Journal of Open Aviation Science (2026), Vol.**4**
doi:10.59490/joas.2026.8738

JOAS

EDITORIAL

Reviews and Responses for Can YouTube Stream Recordings Improve Speech Recognition for Air Traffic Control?

Authors: Niclas Wüstenbecker, Oliver Ohneiser, and Matthias Kleinert

Reviewers: Allan Tart and Max Li

Editor: Junzi Sun

1. Original paper

The DOI for the original paper is https://doi.org/10.59490/joas.2026.8477

2. Review - round 1

2.1 Reviewer 1

Section 2 Related Work – in the current version authors mix the current state of the art of data resources and modelling techniques. It might be beneficial from the readability perspective to divide the two aspects into subsections.

For the multi-model automatic transcription, authors have selected three models. It would be interesting to know which other models were considered and reasons for not selecting. Table 2 indicates that authors had access to other models.

On page 9, authors say: "During analysis, we found that a small portion of videos include players who claim to be professional ATCOs in real life." – authors claim that the fact it does not matter that the ATC speech originates from actual operational ATCOs, it's the phraseology that is the key, then why mention this. Maybe consider removing the sentence.

The quality assessment process in Section 4.2 is rigorous and well described. As minor improvement some references how others are doing the assessment and how the proposed method compares could be added.

The results of the quality assessment are well described in Section 4.2., maybe consider some visual presentation of numerical results to improve the overall readability.

2.2 Reviewer 2

There also exist publicly available channels that feature real ATC communications associated specifically with off-nominal or safety-critical events (e.g., VASAviation). It would be useful for the authors to discuss whether incorporating such sources was considered and, if not, what limitations or trade-offs prevented their inclusion.

The manuscript does not clearly articulate what safeguards exist against LLM hallucination during transcript synthesis, particularly in cases where all upstream ASR models may be uncertain or incor-

rect. A discussion of whether automated consistency checks, syntactic validators, or grammar-based filters were considered would strengthen confidence.

It is not entirely clear from the current description how internally consistent these long recordings are. Do they typically remain within a single facility type, or do they mix multiple domains over time? The manuscript would benefit from a more explicit discussion of potential artifacts introduced by the way these videos are uploaded and curated on YouTube.

Figure 3 and the accompanying discussion regarding post-filtered sample durations remain somewhat vague. Are these short clips treated as independent utterances with no temporal context retained, or is any sequential structure preserved?

3. Response - round 1

3.1 Response to reviewer 1

Section 2 Related Work – in the current version authors mix the current state of the art of data resources and modelling techniques. It might be beneficial from the readability perspective to divide the two aspects into subsections.

Response

We have split Section 2 into two subsections: “Data Resources” and “Modelling Techniques” to improve readability.

For the multi-model automatic transcription, authors have selected three models. It would be interesting to know which other models were considered and reasons for not selecting. Table 2 indicates that authors had access to other models.

Response

We have expanded Section 3.4 to clarify the model selection rationale. The three fusion models were selected based on two criteria: (1) compatibility with vLLM for efficient inference over the 2,000+ hours of data, and (2) complementary error characteristics. This led to the exclusion of Wav2Vec 2.0 models, which lack vLLM support and would have been prohibitively slow for the fusion process at this scale. The additional models shown in the evaluation (Section 4.2) were evaluated post hoc for comparison purposes only and were not part of the fusion pipeline.

On page 9, authors say: “During analysis, we found that a small portion of videos include players who claim to be professional ATCOs in real life.” – authors claim that the fact it does not matter that the ATC speech originates from actual operational ATCOs, it’s the phraseology that is the key, then why mention this. Maybe consider removing the sentence.

Response

We agree with the reviewer and have removed the sentence from the manuscript.

The quality assessment process in Section 4.2 is rigorous and well described. As minor improvement some references how others are doing the assessment and how the proposed method compares could be added.

Response

We have added references establishing the Word Error Rate (WER) as the de facto standard metric for ASR evaluation, citing Errattahi et al. and Ruiz et al. The revised text now explicitly states that we utilize "what is widely accepted as the de facto standard metric for ASR performance evaluation" with supporting citations.

