A Multimodal Approach for Enhancing Decision Support in Remote Digital Tower

Mobyen Uddin Ahmed[†*], Shaibal Barua[†], Mir Riyanul Islam[†], Ricky Stanley D'Cruze[†], Shahina Begum[†], Sara Kebir[‡], Alexandre Veyrie[‡], and Christophe Hurter[‡]
[†]*School of Innovation, Design and Engineering, Mälardalen University*, Västerås, Sweden
[‡]*Ecole Nationale de l'aviation Civile*, Toulouse, France
[*]*Corresponding Author*, Email: mobyen.uddin.ahmed@mdu.se, Phone: +46 21-10 73 69

2025 10th International Conference on Machine Learning Technologies (ICMLT) | 979-8-3315-3672-5/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICMLT65785.2025.11193403

***Abstract*—Trustworthy decision support systems utilizing a multimodal approach (MMA) integrate diverse data modalities to enhance robustness, transparency, and fairness in artificial intelligence (AI) applications. In this study, we present an MMA for decision support in the Air Traffic Management (ATM) domain, particularly within Remote Digital Towers (RDTs). RDTs replace traditional control towers with AI-driven digital solutions, enhancing operational efficiency. Our approach addresses key multimodal challenges—translation, alignment, and co-learning—by implementing (a) an open-vocabulary-based object detection model for video processing and (b) an audio-to-text transcription and semantic word identification model. The YOLO-World deep-learning model is employed for object detection, while audio data analysis takes advantage of a benchmark data set, semantic identification techniques, and explainability. Additionally, the system integrates robust machine learning techniques, including data augmentation and perturbation, to maintain consistent performance across varied operational conditions. This proof-of-concept demonstrates the potential of multimodal AI systems to enhance decision support and improve safety in ATM environments.**

***Index Terms*—Decision Support, Multimodal, Air Traffic Management, ATM, Remote Digital Tower, RDT.**

I. Introduction

The concept of trustworthy decision support systems incorporating a multimodal approach (MMA) has many aspects, encompassing robust and resilient learning, transparency, fairness, and human-machine teaming. MMA is defined when it involves multiple modalities, where each modality contains different statistical properties in the data generation process or in the form of the data. Multimodal data are often heterogeneous, and some challenges dealing with MMAs are representation, alignment, translation, fusion, and co-learning [1]. It is an essential aspect of trustworthy Artificial Intelligence (AI) systems, teaching computers to process and synthesize information from various inputs—visual, auditory, textual, *etc.* Few surveys have been conducted within these domains, *e.g.*, Moujahid *et al.* reviewed multimodal magnetic resonance imaging scans for segmentation [2], Fereidoonian *et al.* presented a study on human activity recognition using multimodal machine learning [3], *etc.*

This study was supported by the following projects: 1) TRUSTY, financed by SESAR JU under the EU's Horizon 2022 Research and Innovation programme, Grant Agreement No. 101114838; 2) Trust_Gen_Z, funded by VINNOVA, Diary No. 2024-01402.

In this paper, we have demonstrated an MMA for decision support in the Air Traffic Management (ATM) domain. In ATM, Air Traffic Control Officers (ATCOs) involve complex multi-activity, which usually impacts psychology [4]. In a study with ATCOs, the authors fused multimodal neurophysiological data to assess the impact of stressful events on them [5]. Modi *et al.* presented a review on intelligence traffic management applying machine learning (ML) algorithms [6]. Remote Digital Towers (RDTs) replace physical towers by incorporating digital technologies as tools to assist tower controllers, for example, runway and taxiway monitoring and adaptive management of critical situations. To improve their abilities, digital towers should be able to take benefit of digital images, such as computer vision, to automatically recognize flying objects (*e.g.*, aircraft, birds, drones, *etc.*) without persistent monitoring by human operators. As the objects are being considered as videos, this method can deal with processing the time frames of the videos. The Single Shot multi-box Detector [7] was used to detect objects which are faster than YOLO [8]. The framework described by Schumann *et al.* is used to detect a moving object by subtraction of the background and frames [9]. Jiang *et al.* mentioned in a review study that these type of methods run faster than R-CNN [10] and YOLO [11], although the accuracy might be lower [8].

Recent advancements have highlighted the critical role of ML in aircraft and drone detection, enabling accurate tracking and improved response times [12]. Studies have also examined the visual implications of digital tower technologies in ATCOs, demonstrating the importance of ensuring safety and clarity in visual displays [13]. An innovative example in RDTs is the integration of visual spectrum and infrared fusion technologies, coupled with optical tracking, to enhance performance under restricted visibility conditions [14]. Automated speech-based service request systems, which recognize callsigns, input commands, and support digital ATC systems, are another AI-driven innovation. These systems reduce ATCOs' workload and improve usability, particularly compared to traditional manual methods [15]. Research into multimodal augmentations has also demonstrated their potential in single remote tower contexts, improving controllers' situational awareness and operational performance under diverse conditions [16].

