Generating Realistic Air Traffic Control Voice Communication using AI-based Text-to-Speech Models

Dao Vu and Peng Wei
Department of Mechanical and Aerospace Engineering
George Washington University
Washington, D.C., USA
Email: {dao.vu, pwei}@gwu.edu

***Abstract*—Recent advances in neural text-to-speech (TTS) and zero-shot voice cloning enable increasingly realistic speech synthesis, introducing dual-use risks for safety-critical aviation communications. In Air Traffic Control (ATC), operational plausibility depends not only on perceptual naturalness but also on strict transcript fidelity, preservation of safety-critical tokens, speaker-identity consistency, and procedural cadence alignment. This work presents an ATC-oriented synthesis and evaluation framework built on an open-source neural TTS backbone with optional speaker conditioning. We introduce a multi-dimensional, aviation-aware evaluation methodology comprising intelligibility via Word Error Rate, critical-token F1, speaker embedding similarity, prosodic alignment, and dynamic time-warped log-mel spectral distance. Across 1000 controller utterances from ATCOSIM, the proposed synthesis framework achieves strong lexical fidelity and preservation of safety-critical tokens, with moderate speaker-embedding similarity and stable spectral-envelope alignment. Prosodic alignment, however, remains less consistent, exhibiting low average pitch-contour correlation and measurable deviations in speaking rate and pause structure. This semantic–prosodic asymmetry suggests that contemporary ATC-style neural synthesis can approach operational plausibility at the lexical and identity levels while retaining systematic temporal artifacts. These artifacts may provide measurable signals for communication resiliency analysis and synthetic-speech detection in safety-critical aviation environments.**

***Index Terms*—Air Traffic Control, Text-to-Speech Model, Voice Cloning, Aviation Cybersecurity, Speech Evaluation**

I. Introduction

Air Traffic Control (ATC) communication is operationally safety-critical and procedurally constrained. Radiotelephony phraseology, message structure, and communication procedures are standardized under ICAO Procedures for Air Navigation Services: Air Traffic Management (PANS-ATM, Doc 4444), which defines globally agreed upon ATC voice communication conventions [1]. Unlike conversational speech, ATC transmissions encode structured operational directives where lexical corruption or misinterpretation may directly affect separation assurance and flight safety.

Recent advances in neural speech generation have substantially improved perceptual naturalness and speaker generalization in text-to-speech, with progress in voice cloning and zero-shot synthesis expanding both capabilities and risks [2], [3]. Speaker-conditioned architectures achieve identity transfer by conditioning synthesis on learned speaker embeddings extracted from reference audio [4]. However, these systems are optimized for conversational quality rather than compliance with domain-specific procedural constraints, leaving their behavior under operational ATC communication insufficiently characterized.

Simultaneously, realistic voice synthesis introduces dual-use implications for safety-critical voice communication. Advances in voice cloning and zero-shot speaker adaptation enable increasingly convincing identity transfer, raising concerns about impersonation and unauthorized instruction injection in aviation contexts [3]. Public demonstrations have illustrated the feasibility of ATC-style voice cloning using widely available tools [5], underscoring the need to rigorously evaluate synthetic speech systems within a communication resiliency framework.

This work reframes ATC-oriented speech synthesis not merely as a quality optimization problem, but as a structured measurement problem under operational and adversarial considerations. We make three primary contributions. First, we formulate an ATC-aligned synthesis framework that explicitly characterizes transcript fidelity, critical-token preservation, speaker conditioning, and procedural cadence as operational constraints. Second, we develop a multi-dimensional, aviation-aware evaluation methodology grounded in established speech metrics, spanning intelligibility, semantic integrity, speaker embedding similarity, prosodic alignment, and spectral-envelope fidelity. Third, we interpret synthesis realism through a communication resiliency perspective, analyzing how these dimensions collectively influence operational plausibility and potential spoofing risk within Air Traffic Management ATM environments.

II. Problem Formulation

Let $x(t)$ denote a ground-truth ATC waveform with aligned transcript $T = (w_1, \ldots, w_N)$, where each token w_i may encode safety-critical semantic information such as altitude, heading, frequency, or call sign identifiers.

