2022a,b; Castillon et al., 2022).

Modeling in general has become more aware of the needs of real-world systems. For example, methods for automatically detecting mind wandering have moved from balanced datasets to heavily imbalanced datasets in acknowledgement of the need for such models to operate in the context of real-world distributions (Kuvar et al., 2022).

What is distinct with this work is that here we focus our analysis on the annotation implications, rather than on attempts to fix issues that arise through machine learning directly. For example, Blanchard et al. (2018) 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 (Graesser et al., 2018), 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 (Stewart et al., 2021; Bradford et al., 2023). The framework defined by Sun et al. (2020) captures CPS at three levels and identifies specific actions that indicate different types of collaborative actions and their impact on group state. In particular, we hope our efforts here facilitate consistency across future CPS modeling efforts and meaningfully contribute to the CPS framework defined by Sun et al. (2020), and in general, we hope to prompt thought about annotation spec design and strategy

in the face of potential uses involving automated preprocessing.

3 Dataset

Our dataset consists of audiovisual recordings of 10 triads performing a shared collaborative task which was developed to promote rich collaboration via multimodal communication. The task is performed by triads at a round table in a laboratory setting. The equipment on the table includes 6 blocks (of varying weight, size, and color), a balance scale, a worksheet demarcated with spaces to place the blocks (indicated with weights in grams), and a laptop on which participants submit their responses to survey questions throughout the task.

Participants are first given a balance scale to determine the weights of five of the colored wooden blocks. They are told that one block weighs 10 grams, but that they have to determine the weights of the rest of the blocks using the balance scale.[1] As the weight of each block is determined, participants place it on the worksheet next to its corresponding weight. The participants also must submit their final answer for the weight of each block to the survey form on the laptop. Once the weights of all five blocks are solved for, participants are given the sixth block and must identifying its weight *without* using the scale (i.e., participants have to deduce the weight based on the pattern observed in the initial block weights). Finally, participants are asked to determine the weight of another mystery block that is *not physically present* and explain how they arrived at the answer. The participants once again submit their answer as a group in the online survey and are given two chances (with a hint after the first guess if it is incorrect).

The total dataset consists of 10 videos, containing 3 participants each, for a total of 170 minutes of video. Participants ranged from 19–35 years old, recruited from a university population. 20% were female while 80% were male. 60% were Caucasian non-Hispanic, 10% were Hispanic/Latino, and 30% were Asian. All volunteers spoke English through the task but spoke a variety of native languages.

Although this data was collected in a lab, the complexity of human-human interaction is appropriately captured in these recordings — participants talk over each other, they speak with disfluencies, they interrupt each other, they engage in long run-on sentences punctuated by only a single em-dash,

and they pause in the middle of sentences before resuming their thought. All of these complications make utterance segmentation quite difficult, and often these ambiguities are only resolved by human annotators with recourse to the visual modality.

4 Preprocessing

4.1 Automatic Segmentation of Speech

Automatic Speech Recognition, or ASR, approaches, must necessarily determine the boundaries of utterances. Each ASR model segments audio in unique ways. This can be either through waiting for any pause in the audio, or waiting until a break of a certain length is encountered. ASR allows for AI to break apart the speech for the listener to in principle break down the amount of empty noise within audio recordings, and different systems using the same ASR component are interoperable on this level.

4.2 Whisper

Whisper (Radford et al., 2022) is a speech recognition system developed by OpenAI that was trained on 680,000 hours of audio to accurately determine and transcribe speech across many different languages. Whisper takes audio files and will listen to the first 30 seconds, or less depending on the length of the file, to determine the language of the speech. It will then segment the audio into full second segments, and will rarely cut off before a single or multiple full seconds have passed. Whisper is also optimized to segment audio into full sentences instead of simply looking for a break in the audio. In principle, this allows for transcription of long audio segments (e.g., lectures or speeches) with a fidelity closer to human transcription.

4.3 Google ASR

Google ASR (Velikovich et al., 2018) is a speech recognition system released by Google, Inc. Google ASR listens for what it assumes to be human speech and attempts to transcribe what it hears. Google also will attempt to segment audio wherever it finds a break in speech. If a word is not picked up by the microphone correctly or is slightly inaudible, then Google will cut off the word and move on with the next segmentation. This could mean cutting off a thought mid-sentence, or removing words entirely from what someone is saying.

5 Annotation Methodology

Videos were first hand-transcribed to ensure the accuracy of the transcriptions. These hand, or ora-

---

[1]The pattern to the weights of the blocks is based on the Fibonacci sequence.

cle transcriptions, were then measured against the transcriptions from both Google ASR and Whisper.

