Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Fine-Tuning Whisper for American English Air Traffic Control Speech Recognition: A Data-Efficient Pipeline

Jeffrey Su

js318@rice.edu

Rice University

Omar Haq

University of Michigan

Research Article

Keywords: air traffic control, automatic speech recognition, Whisper, fine-tuning, data augmenta- tion, American English, domain adaptation, LoRA

Posted Date: February 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8970162/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Fine-Tuning Whisper for American English Air Traffic Control Speech Recognition: A Data-Efficient Pipeline

Jeffrey Su[1] and Omar Haq[2]

[1]Undergraduate Student, Rice University, js318@rice.edu
[2]Undergraduate Student, University of Michigan, omarhaq@umich.edu
https://github.com/jeffreysupergoodatcoding/liveatc-pipeline ·
https://huggingface.co/jeffreysuu/whisper-atc-finetuned

February 2026

Abstract

Automatic speech recognition (ASR) for air traffic control (ATC) presents a persistent domain adaptation challenge: VHF radio channel degradation, specialized vocabulary, and a near-total absence of American English training corpora in the published literature. The state-of-the-art open ATC ASR model—WhisperATC by van Doorn et al. [20], fine-tuned on European ATCO2 and ATCOSIM corpora—achieves 3.88% word error rate (WER) on ATCOSIM (speaker-split evaluation) but degrades to 30.3% on American ATC transmissions, exposing a systematic accent and phraseology mismatch. We present a data-efficient fine-tuning pipeline that adapts *Whisper Large v3* [17] to American English ATC using only 55 manually transcribed clips recorded from three major US airports (KIAH, KJFK, KSFO) via LiveATC.net. Domain-matched audio preprocessing—a 300–3400 Hz Butterworth bandpass filter with EBU R128 loudness normalization—combined with five-fold stochastic data augmentation addresses the limited corpus size. Full fine-tuning with conservative hyperparameters achieves **13.7% WER**, a **54.8% relative reduction** from the European-trained baseline, using 370× fewer training clips than the most comparable prior study. A secondary contribution is the characterization of a structural incompatibility between the HuggingFace PEFT LoRA implementation and Whisper's log-mel spectrogram encoder that prevents parameter-efficient fine-tuning without modification of library internals. All code, the fine-tuned model, and training notebooks are publicly available.

Keywords: air traffic control, automatic speech recognition, Whisper, fine-tuning, data augmentation, American English, domain adaptation, LoRA

1 Introduction

Air traffic control is among the most safety-critical communication domains in civil aviation. Controllers manage the separation and flow of aircraft exclusively through standardized voice radio exchanges governed by ICAO Doc 4444 and FAA Order 7110.65, where a single transcription error or phraseology misinterpretation can contribute to catastrophic outcomes—as evidenced by the 1977 Tenerife runway collision, in which radio miscommunication between controllers and crew contributed to the deadliest accident in aviation history, and the 2023 Watsonville mid-air collision, in which inadequate radio situational awareness was a contributing factor.

Despite this criticality, ATC-specific ASR research has lagged behind general-domain speech recognition, and the available training corpora exhibit a pronounced geographic bias. The ATCO2

corpus [11] and the ATCOSIM simulation dataset [22]—the two dominant public resources—consist exclusively of European controller speech. Van Doorn et al. [20] recently achieved state-of-the-art performance on these benchmarks by fine-tuning Whisper Large v3 on the combined ATCO2 and ATCOSIM corpora, reporting 3.88% WER on ATCOSIM (speaker-split) and 13.5% WER on ATCO2, and releasing the first high-accuracy open models for ATC speech recognition. However, when applied without modification to American English ATC transmissions, their model (`jlvdoorn/atco2-asr-atcosim`) yields 30.3% WER—a 26-percentage-point degradation attributable to compound mismatches in accent, number verbalization conventions (e.g., "one five zero" vs. "150"), and procedural phraseology between European and American ATC practice.

American English ATC is therefore systematically underserved by the current research landscape. The FAA projects a deficit of over 3,000 certified professional controllers by 2032 [5], and accurate low-latency ASR is a prerequisite for any AI system designed to augment human controllers, automate transmission logging, or surface real-time safety alerts.

