Journal of Open Aviation Science (2026), Vol.**4**
doi:10.59490/joas.2026.8685

JOAS

EDITORIAL

Reviews and Responses for
A Methodology for Quantifying Response Times for Deconfliction Actions Through ATC Communications

Authors: Timothé Krauth, Kim Gaume, Xavier Olive, and Junzi Sun

Reviewers: Gabriel Jarry, Allan Tart, and Lucie Smetanová

Editor: Tatiana Polishchuk

1. Original paper

The DOI for the original paper is https://doi.org/10.59490/joas.2026.8462

2. Review - round 1

2.1 Reviewer 1

The paper presents a methodology to estimate pilot response times to Air Traffic Control (ATC) deconfliction instructions. The authors propose a data pipeline that fuses ATC voice recordings (processed via Automatic Speech Recognition and Named Entity Recognition) with ADS-B surveillance data and flight plans. The study utilizes the MUAC DELTA sector as a testbed to identify specific deconfliction maneuvers and align them temporally with voice instructions.

This work addresses a highly relevant topic in Air Traffic Management (ATM). As automation tools for conflict detection and resolution evolve, accurate modeling of human latency (both controller and pilot) is essential for defining safe separation buffers. The paper is well-written, and the proposed pipeline, integrating audio, NLP, and trajectory data, is logical and clearly explained. However, there is a significant discrepancy between the paper's title ("Quantifying...") and the results, which currently serve more as a proof-of-concept than a statistical quantification.

Major Comments

1. **Positioning and Contribution relative to State-of-the-Art.** The authors reference Lutz et al. [4] as a primary predecessor in this domain. The paper states that the proposed method improves upon [4] by using a more sophisticated deconfliction detection algorithm rather than simple thresholding. The paper would benefit significantly from a more explicit delineation of why the proposed NLP-centric pipeline is necessary compared to the approach in [4]. Does the addition of complex ASR/NER significantly improve the accuracy of the timestamping, or does it simply allow for better filtering of non-control communications? *Recommendation:* Please expand Section 2 or 3 to detail the specific limitations of the Lutz et al. approach that this paper solves. If possible, a qualitative and/or quantitative comparison of a scenario processed by both methodologies would strengthen the contribution.

2. **Discrepancy between Title and Results (Quantification vs. Methodology).** The title "Quantifying response times…" implies that the paper will present statistical distributions of response times (e.g., means, variances, distributions). However, the Results section (specifically Section 4.4) focuses on three specific case studies (Figures 6, 7, and 8). While Section 4.3 provides statistics on the matching process (Table 2), there are no aggregate statistics regarding the response times themselves. Consequently, in its current form, the paper reads more as a 'Methodology for Quantifying…' rather than the quantification itself. *Recommendation:* I strongly suggest including a table or histogram showing the distribution of response times for all successfully matched deconfliction events in the dataset. If the number of matched events is too low to be statistically significant, the title should be adjusted to reflect that this is a feasibility study or a methodology proposal.

Minor Comments and Suggestions

1. **Literature Review.** I recommend including Pellegrini, T., et al. (2018). "The Airbus Air Traffic Control speech recognition 2018 challenge" in the literature review. This is a seminal work regarding the difficulties of ATC transcription and callsign detection and provides relevant context for the ASR challenges discussed in Section 3.1.
2. **Generalizability and ANSP Application.** The paper notes that MUAC was chosen partly because it is legally permissible to record ATC audio there. While this constraint limits "large-scale" reproducibility by external researchers in jurisdictions with stricter privacy laws (e.g., France), it presents a specific opportunity for operational stakeholders. I suggest explicitly stating that while open data collection is limited by these regulations, the pipeline is highly generalizable if deployed directly by Air Navigation Service Providers (ANSPs). Since ANSPs already possess legal access to their own voice records, they could implement this pipeline internally to monitor systemic safety performance (e.g., sector complexity analysis) without facing the regulatory hurdles encountered by open-source researchers, provided that appropriate frameworks (e.g., Just Culture protocols, data anonymization) are established to address potential concerns regarding individual performance monitoring.
3. **Operational Application (Means vs. Quantiles).** The introduction correctly notes that separation minima rely on assumptions of human latency. When this methodology is scaled, how should the resulting data be used? From a safety perspective, the average response time is less critical than the tail of the distribution (e.g., the 95th or 99th percentile). While the current analysis identifies individual response times (e.g., 22 seconds in Figure 7), future work should focus on quantifying the "long tail" of pilot reaction times. This is crucial for determining the "maximum probable delay" required for accurate "distance at CPA" risk modeling.