The objective of this study is to investigate and develop a multimodal AI solution that will be trustworthy in terms of

transparency (*i.e.*, the model is interpretable) with the aim of accessibility of user's decision. Here, the MMA addresses 1) translation, 2) alignment, and 3) co-learning aspects through two approaches: a) a vocabulary-based object detection model from video, and b) an audio-to-text transcription and semantic word identification model. Thus, the paper presents a proof-of-concept of an MMA for decision support in the RDTs environment. The system consists of object detection and audio data analysis for two scenarios: enhanced runway and taxiway monitoring, and adaptive management of critical situations. The YOLO-World, a deep-learning pre-trained model, is considered for object detection. For audio-to-text transcription, we have generated a benchmark dataset as an example-based or dictionary model. Different methods, such as Word2Vec [17], LIME (Local Interpretable Model-agnostic Explanations) [18], and SHAP (Shapley Additive Explanations) [19] tools are used for identification of semantic meaning. The study also incorporates the development of robust ML through better calibration, data augmentation and perturbation to ensure consistent performance under normal, long-tail and unusual conditions.

II. Scenario-specific Challenges and Solutions

The MMAs for RDT operations are explored and validated through the TRUSTY[1] project. This solution is designed to support ATCOs by improving situational awareness in runway and taxiway monitoring. Using multimodal data inputs such as video feeds, audio signals, and communication data, the AI processes these streams to detect high-risk situations, such as runway incursions or the impact of adverse weather and promptly directs the ATCO's attention to critical scenarios.

Human operators play a vital role in developing trustworthy AI systems. Hence, we have focused on the human-centred MMA design for XAI, which addresses the following issues: (1) identify assumptions and requirements of the RDT domain, (2) marginalize the requirements as components, (3) rationalize and realize the requirements, and (4) develop explainable ML models to embody the marginalized components. Several workshops with the ATCOs are conducted to design user-acceptable explanations so that the requirements of the XAI design are met. One of the important suggestions from ATCOs is the clarity of AI explanations, *i.e.*, the explanation should not be diluted with too much information.

A. Scenario 1: Enhanced Runway and Taxiway Monitoring

Context: In an RDT environment, ATCOs are required to manage multiple remote airfields simultaneously. This includes monitoring runway and taxiway conditions to ensure safe and efficient aircraft movements.

Challenges: In multi-airfield RDT environments, ATCOs face significant cognitive demands when monitoring runways and taxiways. Unauthorized vehicles, debris, or adverse weather conditions, such as fog or wind shear, can increase the risk of runway incursions. Traditional systems often rely on static sensors or basic video feeds, limiting the accuracy and immediacy of anomaly detection, especially under low visibility or dynamic conditions.

MMA Solution: The TRUSTY system applied MML, integrating high-definition video feeds and audio communication data to enhance monitoring. The AI employs object detection on live video streams, supported by Explainable AI (XAI) technologies that offer detailed reasoning for anomalies. For instance, if an unauthorized vehicle enters the runway, the system cross-references visual inputs with contextual audio cues (*e.g.*, communication logs) to flag the intrusion and pinpoint its location on the ATCO's interface. The robustness of MMA ensures accurate detection even in challenging conditions, such as poor lighting or weather interference.

B. Scenario 2: Adaptive Management of Critical Situations

Context: Different operational events and situations significantly impact air traffic operations, necessitating real-time monitoring and adaptive management to ensure safety. In an RDT setup, ATCOs rely on accurate situational information and clear communication of potential impacts on operations.

Challenges: Critical incidents, such as sudden changes of wind, bird strikes, *etc.*, significantly impacting air traffic operations and posing safety risks. In RDT settings, ATCOs must rely on precise, real-time weather updates, information on critical incidents, and effective communication to quickly adjust protocols. Traditional systems may struggle to synthesize data from pilot communications, leading to delayed responses or misinterpretation of critical information.

MMA Solution: The TRUSTY AI integrates video and audio data to detect and manage sudden wind changes and critical situations proactively. By analyzing pilot communications (*e.g.*, reports of wind shear, bird strikes, engine failures, *etc.*), the system identifies potential risks and generates alerts with detailed explanations of the predicted impact on air traffic operations. These alerts enable ATCOs to adjust take-off and landing procedures proactively and communicate instructions effectively to pilots.

