Inferring Student Engagement via Real-Time Thermal–Visual Voice Activity Detection

Bradley A. Goodman[1,2]
[1]The MITRE Corporation, [2]Brandeis University,

Abstract

We introduce a thermal–visual fusion approach to improve non-invasive Voice Activity Detection (VAD) for classroom engagement monitoring. In noisy multi-speaker classrooms using a single microphone, acoustic-only methods fail to reliably isolate individual speakers. Our method integrates facial thermal signatures—capturing respiratory and speech-related heat patterns—with visual lip-motion cues to provide an acoustic-independent speech signal. This provides a localized, privacy-preserving, and acoustic-independent indicator of speech activity.

This system acts as a visual-diarization front-end, informing Automatic Speech Recognition (ASR) and Natural Language Processing (NLP) systems not only when speech occurs, but precisely which student is speaking. Using up to 19 engineered features, our Thermal-Only Random Forest classifier achieved a Recall of 0.9234 and an F1-score of 0.8105 in subject-independent evaluations, outperforming visual-only baselines. The system was validated as a proof-of-concept on a Raspberry Pi 5 in a controlled laboratory setting, demonstrating real-time feasibility. These results demonstrate that thermal–visual fusion enables more reliable linguistic analysis of collaborative learning and provide critical input for AI agents to facilitate group participation in real-world educational settings that lead to more successful learning outcomes.

1 Introduction and Research Motivation

Voice Activity Detection (VAD) is fundamental to acoustic monitoring and classroom behavioral analysis. Existing solutions, outside of those focused solely on audio, typically rely on visible-spectrum video and suffer from severe degradation in performance due to variability in lighting, occlusions, and person identity changes. The objective of this research was to investigate the efficacy of integrating facial thermal signature data—which captures the distinct heat patterns associated with respiration and articulation—to create a system that is robust and highly generalizable across different scenarios. Thermal cameras, once limited to industrial (Wilson et al., 2023) and medical (Ioannou et al. 2014; Manullang et al. 2021; Kwon et al. 2023) applications, have become increasingly viable for a wider range of uses due to recent reductions in both cost and size. The incorporation of Artificial Intelligence (AI) into classroom settings aims to promote effective peer learning by facilitating better team dynamics and ensuring equitable engagement among students. A critical component of this AI-Student teaming approach is the ability to accurately track active student engagement during collaborative tasks. Previous research into engagement detection in noisy classroom environments has faced significant challenges:

• **Acoustic Limitations**: Efforts relying on speech understanding and speaker identification were hindered by the use of a single microphone and the highly noisy nature of multiple peer groups working concurrently in a classroom. This resulted in relatively poor speech recognition with high Word Error Rates (WER) and an inability to reliably isolate the speaker (Cao et al. 2023; Wang et al. 2024). These acoustic failures often prevent NLP systems from performing accurate group dynamics analysis, knowledge-sharing tracking, or peer-learning modeling, as the input text is either missing or heavily corrupted.

• **Visual Limitations**: Multi-modal gesture recognition (e.g., pointing or looking at a peer) successfully determines engagement but fails to account for instances when a student is actively

[1]Approved for Public Release; Distribution Unlimited. Public Release Case Number 26-0344. The author's affiliation with The MITRE Corporation is provided for identification purposes only and is not intended to convey or imply MITRE's concurrence with, or support for, the positions, opinions, or viewpoints expressed by the author.

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 7–20

speaking but not looking directly at another student or activity-related object (Palmer et al., 2025). Furthermore, attempts at purely visual lip-reading are highly restrictive, requiring the speaker's mouth to be clearly aligned with the camera, which is often impractical in a dynamic classroom. Lip-reading can, however, potentially increase the accuracy of speech recognition (Yoshida et al. 2008; Guan et al. 2024; Guan et al. 2025). Given the inherent limitations of both purely acoustic and constrained visual methods, a more passive and robust detection mechanism is required to reliably determine who is speaking, regardless of noise or precise facial orientation.

2 Related Work

Foundations of Collaboration and Engagement: Early efforts in modeling team dynamics by Goodman et al. (2006) established that individual and group user models could detect faltering collaboration. By utilizing indicators such as dialogue acts, backchannel utterances (e.g., "uh-huh"), and prosodic speech features, they demonstrated that machine learning (ML) can effectively predict listener engagement and shared focus in distributed environments. Similarly, Oztoprak (2022) confirmed that prosodic cues alone—specifically pitch and intensity—are significant predictors of emotional engagement in child-robot interactions, achieving 73.9% accuracy with artificial neural networks. D'Mello et al. (2012) developed a model to explain the dynamics of affective states during learning activities that included states of engagement. Their model establishes engagement as a dynamic, temporally evolving cognitive–affective process, motivating the need for sensing approaches that can capture rapid state transitions during authentic learning activities. While the early research focused on prosodic features and affective states (Cao et al., 2014), more recent work has explored automated scoring of collaborative classroom discussion (Tran et al., 2025).

Audio-Visual and Nonverbal Modalities: Building on these theoretical foundations, subsequent work has operationalized engagement through observable behavioral and physiological signals such as facial expression, gaze, head motion, and speech. Bosch et al. (2016) showed that visible-spectrum video features, including facial expressions and head pose, can infer learner affect in classroom settings, while more recent multimodal approaches (e.g., Sharma et al., 2023) incorporate eye tracking and head movement to improve engagement detection. However, these methods remain sensitive to lighting conditions, occlusion, and expressive variability, limiting robustness in unconstrained educational environments. To address these limitations, researchers have explored multimodal architectures. Tao et al. (2021) proposed TalkNet, an active speaker detection framework that uses audio-visual cross-attention and long-term temporal features to align lip motion with speech, improving performance under overlapping speech and visual noise. It processes raw video instead of running in real time. To empirically assess TalkNet's applicability as a baseline for thermal-domain VAD, we ran TalkNet on the same thermal false-color video used for real-time evaluation. TalkNet detected a stable face track across 3,218 frames (128.7 seconds), yet only 54 frames (1.7%) were classified as speaking, compared to 72% ground-truth speech in the same session. The audio-visual synchrony model relies on visible-spectrum lip texture to align mouth motion with audio; thermal false-color imagery suppresses this texture, rendering the model's visual stream uninformative regardless of acoustic content. This confirms that RGB-trained audio-visual models do not transfer to thermal input without retraining, empirically motivating the purpose-built thermal feature pipeline described in Section 3. Abdrakhmanova et al. (2021) introduced the large SpeakingFaces dataset, combining thermal, visual, and audio recordings of voice commands across 142 subjects; however, its clips range from 2–7 seconds (50–175 frames at 28 fps), each beginning with a brief silence, which is insufficient to populate the 125-frame rolling window required for our spectral breathing features (*Breathing_Power_In_Band_RA*, *Breath_BRSI*) to stabilize — causing both features to evaluate to zero across the entire dataset and producing near-random classification when the trained model was applied to continuous thermal video. Palmer et al. (2025) further argue that speech recognition alone is insufficient in noisy classrooms and advocate integrating nonverbal cues such as gaze, posture, and gesture for context-aware modeling of group interaction and engagement.