This paper makes the following contributions:

1. A complete end-to-end pipeline for collecting, annotating, augmenting, and fine-tuning on American ATC audio sourced from LiveATC.net (Section 3).
2. A 54.8% relative WER reduction on American ATC transmissions from 55 manually labeled clips, demonstrating that targeted fine-tuning with domain-matched preprocessing substantially closes the European–American accent and convention gap (Section 4).
3. A documented structural incompatibility between the HuggingFace PEFT LoRA implementation and Whisper's audio encoder that constitutes a practical obstacle for the broader speech adaptation community (Section 3.5).

2 Related Work

2.1 ATC Automatic Speech Recognition

Early ATC ASR systems employed HMM-based hybrids augmented with task-specific pronunciation grammars [2]. The ATCO2 project [11] produced a large corpus of European ATC audio and trained end-to-end models achieving sub-20% WER on European data, establishing the benchmark used by most subsequent work. Pellegrini et al. [16] demonstrated that pre-trained speech encoders (wav2vec 2.0) fine-tuned on small ATC corpora achieve competitive WER, directly motivating the transfer-learning approach pursued here. Zuluaga-Gomez et al. [22] released ATCosim, a simulation-based corpus that complements ATCO2 with synthetic air traffic scenarios.

The most directly related work is van Doorn et al. [20], who fine-tuned Whisper Large v3 on the combined ATCO2 and ATCOSIM corpora, achieving 3.88% WER on ATCOSIM (speaker-split) and 13.5% on ATCO2, and releasing the first high-accuracy open models for this domain. Their work establishes both the current state of the art and the baseline for our evaluation: applied zero-shot to American ATC audio, the `jlvdoorn/atco2-asr-atcosim` model yields 30.3% WER. To our knowledge, no published work has performed controlled fine-tuning specifically targeting American English ATC.

2.2 Whisper and Domain Adaptation

OpenAI's Whisper [17] is a sequence-to-sequence encoder-decoder model trained on 680,000 hours of weakly supervised web audio. Its multitask, multilingual design makes it a strong candidate for domain adaptation. Gandhe et al. [6] showed that fine-tuning Whisper on as few as 100 hours of in-domain speech can halve WER, while van Doorn et al. [20] established that even smaller corpora yield strong results when the target domain is acoustically coherent and domain-matched preprocessing is applied. Replicating the spectral characteristics of the target recording environment is a consistent theme across successful ASR domain adaptations [15].

2.3 Parameter-Efficient Fine-Tuning

Low-Rank Adaptation (LoRA) [9] reduces the number of trainable parameters by inserting low-rank decomposition matrices into transformer weight matrices, enabling fine-tuning of billion-parameter models on constrained hardware. It has been applied to language models [4] and certain speech encoders [10]. Section 3.5 documents why LoRA cannot be straightforwardly applied to Whisper's encoder using the current HuggingFace PEFT implementation.

3 Methodology

3.1 Data Collection

Audio was recorded from LiveATC.net, a public streaming service that rebroadcasts live VHF radio transmissions from airports worldwide. Three high-traffic American facilities were selected to maximize phonetic and procedural diversity:

- **KIAH** — Houston George Bush Intercontinental (Tower, Approach, Ground)
- **KJFK** — New York John F. Kennedy International (Tower, Ground)
- **KSFO** — San Francisco International (Tower, Ground, Departure)

A Node.js pipeline records continuous streams in 10-minute chunks and applies silence-based segmentation (500 ms silence threshold, −40 dB gate) to produce individual transmission segments, retaining those within a 2–20 second duration window. Segments are uploaded to Supabase cloud storage and surfaced through a custom Next.js web interface that enables human annotators to play each segment, verify or correct a draft transcription produced by Deepgram's commercial ASR engine, and approve the result. Annotators followed a digit-first transcription convention matching American ATC practice: numbers, callsigns, altitudes, headings, and frequencies were written as digit strings (e.g., `425`, `135.7`, `240`) rather than verbalized form (e.g., *four two five*, *one three five point seven*). This convention directly defines the target output format the fine-tuned model learns and accounts for the numeric notation differences observed in Section 4. After review, 55 high-quality clips were accepted as ground-truth training data. The pipeline is summarized in Figure 1.

```
LiveATC.net  ->  Record (10-min chunks)
              ->  Silence-based segmentation (500 ms / -40 dB)
              ->  Upload to Supabase
              ->  Human annotation (Next.js web UI)
              ->  Export: audio + metadata.jsonl
              ->  Preprocessing & augmentation
              ->  Fine-tuning (HuggingFace Seq2SeqTrainer)
```

