Detecting Self-Reported Cognitive-Affective States in Collaborative Problem Solving from Prosodic and Linguistic Features

Videep Venkatesha
Colorado State University

Sifatul Anindho
Colorado State University

Nathaniel Blanchard
Colorado State University

{videep.venkatesha,sifatul.anindho,nathaniel.blanchard} @colostate.edu

ABSTRACT

Detecting cognitive-affective states in collaborative problem-solving is challenging due to their subjective nature, social regulation, and the complexity of multiparty speech. While prior work has largely relied on observer-coded labels in individual learning contexts, self reported states may capture internal experiences that learners regulate and do not outwardly express, precisely the states where intervention could be most valuable. We investigate whether conversational speech alone provides sufficient signal to detect self reported cognitive-affective states during collaborative problem solving. Using retrospective video-cued recall, we obtained labels for seven states (Optimistic, Confused, Curious, Disengaged, Surprised, Frustrated, and Conflicted) from 27 participants (9 groups) completing a collaborative task. We systematically compared prosodic features (eGeMAPS) and semantic embeddings (SentenceTransformers) across multiple temporal window sizes using Leave One Group Out cross validation. Four of seven states were detectable above chance with statistical significance (permutation $p < 0.05$), with AUROCs ranging from 0.570 to 0.618. The remaining three states showed promising trends but did not reach significance, likely due to limited statistical power. No single feature modality dominated across states, and optimal detection windows varied substantially, from 5 seconds for transient states like Confused to 30 seconds for sustained states like Frustrated. These findings suggest that meaningful signal for self-reported affective states exists in conversational speech during collaborative problem-solving, motivating further investigation with larger samples and richer feature representations.

Keywords

Collaborative learning, affective computing, multimodal learning analytics, prosodic features

1. INTRODUCTION

Videep Venkatesha, Sifatul Anindho, and Nathaniel Blanchard. Detecting Self-Reported Cognitive-Affective States in Collaborative Problem Solving from Prosodic and Linguistic Features. In Anthony Botelho, Maria Mercedes T. Rodrigo, Adish Singla, Hiroaki Ogata, Hyojeong So, and Young Hoan Cho (eds.) Proceedings of the 19th International Conference on Educational Data Mining, Seoul, Republic of Korea, June, 2026, pp. 632–636. International Educational Data Mining Society (2026).

© 2026 Copyright is held by the author(s). This work is distributed under the Creative Commons Attribution NonCommercial NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license.

https://doi.org/10.5281/zenodo.21039901

Collaborative problem-solving (CPS) is a learning paradigm where two or more individuals work together to achieve a shared goal [1]. Through CPS, learners engage in social interactions such as joint reasoning, finding common ground, and coordination of strategies, which require the regulation of cognitive-affective states such as confusion and conflict. Prior research has shown that such states influence both group performance and individual learning outcomes [2]. Therefore, automatically detecting cognitive-affective states during collaborative problem-solving is useful for adaptive educational technologies to better support learners in real time.

However, automatic detection of these states in collaborative settings remains challenging. Prior work on affect detection has largely focused on individual learning contexts, where confusion and frustration can be inferred from dialogue in intelligent tutoring systems [3, 4], and multimodal signals including speech, facial expression, and gaze have improved detection [5–7]. Collaborative learning introduces additional complexity: speech overlaps and social masking make it difficult to map behavioral signals to affective states, and observable behavior does not always align with subjective experience. To obtain more ecologically valid labels, Anindho et al. [8] used retrospective cued-recall to capture cognitive-affective labels aligned to collaborative interaction, providing access to internally experienced states that may not be externally observable. Although retrospective labeling may introduce recall bias and some temporal imprecision, it may provide access to subjective states that are not captured by external observation. This brings forth an interesting question about the relationship between subjective states and observable behavior: how does self-reported affect manifest in external signals such as speech in collaborative problem-solving? To address this, we turn to a naturalistic collaborative setting in which both speech and self-reported states can be examined together.