Classroom Implementation and Constraints: Translating these models to the classroom remains a primary challenge. Xu et al. (2023) successfully predicted tutoring engagement using a combination of face exposure and speech duration met-

rics. In more complex group settings, Wang et al. (2024) utilized Whisper ASR and SpeechBrain for speaker diarization, achieving a moderate correlation (≈ 0.62) with human annotations despite the high-noise classroom environment. Baker et al. (2014) examined cost-effective engagement detection.

Research Gap: While these studies leverage prosodic, acoustic, and visual cues, they remain susceptible to background noise, ambient lighting variability, and the physical occlusions or motion artifacts inherent in unconstrained settings. Our research addresses this gap by proposing a passive, multi-modal fusion of thermal and visual signatures, offering an acoustic-independent physiological signal for reliable VAD in real-time.

3 Research Hypothesis and Methodology

Our core hypothesis is that thermal imaging of the mouth and nasal regions can provide a novel, reliable, and passive method to determine speech activity by tracking the localized temperature changes caused by respiration. We further hypothesize that a fusion model integrating this thermal data with robust lip motion data will provide a superior capability to detect speech even when the speaker is not fully facing the camera (e.g., Bennett et al. (2020) showing side-view of respiratory plume using a thermal camera).

To test this hypothesis, we collected thermal videos of individuals engaging in both episodes of speaking and non-speaking in a laboratory setting. The thermal video set included single and multi-party participants (Figure 1). Using the trained Yolov8n face-detection model (Vemulapalli et al., 2023), we isolated the face, mouth and nasal regions in each thermal video frame to permit localized extraction of temperatures. This visual tracking data was combined with automatic audio annotation using a widely-used neural-acoustic VAD tool (Silero VAD (Silero, 2024)) to segment periods of active speech. From this multi-modal data stream, we developed a suite of metrics—including rolling thermal averages, derived thermal deltas, and continuous lip motion kinetics—which were then used to train and evaluate ML classifiers for speech prediction.

Our research focuses on one aspect of student engagement detection, episode-level speech detection, which can be fused with other forms of engagement such as gesture, pointing, gaze, and speech recognition.

Figure 1: Single and multiple participant data collection.

3.1 Dataset and Preprocessing

The final dataset used for model development and evaluation included synchronized audio, thermal video, and their associated feature streams. The raw ground truth was generated using the open-source Silero Voice Activity Detector (VAD) that looked for speech activity versus background noise, labeling each frame with a binary Speech_Present flag (TRUE/FALSE) based on whether or not the probability of speaking exceeded a certain value. As the ground truth source, Silero VAD achieves perfect agreement with its own labels by definition; the thermal classifier's performance reported in Section 4 therefore represents the degree to which thermal physiological signals can replicate an acoustic oracle operating under controlled laboratory conditions. A total of 40,855 frames (approximately 27.2 minutes at 25 FPS) were aggregated from twenty distinct recording sessions, encompassing both single- and multiple-subjects. The single-subject data (17 distinct recording sessions, $\approx 28,000$ frames) formed the foundation for our Generalization Test Set. For rigorous model evaluation, the entire dataset was partitioned using Leave-One-Experiment-Out (LOEO) cross-validation. This methodology ensured strictly subject- and experiment-independent testing by training the model on the data from 16 experiments and validating its performance on the single remaining, unseen experiment. Six model types were considered during LOEO: Fusion Random Forest (RF) (Liaw and Wiener, 2002), Visual-Only RF, Thermal-Only RF, Fusion Multi-Layer Perceptron (MLP) (Chan et al., 2023), Fusion Support Vector Machine (SVM) (Noble, 2006), and Histogram-based Gradient Boosting Classification (HGBC) (Guryanov, 2019).

Frame-level VAD is inherently susceptible to noise, resulting in intermittent suppression of Speech_Present flags during continuous speech segments due to brief drops in vocal energy dur-

ing breath inhalation or soft phoneme articulation. Rather than applying a rolling mean to the binary signal, we addressed this at the source by setting Silero VAD's internal confidence threshold to 0.70, which suppresses low-confidence transient detections while preserving sustained speech intervals. Each detected speech interval was then extended forward by 0.3 seconds (1–2 frames at 5 fps) to capture residual thermal signal that persists briefly after speech cessation. To compensate for an initial estimate of the audio-thermal synchronization lag, labels were shifted 2 frames earlier prior to assignment; subsequent manual alignment revealed the true lag to be 12 frames, as discussed in Section 4.2. Finally, during training, segments shorter than 8 frames (1.6 seconds) were excluded to remove transition-zone frames where the lag correction is insufficient to guarantee correct labeling. Before this segment filter, approximately 77% of frames carried a TRUE label; after filtering, the proportion remained stable at 76–78% across sessions, confirming that the filter removed short segments of both classes proportionally rather than biasing the label distribution.

• Thermal Data: Extracted features from thermal camera at 25 frames per second (AMPBANK M256 thermal camera, 256x196 pixels, 56^o x 42^o Field of View) focused on the thermal gradient changes around the mouth-nasal regions, indicating exhalation and subtle muscle movements during speech and sent to a local Raspberry Pi 5 8GB.

• Audio Data: Audio from a single USB microphone.

• Preprocessing: All data streams were normalized and aligned temporally before being passed to the ML architecture. The YOLOv8 model (Vemulapalli et al., 2023) was used for face, nose, and mouth detection in each thermal video frame as the frames were assembled and aligned with the audio stream. Thermal data was extracted and stored into a CSV file.

3.2 Feature Engineering and Preprocessing

The raw thermal and visual data were rigorously preprocessed to derive the final feature set used in the analysis. Crucially, due to the high volatility of instantaneous thermal readings, we introduced Temporal Rolling Average (RA) features to capture the sustained thermal signature of speech activity. Table 1 lists the thermal and visual features collected and calculated. These features do not just detect activity; they provide the temporal 'gating' required to provide corroborating evidence of ASR events when a specific student is speaking, thereby reducing 'hallucinations' or noise-injection.

3.3 Model Architecture and Optimized Validation

We evaluated the fusion approach using a refined dataset of 17 single-speaker subject-sessions (28,000 frames) to establish a reliable generalization baseline. A rigorous Leave-One-Experiment-Out (LOEO) cross-validation was employed. After systematic hyperparameter tuning of candidate architectures—including Histogram-based Gradient Boosting Classification (HGBC) and Multi-Layer Perceptron (MLP)—the Random Forest (RF) classifier emerged as the optimal architecture. HGBC is a ML algorithm designed to train efficiently on large datasets. It speeds up the training process by grouping continuous feature values into discrete bins, which simplifies how decision trees determine split points (Guryanov, 2019). To maximize temporal stability, a 125-frame Rolling Average (RA) window was applied to the final 19-feature set, as a temporal sweep confirmed this window size provided the highest subject-independent generalization.