III. Multimodal Approach for Remote Digital Towers

The overall MMA approach for RDT solutions is shown in Fig. 1. Here, the tasks start with problem formulation and datafication, *i.e.*, data processing and feature extraction; selecting ML algorithms; classifiers (*e.g.*, object detection model) and evaluation of individual classifiers, *i.e.*, algorithm selection, tuning, training, and validation; and finally, comparison and validation of the output models. For the translation of video data, the YOLO-World model, which is an Open-Vocabulary Object Detection (OVD) model, is considered. For audio, a benchmark dataset of audio transcription is generated as an example-based or dictionary model. Computer vision and deep learning models have been used for implicit alignment, and explicit alignment was achieved through a rule-based approach. Co-learning is complementary information sharing across the different modalities using transfer learning. Using

[1] https://research.dblue.it/trusty/

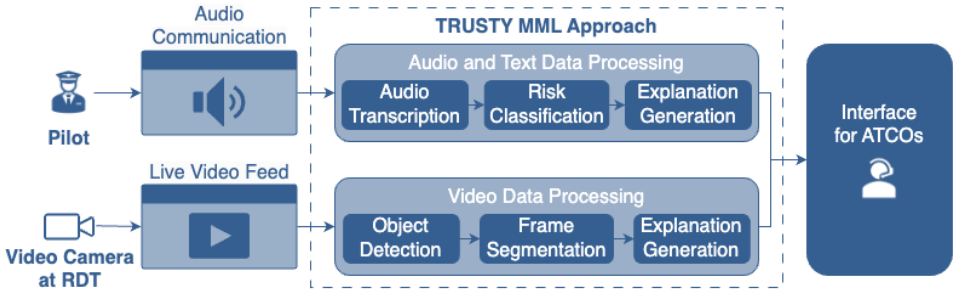

Fig. 1. Overall approach of MMA for RDT operations.

pre-trained models as a transfer learning approach, this study has not only shared the information but also learned.

A. Video and Vocabulary

In the airport surveillance setting, objects can vary significantly in size or be partially hidden or unclear, making detection a difficult task. For object detection, the YOLO-world model is exploited, which is a cutting-edge system built on Ultralytics YOLOv8, designed for OVD [20]. Instead of relying on a fixed list of known objects (like "dog" or "car"), OVD uses descriptive text or general knowledge to recognize objects it has never seen before, which makes it more flexible and capable of handling a wider range of tasks [21]. The study uses a lightweight YOLO architecture, a re-parameterizable vision-language path aggregation network, and a prompt-then-detect paradigm (a strategy for increasing the efficiency of open-vocabulary object detection), allowing for real-time inference, which is easy to deploy. YOLO-world is also pre-trained with region-text contrastive learning on large-scale datasets, including detection, grounding, and image-text data, allowing it to perform well in zero-shot cases on the LVIS dataset (a benchmark for large vocabulary instance segmentation). It can also be fine-tuned for downstream tasks such as traditional object detection and open-vocabulary instance segmentation [20]. The workflow is illustrated in the diagram provided in Fig. 2.

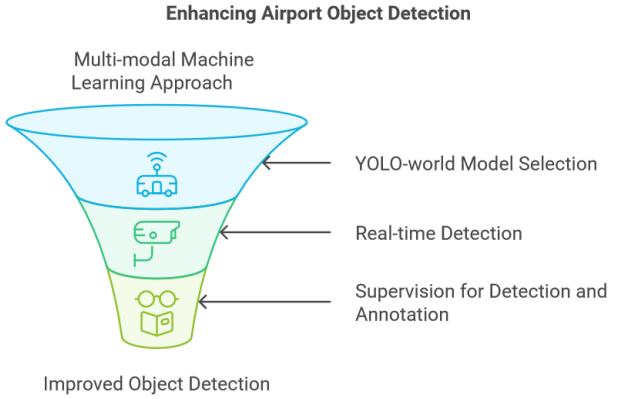

Fig. 2. Workflow of object detection from video data.

For the detection and annotation of objects, we have used *Supervision*[2], which is an open-source Python library designed to streamline and simplify the development of computer vision applications. It comprises many tools and functionalities that enhance the efficiency and effectiveness of working with vision models, particularly in the context of object detection and annotation. Some of its key features are the integration of various object detection models, such as Ultralytics YOLO and Transformers. Developers can easily load and use this pre-trained model without customising it with their own dataset. Rich sets of annotators for visualising and labelling detections, like bounding boxes, segmentation masks, and labelling objects with confidence scores, support for object tracking, and efficient handling of large datasets, *etc.*, are great features of *Supervision*.