This paper proposes a robust technical pipeline for a critical ATM problem. The methodology is sound, but the current results do not yet fully deliver on the promise of "quantification." With the addition of aggregate statistics (even preliminary ones) and a sharper distinction from prior art, this would be a very strong contribution.

2.2 Reviewer 2

This paper proposed a novel methodology for automatically measuring pilot response times to Air Traffic Control (ATC) deconfliction instructions. By combining Automatic Speech Recognition (ASR), Named Entity Recognition (NER), and ADS-B surveillance data, the authors attempt to bridge the gap between voice instructions and physical aircraft maneuvers.

In short, the method uses speech-to-text processing and callsign extraction; the callsigns are associated with ADS-B trajectories. To discover pilots' corrective action, a trajectory extraction method is used. Based on the above, the pilot's response time is estimated.

In addition to describing the proposed methodology, authors extensively discuss the issues and shortcomings of the data processing pipeline. As such, the paper can be viewed as an initial step in proposing a robust method for automatically measuring the pilots response time.

The paper is well structured covering all required aspects.

The reviewer has noted some comments:

1. The title of the paper is "Quantifying response times for deconfliction actions through ATC communications", in the introduction, the authors define the response time as "the elapsed time between the detection of a potential conflict and the initiation of the corresponding action to mitigate it." However, the method only covers the pilots response time; the air traffic controllers time between detection of an issue and issuing a corrective action to pilot is not handled. Paper title, abstract and Introduction should make the scope of the method explicitly clear to the reader.
2. While the 'Results' section provides an exhaustive analysis of the individual processing steps – such as callsign matching rates and conversation reconstruction – it does not present any statistics on the pilot response times themselves. If the scope of the paper is such analysis, it explicitly be highlighted in the paper title and introduction.
3. Figure 4, maybe cumulative distribution function would be more suitable. The long tail (up to 43 communications per callsign) seems to be insignificant?
4. Figure 5. Some days seem to be missing data, it should be commented on in the text, why the data is missing.
5. Maybe some data about the recording quality would be beneficial. (E.g., some statistics about the signal-to-noise ratios, etc.)

2.3 Reviewer 3

The authors of this paper presented a robust pipeline to investigate the ATCO-pilot radio communication using speech and entity recognition algorithms together with ADS-B data fusion and action extraction algorithms with the purpose to detect the pilot maneuver response times after the instruction is given. Furthermore, the authors discuss the possibilities and limitations of the proposed approach.

The paper is well-structured and easy to read. The literature review section is on point and nicely sets this study into context. While reading the paper, some questions arose but majority of them were answered by further reading of the paper which suggests that the authors explained their approach in great detail.

While I think the paper covers the approach details very well, it appeared that the main purpose of the paper should be the quantification of the instruction-maneuver response times. I would expect a separate section providing the overview and distribution of the matched and detected response times. Right now the title seems to be a little misleading.

Also, in the section where the data used are presented, I would like to have some insight into why was this exact time period used. Is there any reason for that apart from availability? What was the traffic density in this period in comparison with other parts of the year? Were there any major disruptions during that period?

In Section 4.2 the authors are presenting some example transcriptions. I think this part is interesting and provides a good insight into the challenges of the matching process. However, some of the discussed parts of conversations are not present in Table 1, for example 23:59:46 or 23:59:38.

Finally, there is a small typo in the Open Data Statement (plateform).