4 Results and Performance Analysis

The comparative classifier evaluation study (Table 2) confirms that thermal-only and thermal-visual fusion significantly outperform the visual-only baseline.

4.1 Discussion of Performance Metrics

The Thermal-Only RF (no *MAR_Variance*) achieved F1=0.8105 and Recall=0.9234 under rigorous LOEO cross-validation, confirming the system captures physiological markers of respiration and articulation rather than superficial environmental cues. The 11.64% F1 gain over the visual baseline (Visual-Only_RF) validates that thermal features provide non-redundant complementary information. While Precision (0.7308) indicates some false alarms from heavy-breathing artifacts, the high Recall confirms thermal signatures are the primary drivers for capturing continuous speech indicators. The Fusion_RF model ranked second, with *MAR_Variance* providing the greatest visual contribution.

Feature	Type	Signal captured	Speech vs. quiet
Optimised temporal thermal cues (rolling averages)			
Avg_Mouth_Temp_RA$_{125}$	Thermal	Mean mouth temp, 125-frame rolling avg	Elevated during speech
Avg_Nostril_Temp_RA$_{125}$	Thermal	Mean nostril temp, 125-frame rolling avg	Lower during speech
Center_Mouth_Temp_RA$_{125}$	Thermal	Center-mouth temp, 125-frame rolling avg	Elevated during speech
Engineered thermal contrast			
Delta_AvgTemp_Mouth_Roll_Avg	Thermal	Rolling mean face−mouth temperature delta	Increases during speech
Breathing_Power_In_Band_RA	Spectral	Nostril spectral power, 0.1–0.5 Hz band	Lower during speech
Breath_BRSI	Spectral	Nostril rhythm stability (100-frame FFT)	Lower during speech
Delta_Mouth_Min	Thermal	Mean mouth temp minus min face temp	Higher during speech
Center_Mouth_Contrast	Thermal	Center-mouth / mean-mouth temperature	Higher during speech
Center_Mouth_Delta_Face	Thermal	Center-mouth minus mean face temperature	Higher during speech
Instantaneous thermal components			
Avg_Mouth_Temp_C	Thermal	Mean mouth ROI temperature (current frame)	Elevated during speech
Avg_Nostril_Temp_C	Thermal	Mean nostril ROI temperature	Lower during speech
Avg_Temp_C	Thermal	Mean temperature across full face ROI	Baseline reference
Center_Mouth_Temp_C	Thermal	Center-mouth temperature (current frame)	Elevated during speech
Min_Temp_C	Thermal	Minimum temperature across face ROI	Lower during speech
Nostril_Temp_Diff_Min	Thermal	Mean nostril temp minus min face temp	Higher during speech
Face_Temp_Spread	Thermal	Max−min temperature across face ROI	Higher during speech
Nostril_Temp_Max_Slope	Thermal	Max $\lvert dT/dt \rvert$ nostril, 100-frame window	Lower during speech
Mouth_Temp_Max_Slope	Thermal	Max $\lvert dT/dt \rvert$ mouth, 100-frame window	Higher at onset
Visual mouth motion			
MAR_Variance	Visual	Mouth aspect ratio variance, 25-frame	Higher during speech

Revised feature set: A subsequent 13-feature revision replaces the three RA$_{125}$ features and redundant instantaneous temperature columns with slope-based dynamics (*Airflow_Dominance*, *Slope_Coupling*, *Mouth_Slope_Accel*) and spectral coupling features (*NM_Xcorr*, *Mouth_Breath_BRSI*). A 12-feature thermal-only version drops *MAR_Variance*. Results using the revised set are reported in Table 2 and discussed in Section 4.3.

Table 1: Thermal and visual features used in the original 19-feature classifier. S_m=Max $|dT/dt|$ mouth, S_n=Max $|dT/dt|$ nostril: max temperature slopes in mouth and nostril ROIs. RA$_{125}$ denotes a 125-frame rolling average.

4.2 Feature Importance Analysis

For the original 19-feature model (Table 1), the Permutation Feature Importance rankings (Table 3) provide a clear "biological" map of how the model makes these decisions. Feature importance analysis substantiates the top thermal cues that drive the model's predictive power. The top feature, *Center_Mouth_Delta_Face*, highlights that the temperature difference between the mouth and the rest of the face is the most critical indicator of speech. This is closely followed by *Breathing_Power_In_Band_RA* which is calculated over a continuous rolling window of 500-frame and then a 125-frame window for smoothing, which confirms that rhythmic respiratory signals captured by the thermal sensor are vital for distinguishing speech from silence. Notably, the many negative values at the bottom of the list (such as *Mouth_Temp_Max_Slope*) reveal multicollinearity: because many thermal features are highly correlated, the model sees them as redundant. These results suggest the model might be simplified by focusing only on the top 5–6 unique thermal and breathing markers without losing significant performance. Correlation analysis reveals that the feature set naturally separates into 3 partially independent clusters corresponding to baseline facial thermal state, thermal gradients associated with respiratory airflow, and breathing rhythm and motion dynamics. This separation suggests that the classifier exploits complementary physiological signals rather than redundant measurements.

The visual cue, *MAR_Variance*, ranked first overall in a Permutation Importance Analysis of the Fusion RF Model (results not shown here). *MAR_Variance* captures continuous lip deformation changes rather than binary motion events. Together, the Permutation findings for the Thermal-Only RF and Fusion RF models demonstrate that both model's performance gains arise from carefully engineered feature sets that utilize stable thermal physiology or refined visual motion descriptors, leveraging the strengths of both modalities.

Manual alignment of VAD labels against thermal frames revealed a systematic 12-frame lag (2.4 s at 5 fps) from Silero VAD onset latency combined with audio-thermal synchronization offset. Correcting this lag in a revised 12-feature thermal-only

Model	F1	Accuracy	Precision	Recall	ΔF1	Training Details
Visual-Only_RF	0.7260	0.6196	0.7413	0.7150	–	500 trees (tuned)
Thermal-Only_RF[a]	0.8105	0.7079	0.7308	0.9234	+11.64%	500 trees (tuned)
Thermal-Only_RF(revised)[b]	0.8300	0.7407	0.7830	0.8967	+14.33%	500 trees (tuned)
Fusion_MLP	0.6574	0.5878	0.7240	0.6286	–9.45%	layer_sizes=(100,50,25), max_iter=500
Fusion_HGBC_Opt	0.6921	0.6455	0.7475	0.7011	–4.67%	100 epochs
Fusion_SVM	0.7048	0.6369	0.7204	0.7385	–2.92%	optimized-based
Fusion_RF	0.8047	0.7000	0.7270	0.9183	+10.84%	500 trees (tuned)

[a] Silero VAD threshold=0.70, 0.3s hysteresis tail, 2-frame onset correction, 8-frame transition-boundary exclusion (76.2% TRUE frames). Without segment filter: F1=0.8133, Accuracy=0.7192, Precision=0.7734, Recall=0.8693.