1) Rule-based Approach: Object detection with a set of constraints approach is used to detect objects in videos employing specific rules or limitations. For this purpose, the three regions of interest, the runway, the taxiway, and the sky, are specified in the videos. Several constraints are employed, which include:

- Spatial Constraints — Detecting objects only within a specific region, *e.g.*, detecting aircraft, animals, or vehicles only within a designated zone.
- Object class Constraints — Detecting specific types of objects, *e.g.*, detecting only aircraft.
- Temporal Constraints — for effectively tracking objects to be present for a certain duration in the scene.
- Environmental Constraints — specific conditions are ignored for detections, *e.g.*, low light, occlusion, *etc.*

Based on these constraints, several dangerous and non-dangerous cases and their associated rules are formulated. Table I represents some examples of these events and rules.

The algorithm for object detection with bounding boxes for dangerous and non-dangerous cases is presented in Algorithm 1.

B. Audio and Text Data

The proposed solution incorporates advanced AI techniques to analyse audio and transcribed communication data, enabling

[2]https://github.com/roboflow/supervision

TABLE I
EXAMPLE OF EVENTS AND RULES ASSOCIATED WITH DANGEROUS AND NON-DANGEROUS CASES FOR OBJECT DETECTION.

Event	Condition	Dangerous Case	Non-dangerous Case
An aircraft stopped on the runway.	No other objects on the runway except an aircraft.	If the aircraft stopped for $\geq$ threshold seconds on the runway, then the bounding box is Red.	If the aircraft is staying at the edge of the runway or the aircraft is moving or the aircraft is holding short ($<$ threshold sec) of the runway, then the bounding box is Green.
An aircraft lands and stops on the runway.	No other objects are on the runway while an aircraft is landing.	If the aircraft stops for $\geq$ threshold seconds either on the runway or at the threshold between the runway and taxiway, then the bounding box is Red.	If the aircraft is moving or the aircraft initiates a go-around, then the bounding box is Green.
Birds or Drones are present at the runway threshold.	The runway is empty, or an aircraft is on the runway, or an aircraft wants to land.	If there are folks of birds or drones at the threshold of the runway, then the bounding box on the birds, drones and/or aircraft on the runway is Red.	If the birds or drones are away from the threshold or flying high in the sky, then the bounding box on the birds, drones and/or aircraft is Green.
Vehicle, animal or human on the runway.	There is a vehicle, animal or human on the runway. There is an aircraft on the taxiway that wants to take off or an aircraft that wants to land.	If there is a vehicle, animal or human on the runway (moving/stopped), then the bounding box on the vehicle, animal, human and/or aircraft is Red.	When the vehicle, animal or human exits the runway to a safety margin, then the bounding box on the vehicle, animal, human and/or aircraft is Green.
A smoke or fire event near the runway.	There is smoke near the runway, and there is an aircraft that wants to land. Or there is an aircraft catching fire and smoke.	Because of the smoke, there is low visibility, so the bounding box of the smoke area is Red. In the case of an aircraft on fire, the bounding box of that aircraft is Red.	If the smoke is outside some threshold distance, the bounding box of the smoke area is Green.

real-time monitoring and situational awareness in RDT operations. By processing pilot and ATCO communications, the system detects critical patterns and anomalies, such as sudden weather changes or operational irregularities, ensuring timely and precise decision-making in complex air traffic scenarios.

The audio and text data processing workflow in the TRUSTY solution with MMA consists of three major stages, as depicted in Fig. 1. The three stages are: i) transcription of pilots' radio messages, ii) classification of transcriptions for high-risk situations, and iii) generation of an explanation for the classified transcription. Each stage is described briefly in the sub-sections below and illustrated in Fig. 3.

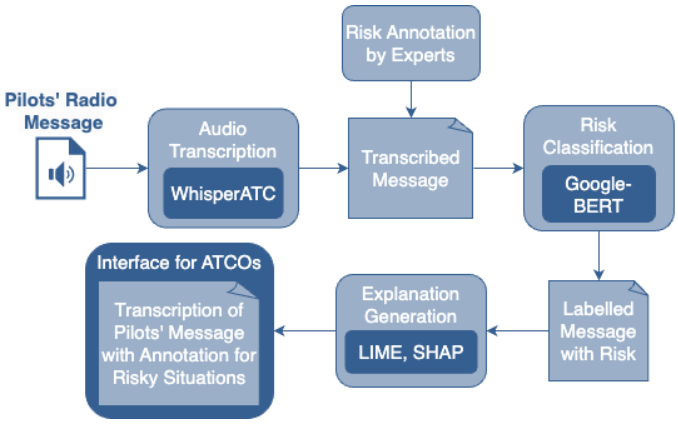

Fig. 3. Workflow of audio transcription and text classification with specific models.

