BOLLD: Body and Oral Language Learning Decoder

Nicole Sorokin
McMaster University
sorokinn@mcmaster.ca

Zuhair Qureshi
McMaster University
quresz23@mcmaster.ca

Julia Brzustowski
McMaster University
brzustoj@mcmaster.ca

Grady Rueffer
McMaster University
ruefferg@mcmaster.ca

Sophia Shantharupan
McMaster University
shants5@mcmaster.ca

***Abstract*—Detecting threatening behavior remains a significant challenge in security and public safety. BOLLD is a multimodal threat detection approach that combines body language analysis, lip reading, and reinforcement learning to assess potential malicious behaviour in real-time. Using MediaPipe for skeletal tracking, a Random Forest classifier and a modified LipNet model, the system evaluates both physical and verbal cues to improve detection accuracy. In testing, BOLLD significantly improved its performance, demonstrating its potential for security applications in environments where audio is unreliable as well as aid individuals with visual impairments by enhancing situational awareness. The project is available at github.com/McMasterAI2024-2025/BOLLD**

I. Introduction

Detecting threatening behavior is a key challenge in security and public safety, but most existing solutions focus on either physical actions or verbal communication rather than both. This project introduces BOLLD (Body and Oral Language Learning Decoder), a system that combines body language analysis and lip reading for real-time threat detection. It uses MediaPipe for skeletal tracking and facial landmark detection, a Random Forest classifier to categorize body poses, and a modified LipNet model to analyze spoken words for potential dangerous actions. A reinforcement learning component further refines detection by integrating physical and verbal cues. This multimodal approach improves real-time processing and could be useful in environments where audio is unreliable, as well as in assistive technology for visually impaired individuals.

A. Motivation

BOLLD takes a multimodal approach to real-time maliciousness detection by combining computer vision, natural language processing, and reinforcement learning. This research is particularly relevant today as AI advancements shape public safety and security measures.

By integrating body language analysis, lip reading, and reinforcement learning, BOLLD detects threats in environments where audio may be unavailable or unreliable, aligning with AI's growing role in cybersecurity and physical security.

Research on AI-powered multimodal search engines demonstrates the effectiveness of combining text, images, audio, and video for situational awareness [1]. Similarly, BOLLD merges visual and verbal cues to assess alarming actions.

Natural language processing (NLP) has proven valuable in Cyber Threat Intelligence by automating large-scale dataset analysis to identify malicious activity [2]. BOLLD extends this by applying NLP to lip-transcribed speech, detecting verbal threats without audio.

Recent studies highlight the predictive potential of AI and NLP in cybersecurity threat detection [3]. These technologies can identify risks early, and BOLLD adapts this capability for physical threat detection by analyzing both body language and spoken content.

Additionally, research into AI-driven tracking and real-time detection in cybersecurity [4] provides a framework that parallels BOLLD's multimodal strategy for identifying physical threats.

By building on these advancements, BOLLD enhances public safety, particularly where audio-based violence detection is ineffective.

B. Related Works

Recent research in multimodal AI has made significant progress in threat detection, particularly in cybersecurity and public safety. This section explores related work and how BOLLD contributes to addressing some of the existing challenges.

Multimodal AI systems integrate different types of data, such as text, images, and geospatial information, to improve the identification of suspicious actions [5]. arge language models (LLMs) like ChatGPT and Gemini have also embraced multimodality, processing and reasoning across text, images, and even audio inputs. By combining multiple sources, these models overcome the limitations of traditional methods, making real-time decision-making more reliable.

AI-driven multimodal search engines have also been explored for cybersecurity applications. These systems use machine learning to analyze security threats from multiple perspectives, but challenges remain in refining their accuracy and efficiency [1].

One recent development is FIRE, a framework designed for few-shot inter-domain threat detection using large-scale multimodal pre-training [6]. This approach helps detect hostility in complex network environments with minimal labeled data, addressing a key issue in cybersecurity.

However, there are still significant challenges in this field, including processing multimodal data in real time, balancing accuracy with computational efficiency, ensuring privacy in surveillance applications, and adapting to evolving aggressive patterns. This integration allows for real-time threat detection in scenarios where audio may be unreliable or unavailable, expanding its potential applications across various security contexts.