This paper investigates the automatic detection of seven self-reported cognitive-affective states: Optimistic, Confused, Curious, Disengaged, Surprised, Frustrated, and Conflicted, from audio recordings of 27 participants (nine groups) completing a collaborative problem-solving task. Using retrospective video-cued self-report labels, we systematically compare prosodic features (eGeMAPS) and sentence embeddings (SentenceTransformers) across four temporal window sizes (5s, 10s, 20s, 30s). Our results establish preliminary baseline detection performance for each state, revealing that

four states show reliable above-chance detection (permutation $p < 0.05$): Optimistic, Curious, Disengaged, and Confused. Three additional states—Frustrated, Conflicted, and Surprised—have confidence intervals above chance, though statistical power limitations suggest these findings are preliminary. We find that no single feature modality dominates across all states, and that optimal temporal windows vary systematically by state (ranging from 5 seconds for Confused to 30 seconds for Frustrated). Despite the inherent difficulty of detecting subjective internal states from speech alone, our results demonstrate that meaningful signal exists in conversational audio for multiple cognitive-affective states, establishing an initial baseline for self-report-driven affect detection in collaborative learning environments.

2. METHODS

2.1 Participants and Task

This study examines 27 participants who were older than 18 and all fluent in English. They were recruited from the author's department with approval from the Institutional Review Board and compensated with 15 USD. The sample included 18 male, 7 female, and 2 non-binary participants, who self-identified as Asian (13), White (12), or Hispanic (2). Participants were organized into nine collaborative groups, with three participants per group. Each group completed the first part of the Weights Task, which is a collaborative problem-solving activity from Khebour et al. [9] in which participants are given five blocks, a balance scale, and told that one block weighs 10 grams. Their task is to determine the weights of the remaining four blocks through collaborative experimentation and reasoning. The task requires coordination, hypothesis testing, and shared decision-making, naturally eliciting a range of cognitive-affective states as participants encounter uncertainty, disagreement, and discovery [9]. Participants provided written consent from each participant and were informed that their audio and video recordings would be used for research purposes. Recordings are stored on secure, access-controlled servers and transcripts were de-identified prior to analysis. Figure 1 shows one of the groups collaborating during the problem-solving activity.

Figure 1: Our experimental setup during a collaborative problem solving session

2.2 Annotation of Cognitive-Affective States

Cognitive-affective states were annotated using retrospective video-cued recall. After completing the task, each participant individually watched a recording of their collaborative session and marked instances where they experienced specific cognitive-affective states. Seven states were adopted from previous research [10] and were made available for selection: Optimistic, Confused, Curious, Conflicted, Frustrated, Surprised, and Disengaged. Participants could report multiple states concurrently if they experienced overlapping cognitive-affective responses. Each annotation generated a timestamp aligned to the video timeline, along with participant and group identifiers. Annotations were completed independently to avoid social influence, and no external coders were involved in labeling. The resulting dataset therefore reflects participants' subjective, post-hoc recall of their internal cognitive–affective experience, rather than behavioral interpretations assigned by observers. An example conversational segment is presented below. Participant 3 labeled this moment as Frustrated.

> **P1:** So what if we tried 40? Which one did we say was 30?
> **P2:** Purple.
> **P1:** Perfect.
> **P1:** Okay, incorrect. So it is 50.

This segment illustrates the challenge of detecting internal states from speech alone: P1's self-reported frustration is not explicitly verbalized, and the prosodic signature must be extracted from brief utterances ("Okay, incorrect. So it is 50") that occur in the context of ongoing collaboration. The detection task requires distinguishing such moments from similar task-focused utterances spoken without frustration. Table 1 presents the distribution of labeled samples across states and window sizes. The distribution is imbalanced, with Optimistic and Confused being the most frequent states and Surprised, Disengaged, and Frustrated each having fewer than 20 samples. Windows are centered on the point of self-report: a 5-second window includes 2.5 seconds before and 2.5 seconds after the reported moment, a 10-second window includes 5 seconds before and after, and so on. We exclude any window without an associated transcript (i.e., periods of group silence). As window size decreases, the number of such samples increases.