The first step in the approach for audio data is to transcribe pilots' radio messages into text using the WhisperATC model [22], which is a variant of OpenAI's Whisper model [23]. This specialized Automatic Speech Recognition (ASR) system was trained on labelled data [23] from the ATCO2[3] project, which provides a diverse dataset of real-world air traffic communications. Given the noisy and domain-specific nature of pilot-ATCO radio message exchanges, WhisperATC was fine-tuned to handle such variability effectively. However, the context of RDT differs from that of physical towers, which is also quite rare in the ATM domain, resulting in scarce pilot-ATCO radio messages and transcriptions. To bridge this gap, synthetic transcriptions were generated using Large Language Models (LLM), ensuring that the RDT-specific context remains well-represented and sufficient data. The synthetic transcriptions were lastly validated and annotated by aeronautical experts and an ATCO to ensure accuracy and operational relevance.

After preparing the pilots' radio message transcriptions, the next step in the MMA is to classify the transcriptions into high- and low-risk situations using a fine-tuned Google-BERT model [24]. A critical component of this process is Named Entity Recognition (NER), which identifies and extracts key entities from transcriptions, such as call signs (*e.g.*, AF123), measurements (*e.g.*, altitude 3000 feet), operational situations (*e.g.*, engine failure or wind shear), and other contextual information critical to air traffic operations. These extracted entities serve as valuable input to understand the operational context and informing the classification task.

The Google-BERT model is trained on a rich dataset combining real-world and synthetic transcriptions. Here, the synthetic transcriptions are generated using LLM and curated by domain experts, covering a wide range of scenarios. Here, the synthetic dataset complements the scarcity of trainable real-world transcriptions in critical situations with all possible variations. These include high-risk situations such as occupied

[3]https://www.atco2.org/

Algorithm 1 Object detection with bounding boxes

input: videos of a specific event and pre-trained YOLO-World model.

output: object(s) detected with bounding boxes indicating dangerous or non-dangerous cases.

1: initialize an empty list for filtered detections.
2: define constraints:
   a) set allowed object classes.
   b) define the region of interest (ROI).
   c) set a confidence threshold.
3: **for** each detected object in the video **do**
4: get the detected object's class, confidence score, and bounding box coordinates.
5: apply constraints
   a) check if the object belongs to the target classes.
   b) verify that the confidence score is above the threshold.
   c) ensure the object is within the defined region of interest.
   d) if all constraints are met, add the object to the filtered list.
6: draw bounding boxes, labels, and colours (red or green)
   for valid objects in the image.
7: **end for**
8: save the coordinates of the boxes and display the processed video.

return list of valid detections, *i.e.*, bounding boxes with green and red colors.

runways, low fuel alerts, and mechanical failures, as well as low-risk conditions like touch-and-go manoeuvres or standard communication exchanges. By incorporating NER, the model can identify and focus on these critical elements, ensuring that even subtle cues in the text contribute to accurate classification outcomes.

The classification is evaluated using standard accuracy metrics, with higher accuracy indicating the system's ability to distinguish between high- and low-risk situations reliably. By integrating NER into the pipeline, the Google-BERT model can also capture nuanced patterns in the data, such as specific call signs associated with urgent messages or measurements indicating unsafe conditions. This approach not only improves the robustness of the classification process but also provides ATCOs with more contextually relevant information, aligning the AI outputs with the complex and dynamic needs of the ATM domain.

To improve the transparency and explainability of the classification process, explanation techniques such as LIME [18] and SHAP [19] are employed. These methods identify and highlight the most relevant words or phrases in the transcriptions that contribute to the classification of high- or low-risk situations. For example, in a high-risk scenario involving wind shear, terms like wind and shear would be annotated as critical contributors. By providing these explanations, the system not only intends to enhance user trust but also facilitates validation and feedback from domain experts, ensuring alignment with operational expectations [25].

IV. Results and Evaluation

MMAs, both for video and audio data, are evaluated separately with quantitative evaluation methods. The output of the exploited models and the evaluation results are presented in the following sections.

A. Evaluation on Video Dataset

Object detection and annotation of supplied videos using the YOLO-world model gave us a good result, as some of the samples of the resulting videos are presented in Fig. 4.

For the evaluation, several metrics have been considered, *i.e.*, Intersection over Union (IoU), Mean Average Precision (mAP), Precision and Recall. The IoU is a metric used to measure the overlap between the predicted bounding boxes and the ground truth bounding boxes drawn using the Computer Vision Annotation Tool. To calculate the IoU in this context the following steps are considered:

1) For each detected object, the YOLO-world model generates a predicted bounding box.
2) The predicted bounding box is compared to the ground truth bounding box for that object.
3) The IoU is calculated as the area of overlap between the predicted and ground truth boxes divided by the total area covered by both boxes.