[b] Revised 12-feature thermal-only set, 18 sessions including one balanced session (37% speech); *NM_Xcorr* and *Mouth_Breath_BRSI* rise to top-2 permutation importance after 12-frame VAD lag correction (see Section 4.2); adding *MAR_Variance* gives the 13-feature Fusion RF variant (F1 = 0.8267).

Table 2: Model performance comparison.

model elevated *NM_Xcorr* (nostril–mouth temperature correlation, which decouples during speech as oral exhalation displaces nasal breathing) to rank 1 (+0.030) and *Mouth_Breath_BRSI* (oral breathing rhythm stability, disrupted by phonation) to rank 2 (+0.028), with 11 of 12 features positive; the sole exception, *Nostril_Temp_Max_Slope* (-0.005), is subsumed by *NM_Xcorr* and *Airflow_Dominance* which capture nostril-mouth dynamics more directly. Both features had shown negative importance in prior runs attributed to insufficient quiet data; the lag correction reveals systematic label misalignment as the true cause, validating thermal channel decoupling as the dominant physiological marker of speech. Revised results are shown in Table 2 and discussed in Section 4.3.

4.3 Future Work and Generalization

Iterative Feature Refinement. Correlation analysis of the original 19-feature set (Table 1) revealed substantial redundancy among the three 125-frame rolling average temperature features and the instantaneous temperature columns from which they are derived, as these cluster into a single highly-correlated group encoding overall facial thermal state rather than speech-specific dynamics. A revised 12-feature thermal-only set replaces these redundant columns with slope-based dynamics (*Airflow_Dominance*, *Slope_Coupling*, *Mouth_Slope_Accel*) and spectral coupling features (*NM_Xcorr*, *Mouth_Breath_BRSI*) that more directly characterize the physiological contrast between oral speech exhalation and quiet nasal breathing. Evaluated on 18 sessions—including one session with extended alternating quiet and speech periods (37% speech, compared to a mean of 77% across the original 17 sessions)—and with an 8-frame minimum segment filter and 12-frame VAD label correction applied at training time, the revised Thermal-Only RF achieved F1 = 0.8300, Accuracy = 0.7407, Precision = 0.7830, and Recall = 0.8967 (Table 2). The 12-frame offset (2.4 s at 5 fps) reflects the combined effect of Silero VAD's onset latency and audio-thermal synchronisation delay, as identified through manual annotation alignment and discussed in Section 4.2. Transitioning this approach into a high-confidence classroom solution requires refinement in three key areas.

4.3.1 Feature and Tracking Robustness

The success of the temperature differences and rhythmic respiratory signals confirms that temporal context is paramount. Future iterations will explore advanced time-series features to better characterize the frequency of speech expulsion. Furthermore, replacing the current 5-point YOLOv8n landmarks with high-density facial models (e.g., 68-point landmarks) could increase the predictive weight of visual features like *MAR_Variance*. Finally, to handle multi-student interactions, we intend to replace our basic tracker with Deep SORT (Wojke et al., 2017) to maintain subject identity across thermal transients and occlusions. An initial effort with Deep SORT maintains identity through head turns via joint appearance embedding and Kalman-filter position prediction, holding tracks alive across brief occlusions rather than creating new identifiers when a face temporarily exits the detector.

4.3.2 Dataset Expansion and Kinematics

By expanding the dataset to include longer data collection sessions, a broader population, and collecting data across various speech styles and ambi-

Rank	Feature	Avg.	Rank	Feature	Avg.
1	*Center_Mouth_Delta_Face*	0.0031	10	*Center_Mouth_Temp_C*	-0.0014
2	*Breathing_Power_In_Band_RA*	0.0030	11	*Delta_AvgTemp_Mouth_Roll_Avg*	-0.0024
3	*Nostril_Temp_Diff_Min*	0.0018	12	*Avg_Temp_C*	-0.0026
4	*Avg_Nostril_Temp_C*	0.0017	13	*Min_Temp_C*	-0.0027
5	*Delta_Mouth_Min*	0.0012	14	*Avg_Mouth_Temp_C*	-0.0035
6	*Breath_BRSI*	0.0004	15	*Center_Mouth_Temp_125Frame_RA*	-0.0038
7	*Center_Mouth_Contrast*	-0.0006	16	*Avg_Nostril_Temp_125Frame_RA*	-0.0041
8	*Nostril_Temp_Max_Slope*	-0.0011	17	*Avg_Mouth_Temp_C_125Frame_RA*	-0.0053
9	*Face_Temp_Spread*	-0.0012	18	*Mouth_Temp_Max_Slope*	-0.0067

Table 3: Permutation feature importance analysis (Thermal RF model). *MAR_Variance* (visual) excluded from Thermal-Only RF; 18 remaining features shown.

ent temperatures, the model will be forced to learn the universal physical principles of speech kinematics and thermal signatures. This is essential for achieving a high-confidence, deployable solution that generalizes across multiple individuals and conditions.

4.3.3 Integrated Multimodal Framework for Classroom Dynamics

Our long-term goal is to transition from a supplementary thermal VAD to a fully integrated framework for analyzing peer-learning dynamics such as Palmer et al. (2025). Thermal signatures will be fused with gaze, posture, and gesture recognition (VanderHoeven et al., 2023) to triangulate referential intent, and coupled with lip-reading (Guan et al., 2025) and diarization (Wang et al., 2025) to improve acoustic resilience in high-noise classrooms. This combined approach should achieve better results at speech and engagement detection.

5 Towards a Real-Time Implementation and Deployment – Engagement Detection System

5.1 Edge-Computing Architecture

The objective of this real-time system is to assess whether a thermal-visual VAD can reliably identify one form of student engagement, defined here as the presence of speaking activity overlapping with true speaking episodes. The system is not intended to perform precise speech segmentation; instead, it aims to robustly detect whether a student speaks during an interaction which can indicate active engagement in a peer-learning activity. The evaluation of the system focuses on episode-level detection and engagement-oriented metrics rather than frame-level accuracy that was pursued during classifier training.

To evaluate the practical utility of the thermal-visual fusion model, we developed a custom Python runtime on a local and cost-effective Raspberry Pi 5. The system utilizes an AMPBANK M256 thermal camera. Facial landmarks are isolated using a YuNet/YOLOv8-face algorithm, while thermal statistics from the mouth and nasal regions are extracted and summarized over the 125-frame rolling window.