Table 1: Number of positive samples per cognitive-affective state across four temporal window sizes.

State	5s	10s	20s	30s
Optimistic	73	81	88	90
Confused	57	58	58	59
Curious	42	49	52	54
Conflicted	18	19	19	19
Frustrated	16	16	16	16
Surprised	15	15	16	16
Disengaged	15	18	20	21
Total labeled	**236**	**256**	**269**	**275**

2.3 Feature Extraction

Prosodic features. Each participant wore an individual microphone, providing speaker-separated audio. We extracted

88 features using the eGeMAPS feature set via openSMILE [11], including F0 statistics, loudness, spectral features, voice quality measures (jitter, shimmer, HNR), and temporal descriptors. Features were computed over each temporal window.

Transcript features. Speech was transcribed using Whisper [12] and segmented into temporal windows. We obtained 768-dimensional sentence embeddings using `all-mpnet-base-v2` from SentenceTransformers [13], which outperformed BERT `[CLS]` embeddings across states.

Combined features. We additionally concatenated prosodic and transcript vectors (856 dimensions).

2.4 Classification Framework

We framed detection as a set of binary one-vs-rest classification problems, training a separate classifier for each of the seven cognitive-affective states. For each state, the positive class consisted of all windows labeled with that state (using the primary label only), and the negative class was formed by randomly sampling from all non-target windows at a 1:1 ratio. The negative class for each state consists of randomly sampled non-target windows, including both unlabeled windows and windows labeled with other states, to allow for a balanced dataset. This formulation tests whether each state has a discriminative signature that distinguishes it from both neutral collaboration and other affective states, a pragmatic choice for real-world affect detection systems that must distinguish target states from heterogeneous background activity. To reduce variance from negative sampling, we repeated the random undersampling with five different random seeds and averaged results across seeds. We evaluated Logistic Regression (L2, balanced weights), Random Forest (200 trees, balanced weights), and SVM with RBF kernel (balanced weights, Platt scaling). Features were standardized with parameters fit on training data only.

We used Leave-One-Group-Out (LOGO) cross-validation, holding out one group at a time, and report AUROC as the primary metric.

3. RESULTS

We report classification performance using AUROC as the primary metric, averaged across five negative resampling seeds. We report 95% bootstrap confidence intervals (2,000 iterations on pooled LOGO predictions) and permutation test p-values (1,000 iterations) to assess whether observed performance exceeds chance. Logistic Regression was the best or near-best model for the majority of configurations; we focus on Logistic Regression unless otherwise noted.

3.1 Detection Performance Across States

Table 2 reports the best AUROC per state across all window sizes and modalities, along with the optimal modality/window and permutation p-values. Three states met a significance criterion of CI excluding 0.50 and permutation $p < 0.05$: Optimistic (0.607, $p = 0.004$), Curious (0.600, $p = 0.016$), and Disengaged (0.618, $p = 0.023$). Confused reached significance at 5 seconds (0.570, $p = 0.049$). The remaining states of Frustrated (0.634), Conflicted (0.623), and Surprised (0.613), had CIs above chance but permutation p-values between 0.10 and 0.14, reflecting limited power with a 9-fold LOGO design. We additionally evaluated late fusion (averaging classifier probabilities from prosodic-only and transcript only models) and learned fusion (training a meta-classifier on concatenated probability outputs). Neither approach improved over the better single modality (mean AUROC difference: -0.01 for late fusion, $+0.00$ for learned fusion), suggesting that prosodic and linguistic features are largely redundant rather than complementary for most states.