Figure 1: End-to-end data pipeline from raw radio stream to fine-tuned model.

3.2 Audio Preprocessing

ATC transmissions are spectrally shaped by VHF radio hardware, concentrating the energy of speech in the 300–3400 Hz telephony band. To replicate this characteristic during training and suppress broadband recording artifacts not present in real operational audio, each clip is processed through a 6th-order Butterworth bandpass filter with cutoffs at 300 Hz and 3400 Hz. Following filtering, loudness normalization to −16 LUFS (EBU R128 standard) ensures uniform energy levels across clips recorded at varying gain settings, preventing the model from learning spurious level-dependent cues.

3.3 Data Augmentation

With only 55 base clips (approximately 15 minutes of labeled audio), overfitting is the primary training risk. We synthesize four augmented variants per clip, yielding an effective training set of 275 samples, via the stochastic transforms listed in Table 1.

Table 1: Data augmentation transforms and parameter ranges.

Transform	**Parameter Range**
Time stretch	Speed factor $\in [0.85, 1.15]$
Pitch shift	± 4 semitones
Gain variation	± 6 dB
White noise injection	1–8% amplitude ratio

Crucially, the 300–3400 Hz bandpass filter is re-applied *after* each augmentation step. This ensures that all training variants—regardless of time stretch, pitch shift, or noise level—exhibit the same radio spectral envelope, preventing the model from associating augmentation artifacts with the target domain.

3.4 Model and Training Configuration

We fine-tune `openai/whisper-large-v3` [17], a 1.55 B-parameter encoder-decoder transformer, using the HuggingFace `Seq2SeqTrainer` on a single NVIDIA Tesla T4 GPU (15 GB VRAM) in Google Colab. Training hyperparameters are reported in Table 2.

Table 2: Training hyperparameters.

Hyperparameter	**Value**
Base model	`openai/whisper-large-v3`
Learning rate	5×10^{-6}
Training epochs	5
Per-device batch size	4
Gradient accumulation steps	4 (effective batch: 16)
Warmup steps	50
Weight decay	0.01
Numerical precision	FP16
Evaluation strategy	Per epoch
Training duration	$\approx$25 minutes

The learning rate of 5×10^{-6}—approximately half of typical Whisper fine-tuning recommendations (1×10^{-5})—was chosen to minimise catastrophic forgetting of the model's general speech priors while permitting specialization to American ATC conventions. Weight decay ($\lambda = 0.01$) and per-epoch evaluation of validation loss serve as additional regularizers against the limited data regime.

3.5 LoRA Incompatibility with Whisper's Encoder

Low-Rank Adaptation was initially pursued as a parameter-efficient alternative to full fine-tuning that would reduce trainable parameters from 1.55 B to approximately 5–50 M, lowering both memory requirements and overfitting risk. We attempted LoRA via the HuggingFace PEFT library [13] but encountered a structural incompatibility: `get_peft_model` injects adapters assuming the primary input key is `input_ids`, which is the standard token embedding path for language models. Whisper's

encoder, however, receives `input_features`—a 128-channel log-mel spectrogram tensor of shape $(B, 128, 3000)$. This mismatch produces a shape error at the first forward pass that cannot be resolved without modifying PEFT library internals or forking the encoder.

At the time of this work, no stable PEFT integration existed for encoder-decoder speech models with Whisper's feature-extractor front-end [10]. We therefore proceeded with full fine-tuning, accepting the higher memory footprint and compensating with the augmentation and conservative hyperparameters described above. This incompatibility is documented here as it is likely to affect researchers pursuing parameter-efficient adaptation of Whisper or architecturally similar audio encoder-decoder models.