3. Response - round 1

3.1 Response to reviewer 1

Response 1

The authors reference Lutz et al. [4] as a primary predecessor in this domain. The paper states that the proposed method improves upon [4] by using a more sophisticated deconfliction detection algorithm rather than simple thresholding. The paper would benefit significantly from a more explicit delineation of why the proposed NLP-centric pipeline is necessary compared to the approach in [4]. Does the addition of complex ASR/NER significantly improve the accuracy of the timestamping, or does it simply allow for better filtering of non-control communications?

Recommendation: Please expand Section 2 or 3 to detail the specific limitations of the Lutz et al. approach that this paper solves. If possible, a qualitative and/or quantitative comparison of a scenario processed by both methodologies would strengthen the contribution.

Response

The work of [1] served as the primary reference for the development of the present methodology. Their study demonstrated that applying ASR to ATC voice communications can enable the automated estimation of maneuver initiation times. While our pipeline builds on a similar overall backbone, it explicitly addresses several limitations identified in that earlier work. In particular, the proposed methodology introduces the following key contributions:

- the use of a machine learning–based NER module for callsign detection, rather than rule-based keyword identification, in order to reduce sensitivity to transcription errors and non-standard or unconstrained phraseology;
- an enhanced callsign–flight association strategy based on semantic similarity, replacing simple string-based matching, with the goal of increasing the robustness and overall matching rate between transcribed communications and ADS–B callsigns;
- an improved maneuver initiation detection algorithm based on a flight-plan–aware method developed in previous work [2], enabling more precise detection and reducing false positives compared with threshold-based approaches.

An additional explanatory paragraph has therefore been added at the beginning of Section 3 to clarify these methodological differences and position the proposed pipeline with respect to existing work.

In our view, the use of ASR and NER does not, in itself, improve the accuracy of maneuver timestamping. Their primary role is to enable reliable association between voice communications and the correct aircraft trajectories by identifying callsigns in the audio stream, and by filtering out non-operational or irrelevant communications. In this sense, ASR and NER act as enablers for track-data association rather than as direct contributors to temporal precision. The accuracy of the response-time estimation instead critically depends on the maneuver initiation detection algorithm applied to the surveillance data. This step is high-stakes, as it determines the exact moment at which the aircraft begins to react to an ATC instruction. For this reason, a central objective of the present work is to improve upon the approach of [1] by replacing threshold-based detection methods with a more sophisticated, flight-plan–aware algorithm. This design choice directly targets the precision and reliability of maneuver timing, which ultimately governs the quality of the computed response times.

Response 2

The title “Quantifying response times…” implies that the paper will present statistical distributions of response times (e.g., means, variances, distributions). However, the Results section (specifically Section 4.4) focuses on three specific case studies (Figures 6, 7, and 8). While Section 4.3 provides statistics on the matching process (Table 2), there are no aggregate statistics regarding the response times themselves. Consequently, in its current form, the paper reads more as a 'Methodology for Quantifying…' rather than the quantification itself.

Recommendation: I strongly suggest including a table or histogram showing the distribution of response times for all successfully matched deconfliction events in the dataset. If the number of matched events is too low to be statistically significant, the title should be adjusted to reflect that this is a feasibility study or a methodology proposal.

Response

Thank you for this comment. This discrepancy was indeed identified by all reviewers. Unfortunately, at the current stage of the project, it is not yet possible to apply the proposed methodology in a fully automated manner to a sufficiently large set of trajectories to derive statistically significant response-time distributions. This limitation primarily stems from the current implementation of the maneuver initiation detection algorithm, which requires access to accurate flight plans and is presently restricted to lateral deconfliction actions. As a consequence, matching trajectories that simultaneously satisfy the deconfliction criteria and have sufficiently complete associated ATC communications remain relatively rare and must be manually validated. We have explicitly acknowledged and discussed this limitation in the Discussion section. We have modified the title for "A Methodology for Quantifying Response Times for Deconfliction Actions Through ATC Communications."

Response 3

I recommend including Pellegrini, T., et al. (2018). "The Airbus Air Traffic Control speech recognition 2018 challenge" in the literature review. This is a seminal work regarding the difficulties of ATC transcription and callsign detection and provides relevant context for the ASR challenges discussed in Section 3.1.