Table 2: Best AUROC per state, with the corresponding optimal modality/window and permutation p-values.

State	n	AUROC	Modality	Window	p
Frustrated	16	0.634	Prosodic	30s	.108
Conflicted	18–19	0.623	Combined	30s	.136
Disengaged	15–21	0.618	Transcript	30s	.023
Surprised	15–16	0.613	Transcript	30s	.103
Optimistic	73–90	0.607	Transcript	10s	.004
Curious	42–54	0.600	Combined	20s	.016
Confused	57–59	0.570	Prosodic	5s	.049

Each state's optimal configuration differed in both temporal window size and feature modality. Across all 28 state–window combinations, transcript features were optimal 13 times, prosodic features 9 times, and combined features 6 times.

3.2 Effect of Temporal Window Size

Table 3 reports the best AUROC per state at each window size with bootstrap CIs. The optimal window varied substantially by state as seen by Figure 2. Four states achieved their best performance with 30 second windows (Frustrated, Conflicted, Disengaged, Surprised), while Optimistic peaked at 10 seconds and Confused at 5 seconds. Curious improved steadily from 5 to 20 seconds, then plateaued.

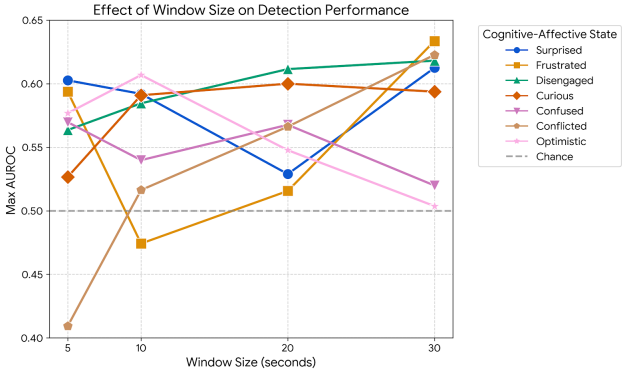

Figure 2: Interaction between temporal window size and detection performance. Max AUROC achieved for each state across 5s, 10s, 20s, and 30s windows. The crossing trajectories highlight distinct temporal dynamics: Optimistic and Confused decay with longer windows (suggesting transient expression), while Disengaged and Frustrated improve (suggesting sustained states).

The temporal trajectory differed qualitatively across states. Optimistic peaked early and declined ($0.577 \rightarrow 0.607 \rightarrow 0.548 \rightarrow 0.504$), with CIs excluding chance only at 5, 10, and 20 seconds. Curious showed consistent above-chance detection from 10 seconds onward, with significant permutation

Table 3: Best AUROC [95% CI] per state at each window size. Bold indicates the highest AUROC per state. † CI excludes 0.50; * permutation $p < 0.05$.

State	5s	10s	20s	30s
Frustrated	.594 [.51, .68]†	.474 [.39, .57]	.516 [.42, .61]	**.634** [.56, .72]†
Conflicted	.409 [.32, .49]	.516 [.44, .60]	.566 [.48, .63]	**.623** [.55, .70]†
Disengaged	.564 [.48, .66]	.585 [.50, .67]	.611 [.52, .68]†	**.618** [.54, .69]†*
Surprised	.603 [.51, .69]†	.592 [.49, .67]	.529 [.45, .63]	**.613** [.51, .69]†
Optimistic	.577 [.54, .62]†	**.607** [.57, .65]†*	.548 [.51, .58]†	.504 [.46, .54]
Curious	.527 [.47, .58]	.591 [.54, .64]†	**.600** [.55, .65]†*	.594 [.55, .64]†*
Confused	**.570** [.52, .62]†*	.540 [.49, .58]	.568 [.52, .62]†	.521 [.47, .56]

p-values at 20s ($p = 0.016$) and 30s ($p = 0.031$). Disengaged improved monotonically with window size but only reached significance at 30 seconds ($p = 0.023$).