The mAP is a commonly used metric to evaluate the overall performance of an object detection model. It considers both the precision and recall of the model across different IoU thresholds. To calculate mAP, the following steps are considered:

1) For each IoU threshold, the model's precision and recall are computed.
2) The average of these precision values is taken to get the Average Precision (AP) for that IoU threshold.
3) The mAP is then calculated by taking the mean of the APs across all the IoU thresholds.

Precision measures the fraction of predicted detections that are true positives, and Recall measures the fraction of ground truth objects that are correctly detected. The standard deviations (σ) of both the Precision ($\sigma_{Precision}$) and Recall (σ_{Recall}) values provide a measure of how stable these metrics are across the different IoU thresholds.

TABLE II
Class distribution between ground truth and detected objects.

Category	Truck	Aeroplane
Ground Truth Class Distribution	992	1084
Detected Object Class Distribution	992	1086

Here, the evaluation data and results for the object detection analysis are presented in Tables II and III for one original

Fig. 4. Result of object detection within the region of interests.

and one object-detected video. These metrics help evaluate the overall quality and robustness of the object detection model. In addition, the object detection model performed with a high mAP score of 0.9829 (98.29%) that indicates the model is performing very well at almost accurately detecting and classifying the objects of interest.

TABLE III
SUMMARY RESULTS OF IOU AND OTHER METRICS OF OVERALL BOUNDING BOXES.

IoU Threshold	Precision	$\sigma_{Precision}$	Recall	σ_{Recall}
0.1	0.9940	0.0707	0.9960	0.0634
0.2	0.9940	0.0707	0.9960	0.0634
0.3	0.9936	0.0715	0.9956	0.0642
0.5	0.9837	0.0989	0.9857	0.0940
0.75	0.9493	0.1386	0.9513	0.1357

B. Evaluation on Audio Dataset

The MMA for audio-to-text transcription and classification incorporates quantitative evaluations to ensure its reliability and usability. For transcription, the Word Error Rate (WER) is used as the quantitative metric, measuring the accuracy of transcribed text against the ground truth. WER quantifies errors by comparing the trained model's predicted transcription with the reference (ground truth) transcription and represents these errors as a percentage of the total number of words. The formula for calculating WER is structured as (1) [22]:

$$WER = \frac{Substitutions + Deletions + Insertions}{Total\ Words\ in\ Reference} \quad (1)$$

Here, the three different errors are addressed:

- *Substitution*: A word from the reference transcription is replaced by an incorrect word in the predicted transcription.
- *Deletion*: A word present in the reference transcription is missing from the predicted transcription.
- *Insertion*: An extra word appears in the predicted transcription that does not exist in the reference transcription.

This metric provides a comprehensive assessment of transcription accuracy by considering all possible error types. For classification, regular accuracy is the performance metric, reflecting the model's ability to correctly label transcriptions into high- and low-risk categories. These evaluations provide a robust assessment of the system's overall performance.

The first evaluation was done for audio transcriptions with the WhisperATC model for transcribing pilots' radio messages WER as the primary metric. The results indicated a WER of 25.67%, meaning that 25.67% of the words in the transcriptions differed from the reference transcripts due to substitutions, deletions, or insertions. While this WER suggests room for improvement, it is within an acceptable range for handling noisy and domain-specific air traffic communications, where overlapping speech, radio static, and varying accents can affect ASR performance. The WhisperATC model, trained on ATCO2[4] corpus, demonstrated robustness in handling real-world pilot-ATCO exchanges despite these challenges. Future improvements, such as domain-specific fine-tuning or additional training on augmented datasets, could further reduce the WER and enhance transcription reliability for downstream classification tasks.

As the risk classifier, the Google-BERT model was evaluated on a dataset consisting of 38 selected audio transcriptions, with 30 samples used for training and 8 samples reserved for testing. The model achieved an accuracy of 75.00%, correctly classifying 5 high-risk cases and 1 low-risk case while misclassifying one low-risk case as high-risk and one high-risk case as low-risk. The precision of the model was 83.33%, indicating that 83.33% of the instances predicted as high-risk were high-risk. The recall was also 83.33%, meaning the model successfully detected 83.33% of all actual high-risk cases without missing any. Additionally, the F_1-score, which balances precision and recall, was recorded at 83.33%, highlighting a well-balanced performance. The confusion matrix is illustrated in Fig. 5. These results suggest that the model effectively identifies high-risk situations while maintaining a moderate level of false positives. However, further improvements in model generalization may be needed with a larger dataset and additional fine-tuning.