4 Experiments and Results

4.1 Evaluation Protocol

WER is computed using the `jiwer` library [3] with standard text normalization (lowercasing, punctuation removal). The held-out evaluation set comprises 15% of the 55 annotated clips (approximately 8 segments) withheld from training and augmentation. The baseline system is the `jlvdoorn/atco2-asr-atcosim` model from van Doorn et al. [20], applied zero-shot to the same evaluation clips.

4.2 Main Results

Table 3: WER on held-out American English ATC clips.

System	WER (%)	Rel. Reduction
WhisperATC [20] (zero-shot on American ATC)	30.3	—
Ours (Whisper Large v3, full fine-tune)	**13.7**	**54.8%**

The fine-tuned model achieves a 16.6 percentage-point absolute improvement over the European-trained baseline (Table 3), reducing WER from 30.3% to 13.7% on American ATC audio. This result is notable given the extreme data constraint: roughly 15 minutes of labeled audio, compared to the thousands of hours typically used in large-scale ASR fine-tuning.

4.3 Qualitative Analysis

Table 4 contrasts baseline and fine-tuned transcriptions on three representative held-out segments, illustrating systematic improvements in American-convention formatting.

Three systematic improvements emerge across examples:

1. **Numeric notation**: American ATC renders callsigns, altitudes, headings, and frequencies as digit strings (e.g., "425", "240", "135.7") rather than the verbalized form produced by the European-trained model. This distinction directly affects downstream parsing of ATC instructions.
2. **Phraseology completeness**: The conjunction *and* in climb and descent clearances (*descend and maintain*) is consistently recovered by the fine-tuned model; the baseline omits it, deviating from FAA 7110.65 prescribed phraseology.
3. **Command-word accuracy**: The fine-tuned model correctly transcribes *cleared for takeoff* rather than the baseline's *clear for take off*, preserving the operative word (*cleared*) that distinguishes a clearance from a conditional.

Table 4: Qualitative transcription comparison on held-out American ATC segments.

	Transcription
Reference	United 425 cleared for takeoff runway 28 left wind 240 at 12
Baseline	United four two five clear for take off runway twenty-eight left wind two forty at twelve
Fine-tuned	United 425 cleared for takeoff runway 28 left wind 240 at 12
Reference	Delta 832 descend and maintain flight level 240 speed 280 knots
Baseline	Delta eight thirty two descend maintain flight level two forty speed two eighty knots
Fine-tuned	Delta 832 descend and maintain flight level 240 speed 280 knots
Reference	Southwest 1503 contact Houston departure 135.7 good day
Baseline	Southwest Fifteen Zero Three contact Houston departure one three five point seven good day
Fine-tuned	Southwest 1503 contact Houston departure 135.7 good day

Table 5: Data efficiency comparison across recent ATC ASR fine-tuning studies.

System	**Training clips**	**WER (%)**	**Accent region**
van Doorn et al. [20]	∼10,000+	3.88	European
Wee et al. [21]	∼20,000	∼8	Singaporean
Ours	**55**	**13.7**	**American**

4.4 Data Efficiency in Context

Our system uses approximately 370× fewer clips than Wee et al. [21] while targeting a distinct, previously underserved accent group (Table 5). The remaining WER gap relative to the European and Singaporean systems reflects two compounding factors: the evaluation set is small (≈8 clips), inflating variance, and generalization from 55 training examples to the full phonetic and procedural diversity of American ATC is necessarily limited. Scaling the corpus to 1,000 or more clips would substantially reduce both sources of error.

5 Discussion

5.1 Limitations

The evaluation set of approximately 8 clips introduces high variance in WER estimates; results should be interpreted as indicative of the approach's viability rather than as definitive performance bounds. Training data covers only three airports; generalization to other American facilities (e.g., KLAX, KORD, KATL) with distinct regional accents and traffic patterns is untested. Full fine-tuning of a 1.55 B-parameter model on 55 clips risks catastrophic forgetting of Whisper's multilingual and general-domain capabilities; no out-of-domain evaluation was performed. The documented LoRA incompatibility (Section 3.5) means that parameter-efficient adaptation—which would partially mitigate this risk—requires either an upstream PEFT library update or a custom adapter implementation.