4. DISCUSSION

This study investigated whether conversational speech contains sufficient signal to detect self-reported cognitive affective states during collaborative problem-solving. The central finding is encouraging: despite the inherent subjectivity of internal states and the social pressures that shape expression during collaboration, detectable signal exists for multiple states. We discuss the implications of this finding and directions for building on it.

4.1 Signal Exists for Subjective Internal States

We used *self-reported* labels capturing what participants said they internally experienced, rather than observer-coded labels capturing externally visible behavior. In collaborative settings, participants may regulate outward expression to maintain group harmony while still experiencing the underlying state [14, 15]. Our ability to still identify the signal despite this regulation suggests that speech-based detection of internal affective states during collaboration is feasible and worth further investment.

4.2 State-Specific Detection Configurations

A consistent finding across our systematic comparison is that each cognitive-affective state has a distinct optimal detection configuration. No single modality dominated: transcript embeddings were optimal for some states (Disengaged, Surprised, Optimistic), prosodic features for others (Frustrated, Confused), and the combination for two (Curious, Conflicted). Similarly, optimal temporal windows ranged from 5 seconds (Confused) to 30 seconds (Frustrated, Conflicted, Disengaged, Surprised).

The temporal patterns are particularly notable. Confused was best detected at short windows and degraded with longer context, while Frustrated and Disengaged improved monotonically up to 30 seconds. These patterns align with theoretical accounts distinguishing transient states from sustained episodes [16]—confusion as brief cognitive impasse that either resolves or escalates, frustration as accumulated difficulty over time. The state-specific temporal signatures suggest that affect detection systems for collaborative learning may need adaptive windowing strategies rather than fixed parameters.

The failure of multimodal fusion to improve over the best single modality (mean AUROC difference <0.01 across early, late, and learned fusion) indicates that prosodic and linguistic features are largely redundant for these states. This contrasts with typical findings in acted-speech emotion recognition [17], suggesting that the relationship between modalities may differ for self-reported states in naturalistic settings.

4.3 Limitations

The sample of nine collaborative groups limits both the reliability of leave-one-group-out estimates and the power of permutation tests, in addition to limiting generalization .

The retrospective self-report methodology introduces temporal imprecision and potential recall biases [18]. Participants may misalign labels when marking video, or preferentially recall salient moments. The binary one-vs-rest framing creates heterogeneous negative classes that include both unlabeled windows and windows labeled with other states.

Feature representations. Transcript embeddings were computed by concatenating all utterances within each window into a single string, which may not optimally capture turn-taking dynamics, overlapping speech, or the sequential structure of collaborative dialogue. Additionally, prosodic features were aggregated over fixed temporal windows, potentially missing finer-grained affective dynamics within windows.

Generalizability. Our findings are based on a single collaborative task (the Weights Task) with groups of three participants in a controlled laboratory setting. Task characteristics such as problem structure, duration, and social dynamics may influence which states emerge and how they are expressed. Detection performance may differ for other collaborative learning contexts (e.g., open-ended projects, peer tutoring, online collaboration) or different group sizes.

5. CONCLUSION

In this work, we used conversational speech to automatically detect self-reported cognitive-affective states during a collaborative problem-solving task, employing a retrospective video-cued recall approach to obtain affective labels without disrupting the task itself. Four of seven states were detectable above chance from prosodic and linguistic features alone: an encouraging result given the inherent subjectivity of internal states and the social regulation that characterizes collaborative settings. These results motivate future work incorporating richer features, additional modalities, and larger samples. We hope they inspire continued investment in self-report-driven approaches to affect detection in educational technology. We note that the intended application of this work is to assist educators and AI tutoring

systems in providing timely support; any deployment should ensure that inferred states are used to benefit the individual learner rather than for surveillance or social comparison.