Response

Thank you for bringing this additional material to our attention, which we were not previously aware of. This study is highly relevant to our literature review and, in our view, further supports the growing role of NLP techniques in ATC audio processing. In particular, it highlights the central importance of data-driven language models for future large-scale analysis of ATC–pilot communications. We have therefore incorporated the following paragraph into the manuscript to reflect this contribution (lines 125–133):

"The application of NLP techniques to the analysis of ATC–pilot communications can be traced back to 2018 with the *Airbus Air Traffic Control Speech Recognition Challenge* [3]. In this context, participating teams demonstrated that, on a small-scale dataset, Automatic Speech Recognition and callsign detection could achieve strong performance, with word error rates below 8% for ASR and callsign-detection F1 scores exceeding 80%. These results should nonetheless be interpreted with caution, as the models were trained on approximately 40 hours of high-quality audio collected in a single, well-defined airspace. Even so, they indicate that automatic transcription of ATC communications can reach a level of accuracy sufficient to support downstream tasks such as response-time estimation."

Response 4

The paper notes that MUAC was chosen partly because it is legally permissible to record ATC audio there. While this constraint limits "large-scale" reproducibility by external researchers in jurisdictions with stricter privacy laws (e.g., France), it presents a specific opportunity for operational stakeholders. I suggest explicitly stating that while open data collection is limited by these regulations, the pipeline is highly generalizable if deployed directly by Air Navigation Service Providers

(ANSPs). Since ANSPs already possess legal access to their own voice records, they could implement this pipeline internally to monitor systemic safety performance (e.g., sector complexity analysis) without facing the regulatory hurdles encountered by open-source researchers, provided that appropriate frameworks (e.g., Just Culture protocols, data anonymization) are established to address potential concerns regarding individual performance monitoring.

Response

This is a very interesting point that we have stated in the conclusion. Thank you for this comment.

"While constraints on ATC audio data quality and availability currently limit large-scale reproducibility in airspaces governed by stricter privacy regulations (e.g., France), this work presents opportunities for operational stakeholders. In particular, the pipeline could be highly generalizable if deployed directly by ANSPs, which legally possess access to high-quality ATC voice communications. Such stakeholders could implement the methodology internally to monitor systemic safety performance (for example, through sector complexity or workload analyses) without encountering the regulatory barriers faced by open-source researchers. This would, however, require the establishment of appropriate governance frameworks, such as Just Culture principles and data anonymization procedures, to address legitimate concerns related to individual performance monitoring."

Response 5

The introduction correctly notes that separation minima rely on assumptions of human latency. When this methodology is scaled, how should the resulting data be used? From a safety perspective, the average response time is less critical than the tail of the distribution (e.g., the 95th or 99th percentile). While the current analysis identifies individual response times (e.g., 22 seconds in Figure 7), future work should focus on quantifying the "long tail" of pilot reaction times. This is crucial for determining the "maximum probable delay" required for accurate "distance at CPA" risk modeling.

Response

This observation is entirely valid, as safety assessments typically focus on worst-case scenarios rather than average behavior. For example, ICAO bases its estimation of human response latency on conservative assumptions reflecting upper-tail behavior rather than mean values. One key advantage of the proposed methodology, once scaled, is its ability to collect a sufficiently large set of observed response times to statistically estimate their underlying distribution. Such large-scale empirical characterization would enable accurate estimation of high-order quantiles, and thus allow the computation of the probability that a pilot response time exceeds a given threshold. These probabilities can then be directly integrated into collision risk models, where response-time uncertainty plays a critical role. By adjusting operational thresholds accordingly, it becomes possible to ensure that the resulting accident probability within a given airspace remains consistent with the target level of safety.

3.2 Response to reviewer 2

Response 1

The title of the paper is "Quantifying response times for deconfliction actions through ATC communications", in the introduction, the authors define the response time as "the elapsed time between the detection of a potential conflict and the initiation of the corresponding action to mitigate it." However, the method only covers the pilots response time; the air traffic controllers time between detection of an issue and issuing a corrective action to pilot is not handled. Paper title, abstract and Introduction should make the scope of the method explicitly clear to the reader.