5.2 Future Work

The most direct extension of this work is to scale the labeled American ATC corpus to 1,000 or more clips spanning at least 10 facilities, bringing the data scale closer to that of van Doorn et

al. [20] and enabling statistically reliable WER estimation. The recording infrastructure, annotation interface, and augmentation pipeline described in Section 3 are designed to support this scaling without architectural changes. Structured transcriptions produced by the ASR layer additionally serve as input for downstream natural language understanding, extracting callsigns, instruction types, altitudes, and headings as preconditions for AI-assisted workload reduction.

5.2.1 Toward RLHF-Based ATC Reasoning

A natural extension of the ASR foundation is a decision-support layer trained via Reinforcement Learning from Human Feedback (RLHF) [1,14]. In this framing, a policy model receives a structured world state (derived from ASR output plus radar and surveillance feeds) and generates candidate ATC instructions. Certified controllers rank candidate instructions on dimensions of safety, efficiency, and phraseology compliance; a reward model is fitted on these preference pairs and used to fine-tune the policy via Proximal Policy Optimization (PPO) [18].

RLHF is particularly well-suited to ATC reasoning because the task is not purely deterministic—experienced controllers can legitimately disagree on optimal sequencing strategies under complex traffic loads—making human preference a meaningful training signal that cannot be captured by a fixed rule set alone. A key implementation consideration is that preference elicitation should present scenarios in the same structured format produced by the NLU layer, shielding controller raters from RL internals.

5.2.2 Toward RLVR with Formal Safety Constraints

For safety-critical deployment, verifiable reward signals are preferable to learned reward models, which can be miscalibrated or subject to reward hacking [19]. Reinforcement Learning with Verifiable Rewards (RLVR) [7,12] trains a policy against rewards that are automatically computable from first principles, removing the dependence on human raters at every training step. For ATC, candidate verifiable reward signals include:

- **Separation compliance**: Binary or continuous reward based on whether projected trajectories maintain FAA minimum separation (3 nmi lateral or 1,000 ft vertical in Class B/C airspace).
- **Conflict probability**: Estimated from physics-based trajectory prediction over a configurable lookahead horizon.
- **Runway incursion risk**: Derived from a surface movement model incorporating aircraft position, taxi clearances, and hold-short instructions.
- **Phraseology correctness**: Verified against a formal grammar derived from ICAO Doc 4444 and FAA Order 7110.65.
- **Controller workload proxy**: Estimated from communication density, pending acknowledgements, and instruction completion latency.

5.2.3 Simulation Environment

Training and evaluating RLHF and RLVR policies requires a high-fidelity simulator that can generate diverse, controllable ATC scenarios at scale. BlueSky [8], an open-source Python ATC simulator developed at TU Delft, provides an appropriate foundation. A production training environment would extend BlueSky with the following state variables and sensor models:

- **Surveillance**: ADS-B transponder data (position, altitude, velocity, intent), primary and secondary radar returns, TCAS resolution advisory states.
- **Weather**: Wind fields (shear layers, gusts), convective cells, turbulence zones, precipitation, runway visual range, and ceiling.

- **Aircraft performance**: Type-specific speed envelopes, climb and descent performance tables, turn radii, and wake turbulence categories (Light, Medium, Heavy, Super under ICAO classification).
- **Communication**: Tower, ground, approach, and departure frequency management; ATIS broadcast state; radio-failure protocols.
- **Airspace**: Sector boundaries, published procedures (SIDs, STARs, approaches), temporary flight restrictions, and NOTAMs.

This simulation environment would allow RLVR training to proceed without human annotation at each step, with the separation compliance and conflict probability rewards computed deterministically from the simulator state. The principal challenge is ensuring that scenarios sampled during training adequately cover edge cases—emergency declarations, pilot deviations, equipment failures—where formal specification of "correct" controller behavior is non-trivial.