A pretrained neural TTS backbone generates a synthetic waveform $\hat{x}(t)$ conditioned on transcript T, optional speaker embedding e_s, and inference parameters θ controlling prosodic and acoustic characteristics, i.e., $\hat{x}(t) = \mathcal{G}(T, e_s, \theta)$. In our implementation, $\mathcal{G}$ corresponds to the open-source Chatterbox neural TTS model [6], with inference-time control over speaker embeddings and prosodic parameters.

In contrast to open-domain TTS systems that prioritize perceptual naturalness, ATC-oriented synthesis is inherently multi-constrained. Specifically, generation must satisfy:

1) **Transcript Fidelity**: Minimization of lexical deviation relative to T.
2) **Critical Token Preservation**: Exact retention of safety-critical semantic elements contained in $\{w_i\}$.
3) **Identity Consistency**: Controlled similarity between synthesized speech $\hat{x}(t)$ and a target speaker embedding e_s when conditioning is enabled.
4) **Procedural Cadence Consistency**: Alignment of speaking rate, pause structure, and pitch magnitude with operational ATC norms.

Accordingly, synthesis performance should be evaluated not by a single perceptual objective but through structured measurement across semantic, identity, spectral, and prosodic dimensions.

III. Threat Model and Resiliency Context

We consider an adversary capable of generating synthetic ATC-style audio conditioned on publicly available recordings of controller speech. The attacker's objective is to transmit plausible but unauthorized instructions over unencrypted Very High Frequency (VHF) voice channels, exploiting the absence of cryptographic authentication and encryption in legacy VHF voice communication architectures, which are transmitted as unprotected analog broadcasts accessible to anyone with appropriate radio equipment [7], [8].

Public demonstrations have shown that ATC-style voice cloning is technically feasible using widely available tools [5]. At the same time, anti-spoofing evaluations such as ASVspoof [9] highlight the ongoing difficulty of reliably distinguishing synthetic from authentic speech as generative models improve.

Within an ATM framework, spoofing risk is not determined solely by perceptual naturalness. Operational plausibility depends on a conjunction of factors including semantic correctness, preservation of safety-critical tokens, speaker-identity similarity, acoustic consistency, and procedural cadence alignment. A transmission that satisfies these dimensions may appear credible even if minor prosodic deviations are present.

Accordingly, ATC-oriented TTS evaluation must be framed as a communication resiliency problem. The central question is not only whether synthesized speech sounds natural, but under what measurable conditions it approaches operational plausibility in safety-critical environments.

IV. Evaluation Framework and Metrics

The preceding sections define synthesis as a multi-constraint problem under operational and adversarial considerations. Evaluation must therefore quantify performance along these constrained dimensions rather than relying on a single perceptual score.

We adopt a structured, multi-metric framework grounded in established speech processing methodologies. Each metric operationalizes one component of synthesis realism identified in the system formulation and threat model. Metrics are analyzed both independently and jointly to assess not only signal-level quality but also operational plausibility within safety-critical aviation environments.

The following subsections define the individual metric components and their interpretations.

A. Word Error Rate

ASR decoding is performed using a Whisper Large-v2 model fine-tuned for ATC-domain transcription [10]. Let T denote the reference transcript and $T_{\text{ASR}} = \text{ASR}(\hat{x}(t))$ the transcript obtained by decoding the synthesized waveform.

Word Error Rate (WER) quantifies transcript intelligibility by measuring the minimum edit distance between T and T_{ASR} [11], [12]. For each utterance,

$$\text{WER} = \frac{S + D + I}{N},$$

where S, D, and I denote substitutions, deletions, and insertions, respectively, and N is the number of words in the reference transcript. Metrics are computed per utterance and aggregated across the dataset.

Operationally, WER captures overall lexical intelligibility but does not distinguish between benign substitutions and corruption of safety-critical instructions. Consequently, WER constitutes a necessary but insufficient condition for operational plausibility in ATC communication.