Response

Thank you for this comment. You are correct, this inconsistency resulted from an imprecise use of terminology on our side, where the terms reaction time and response time were mixed.

The overall reaction time process in a deconfliction scenario can be decomposed into multiple components: (i) the controller's response time, defined as the interval between conflict detection and issuance of an instruction; (ii) the pilot's response time, defined as the interval between receipt of the controller's clearance and initiation of the maneuver; and (iii) the maneuver completion time, corresponding to the time between the ATC instruction and full execution of the maneuver.

In the present paper, we focus specifically on the pilot response time. The methodology could, in principle, be extended to estimate maneuver completion times as well. However, estimating the controller's reaction time is considerably more challenging, as the exact moment at which a conflict is cognitively detected by the controller cannot be directly inferred from operational data. Addressing this question would likely require complementary sources of information, such as physiological measurements or eye-tracking data.

We have revised the abstract and introduction to clarify these definitions and ensure consistent terminology throughout the manuscript.

Response 2

While the 'Results' section provides an exhaustive analysis of the individual processing steps – such as callsign matching rates and conversation reconstruction – it does not present any statistics on the pilot response times themselves. If the scope of the paper is such analysis, it explicitly be highlighted in the paper title and introduction.

Response

Thank you for this comment. This concern is entirely valid and was indeed raised by all reviewers. While the ultimate objective of the proposed methodology is to collect a sufficiently large number of pilot response times to estimate their distribution, and subsequently integrate the associated probabilities into global collision risk models, the limitations discussed in the paper currently prevent fully automated large-scale deployment. At this stage, the methodology still requires manual verification of each identified case, primarily due to constraints in the conflict resolution action detection algorithm. Further improvements to this component will be necessary to enable scalable and fully automated response time estimation. To better reflect the current scope and maturity of the work, we have revised the title of the paper to: "A Methodology for Quantifying Response Times for Deconfliction Actions Through ATC Communications."

Response 3

Figure 4, maybe cumulative distribution function would be more suitable. The long tail (up to 43 communications per callsign) seems to be insignificant?

Response

The term cumulative distribution function would imply a representation of the cumulative proportion (or count) of communication threads with a size less than or equal to a given value. For example, at $x = 3$, a CDF would represent the number of callsign-associated threads containing at most three voice communications. However, the figure in question does not display a cumulative distribution. Instead, it represents the empirical distribution itself: for instance, approximately 2,000 callsigns are associated with exactly three voice communications. Our intention with this representation was to highlight uneven associations. In an ideal scenario, each callsign should be linked to an even number of communications (i.e., controller instruction + pilot readback). The non-cumulative form makes such imbalances more immediately visible. To avoid confusion, we have clarified the figure caption.

We agree that, given the scale chosen for the figure, the long tail above 10 communications appears visually insignificant. Nevertheless, we chose to display the full range up to 43 (the maximum observed value) in order to highlight the presence of unrealistically long communication threads. In practice, it is highly unlikely that a single pilot–controller exchange would legitimately consist of 43 transmissions. The fact that our algorithm identifies threads of this length points to limitations in the callsign association process, particularly in cases where distinct exchanges are incorrectly merged. Retaining the full range therefore serves to illustrate these edge cases and makes the limitations of the current approach more transparent. Figure 4 should be interpreted in conjunction with Figure 5, which provides a complementary perspective. Although the number of threads exceeding 10 communications is relatively small, Figure 5 shows that such cases are not negligible and warrant consideration when assessing the robustness of the association pipeline.

Response 4

Figure 5. Some days seem to be missing data, it should be commented on in the text, why the data is missing.

Response

We stated in Section 4.1 that, during the acquisition period (25.08.26–17.09.26), several days are missing from the analysis (5–8.09.26 and 13.09.26). These gaps are due to either audio data loss caused by stability issues with the VHF dongle or to missing ADS–B data. Since temporal continuity is not required for the objectives of this study, these days were excluded from the analysis. To avoid ambiguity, we have also added a clarifying sentence in the caption of Figure 5.