C. Problem Definition

Most threat detection systems focus either on physical actions or speech, but rarely consider the connection between body language and spoken words. This gap can make them less effective in real-world situations where audio is unreliable or unavailable, such as in noisy environments, security footage without sound, or meetings where microphones fail.

BOLLD is designed to address these challenges by combining computer vision-based body language recognition with real-time lip reading. By analyzing skeletal motion alongside transcribed speech, BOLLD aims to improve the accuracy of identifying malicious actions. It also incorporates reinforcement learning to refine its predictions over time. Given a sequence of upper body movements and lip motions (X), the system predicts a threat/violence score (y), adjusting dynamically based on behavioral patterns.

By linking physical and verbal cues, BOLLD could be useful in security applications where audio isn't available and in assistive technology for visually impaired individuals who rely on real-time alerts through wearable devices. This approach offers a step toward more adaptive and effective multimodal violence detection, addressing the limitations of systems that rely on a single data source.

II. Methodology

This section outlines the approach used to develop BOLLD, a system that combines body language analysis and lip reading for real-time threat detection. The process includes data collection, model training, evaluation, and continuous refinement to improve accuracy.

BOLLD is built using Streamlit and integrates AI models for analyzing both body language and lip movements in live video. Figure 1 illustrates the data flow. When the system is activated, it initializes session state variables to track violence levels, actions, rewards, and video frames. It also loads pretrained models for body language recognition and lip reading.

The system processes live video input, extracting facial and body landmarks using MediaPipe and dlib for feature normalization. Lip movements are detected, converted to grayscale, and passed through a lip-reading model that transcribes speech roughly every 75 frames. A predefined dictionary assigns violence scores to transcribed words, and a reinforcement learning (RL) agent determines an action, either "all good" or "de-escalate," based on detected danger levels. The RL model continuously updates its Q-table by evaluating past rewards and adjusting its predictions accordingly. Real-time metrics, including threat levels and rewards, are visualized with Plotly, while the video feed is displayed on Streamlit's UI. The system runs continuously, analyzing each incoming frame, until it is manually stopped.