Lastly, explanations are generated for both high- and low-

[4]https://github.com/idiap/atco2-corpus

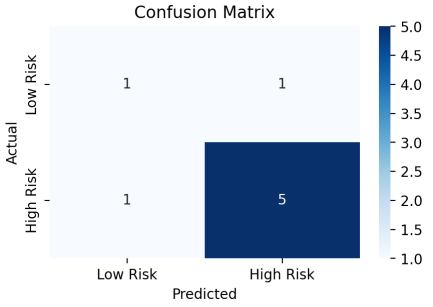

Fig. 5. Confusion matrix for risk classification of transcribed audio messages.

risk situations from the classified transcriptions. Two examples of explanations on classified transcriptions are illustrated in Fig. 6.

Muret Ground, Delta Papa three three four, TB20, Holding Runway 30L, wind shear ongoing. Wait for information updates.

(a) High Risk

Muret Ground, Victor Zoulou seven four nine, DR400, Request take-off clearance Runway 30L.

(b) Low Risk

Fig. 6. Examples of explanations on the classification of audio transcriptions.

V. Conclusion

The main objectives of this study were to investigate and develop an MMA for AI solutions in RDTs. The study has also considered the trustworthiness of the MMA in terms of transparency (*i.e.*, the model is interpretable), and human-centred explainability. Two operational use cases are considered during the development: 1) enhanced runway and taxiway monitoring and 2) adaptive management of critical situations. For the first use case, object detection and event annotation from video is employed. The audio data transcription and semantic meaning identification model is developed for the second use case. For both use cases, state-of-the-art deep learning models are considered with a rule-based approach for bounding box detection through computer vision and explainable AI algorithms.

The proposed MML approach incorporates video and audio modalities through an interface for ATCOs as shown in Fig. 1. One possible improvement could be multimodal integration using late fusion approaches by weighting visual and audio inputs differently based on context. In this work, YOLO-World is used since it is an efficient and practical model with zero-shot learning capability. Further, investigations will be carried out to compare YOLO-World with Vision Transformers. Also, transformer-based multimodal architectures such as CLIP or Perceiver will be investigated for visual and audio feature alignment.

ATM is a safety- and time-critical domain where ATCOs do not want to be overwhelmed with information [25]. Hence, for both video and audio modalities, simplified explanations are generated with the bounding box and highlighted texts. The system's trustworthiness can be improved by expanding on explainability techniques, *e.g.*, counterfactual explanations; however, counterfactual explanations are not favourable in operational settings. Explanations with bounding boxes are simple explanations for ATCO, whereas AI system developers may be interested in understanding model decision-making based on the heat-map-based attention method and counterfactuals.

The initial experiment shows promising results with a high score of 83% according to the evaluation. It is noteworthy that the videos and the audios were developed through brainstorming sessions with air traffic controllers. Subsequently, these datasets were internally validated by aeronautical experts, with a final review conducted by an ATCO. The proposed MMA shows a high potential for developing AI solutions in RDTs. However, one limitation is that the LLM used in this study was not specific to aviation and might require other specific adjustments for remote tower operations. Also, there is a scarcity of video data in real RDT environments, and these datasets are often not publicly accessible. Hence, our contribution is the benchmark dataset that we have generated through the TRUSTY project. In the future, environmental conditions such as fog, night-time, glare, *etc.*, will be integrated to ensure dataset diversity. These conditions will help to avoid AI bias in RDT environments across different weather conditions and operational scales, as well as impact detection accuracy in predicting potential hazards such as runway incursions. Thus, this paper presented a proof of concept based on benchmark evaluation, with a follow-up work focusing on user evaluation.

Acknowledgment

The authors would like to acknowledge Elizabeth Humm, Giulia Cartocci, and Pietro Aricó for reviewing the manuscript.

References

[1] A. Barua, M. U. Ahmed, and S. Begum, "A Systematic Literature Review on Multimodal Machine Learning: Applications, Challenges, Gaps and Future Directions," *IEEE Access*, vol. 11, pp. 14 804–14 831, 2023.

[2] H. Moujahid, B. Cherradi, and L. Bahatti, "Convolutional Neural Networks for Multimodal Brain MRI Images Segmentation: A Comparative Study," in *Smart Applications and Data Analysis*, M. Hamlich, L. Bellatreche, A. Mondal, and C. Ordonez, Eds., vol. 1207, Cham: Springer International Publishing, 2020, pp. 329–338.

[3] F. Fereidoonian, F. Firouzi, and B. Farahani, "Human Activity Recognition: From Sensors to Applications," in *2020 International Conference on Omni-layer Intelligent Systems (COINS)*, 2020, pp. 1–8.