B. Critical Token Accuracy

Let $\mathcal{C}_{\text{ref}}$ denote the set of critical tokens extracted from the reference transcript T, and let $\mathcal{C}_{\text{ASR}}$ denote the corresponding set extracted from T_{ASR}. Precision and recall are defined in terms of true positive matches between these sets [11]:

$$\text{Precision} = \frac{|\mathcal{C}_{\text{ref}} \cap \mathcal{C}_{\text{ASR}}|}{|\mathcal{C}_{\text{ASR}}|}, \quad \text{Recall} = \frac{|\mathcal{C}_{\text{ref}} \cap \mathcal{C}_{\text{ASR}}|}{|\mathcal{C}_{\text{ref}}|}.$$

The critical-token F1 score (i.e., the F-measure with $\beta = 1$) is the harmonic mean of precision and recall:

$$\text{F1}_{\mathcal{C}} = \frac{2\,\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Critical tokens are extracted via a rule-based, pattern-matching parser applied to normalized text. Normalization lowercases the transcript, removes punctuation (retaining decimal points for frequency parsing), and collapses whitespace. The parser tokenizes the result and scans left-to-right, attempting to match four ATC-specific token types at each position: flight levels (e.g., "flight level two nine zero" $\rightarrow$ `FL:290`), headings (e.g., "heading one one five" $\rightarrow$ `HDG:115`), VHF frequencies (e.g., "one three three decimal four" $\rightarrow$ `FREQ:133.4`), and runway designators (e.g.,

"runway two seven" → `RWY:27`). Each extractor consumes the matched tokens and advances the scan index, preventing double-counting. An optional callsign proxy heuristic captures airline-word plus digit sequences at the utterance start (e.g., `CS:alitalia117`). Number words are resolved via a fixed digit-word vocabulary covering singles, teens, and tens, supporting both spoken-form digit sequences and compact numeric tokens. Token matching between $\mathcal{C}_{\text{ref}}$ and $\mathcal{C}_{\text{ASR}}$ is performed as exact set intersection over the extracted token strings. This deterministic, lexicon-driven approach is intentionally conservative: it captures structured operational fields with high precision while avoiding spurious matches from free-form numeric context.

F1 is computed per utterance and aggregated across the dataset via macro-averaging. If both $\mathcal{C}_{\text{ref}}$ and $\mathcal{C}_{\text{ASR}}$ are empty for a given utterance, F1 is defined as 1; if $\mathcal{C}_{\text{ref}}$ is nonempty and $\mathcal{C}_{\text{ASR}}$ is empty, F1 is 0.

C. Speaker Embedding Similarity

Speaker identity similarity is evaluated using ECAPA-TDNN embeddings [13] implemented via SpeechBrain [14]. Both ground-truth and synthesized waveforms are resampled to 16 kHz prior to embedding extraction.

Let $e(x)$ and $e(\hat{x})$ denote the ECAPA-TDNN embedding vectors of ground-truth and synthesized speech, respectively. Speaker similarity is computed per utterance via cosine similarity:

$$\rho_{\text{spk}} = \frac{e(x) \cdot e(\hat{x})}{\|e(x)\| \, \|e(\hat{x})\|}.$$

Values approach 1 for highly similar speaker representations. High embedding similarity indicates closer alignment in learned speaker-representation space and may increase impersonation plausibility in systems that rely on speaker verification embeddings.

D. Prosodic Alignment

Prosody in ATC communication emphasizes procedural cadence, controlled pitch range, and structured pause placement rather than expressive variability. Unlike conversational speech, ATC transmissions prioritize clarity, regular pacing, and standardized intonation contours.

Prosodic alignment is evaluated along three dimensions: (i) pitch magnitude deviation, (ii) pitch contour similarity, and (iii) cadence consistency.

1) Pitch Deviation and Contour Similarity: Fundamental frequency (F_0) trajectories are extracted using the probabilistic YIN (pYIN) estimator implemented in Librosa. Signals are resampled to 16 kHz and analyzed using a 1024-sample window with a 10 ms hop size. The pitch search bounds of 70–500 Hz were selected to encompass typical adult speech while excluding subharmonic and high-frequency tracking artifacts.

Let $F_0[k]$ and $\hat{F}_0[k]$ denote frame-level pitch estimates for the ground-truth and synthesized signals, respectively. We restrict the analysis to mutually voiced frames $\mathcal{K} = \{k : F_0[k] \text{ and } \hat{F}_0[k] \text{ are voiced}\}$, with $K = |\mathcal{K}|$. If $K = 0$, the utterance is excluded from pitch-based analysis.