5.2 Adaptive Calibration and Hysteresis

To improve classification robustness across individuals and recording environments—and to account for deployment conditions that may differ from those encountered during classifier training—we introduced a silent buffering period followed by an interactive calibration phase at the beginning of each session. The buffering period, during which participants remained silent, lasted 20 or more seconds (depending on the frame rate achieved by the Raspberry Pi 5 processor) and served to stabilize the thermal camera and determine participants' breathing rates while allowing participants to acclimate to the data collection environment. During the subsequent calibration phase, participants first remained silent for 5 seconds and then spoke continuously for 5 seconds. Median classifier prediction probabilities from these two segments were used to compute an adaptive decision threshold T_{base}.

This buffering and adaptive calibration strategy successfully stabilized speaking-state detection for a single speaker across short (one to three minute) test sessions under varying ambient conditions. However, thermal speech detection remains sensitive to head and upper-body motion: movement can displace the air surrounding the face, mouth, and nostrils, causing thermal signatures to lag behind physical landmarks and resulting in transient temperature artifacts. Such effects can lead to false-positive speech detections.

To address this, the speech detection pipeline fol-

lows a staged decision process designed to enforce temporal stability under noisy thermal conditions. First, the Random Forest classifier produces a raw frame-level speech probability. This signal is temporally smoothed using a rolling median filter to suppress short-lived fluctuations. Next, a motion-gating mechanism (Meyers et al., 2022) computes the Euclidean distance between the tracked face bounding box centers in consecutive frames. If the displacement exceeds a predefined threshold, a Motion Lock is triggered, and the current thermal contribution is suppressed by substituting a neutral probability value. This prevents thermal transients caused by head movement from influencing the speech decision.

The stabilized probability is then passed through a hysteresis-based state machine, a standard control strategy for preventing rapid state oscillations (Meister et al., 2015), with separate entry and exit thresholds derived from T_{base}. This mechanism prevents rapid state oscillations near the decision boundary—analogous to thermostat control—by requiring stronger evidence to enter the speaking state than to remain in it.

5.3 Engagement Confidence Index

Because the objective of the classroom system is to identify active peer interaction rather than precise speech segmentation, we developed the Engagement Confidence Index (ECI). ECI is a composite score that estimates the likelihood and strength of a user's future engagement based on observed behavioral signals and interaction patterns. This metric prioritizes episode-level detection over frame-level accuracy. ECI and a subset of related metrics we used are defined in Table 4.

Metric	Definition
Engagement Confidence Index (ECI)	Multi-objective engagement score over a temporal window; clamped to $[0, 1]$ with weights $w_1 = 0.5, w_2 = 0.3, w_3 = 0.2$: $\text{clamp}\left(\begin{smallmatrix} w_1 R_{\text{ep}} + w_2 C_{\text{speech}} \\ + w_3 S_{\text{temp}}, 0, 1 \end{smallmatrix}\right)$
Episode Recall (R_{ep})	Fraction of GT speech episodes matched by a predicted ENGAGED episode (IoU ≥ 0.5, one-to-one, ± 1 frame tolerance): $R_{\text{ep}} = N_{\text{matched}}/N_{\text{GT}}$
Coverage (C_{speech})	Temporal overlap between predicted engagement and GT speech: $\sum_i (\text{Engaged}_i \cdot \text{GT}_i) / \sum_i \text{GT}_i$
Temporal Stability (S_{temp})	Penalty for rapid state flicker where $F = \text{Transitions}/\text{Duration}_{\text{sec}}$: $e^{-0.5F}$

Table 4: Engagement Confidence Index (ECI) and component metrics.

5.4 Runtime Results and Discussion

Only a small initial test has been conducted using three different subjects and five thermal video sessions. Figure 2 shows probability trajectories and adaptive decision boundaries during runtime for a single subject. For the evaluated session, Engagement Recall was 0.620, Engagement Precision was 0.795, temporal stability score was 0.826, Episode Recall was 0.400, and ECI was 0.551. The gap between frame-level Recall (0.923 under LOEO cross-validation; 0.788 in the Figure 2 runtime session) and Episode Recall (0.400) reflects that short episodes fall below the temporal smoothing threshold or are interrupted by Motion Lock; the conservative calibration strategy favors precision over episode-onset sensitivity. These results suggest the acoustic-independent thermal stream, stabilized by temporal smoothing, provides a useful engagement signal in real time.

The Episode Recall of 0.400 warrants interpretation in the context of the deployment pipeline. A speech episode is counted as detected only if a predicted ENGAGED interval overlaps the ground-truth episode by at least 50% (IoU ≥ 0.5, where IoU is the ratio of the temporal intersection to the union of predicted and ground-truth episode boundaries). Three pipeline stages each independently reduce the probability of achieving this threshold. First, the 125-frame rolling window introduces an onset delay before features stabilize at the start of a new speaking episode, causing the classifier to respond slowly at speech boundaries. Second, the Motion Lock mechanism suppresses thermal contributions during head movement, which frequently co-occurs with speech onset as a speaker turns to address a peer. Third, the conservative hysteresis entry threshold—set to favor precision over recall to minimize false-positive alerts to the AI agent—requires sustained evidence before transitioning to the ENGAGED state, penalizing short episodes. Together these stages create a system that detects the interior of long speaking episodes reliably (reflected in the 0.923 frame-level Recall) while frequently missing the onset of short or movement-accompanied episodes (reflected in the 0.400 Episode Recall). For the classroom engagement use case this tradeoff is acceptable: the system is designed to confirm sustained engagement rather than react to every utterance, and false-

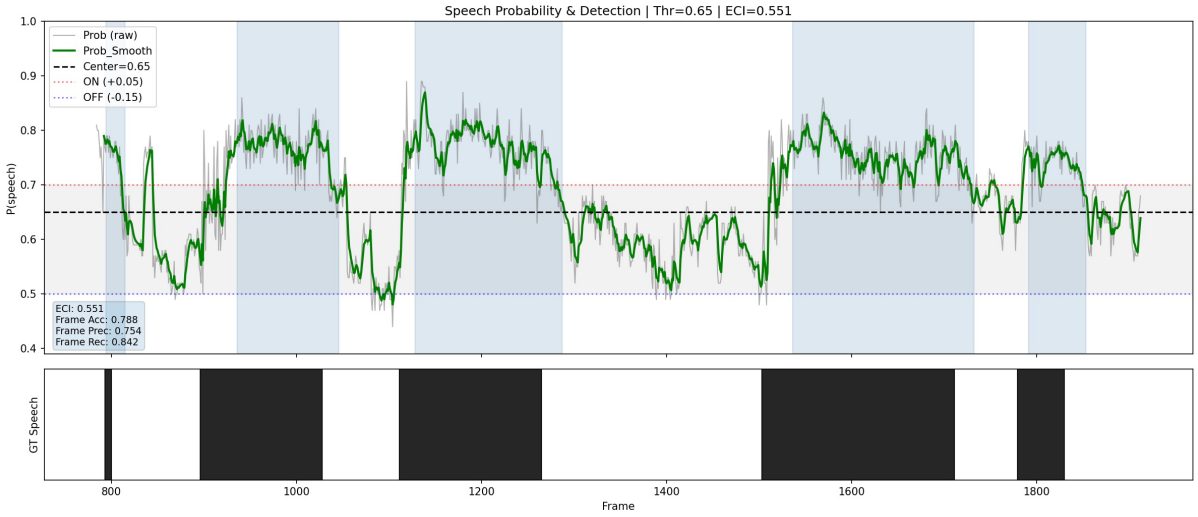

Figure 2: Real-time speech probability trajectories for a single subject. Gray: raw RF output; green: 5-frame smoothed probability; black dashed: calibrated threshold T_{base}; red/blue dotted: hysteresis ON/OFF boundaries; light blue shading: predicted speaking intervals. Ground truth shown below (black = speaking, white = quiet).

positive interventions by an AI agent are more disruptive than missed short episodes.