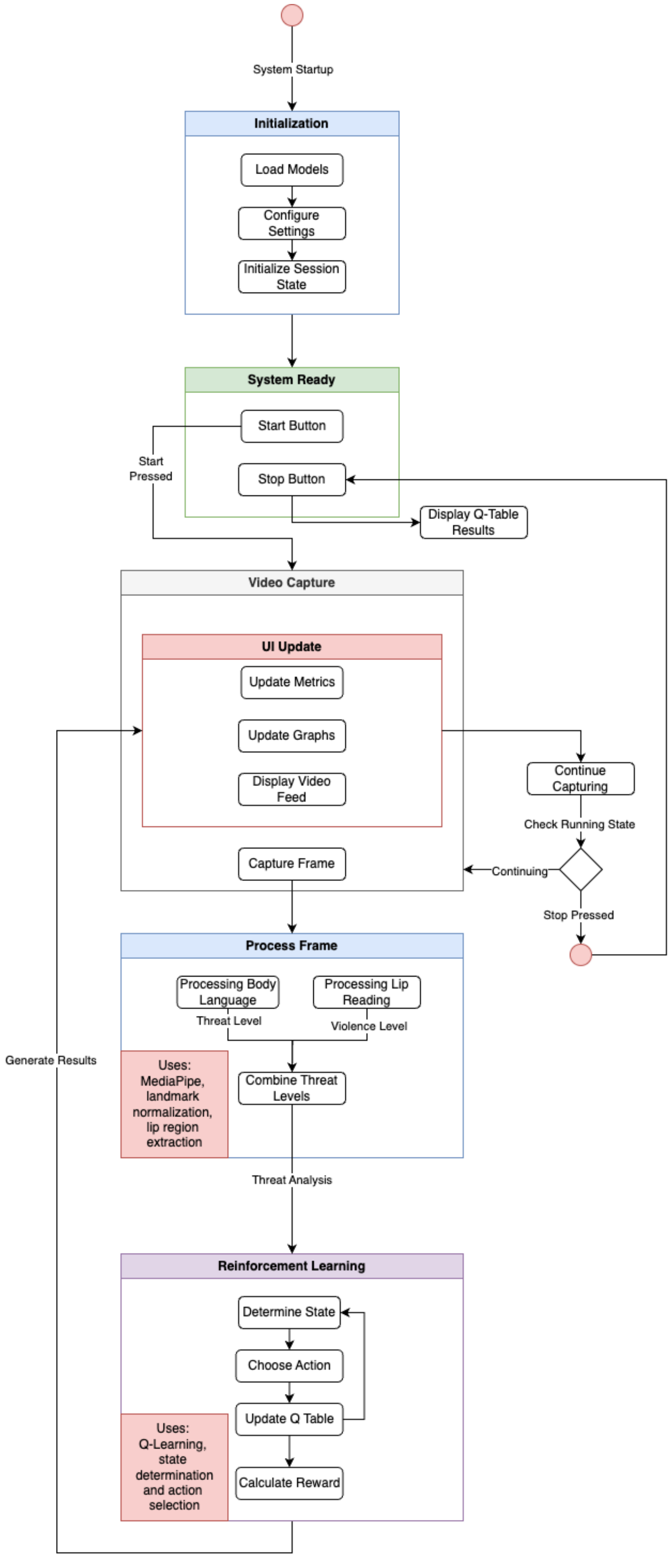


Fig. 1. Process flow diagram.

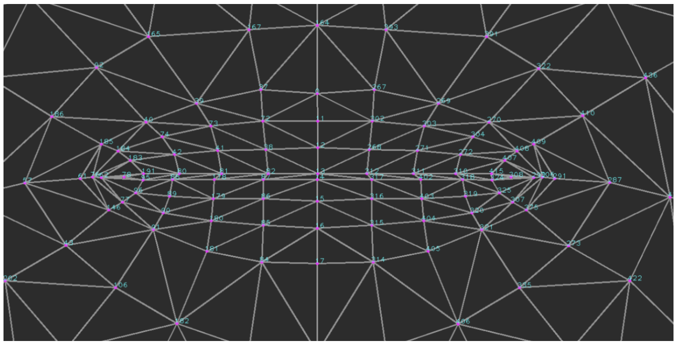


Fig. 2. The MediaPipe mouth/lip area mappings.

A. Body Language Component

The body language model relies on skeletal tracking data generated through Google's MediaPipe computer vision framework (Figure 2). The MediaPipe framework consists of abstracted pre-trained deep learning models that can readily identify skeletal landmarks on the human body in video frames.

Participants were recorded performing various poses. Using MediaPipe, the coordinates of a participant's face, pose, and hand landmarks across each frame in a recording session were written to a dataset with a pre-set label of ”threatening” or ”non-threatening.” The frame coordinates and their labels were stored in a CSV file and used to train a Random Forest classifier, which learns to distinguish between threatening and non-threatening postures based on skeletal coordinates.

To ensure consistency, normalization techniques were applied, adjusting landmark positions relative to a reference body point to minimize position-based distortions. This helped reduce misclassifications caused by variations in user positioning in front of the camera. Inter-landmark normalization was also applied, using the distance between the shoulders as a relative scale.

Additionally, a rolling average mechanism was introduced to smooth fluctuations in predictions across frames and contextualize the instantaneous threat score. The trained Random Forest model was exported using Pickle and deployed for real-time classification. As the system processes video input, predictions are continuously updated and displayed on the UI.

B. Lip Reading Component

The lip-reading model is based on LipNet [7], [8], a deep learning framework that uses convolutional and recurrent neural networks to transcribe speech from visual input. Unlike models that classify individual words, LipNet processes entire sequences, improving accuracy by capturing context over time. The model was modified to work with live video input by storing the 75 most recent frames and using them as input. The oldest 15 frames are then removed, making room for 15 new frames.

The model consists of three spatio-temporal convolutional layers, each followed by a max-pooling layer, which extract spatial and temporal features from lip movement sequences. These features are then processed by two recurrent neural networks, which analyze sequential dependencies. Finally, the Connectionist Temporal Classification (CTC) loss function helps align predicted sequences with the transcribed text, accounting for natural variations in speech (Figure 3).