Mean absolute pitch deviation is defined as

$$\Delta F_0 = \frac{1}{K} \sum_{k \in \mathcal{K}} |F_0[k] - \hat{F}_0[k]|.$$

Pitch contour similarity is quantified using the Pearson correlation coefficient over mutually voiced frames:

$$\rho_{F_0} = \operatorname{corr}\big(F_0[k], \hat{F}_0[k]\big), \quad k \in \mathcal{K}.$$

Here, ρ_{F_0} measures linear association between aligned pitch trajectories and therefore captures contour-shape similarity and temporal consistency.

2) Speaking Rate Deviation: Speaking rate is approximated as

$$R = \frac{N}{D},$$

where N is the transcript word count and D is utterance duration (seconds). Because synthesized and reference utterances share identical lexical content, rate deviation isolates duration-based pacing differences:

$$\Delta R = |R_x - R_{\hat{x}}|.$$

3) Pause Structure Consistency: Pause segments are detected using short-time RMS energy thresholding with a minimum silence-duration constraint. For each utterance, we compute

$$\text{Pause Ratio} = \frac{\text{Total Pause Duration}}{\text{Total Utterance Duration}},$$

as well as mean pause duration. Absolute deviations between ground-truth and synthesized signals quantify differences in silence placement and timing.

Collectively, these measures operationalize procedural timing consistency in ATC speech. Moderate pitch magnitude deviations may remain operationally acceptable, whereas substantial cadence or pause irregularities may reduce perceived authenticity despite correct lexical content.

E. Log-Mel Spectral Distance

To quantify acoustic envelope similarity, we compute 80-bin log-mel spectrograms from ground-truth and synthesized waveforms sampled at 16 kHz using a 1024-point FFT and a 10 ms hop size. The mel representation emphasizes perceptually relevant spectral structure while suppressing fine-grained phase information. Because utterance durations may differ, spectrograms are temporally aligned using dynamic time warping (DTW) in mel-feature space to compensate for timing and cadence mismatches.

Let $\mathbf{M}_t \in \mathbb{R}^{80}$ and $\hat{\mathbf{M}}_{t'} \in \mathbb{R}^{80}$ denote log-mel feature vectors (in dB) from the ground-truth and synthesized signals, respectively. DTW produces an alignment path $\pi = \{(t_i, t'_i)\}_{i=1}^{L}$ mapping frame indices between the two sequences. The frame-wise spectral error along this alignment path is defined as $d_i = \|\mathbf{M}_{t_i} - \hat{\mathbf{M}}_{t'_i}\|_2$.

The log-mel spectral distance is then computed as the mean Euclidean distance over aligned frames:

$$D_{\text{mel}} = \frac{1}{L} \sum_{i=1}^{L} d_i.$$

This metric measures average spectral-envelope deviation in log-amplitude space after compensating for temporal differences. Because D_{mel} depends on feature dimensionality, log scaling, and DTW alignment, its magnitude should be interpreted comparatively within this experimental configuration. Lower values indicate closer spectral-envelope alignment between synthesized and ground-truth speech, whereas larger values reflect increased acoustic deviation.

V. Experimental Results

A. Dataset and Setup

Evaluation was conducted on 1000 controller-side utterances randomly sampled without replacement from the ATCOSIM corpus [15], which contains authentic ATC radiotelephony transmissions recorded under operational radio conditions with aligned transcripts. The subset used in this study was accessed via a publicly available HuggingFace mirror of the corpus [16].

Corrupted audio files were excluded. No acoustic filtering was applied to the ground-truth recordings; however, transcripts were text-normalized prior to synthesis and evaluation. No additional preprocessing was performed. Ground-truth utterance durations averaged 4.18 seconds.

Transcript–audio pairs were synthesized using the proposed TTS framework. All waveforms were resampled to 16 kHz prior to metric computation to ensure consistency across evaluation procedures.

All metrics were computed per utterance and aggregated across the dataset. Pearson correlations were computed over utterance-level metric values to analyze relationships among semantic, spectral, speaker, and prosodic dimensions.