5.1 Annotation Procedure

When annotating the oracle files, speech start and end times would be marked down to the hundredth of a second. Each audio file would then be segmented into proper sentences or thoughts if the sentences were not completed. If people within the audio recording spoke over each other, each person's sentence was recorded as closely as possible, each with its own beginning and ending timestamp. This was done for each audio file from the 10 separate groups. Each segmented utterance was then coded by expert annotators using an updated version of the framework developed by Sun et al. (2020). The annotators initially annotated all 10 videos separately, to get familiar with the framework, then were trained by experts in the framework on one video, while discussing how each CPS indicator would align with the weights task. The experts then annotated another video with a Fleiss' kappa score of 0.62 (agreement over 96% of the number of subjects to be coded).

6 Transferring Annotations from Oracle to Automatically-Segmented Utterances

Once oracle utterances are labeled, we map those labels to the automatically segmented utterances. The approach for that mapping depends on the task at hand and the type of labels we see. In the case of labeling collaborative problem solving (CPS), the multiclass binary labels can be inherited from the oracle segments to the automatic segments using overlap in timestamps. This is because the labels all still exist during that period. However, we lose label accuracy when we lose the exact timestamp where the label occurred. Another option is to only apply labels that occur in every oracle included in the segment; however, with CPS, this would rarely occur and we would lose most of our labels.

7 Effects of Oracle vs. Automatic Utterance Segmentation

Count of utterances Table 1 shows the different number of utterances segmented out by each method for each group.

Almost uniformly, Whisper segments more individual utterances than occur in the oracle transcripts, due to breaking up single oracle utterances into multiples (exceptions are groups 7 and 10). Across all groups, Google segmentation creates fewer (sometimes far fewer) utterances than exist

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

in the oracle, due to dropping utterances entirely or mistaking speech for background noise. Google ASR does perform very well removing empty space from audio files compared to Whisper.

Intrinsic ASR metrics Evaluation of the automatic speech transcription itself after automatic segmentation can be used as a proxy for information lost in part due to the segmentation process. Since error rates must be calculated with respect to the same set of utterances in order to be directly comparable, we focused this analysis on the transcription of Google-segmented utterances. Given an oracle transcript with assumed insertion, deletion, substitution, and total word error rates of 0, we observe that while overall word error rate (WER) is similar using Google and Whisper (Google: 0.573; Whisper: 0.542), Google has higher rates of substitutions (words in the oracle swapped for a different word) and deletions (words in the oracle removed by automated transcription), while Whisper has a significantly higher rate of insertions (words in the automated transcript not in the oracle). See Table 2.

We investigated why Whisper had far more insertions and found it was linked to Google utterances that did not contain any speech. Occasionally, when listening to the audio files Google will hear empty noise as speech and create a segment for it. When feeding Whisper an audio segment containing only background noise, it would generate its own sentence to fill the void, and would occasionally choose a random language to generate the utterance in as well. This does not pose an issue in most situations, since the main purpose of ASR and transcription software would be to transcribe and recognize actual speech in audio files. Thus, the WER of Whisper seems to be partially be a product of our decision to use Google utterances. An appropriate method to filter out such segments, or, the use of Whisper's own segmentation would likely substantially lower the WER of whisper.

We also noticed Whisper would insert words when there was no speech recognized in the audio

	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

Table 2: WER, substitution rate, deletion rate, and insertion rate by group.

clip. Typically words that had been previously transcribed would be repeated during void sections, and only one word or phrase would be repeated. This word or phrase would also be repeated for each second during the break, which created a large amount of insertions and threw off some data while testing. Finally, we found Google ASR would also hear words incorrectly and misinterpret what the speaker was saying, replacing the intended words with homonyms or phonological near-neighbors.

Difference in resulting annotation labels The difference in labels when going from automatic segments to oracle segments can be significant. A particular case is the annotation of interruptions, one of the CPS indicators in question. Relying on the automatic segments only may split or lump utterances separated by an interruption, which may cause annotators to miss the interruption entirely (because they are only coding utterance by automatically segmented utterance), or lumping an "interruption" annotation with annotations of other meaningful indicators in a single, multi-speaker "utterance." For tasks like this, each utterance is important as it can be the one where the correct solution has been proposed. Interestingly, Bradford et al. (2023) found that prosodic features were essential for identifying interruptions when using automatically segmented utterances, indicating there may be times automatic segmentation methods capture information not present in oracle contexts.

One example of label difference can be seen in the utterances shown in Fig. 1, with the different segmentations provided by oracle and automatic segmentation. The utterances "Weren't those both

Figure 1: Overlap between oracle (top), Google (middle), and Whisper (bottom) segments. Right column shows the CPS indicator annotated for each utterance.

thirty or no only one of them twenty and thirty", "No this is twenty you're off the team" and "Twenty and then" (which were each spoken by a different person), are combined into one segment by Google voice activity detection (VAD). The first utterance should have the label *confirms understanding*, the second utterance should have the labels *interrupts* and *initiates off-topic conversation*, and the third utterance should have no label. However, when these are combined, all of the labels are inherited and the distinction between the different content supplied by each utterance is lost. Whisper segments split up continuous utterances by a single person, and thus person 2's *interrupts* and *initiates off-topic*

conversation indicators are applied to two separate segments. Both of these cases can cause confusion in downstream semantic classification tasks like classifying CPS indicators from linguistic features, as in Bradford et al. (2023), if the target label for training is not clear.