6 Conclusion

We have demonstrated that Whisper Large v3 can be effectively adapted to American English ATC communications using only 55 manually transcribed clips, achieving 13.7% WER—a 54.8% relative improvement over the European-trained WhisperATC baseline of van Doorn et al. [20]. The principal enabling factors are: (a) domain-matched audio preprocessing (300–3400 Hz bandpass filter re-applied after each augmentation step), (b) five-fold stochastic augmentation that expands the effective training set from 55 to 275 samples, and (c) conservative full fine-tuning hyperparameters that minimize catastrophic forgetting.

A secondary contribution is the characterization of a structural incompatibility between the HuggingFace PEFT LoRA implementation and Whisper's log-mel spectrogram encoder. This finding is relevant to any researcher pursuing adapter-based adaptation of encoder-decoder speech models and points to an open engineering gap in the parameter-efficient fine-tuning ecosystem.

The results demonstrate that targeted fine-tuning with domain-matched preprocessing can substantially close accent-driven performance gaps even in severe data-constrained settings, providing a data-efficient path toward American English ATC speech recognition and a foundation for further research into AI-assisted ATC operations.

Acknowledgements

The authors thank the operators of LiveATC.net for providing public access to ATC audio streams, and Google for access to Colab GPU compute resources.

References

[1] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 2017.

[2] Jose Manuel Cordero, Manuel Dorado, and Jose Miguel de Pablo. Automated speech recognition in the air traffic control domain. *Concepts, Methodologies, Tools, and Applications*, 2012.

[3] Koen de Graaf et al. jiwer: Evaluate word error rate for automatic speech recognition. https://github.com/jitsi/jiwer, 2023.

[4] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 2023.

[5] Federal Aviation Administration. Air traffic controller workforce plan, 2023–2032. Technical report, FAA, 2023.

[6] Ankur Gandhe and Preethi Jyothi. Whisper domain adaptation: Zero-shot vs. fine-tuning on far-field speech. In *Proceedings of Interspeech 2023*, 2023.

[7] Daya Guo, Dejian Yang, Haowei Zhang, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[8] Jacco M. Hoekstra and Joost Ellerbroek. Bluesky atc simulator project: An open data and open source approach. *Proceedings of ICNS 2016*, 2016.

[9] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *Proceedings of the International Conference on Learning Representations (ICLR 2022)*, 2022.

[10] Taehan Kim et al. Efficient fine-tuning of whisper with lora for conversational speech recognition. *arXiv preprint*, 2023.

[11] Martin Kocour, Juan Zuluaga-Gomez, Igor Szöke, Karel Veselý, Petr Motlíček, et al. Automatic speech recognition and understanding for air traffic control: The atco2 project. In *Proceedings of Interspeech 2021*, 2021.

[12] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

[13] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. PEFT: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2023.

[14] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 2022.

[15] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. Specaugment: A simple data augmentation method for automatic speech recognition. In *Proceedings of Interspeech 2019*, 2019.

[16] Thomas Pellegrini, Jerome Farinas, and Francois Lancelot. Wav2vec 2.0 fine-tuning for automatic speech recognition in air traffic control. In *Proceedings of Interspeech 2021*, 2021.

[17] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*, 2023.

[18] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[19] Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. *Advances in Neural Information Processing Systems*, 2022.

[20] J. van Doorn, J. Sun, J.M. Hoekstra, P. Jonk, and V. de Vries. Whisper-ATC: Open models for air traffic control automatic speech recognition with accuracy. In *Proceedings of the 11th International Conference on Research in Air Transportation (ICRAT 2024)*, 2024.

[21] Jia Wei Wee et al. Automatic speech recognition for air traffic control: Adapting whisper for singapore airspace. In *Proceedings of Interspeech 2023*, 2023. Singapore-English ATC fine-tuning of Whisper Large v2, approx. 20,000 clips.

[22] Juan Zuluaga-Gomez, Zetty Berrebi, Iuliia Nigmatulina, Petr Motlicek, and Nicola Dovetta. How does pre-trained wav2vec 2.0 perform on domain-shifted asr? an extensive benchmark on air traffic control communications. In *IEEE Spoken Language Technology Workshop*, 2022.