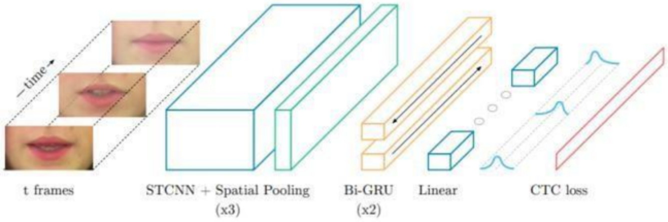


Fig. 3. LipNet architecture [9].

To identify speech that is not friendly, a dictionary of violent keywords was created, assigning each word a predefined violence score. The lip-reading model transcribes speech, and each word is checked against this dictionary. If threatening words appear consistently, the highest detected violence value is passed to the RL model for further decision-making, every 15 frames.

C. Reinforcement Learning Component

To refine threat classification, a reinforcement learning (RL) framework was implemented, allowing the system to adapt to new behaviors over time. The RL model learns to associate physical movements and speech patterns with hostile levels, adjusting its predictions dynamically.

The reinforcement learning environment/approach in this code is simulated using a Q-learning approach to classify behaviors as threatening or non-threatening based on speech and body language. The system maintains a Q-table, which maps states (determined by the detected threat level) to actions (”all-good” or ”de-escalate”). It updates the Q-values using the formula:

$$Q(s,a) = Q(s,a) + \alpha(r + \gamma \max_{a'} Q(s',a') - Q(s,a))$$

where α is the learning rate, γ is the discount factor, r is the received reward, and $\max_{a'} Q(s',a')$ represents the estimated future reward. The system balances exploration and exploitation using an epsilon-greedy strategy, which sometimes chooses random actions to improve learning.

During execution, the model receives input from lip-reading and body language analysis to classify behaviors. If a threatening word is detected in speech or if body language suggests aggression, the system assigns a threat level and chooses an action. After an action is taken, a reward is assigned based on correctness, and the Q-table is updated accordingly. The model saves and loads the best-performing Q-table to improve over time. By continuously updating based on real-time data, the system learns to detect threats more accurately and respond effectively.

The RL components: **State:** Real-time threat conditions, including body language patterns and speech analysis phases. **Action:** Choosing between ”all-good” or ”de-escalate” responses.**Reward:** A combination of correct threat assessment time and appropriate response outcomes.

III. Results

This section presents the model’s performance in detecting threatening and non-threatening behavior. The results are analyzed from different perspectives to evaluate the effectiveness of the approach. We also discuss key findings, improvements made during development, and the impact of specific methodological choices.

A. Parameter Optimization

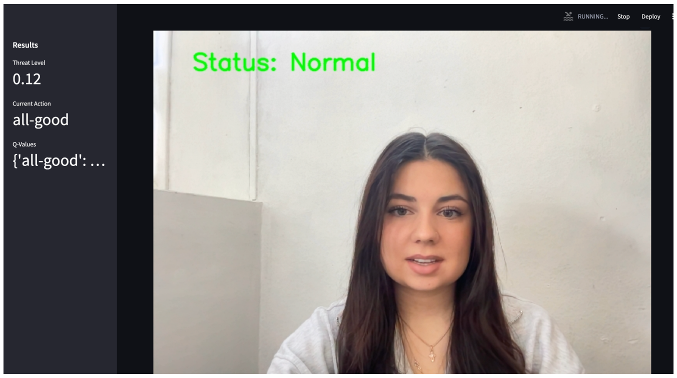
Extensive testing was conducted to fine-tune the reinforcement learning model’s parameters.

For the learning rate, a range of 0.15–0.20 provided a balanced approach with moderate value fluctuations, allowing the model to adjust gradually based on the data, preventing possible overshooting of optimal solutions while learning from the outcomes. This range reduces volatility and adjusts weights appropriately in accordance with the environment. Alternatively, a range of 0.25–0.35 resulted in more stable outcomes with improved interpretability. In this range, the model was capable of converging faster by making larger updates to the weights, resulting in a smoother learning process and reducing random fluctuation, albeit at a marginal cost to responsiveness. Based on this, the optimal learning rate was set at 0.28, providing a reasonable balance into a smoother operation with respect to the adaptability of the model. At 0, the model would not learn from new experiences, while at 1, it would completely overwrite previous learning with each new experience.