Figure 3 summarizes the overall system architecture, highlighting the interaction between offline model training and the real-time inference pipeline incorporating buffering, adaptive calibration, motion gating, and hysteresis.

6 Conclusion and Future Impact

This research moves towards a high-confidence, acoustic-independent solution for detecting student speech in noisy, multi-peer environments. By successfully validating the fusion concept on a Raspberry Pi 5, we demonstrate the feasibility of an edge-computing architecture for classroom deployment. Future work will transition this proof-of-concept into a split-computing framework where the Raspberry Pi performs real-time feature extraction while a remote pipeline (Zhu et al., 2026) on an HPC platform fuses multi-modal components—gaze, posture, gesture, speaker diarization, and spoken language understanding—to build educational applications. The future work will integrate the thermal-visual front-end with large language models (LLMs) to provide real-time feedback on group dynamics, enabling AI agents to facilitate more equitable and linguistically-rich peer learning interactions.

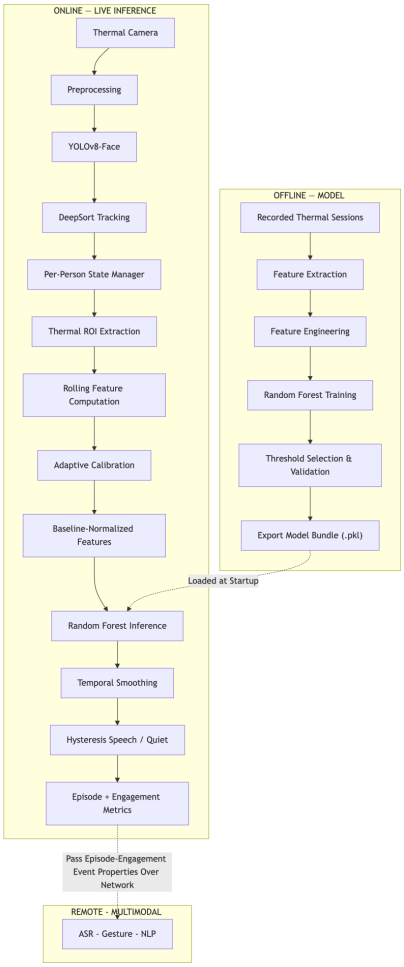

Figure 3: Online live inference (real-time) engagement detection system, remote multimodal processing, and offline model training system.

Limitations

The current focus of this research was modeling each person's thermal properties in a classroom peer group to determine whether or not someone was speaking and whether or not they were engaged in the learning activity. Some limitations exist in the current implementation:

- Thermal camera resolution and field-of-view become an issue as the scene includes more people. The ROIs will cover fewer pixels which reduces the quality of thermal data. We must explore how this affects our models if we are to continue to utilize inexpensive thermal cameras.
- A simplistic face tracker was employed but it was not effective in continuously distinguishing each person as the same person especially when one turned his or her head quickly or moved. For that reason, the current research results focused on one person to better model thermal characteristics of speaking or not speaking. Future work must examine multiple participants in a scene since that is what one expects in a classroom peer-learning event. Better face tracker algorithms, such as Deep SORT, are available. We are currently modifying a version of Deep SORT to handle not only multiple participants but to avoid confusion when someone turns their head or looks away, thinking they are now someone else. We have plans to collect multiple participant data once it is working.
- Our real-time component requires per-person calibration at the beginning of a session. This is time-consuming and provides a poor user-experience but necessary given different ambient conditions. A more user-friendly approach is needed.
- The current training data was collected in a laboratory and not in a classroom where different room temperature characteristics and ventilation conditions could be expected (e.g., fans running in the classroom blowing air towards the peer group). A broader and much larger dataset, with more participants and varied collection environments, is necessary for full validation of our approach. We have plans to collect such data.
- The data collected for training was only from adults and not the targeted middle school students due to privacy concerns yet school-age children's thermal properties are likely to differ from those of adults. The face/mouth/nostril region detection was computed on individually-stored thermal video frames in a separate detection phase after the data collection phase to preserve the 25 fps thermal camera capture rate. The addition of a Raspberry Pi AI module could increase the frame rate substantially for real-time detection so that no potentially personally-identifiable information (namely the thermal video frames of a student) would need to be retained. Instead, only the thermal temperature data and the coordinates of the face/mouth/nostril regions would need to be stored for use in calculating the features used for training the classifier. Similarly, the real-time thermal VAD/engagement system could perform all calculations without retaining the thermal video frames. We are testing the Raspberry Pi AI module and have had great results for face/mouth/nostril detection achieving 25 fps. However, the AI module does not speed up the trained classifier limiting the system to 5 fps.
- Currently, we have only employed our Visual-Only_RF model as a baseline but recognize another baseline could provide valuable insight. We tested TalkNet as a baseline and found it does not run in real-time and cannot detect facial regions in thermal false-color imagery, as the audio-visual synchrony model relies on visible-spectrum lip texture absent in thermal imaging. Future work will add a parallel RGB camera to enable a proper audio-visual comparison on aligned video footage to provide another basesline.
- We currently only explore a laboratory setup as proof-of-concept. We will perform additional validation in more realistic settings once the full iSAT pipeline is tested.

Acknowledgments

I would like to acknowledge the Brandeis iSAT team for helpful comments on my draft. This research was conducted as a volunteer on the NSF AI Institute for Student AI Teaming (iSAT) grant, NSF DRL Award 2454151. *AI assistance disclosure: Claude AI, ChatGPT, and Google Gemini assisted in software development and editing this document to fit page limitations.*

References

M. Abdrakhmanova, A. Kuzdeuov, S. Jarju, Y. Khassanov, M. Lewis, and H. A. Varol. 2021. *SpeakingFaces: A Large-Scale Multimodal Dataset of Voice Commands with Visual and Thermal Video Streams.* *Sensors*, 21(10).

R. S. Baker and J. Ocumpaugh. 2014. *Cost-effective, actionable engagement detection at scale.* In *Educational Data Mining*.