B. Aggregate Performance

TABLE I
Aggregate Evaluation Metrics

Metric	Mean	Std	Median	P25	P75
WER (%)	3.9	12.5	0.0	0.0	0.0
$\text{F1}_{\mathcal{C}}$	0.958	0.176	1	1	1
ρ_{spk}	0.726	0.102	0.752	0.671	0.799
ΔF_0 (Hz)	45.36	91.07	15.26	11.00	27.45
ρ_{F_0}	0.082	0.429	0.098	-0.220	0.386
ΔR (words/s)	0.456	0.378	0.366	0.172	0.641
ΔPause Ratio	0.096	0.090	0.073	0.033	0.131
ΔMean Pause (s)	0.151	0.120	0.113	0.050	0.207
D_{mel}	8.91	1.34	8.69	8.06	9.57

*Pitch-based metrics (ΔF_0, ρ_{F_0}) are reported for 890 and 821 utterances respectively, excluding those with insufficient voiced frames for reliable pitch estimation.

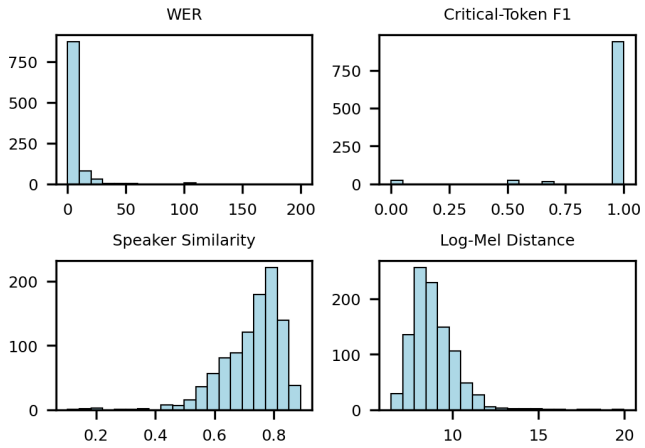

Fig. 1. Per-utterance distributions for WER, $\text{F1}_{\mathcal{C}}$, ρ_{spk}, and D_{mel} over the 1000-utterance evaluation set.

Table I summarizes aggregate evaluation metrics across the 1000-utterance test set and highlights a separation between lexical fidelity and prosodic stability. Lexical preservation was strong. WER is reported as a percentage and averaged 3.9% ($\pm 12.5\%$), with a median of 0% and $P75 = 0\%$. The 75th percentile of zero indicates that three-quarters of utterances had no word-level errors; the remaining 25% exhibit an average WER of approximately 15.7%, dominated by a small number of high-error outliers (e.g., complete transcription failures on very short utterances).

Analysis of the minority of utterances with imperfect critical-token preservation ($\text{F1}_{\mathcal{C}} < 1$) reveals that most failures involve single-phoneme substitutions (e.g., "ukay" → "okay") or single-digit frequency errors (e.g., "125.8" → "135.8"). This pattern suggests that while the TTS model generally preserves semantic content, occasional mispronunciations of critical tokens in specific phonetic contexts can still fool the ASR pipeline. This proposes a consideration for spoofing detection given that such minor errors may be perceptually masked in noisy VHF channels.

Figure 1 visualizes the corresponding per-utterance distributions. WER and $\text{F1}_{\mathcal{C}}$ are strongly concentrated near their optimal values, whereas ρ_{spk} and D_{mel} exhibit broader dispersion, consistent with the aggregate statistics in Table I.

Spectral alignment was quantified using D_{mel}, defined as the mean Euclidean distance between 80-dimensional log-mel spectrogram vectors (expressed in dB) after dynamic time warping alignment. This metric measures spectral-envelope deviation in log-amplitude mel feature space rather than absolute acoustic error in physical units. Across the evaluation set, D_{mel} yielded a mean of 8.91 (± 1.34) with a relatively narrow interquartile range ($P25 = 8.06$, $P75 = 9.57$). Because D_{mel} depends on mel dimensionality, log scaling, and alignment configuration, its magnitude should be interpreted comparatively within this experimental setup. The tight dispersion suggests consistent spectral-envelope similarity between synthesized and ground-truth signals across most utterances.

Speaker embedding similarity ρ_{spk} averaged 0.726 (± 0.102), with median 0.752 and interquartile range [0.671, 0.799]. This distribution indicates generally stable alignment in learned speaker-representation space, with moderate variability and a limited number of lower-similarity outliers rather than systematic identity collapse across the dataset.