The discount rate was tested across multiple ranges. Values between 0.60–0.70 led to faster response times, whereas 0.75–0.85 improved stability and reliability. On the lower range, the model prioritizes more immediate reward as a faster response, with the trade-off of being less reliant on long term interpretation. On the higher range, the model alternatively prioritizes future reward, leading to cautious decision making at the cost of speed. Given the use case of the model, the best trade off was found at 0.71, where balance was shifted more towards quick interpretation.

For the exploration rate, values between 0.25–0.40 were analysed. Higher values in this range led to better responsiveness but introduced slight latency, as the model explores a broader set of actions at a cost of excessive exploration and increased randomness.

Lower values produced more consistent but less adaptive results, as the model relies more on the exploitation of known actions, which provides more consistent results while lowering the models adaptability. The optimal setting was determined to be 0.39, striking a balance between exploration and exploitation with an emphasis on the ability to adapt to human behaviour.



Results

Threat Level

0.12

Current Action

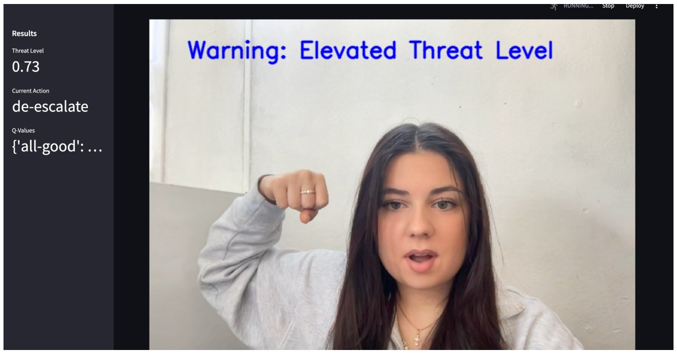
all-good

Q-Values

{'all-good': …

Status: Normal

Fig. 4. An example of model identifying non-threatening behaviour.



Results

Threat Level

0.73

Current Action

de-escalate

Q-Values

{'all-good': …

Warning: Elevated Threat Level

Fig. 5. An example of model picking up on threatening behaviour.

B. Model Performance and Improvements

Early tests revealed inconsistencies in the body language model’s ability to identify threats, mainly due to variations in the user’s distance from the camera. This positional bias affected classification accuracy, leading to a high rate of false positives and false negatives.

To address this, a normalization process was implemented to account for differences in on-screen position. This adjustment reduced classification errors caused by distance by 90%, ensuring that maliciousness detection was based on actual behavior rather than a user’s relative position.

Another major improvement was the introduction of a rolling average for threat scores. Initially, the model classified aggressive actions on a frame-by-frame basis, which caused unpredictable fluctuations. By averaging threat scores over multiple frames, sudden spikes and inconsistencies were significantly reduced, making the predictions more stable and reliable.

For the lip-reading model, increasing the number of frames analyzed per sequence from 50 to 75 resulted in noticeable improvements. It was a challenging task to find the ideal number of frames such that the model gets enough context but where it’s also not taking too long to update the frames. This adjustment provided the model with greater temporal context, enhancing its ability to recognize speech patterns and improving transcription accuracy. Currently, the lip transcription accuracy remains limited, as this is the first lip-reading AI designed for live video input. It is currently a challenge to detect whole words quickly, accurately and consistently. However, to address this, the current approach leverages phoneme-based analysis. An algorithm was developed to identify specific phonetic patterns, matching them against a predefined dictionary. The LipNet model was not originally trained on live video feed as well as violent, threatening, or profane language, which presents a challenge. Thus, this is breaking into cutting-edge territory where future work will focus on training a new model specifically on violent, threatening, or profane language, which alone should significantly improve transcription.

These refinements collectively led to a 90% reduction in false positives. A major factor behind this improvement was eliminating screen position biases where previously, the model struggled to classify a fist as threatening if it appeared in certain positions on the screen. With positional normalization and rolling average adjustments, gesture recognition became much more accurate.

Figures 4 and 5 demonstrate the models response to threatening and non-threatening behaviour, respectively. It can be clearly seen that a fist and angrily interpreted mouth position causes the model to flag the behaviour as threatening. Whereas a neutral positioned face with no extra hand cues leads the model to identify a non-threatening state.

At the moment, the RL model performance score, which is calculated from the recent reward history, is 72%. This is a good performance because a score over 0.5 means that the model is making more correct decisions than incorrect ones.