D. Bennett, S. Cahlan, and D. Taylor. 2020. *Military-grade camera shows risks of airborne coronavirus spread.* *Washington Post* (December 11, 2020).

N. Bosch, S. D'Mello, J. Ocumpaugh, R. Baker, and V. Shute. 2016. *Using video to automatically detect learner affect in computer-enabled classrooms.* *ACM Transactions on Interactive Intelligent Systems*, 6.

H. Cao, Š. Beňuš, R. C. Gur, R. Verma, and A. Nenkova. 2014. *Prosodic cues for emotion: Analysis with discrete characterization of intonation.* In *Proceedings of Speech Prosody 2014*, pages 130–134.

J. Cao, A. Ganesh, J. Cai, R. Southwell, E. M. Perkoff, M. Regan, K. Kann, J. H. Martin, M. Palmer, and S. D'Mello. 2023. *A comparative analysis of automatic speech recognition errors in small group classroom discourse.* In *Proceedings of UMAP 2023*, pages 1–13.

K. Y. Chan, B. Abu-Salih, R. Qaddoura, A. M. Al-Zoubi, V. Palade, D.-S. Pham, J. Del Ser, and K. Muhammad. 2023. *Deep neural networks in the cloud: Review, applications, challenges and research directions.* *Neurocomputing*, 545:126327.

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

B. A. Goodman, J. Drury, R. D. Gaimari, L. Kurland, and J. Zarrella. 2006. *Applying user models to improve team decision making.* Technical Report MTR 060150, MITRE Corporation, Bedford, MA.

Y. Guan, V. A. Trinh, V. Voleti, and J. Whitehill. 2024. *Multi-modal speech transformer decoders: When do multiple modalities improve accuracy?* *arXiv:2409.09221*.

Y. Guan, V. A. Trinh, V. Voleti, and J. Whitehill. 2025. *MLLM-based speech recognition: When and how is multimodality beneficial?* *arXiv:2507.19037*.

A. Guryanov. 2019. *Histogram-Based Algorithm for Building Gradient Boosting Ensembles of Piecewise Linear Decision Trees.* In *Analysis of Images, Social Networks and Texts: AIST 2019*, pages 39–50.

S. Ioannou, V. Gallese, and A. Merla. 2014. *Thermal infrared imaging in psychophysiology: Potentialities and limits.* *Psychophysiology*, 51(10):951–963.

J. Kwon, O. Kwon, K. Taek Oh, J. Kim, and S. K. Yoo. 2023. *Breathing-Associated Facial Region Segmentation for Thermal Camera-Based Indirect Breathing Monitoring.* *IEEE Journal of Translational Engineering in Health and Medicine*, 11:505–514.

A. Liaw and M. Wiener. 2002. *Classification and regression by randomForest.* *R Journal*, 2(3).

M. C. T. Manullang, Y.-H. Lin, S.-J. Lai, and N.-K. Chou. 2021. *Implementation of thermal camera for non-contact physiological measurement: A systematic review.* *Sensors*, 21(23):7777.

A. Meister, S. Büchner, and A. Amthor. 2015. *State machine based nonlinear hysteresis model.* *Mechatronics*, 31:215–221.

S. M. Meyers, K. Kisling, T. F. Atwood, and X. Ray. 2022. *A standardized workflow for respiratory-gated motion management decision-making.* *Journal of Applied Clinical Medical Physics*, 23(8):e13705.

W. S. Noble. 2006. *What is a support vector machine?* *Nature Biotechnology*, 24(12):1565–1567.

M. Oztoprak. 2022. *Engagement detection using prosodic cues.* Master's thesis, Northern Illinois University.

D. Palmer, Y. Zhu, K. Lai, H. VanderHoeven, M. Bradford, I. Khebour, C. Mabrey, J. Fitzgerald, N. Krishnaswamy, M. Palmer, and J. Pustejovsky. 2025. *Speech is not enough: Interpreting nonverbal indicators of common knowledge and engagement.* In *Proceedings of AAAI 2025*, 39(28):29676–29678.

P. Sharma, J. Shubham, S. Gautam, S. Maharjan, S. R. Khanal, M. C. Reis, J. Barroso, and V. M. J. Filipe. 2023. *Student engagement detection using emotion analysis, eye tracking and head movement with machine learning.* *arXiv:1909.12913*.

Silero Team. 2024. *Silero VAD: Pre-trained enterprise-grade voice activity detector.* GitHub repository.

R. Tao, Z. Pan, R. K. Das, X. Qian, M. Z. Shou, and H. Li. 2021. *Is someone speaking? Exploring long-term temporal features for audiovisual active speaker detection.* In *Proceedings of ACM Multimedia 2021*, pages 1–10.

N. Tran, D. Litman, B. Pierce, R. Correnti, and L. C. Matsumura. 2025. *Improving in-context learning example retrieval for classroom discussion assessment.* In *Proceedings of BEA 2025*.

H. VanderHoeven, N. Blanchard, and N. Krishnaswamy. 2023. *Robust motion recognition using gesture phase annotation.* In *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management*, pages 592–608. Springer.

N. S. Vemulapalli, P. Paladugula, G. S. Prabhat, S. Abhishek, and T. Anjali. 2023. *Face detection with landmark using YOLOv8.* In *Proceedings of ICEFEET 2023*, pages 1–5. IEEE.

J. Wang, S. Dudy, X. He, Z. Wang, R. Southwell, and J. Whitehill. 2024. *Speaker diarization in the classroom: How much does each student speak in group discussions?* In *Proceedings of EDM 2024*, pages 360–367.

J. Wang, S. Dudy, X. He, Z. Wang, R. Southwell, and J. Whitehill. 2025. *Optimizing speaker diarization for the classroom.* *Journal of Educational Data Mining*, 17(1):98–125.

A. N. Wilson, K. A. Gupta, B. H. Koduru, A. Kumar, A. Jha, and L. R. Cenkeramaddi. 2023. *Recent advances in thermal imaging and its applications using machine learning: A review.* *IEEE Sensors Journal*, 23(4):3395–3407.

N. Wojke, A. Bewley, and D. Paulus. 2017. *Simple online and realtime tracking with a deep association metric.* *arXiv:1703.07402*.

X. Xu, D. M. Dugdale, X. Wei, and W. Mi. 2023. *Leveraging artificial intelligence to predict young learner online learning engagement.* *American Journal of Distance Education*, 37(3):185–198.

T. Yoshida, E. Yamazaki, and S. Hangai. 2008. *Spoken Word Recognition from Side of Face Using Infrared Lip Movement Sensor.* In E. André et al. (eds.), *Perception in Multimodal Dialogue Systems*, Lecture Notes in Computer Science, vol. 5078, pages 1–6. Springer.