In some instances, the participant pauses mid-sentence and the true utterance gets split into two, but for lack of context only one gets assigned a CPS indicator. For example, a participant says "Think it just feels like it's," pauses for 0.3 seconds, and says, "A lot heavier because it's denser and like just carrying that.", the whole sentence should get coded as *discussing results*, but since the automatic segmentation splits the sentence into two utterances, only the second utterance would be coded as such.

8 Discussion, Recommendations, and Conclusion

One important point to emphasize is the manual cost of annotating for collaborative problem solving (CPS). CPS is a difficult annotation scheme to master (training can last as long as 6 months, depending on how much time a coder is putting toward learning). Although this paper largely focuses on the automated processes of segmenting audio, annotations themselves require complete multimodal context including viewing of video, listening to intonation, and the inclusion of temporal context. If these annotations, performed with access to multimodal information, are subsequently applied to automatically-segmented audio, then the information lost can be expected to impact downstream tasks trained or evaluated over the annotated data, thus potentially wasting the time taken to train annotators properly.

While automatic segmentation of utterances for various semantic annotation tasks certainly saves time and annotator effort, it comes at a potentially significant cost to the quality of annotations for downstream tasks. Particularly, automatic segmentation and transcription methods certainly segment utterances differently from a human oracle transcriber, and different ASR methods perform segmentation drastically differently with profoundly divergent results. This may result in utterances being missed by the automatic segmenter or invented out of whole cloth, which would cause annotators annotating at the automatically-segmented utterance level to likewise omit annotations, or to encounter "hallucinated" segments that are either un-annotatable or, if annotated, introduce seman-

tic noise into the data. Beyond the obvious, we have shown that annotating at the oracle utterance level but then transferring those utterances to the automatically-segmented utterance level may obscure the semantic information originally captured at the oracle utterance level. Even taking the more labor-intensive step of generating oracle transcriptions before annotating is less useful if annotation is not performed at the same level. This backs up previous conclusions in semantic annotation over text-only corpora, such as the need for annotators to come to consensus on both spans and annotations (Pustejovsky and Stubbs, 2012), and shows that they also apply to multimodal use cases.

However, as multimodal AI develops and becomes more integrated with everyday life, inference will necessarily be performed over automatically segmented and transcribed inputs. Therefore, future models will benefit from annotation specs themselves that are task-aware and can take into account potential noise introduced by imperfect automated transcription and adjust accordingly. For instance, if multiple labels are not allowed, should certain labels "dominate" others in case multiple labels are squeezed into the same segment? Future semantic annotation schemes, specifications, and languages, particularly over multimodal data, will need to take into account these requirements to more effectively use automated techniques like ASR as part of larger annotation and inference pipelines.

Acknowledgments

References

Jessica Andrews-Todd and Carol M. Forsyth. 2020. Exploring social and cognitive dimensions of collaborative problem solving in an open online simulation-based task. *Computers in human behavior*, 104:105759.

Nathaniel Blanchard, Patrick J. Donnelly, Andrew M. Olney, Borhan Samei, Brooke Ward, Xiaoyi Sun,

Sean Kelly, Martin Nystrand, and Sidney K. D'Mello. 2016. Semi-automatic detection of teacher questions from human-transcripts of audio in live classrooms. *International Educational Data Mining Society*.

Nathaniel Blanchard, Daniel Moreira, Aparna Bharati, and Walter Scheirer. 2018. 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.

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

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

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

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

Mutlu Cukurova, Rose Luckin, Eva Millán, and Manolis Mavrikis. 2018. The NISPI framework: Analysing collaborative problem-solving from students' physical interactions. *Computers & Education*, 116:93–109.

Pierre Dillenbourg and David Traum. 2006. Sharing solutions: Persistence and grounding in multimodal collaborative problem solving. *The Journal of the Learning Sciences*, 15(1):121–151.

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

Vishal Kuvar, Nathaniel Blanchard, Alexander Colby, Laura Allen, and Caitlin Mills. 2022. Automatically detecting task-unrelated thoughts during conversations using keystroke analysis. *User Modeling and User-Adapted Interaction*, pages 1–25.

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

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. " O'Reilly Media, Inc.".

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Jeremy Roschelle and Stephanie D. Teasley. 1995. The construction of shared knowledge in collaborative problem solving. In *Computer supported collaborative learning*, pages 69–97. Springer.

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

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

Leonid Velikovich, Ian Williams, Justin Scheiner, Petar S. Aleksic, Pedro J. Moreno, and Michael Riley. 2018. Semantic lattice processing in contextual automatic speech recognition for google assistant. In *Interspeech*, pages 2222–2226.