Prosodic deviations exhibited substantially greater dispersion than lexical or spectral metrics. Pitch-based measures were computed only on utterances containing valid mutually voiced frames (890 for ΔF_0, 821 for ρ_{F_0}), as some short or weakly voiced transmissions did not yield stable pitch tracks or well-defined correlations. While the median pitch deviation was 15.26 Hz, the mean (45.36 Hz) and large standard deviation (91.07 Hz) indicate a right-skewed, heavy-tailed distribution, reflecting substantial pitch mismatches in a minority of utterances.

Pitch contour similarity was low on average ($\rho_{F_0} = 0.082$) and broadly distributed ($P25 = -0.220$, $P75 = 0.386$), suggesting weak linear association between ground-truth and synthesized pitch trajectories after alignment. This indicates inconsistent preservation of fine-grained intonation patterns despite strong transcript fidelity. Speaking-rate deviation (ΔR) and pause-based deviations were moderate in magnitude, further indicating that temporal pacing and silence structure are less consistently reproduced than lexical content. Collectively, these results reveal a semantic–prosodic asymmetry: transcript-level correctness is frequently preserved, while temporal and intonational structure exhibit greater variability.

Having established the aggregate performance characteristics across individual dimensions, we now examine the relationships among these metrics. The degree of independence between lexical, spectral, speaker, and prosodic factors has direct implications for spoofing detection: if realism requires simultaneous optimization across multiple uncorrelated dimensions, then synthetic speech must satisfy multiple independent constraints to achieve operational plausibility.

C. Cross-Metric Interactions

Operational plausibility in ATC communication arises from the joint behavior of semantic, identity, spectral, and prosodic factors. High lexical intelligibility alone does not imply impersonation capability, nor does strong speaker similarity guarantee semantic correctness.

Figure 2 shows that realism dimensions form partially independent axes rather than collapsing into a single quality measure. Lexical accuracy exhibits weak association with both spectral and speaker metrics (e.g., WER vs. D_{mel}: $r = 0.25$; WER vs. ρ_{spk}: $r = -0.28$), indicating that transcript fidelity is largely decoupled from acoustic-envelope alignment and embedding-based identity similarity.

In contrast, spectral distance and speaker similarity display a moderate negative correlation (D_{mel} vs. ρ_{spk}: $r = -0.41$), suggesting that larger spectral-envelope mismatch is associated with weaker speaker-embedding alignment. Prosodic metrics demonstrate internal coherence: pause ratio and mean pause

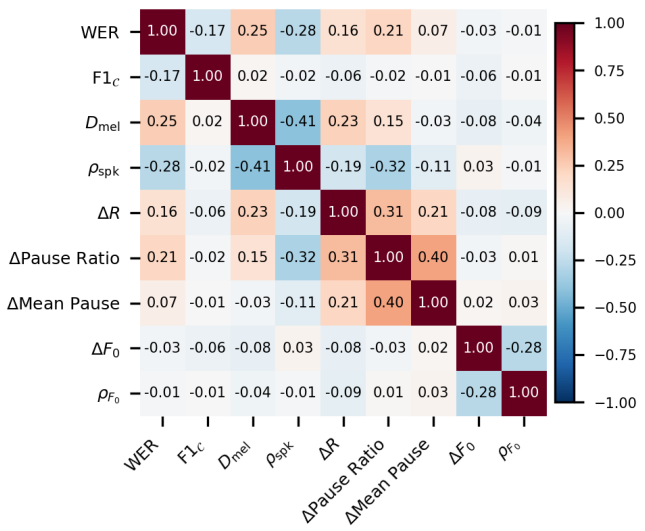

Fig. 2. Pearson correlation matrix across semantic, spectral, speaker, and prosodic metrics. Moderate coupling is observed between D_{mel} and ρ_{spk} ($r = -0.41$) and between pause-related timing deviations ($r = 0.40$), while lexical accuracy is weakly associated with other dimensions.

deviations are positively correlated ($r = 0.40$), and pitch magnitude deviation is negatively associated with pitch contour similarity (ΔF_0 vs. ρ_{F_0}: $r = -0.28$), consistent with expected coupling between timing and intonation deviations.

Collectively, these results support a multi-dimensional view of ATC-oriented synthesis realism: semantic, identity, spectral, and prosodic factors vary semi-independently. From a communication resiliency perspective, spoofing plausibility depends on the simultaneous satisfaction of multiple constraints rather than thresholding any single metric.