6. ACKNOWLEDGMENTS

This material is based in part upon work supported by the U.S. National Science Foundation (NSF) under award DRL 2454151 (Institute for Student-AI Teaming). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

7. REFERENCES

[1] Friedrich Hesse, Esther Care, Juergen Buder, Kai Sassenberg, and Patrick Griffin. A framework for teachable collaborative problem solving skills. In *Assessment and teaching of 21st century skills: Methods and approach*, pages 37–56. Springer, 2014.

[2] Sidney D'Mello and Art Graesser. Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2):145–157, 2012.

[3] Sidney D'Mello and Art Graesser. Affect detection from human-computer dialogue with an intelligent tutoring system. In *Intelligent Virtual Agents (IVA 2006)*, volume 4133 of *Lecture Notes in Computer Science*, pages 54–67. Springer, 2006.

[4] Ryan S. J. d. Baker, Sujith M. Gowda, Michael Wixon, Jessica Kalka, Angela Z. Wagner, Aatish Salvi, Vincent Aleven, Gail W. Kusbit, Jaclyn Ocumpaugh, and Lisa Rossi. Towards sensor-free affect detection in cognitive tutor algebra. In *Proceedings of the 5th International Conference on Educational Data Mining (EDM)*, pages 126–133, 2012.

[5] Rafael A Calvo and Sidney D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing*, 1(1):18–37, 2010.

[6] Nigel Bosch, Sidney K D'Mello, Ryan S Baker, Jaclyn Ocumpaugh, Valerie Shute, Matthew Ventura, Lubin Wang, and Weinan Zhao. Detecting student emotions in computer-enabled classrooms. In *IJCAI*, volume 16, pages 4125–4129, 2016.

[7] Joseph Grafsgaard, Joseph Wiggins, Kristy Elizabeth Boyer, Eric Wiebe, and James Lester. Predicting learning and affect from multimodal data streams in task-oriented tutorial dialogue. In *Educational Data Mining 2014*, 2014.

[8] Sifatul Anindho, Videep Venkatesha, and Nathaniel Blanchard. A methodological framework for capturing cognitive-affective states in collaborative learning. *arXiv preprint arXiv:2507.01166*, 2025.

[9] Ibrahim Khebour, Richard Brutti, Indrani Dey, Rachel Dickler, Kelsey Sikes, Kenneth Lai, Mariah Bradford, Brittany Cates, Paige Hansen, Changsoo Jung, Brett Wisniewski, Corbyn Terpstra, Leanne Hirshfield, Sadhana Puntambekar, Nathaniel Blanchard, James Pustejovsky, and Nikhil Krishnaswamy. When text and speech are not enough: A multimodal dataset of collaboration in a situated task. *Journal of Open Humanities Data*, 2024.

[10] Sifatul Anindho, Videep Venkatesha, Mariah Bradford, Anne M Cleary, and Nathaniel Blanchard. An exploration of internal states in collaborative problem solving. In *International Conference on Human-Computer Interaction*, pages 135–150. Springer, 2025.

[11] Florian Eyben, Martin Wöllmer, and Björn Schuller. opensmile – the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 1459–1462. ACM, 2010.

[12] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 28492–28518, 2023.

[13] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3982–3992. Association for Computational Linguistics, 2019.

[14] Emily A Butler, Boris Egloff, Frank H Wilhelm, Nancy C Smith, Elizabeth A Erickson, and James J Gross. The social consequences of expressive suppression. *Emotion*, 3(1):48–67, 2003.

[15] Hanna Järvenoja, Sanna Järvelä, and Jonna Malmberg. Supporting groups' emotion and motivation regulation during collaborative learning. *Learning and Instruction*, 63:101211, 2019.

[16] Sidney D'Mello and Art Graesser. Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2):145–157, 2012.

[17] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proceedings of INTERSPEECH*, pages 148–152, 2013.

[18] Michael D Robinson and Gerald L Clore. Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin*, 128(6):934–960, 2002.