[4] I. Arminen, I. Koskela, and H. Palukka, "Multimodal Production of Second Pair Parts in Air Traffic Control Training," *Journal of Pragmatics*, vol. 65, p. 46, 2014.

[5] G. Borghini *et al.*, "A Multimodal and Signals Fusion Approach for Assessing the Impact of Stressful Events on Air Traffic Controllers," *Sci Rep*, vol. 10, no. 1, p. 8600, 2020.

[6] Y. Modi, R. Teli, A. Mehta, K. Shah, and M. Shah, "A Comprehensive Review on Intelligent Traffic Management using Machine Learning Algorithms," *Innov. Infrastruct. Solut.*, vol. 7, no. 1, p. 128, 2021.

[7] W. Liu *et al.*, "SSD: Single Shot MultiBox Detector," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham: Springer International Publishing, 2016, pp. 21–37.

[8] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A Review of Yolo Algorithm Developments," *Procedia Computer Science*, The 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 & 2021): Developing Global Digital Economy after COVID-19, vol. 199, pp. 1066–1073, 2022.

[9] A. Schumann, L. Sommer, J. Klatte, T. Schuchert, and J. Beyerer, "Deep Cross-domain Flying Object Classification for Robust UAV Detection," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2017, pp. 1–6.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[12] V.-P. Thai, W. Zhong, T. Pham, S. Alam, and V. Duong, "Detection, Tracking and Classification of Aircraft and Drones in Digital Towers Using Machine Learning on Motion Patterns," in *2019 Integrated Communications, Navigation and Surveillance Conference (ICNS)*, Herndon, VA, USA: IEEE, 2019, pp. 1–8.

[13] L. Meyer, M. Peukert, B. Josefsson, and J. Lundberg, "Validation of an Empiric Method for Safety Assessment of Multi Remote Tower," in *13th USA/Europe Air Traffic Management Research and Development Seminar*, EUROCONTROL, 2019.

[14] M. Hagl *et al.*, "Augmented Reality in a Remote Tower Environment Based on VS/IR Fusion and Optical Tracking," en, in *Engineering Psychology and Cognitive Ergonomics*, D. Harris, Ed., vol. 10906, Cham: Springer International Publishing, 2018, pp. 558–571.

[15] O. Ohneiser *et al.*, "Assistant Based Speech Recognition Support for Air Traffic Controllers in a Multiple Remote Tower Environment," *Aerospace*, vol. 10, no. 6, p. 560, 2023.

[16] M. Reynal *et al.*, "Investigating Multimodal Augmentations Contribution to Remote Control Tower Contexts for Air Traffic Management:" in *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, Prague, Czech Republic: SCITEPRESS - Science and Technology Publications, 2019, pp. 50–61.

[17] K. W. Church, "Word2Vec," *Nat. Lang. Eng.*, vol. 23, no. 1, pp. 155–162, 2017.

[18] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, San Francisco, CA, USA: ACM, 2016, pp. 1135–1144.

[19] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, ser. NIPS'17, 2017, pp. 4768–4777.

[20] T. Cheng *et al.*, *YOLO-World: Real-Time Open-Vocabulary Object Detection*, arXiv preprint, arXiv:2401.17270v3 [cs.CV], 2024.

[21] J. Li, C. Xie, X. Wu, B. Wang, and D. Leng, *What Makes Good Open-Vocabulary Detector: A Disassembling Perspective*, arXiv preprint, arXiv:2309.00227v1 [cs.CV], 2023.

[22] J. Tol, *Whisper Medium EN Fine-Tuned for Air Traffic Control (ATC) - Faster-Whisper Optimized*, Repository, 2024. [Online]. Available: https : / / huggingface . co / jacktol/whisper-medium.en-fine-tuned-for-ATC-faster-whisper (visited on 04/04/2025).

[23] A. Radford *et al.*, "Robust Speech Recognition via Large-Scale Weak Supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML'23, vol. 202, Honolulu, Hawaii, USA: JMLR.org, 2023, pp. 28 492–28 518.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, arXiv preprint, arXiv:1810.04805v2 [cs.CL], 2018.

[25] W. Jmoona *et al.*, "Explaining the Unexplainable: Role of XAI for Flight Take-Off Time Delay Prediction," in *Artificial Intelligence Applications and Innovations. AIAI 2023. IFIP Advances in Information and Communication Technology*, I. Maglogiannis, L. Iliadis, J. MacIntyre, and M. Dominguez, Eds., vol. 676, Léon, Spain: Springer Nature Switzerland, 2023, pp. 81–93.