The results of the quality assessment are well described in Section 4.2., maybe consider some visual presentation of numerical results to improve the overall readability.

Response

We have added a bar chart (Figure 5) presenting the WER results for all evaluated models, separated by speaker role (controller vs. pilot), to complement the textual discussion and improve readability.

3.2 Response to reviewer 2

There also exist publicly available channels that feature real ATC communications associated specifically with off-nominal or safety-critical events (e.g., VASAviation). It would be useful for the authors to discuss whether incorporating such sources was considered and, if not, what limitations or trade-offs prevented their inclusion.

Response

We have added a dedicated paragraph at the end of Section 3.1 addressing this point. We deliberately excluded real-world ATC recordings because our paper aims to validate the general feasibility of the proposed approach without the additional challenges of degraded VHF radio audio quality. Such degraded audio would likely introduce substantially more transcription errors during both the multi-model ASR and LLM fusion stages, making it difficult to isolate the contribution of the pipeline itself. Incorporating real-world recordings with tailored audio preprocessing is discussed as a promising direction for future work in both Section 3.1 and the Conclusion.

The manuscript does not clearly articulate what safeguards exist against LLM hallucination during transcript synthesis, particularly in cases where all upstream ASR models may be uncertain or incorrect. A discussion of whether automated consistency checks, syntactic validators, or grammar-based filters were considered would strengthen confidence.

Response

We have added a new paragraph at the end of Section 3.5 that discusses multiple layers of hallucination mitigation present in our approach:

- Structured JSON output format enables automatic validation; non-conforming responses are discarded and reprocessed.
- The system prompt was developed on a small held-out subset of approximately 100 samples to ensure generalization without overfitting to the dataset.
- The multi-model input constrains the LLM to selecting and combining existing ASR hypotheses rather than generating transcripts from scratch, limiting unconstrained hallucination.
- Word frequency analysis (Section 4.1) empirically validates that the output distribution matches real-world ATC patterns.
- Manual verification on a 120-minute evaluation set (Section 4.2) provides direct quality assessment.

- The downstream model effectively averages out residual noise during training, as evidenced by the fine-tuned model surpassing even the pseudo-label quality.

We further acknowledge in the Conclusion that additional automated safeguards – such as ATC grammar validators or multi-LLM consensus mechanisms – represent a promising direction for future work.

It is not entirely clear from the current description how internally consistent these long recordings are. Do they typically remain within a single facility type, or do they mix multiple domains over time? The manuscript would benefit from a more explicit discussion of potential artifacts introduced by the way these videos are uploaded and curated on YouTube.

Response

We have added a new paragraph in Section 4.1 addressing both aspects. Regarding internal consistency, the working position classification was derived from each video's title and description metadata. Virtual ATC streams on VATSIM typically feature a single controller operating one position per session, as indicated by the session title. Since our pipeline segments each video into independent short utterances (1–20 seconds), any within-video domain mixing is mitigated at the individual sample level: each training sample represents a self-contained utterance regardless of the broader session context. Regarding YouTube-specific artifacts, we now explicitly reference the comprehensive quality assessment in Section 4.2, which systematically characterizes artifact types including intermittent streamer commentary, background music, breaks in operational realism, and synthetic voices from simulation software, along with their prevalence rates in the evaluation subset.

Figure 3 and the accompanying discussion regarding post-filtered sample durations remain somewhat vague. Are these short clips treated as independent utterances with no temporal context retained, or is any sequential structure preserved?

Response

We have added a clarifying passage at the end of Section 3.3. Each resulting clip is treated as an independent utterance throughout all subsequent pipeline stages, including transcription, fusion, and model training. No sequential context or temporal ordering between adjacent segments is preserved. While this simplifies the pipeline and aligns with standard ASR training practices, we acknowledge in the revised text and in the Conclusion that preserving and leveraging sequential context – such as the structure of controller–pilot exchange sequences – represents a potential avenue for future improvement.