Response 5

Maybe some data about the recording quality would be beneficial. (E.g., some statistics about the signal-to-noise ratios, etc.)

Response

Thank you for this suggestion. While audio quality is indeed important, classical SNR metrics are difficult to compute meaningfully for non-stationary VHF ATC recordings with heterogeneous noise sources. We therefore rely on downstream performance indicators (ASR accuracy, callsign matching rates) as operational proxies for recording quality. A more detailed acoustic analysis is left for future work.

3.3 Response to reviewer 3

Response 1

While I think the paper covers the approach details very well, it appeared that the main purpose of the paper should be the quantification of the instruction-maneuver response times. I would expect a separate section providing the overview and distribution of the matched and detected response times. Right now the title seems to be a little misleading.

Response

Thank you for this comment. All three reviewers correctly identified a discrepancy between the scope suggested by the original title and the results presented in the manuscript. As discussed in the paper, the current limitations of the methodology (particularly those related to the deconfliction maneuver initiation detection algorithm) mean that response time estimation still requires manual validation of each identified case. While the framework enables structured quantification, it is not yet fully automated at scale. To better reflect the actual scope and maturity of the work, we have revised the title to: "A Methodology for Quantifying Response Times for Deconfliction Actions Through ATC Communications."

Response 2

Also, in the section where the data used are presented, I would like to have some insight into why was this exact time period used. Is there any reason for that apart from availability? What was the traffic density in this period in comparison with other parts of the year? Were there any major disruptions during that period?

Response

Thank you for this comment. The primary reason for selecting this specific time period was data availability. We did not have access to historical recordings of ATC voice communications and therefore conducted our own recording campaign. The selected period corresponds to the time during which data acquisition was operational and the recordings were of sufficient quality for analysis. The first few days of the campaign were excluded, as we were still refining the VHF recording setup and the audio quality was not yet stable. We aimed to collect data over a sufficiently long duration to ensure diversity in operational conditions and traffic patterns. A few days (5–8 September and 13 September) were subsequently excluded due to VHF signal loss or missing ADS–B data. For the purposes of this study, we did not consider it necessary to select specific time periods based on traffic density or seasonal patterns, as the objective was only to gather a sample of ATC voice communications. Furthermore, temporal continuity was not required for the methodological evaluation conducted here.

We also explored the use of LiveATC audio archives. However, archived data are only available for approximately seven days, and the recording quality is not consistently reliable. In addition, the LiveATC feeders proved insufficiently stable for downloading and storing large volumes of en-route communications. For these reasons, we opted to implement our own dedicated recording solution.

Response 3

In Section 4.2 the authors are presenting some example transcriptions. I think this part is interesting and provides a good insight into the challenges of the matching process. However, some of the discussed parts of conversations are not present in Table 1, for example 23:59:46 or 23:59:38.

Response

Thank you for pointing this out. The original version of the table was slightly longer, but we decided to shorten it because the information related to ASR performance was redundant with results presented elsewhere in the manuscript. In doing so, we inadvertently retained comments referring to communications that had been removed from the table. We have now corrected the text to ensure consistency with the revised table, without omitting any relevant information.

References

[1] Michael Lutz, Gano Broto Chatterji, and Husni R Idris. “Characterization of response times based on voice communication and traffic surveillance data”. In: *Proceedings of the AIAA AVIATION 2022 Forum.* 2022. DOI: 10.2514/6.2022-3762.

[2] Kim Gaume, Richard Alligier, Nicolas Durand, David Gianazza, and Xavier Olive. “Extracting Lateral Deconfliction Actions from Historical ADS-B data with Median Regression”. In: *International Conference on Research in Air Transportation.* 2024. URL: https://drive.google.com/file/d/1F6FOorwkBUkjuKdGE4i2uiYnBODL4Y0d/view?usp=sharing.

[3] Thomas Pellegrini, Jérôme Farinas, Estelle Delpech, and François Lancelot. “The airbus air traffic control speech recognition 2018 challenge: Towards ATC automatic transcription and call sign detection”. In: *arXiv preprint arXiv:1810.12614* (2018).