Y. Zhu, I. Dey, K. Lai, and J. Pustejovsky. 2026. *Nonverbal Behavior Recognition in Multimodal Interactions from RGB Video.* In A. Lücking and A. Mehler (eds.), *Behavioromics*. Springer.

A Appendix

A.1 Updated Tables and Figure for Revised Feature Set

Table 5 shows permutation feature importance for the revised 12-feature model after 12-frame VAD label correction; *NM_Xcorr* (+0.030) and *Mouth_Breath_BRSI* (+0.028) rise to rank 1–2, with 11 of 12 features positive, validating thermal channel decoupling as the dominant physiological marker of speech. Table 6 defines this revised 12-feature thermal-only set, replacing the original 125-frame rolling averages and redundant instantaneous temperature columns with slope-based dynamics and spectral coupling features. Figure 4 shows one sample result applying the revised 12-feature model to detect speech and quiet segments in real time. A per-session logistic recalibration layer, fitted on a short calibration sequence and applied with baseline correction for slow-responding breathing features, compensates for inter-subject variability in the RF classifier and achieves an ECI of 0.724 with mean speech-offset detection latency of 3.4 seconds on this example session. Much more testing will be required to determine if this holds up in practice with different subjects and ambient conditions.

Feature	**Importance**	**Rank**
NM_Xcorr	+0.0304	1
Mouth_Breath_BRSI	+0.0281	2
Breath_BRSI	+0.0096	3
Center_Mouth_Contrast	+0.0058	4
Mouth_to_Face_Delta	+0.0044	5
Nostril_to_Face_Delta	+0.0031	6
Mouth_Slope_Accel	+0.0010	7
Slope_Coupling	+0.0010	8
Mouth_Temp_Max_Slope	+0.0006	9
Airflow_Dominance	+0.0004	10
Face_Temp_Spread	+0.0002	11
Nostril_Temp_Max_Slope	−0.0048	12

Table 5: Permutation feature importance for the revised 12-feature Thermal-Only RF model (18 LOEO folds, 12-frame VAD label correction applied). Importance is mean F1 decrease when feature is permuted. Compare with Table 3 (original 19-feature model).

Feature	Type	Signal captured	Speech vs. quiet
Thermal airflow gradients			
Mouth_to_Face_Delta	Thermal	Z-scored mouth temp minus mean face temp	Higher during speech
Nostril_to_Face_Delta	Thermal	Z-scored nostril temp minus mean face temp	Lower during speech
Face_Temp_Spread	Thermal	Max−min temperature across face ROI (°C)	Higher during speech
Center_Mouth_Contrast	Thermal	Center-mouth / mean-mouth temperature ratio	Higher during speech
Airflow velocity dynamics			
Mouth_Temp_Max_Slope	Thermal	Max $\lvert dT/dt \rvert$ mouth ROI, 100-frame window	Higher at speech onset
Nostril_Temp_Max_Slope	Thermal	Max $\lvert dT/dt \rvert$ nostril ROI, 100-frame window	Lower during speech
Airflow_Dominance	Thermal	$\lvert S_m \rvert \,/\, (\lvert S_m \rvert + \lvert S_n \rvert + \varepsilon)$	Approaches 1 in speech
Slope_Coupling	Thermal	$\min(\lvert S_m \rvert, \lvert S_n \rvert) \,/\, (\max(\lvert S_m \rvert, \lvert S_n \rvert) + \varepsilon)$	Low during speech
Mouth_Slope_Accel	Thermal	Second derivative of mouth temperature	Peaks at onset/offset
Breathing rhythm and spectral coupling			
Breath_BRSI	Spectral	Nostril rhythm stability index (100-frame FFT)	Lower during speech
Mouth_Breath_BRSI	Spectral	Mouth rhythm stability index (100-frame FFT)	Lower during speech
NM_Xcorr	Spectral	Nostril–mouth temperature Pearson correlation	Decouples during speech

Table 6: Revised 12-feature thermal-only set used in the updated classifier. S_m, S_n: maximum temperature slopes in mouth and nostril ROIs; ε: stability constant. Z-scoring of the 2 delta features (normalized by a per-session mean and standard deviation, then clipped to ±3) helps handle session-to-session room temperature variation. Other features like *Face_Temp_Spread* and the slope features might also benefit from z-scoring. *MAR_Variance* (visual) is excluded from the Thermal-Only RF but included in the Fusion RF variant (13 features, F1 = 0.8267).

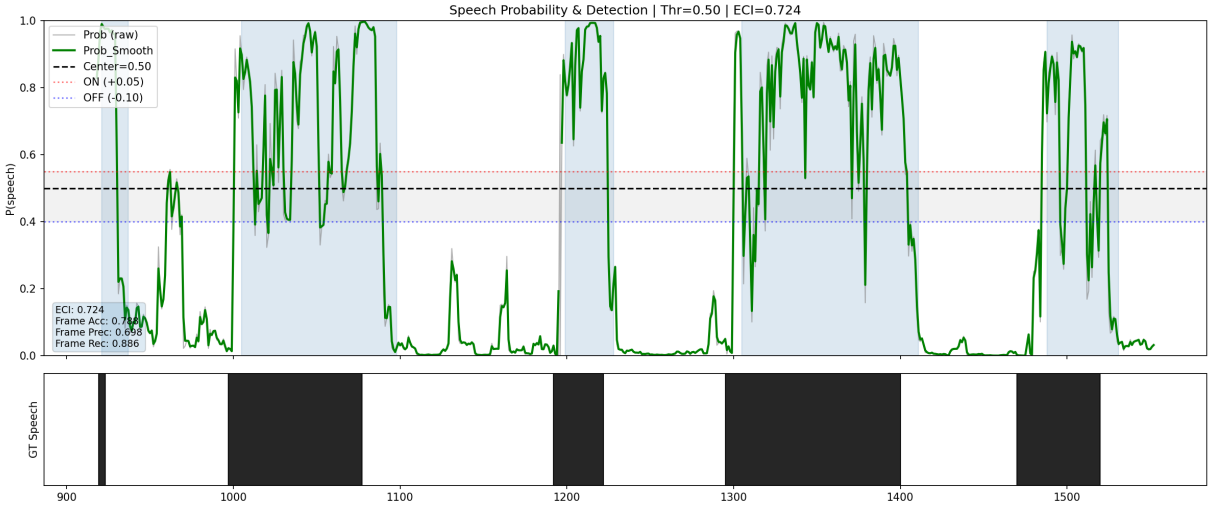

Figure 4: Real-time speech probability trajectories for a single subject using 12-feature thermal-only model. ECI=0.724; Episode Recall=1.0; Frame Precision=0.698; Frame Recall=0.886. Gray: raw RF probability output; green: 5-frame smoothed probability; black dashed: calibrated threshold T_{base}; red/blue dotted: hysteresis ON/OFF boundaries; light blue shading: predicted speaking intervals. Ground truth (GT) shown below (black = speaking, white = quiet).