TABLE I
Q-VALUES FOR DIFFERENT THREAT LEVELS AND ACTIONS

Threat Level	All-Good	De-escalate
Low	3.45	1.45
Medium	0.18	2.72
High	0.00	2.09

The final Q-Table produced when the model is 72% effective can be seen in Table I. Some things to note for the ”low” threat state, a score of 1.45 for ”de-escalate” is good because that means the model favors the action ”all-good” when the threat is low. Continuing to look at ”de-escalate” for ”medium” threat we notice it is a higher score of 2.72, which is also good because that indicates the model prefers ”de-escalate” even when there is a moderate threat. Finally, a score of 2.09 for ”de-escalate” of ”high” threat is also good because it shows that the model avoids using the ”all-good” state during situations of high threat.

We notice the model starts to degrade if the exploration rate gets too low, reward history gets taken over by one type of actions. Some signs of degradation include when performance score is below 50%, and Q-values start becoming very similar between actions.

While these improvements have significantly enhanced the model’s performance, further refinements could still be explored. However, normalizing position data and smoothing predictions have made the process much more reliable for real-time detection of suspicious actions.

IV. Conclusion

BOLLD marks a major step forward in multimodal detection of alarming cues by combining body language analysis, lip reading, and reinforcement learning. Unlike traditional systems, it provides a more adaptable approach to real-time threat assessment, especially in situations where audio is unreliable or unavailable.

Some key accomplishments include developing an advanced multimodal framework and implementing normalization techniques to minimize positional bias. The reinforcement learning model continuously adapts to behavioral patterns, making the system more responsive and effective.

Future improvements could include refining motion-based detection criteria, expanding the violence keyword dictionary, and incorporating adaptive thresholds that adjust based on environmental conditions. Further work is also needed to enhance reinforcement learning strategies and conduct large-scale real-world testing across diverse scenarios.

References

[1] Ramalingam, G. K., & Pattabiraman, S. (2024, August 26). AI-enhanced multimodal search engines for Cybersecurity Threat Detection. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4903692.

[2] Natural language processing in cyber threat intelligence: A major asset for infrastructure threat detection and response. Egis. (n.d.). https://www.egis-group.com/all-insights/natural-language-processing-in-cyber-threat-intelligence-a-major-asset-for-infrastructure-threat-detection-and-response.

[3] Ismail, W. S. (2024, February 29). Threat Detection and Response Using AI and NLP in Cybersecurity. Abu Dhabi; Business Information Technology Department, Liwa College, Abu Dhabi.

[4] Singhal, S. (2024, January). Real Time Detection, And Tracking Using Multiple AI Models And Techniques In Cybersecurity. NJ, USA; Infosys, US Engineering.

[5] OARJST, E. (2024, December 20). Developing multimodal AI systems for comprehensive threat detection and geospatial risk mitigation. Open Access Research Journal of Science and Technology. https://oarjst.com/content/developing-multimodal-ai-systems-comprehensive-threat-detection-and-geospatial-risk.

[6] Li, Y., Li, J., Cao, J., Xie, R., Wang, Y., Xu, M., Yizhi LiINSC, T. U., Jiang LiZhongguancun Laboratory, B., Jiahao CaoINSC, T. U., Renjie XieINSC, T. U., Yangyang WangINSC, T. U., & Mingwei XuINSC, T. U. (2024, December 9). Poster: Few-shot inter-domain routing threat detection with large-scale multi-modal pre-training: Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. ACM Conferences. https://dl.acm.org/doi/10.1145/3658644.3691402.

[7] M. Assael, Y., Shillingford, B., Whiteson, S., & de Freitas, N. (2016, December 16). LIPNET: END-TO-END SENTENCE-LEVEL LIPREADING. Department of Computer Science, University of Oxford, Oxford, UK Google DeepMind, London, UK CIFAR, Canada .

[8] codenigma1. (n.d.). CODENIGMA1/lipappnet: This is LipNet network where model learn from lip movement and predict text without voice. GitHub. https://github.com/codenigma1/LipAppNet

[9] Wang, J., Wang, Y., Liu, A., & Xiao, J. (1970, January 1). Assistance of speech recognition in noisy environment with sentence level lip-reading. SpringerLink. https://link.springer.com/chapter/10.1007/978-3-319-69923-3_64.