VI. Discussion

The experimental results demonstrate that domain-constrained synthesis preserves safety-critical semantic content with high reliability while maintaining moderate-to-high speaker embedding similarity and stable spectral-envelope alignment. Same-speaker verification typically achieves ≥ 0.8 similarity for same utterance, so 0.726 across different utterances is substantial. These properties support the operational utility of synthetic ATC speech for simulation and training applications under controlled conditions.

Speaker-embedding similarity, however, must be interpreted in light of reference-audio duration. Neural speaker encoders such as ECAPA-TDNN rely on sufficient temporal context to produce stable and discriminative identity representations [13]. Longer reference clips typically yield more robust embeddings and improved identity conditioning during synthesis. In the present study, controller utterances were relatively short (mean duration 4.18 s), limiting the available acoustic context for speaker representation. Accordingly, the observed embedding similarity values reflect both synthesis performance and duration-constrained identity estimation, rather than purely architectural limitations of the TTS model.

In contrast to strong lexical preservation and stable spectral alignment, prosodic analysis reveals consistent variability in cadence and pitch-contour alignment. Although the median pitch deviation is modest (15.26 Hz), the heavy-tailed distribution and low mean contour correlation ($\rho_{F_0} = 0.082$) indicate that fine-grained temporal intonation structure is not consistently reproduced. Speaking-rate and pause-based statistics further demonstrate measurable timing drift relative to authentic transmissions, suggesting that temporal dynamics remain more difficult to match than lexical or spectral characteristics.

From a communication-resiliency perspective, these findings highlight a structured asymmetry: synthetic speech can approach semantic and identity realism while retaining systematic prosodic discrepancies. The relative independence of lexical, spectral, speaker, and prosodic metrics reinforces the necessity of multi-dimensional evaluation when assessing both simulation fidelity and spoofing plausibility. Realism in safety-critical ATC communication does not collapse to a single scalar measure; rather, it emerges from the joint satisfaction of partially independent constraints.

Limitations

Evaluation was performed under clean waveform synthesis conditions without explicit VHF channel modeling, radio compression artifacts, or environmental noise injection. Real-world ATC transmissions are subject to bandwidth constraints, distortion, and channel variability that may alter spectral and prosodic characteristics and potentially affect both spoofing plausibility and detection cues. Incorporating controlled VHF channel simulation would provide a more operationally faithful assessment of realism and resiliency.

Lexical evaluation relied on an ASR-based transcription pipeline. While domain fine-tuning mitigates transcription bias, WER and critical-token F1 remain partially conditioned on ASR model behavior rather than direct human judgment. Similarly, speaker similarity was quantified using ECAPA-TDNN embedding cosine similarity, which reflects representation-space proximity but does not directly measure perceptual identity similarity or human impersonation detection sensitivity.

Finally, evaluation was restricted to controller-side utterances drawn from ATCOSIM. Broader role diversity and longer-duration reference segments may influence both identity conditioning and prosodic stability.

Future work should incorporate channel modeling, perceptual evaluation with certified ATC professionals, and adversarial playback experiments to more fully contextualize synthesis realism under operational transmission conditions.

VII. Conclusion

This work developed an ATC-oriented neural TTS synthesis and evaluation framework that treats realism as a multi-dimensional, operationally constrained measurement problem rather than a single perceptual score. Using 1000 controller-side utterances from ATCOSIM, results show strong transcript fidelity and near-perfect preservation of safety-critical tokens, with moderate-to-high speaker-embedding similarity and stable spectral-envelope alignment under DTW. In contrast, prosodic alignment remains less consistent, with low pitch-contour correlation and measurable cadence deviations, revealing a semantic–prosodic asymmetry in contemporary ATC-style synthesis.

These findings suggest that ATC spoofing plausibility is governed by the joint satisfaction of partially independent constraints spanning semantic integrity, identity plausibility, acoustic consistency, and procedural cadence. Future work will incorporate VHF channel effects and perceptual evaluation with certified ATC professionals to better contextualize synthesis realism and detection cues under operational transmission conditions.

Beyond channel modeling and perceptual evaluation, a promising direction is the development of formal operational plausibility standards that map metric thresholds across the evaluated dimensions to graded risk levels within an ATM safety framework. Such standards would require collaboration with domain regulators and certified ATC professionals to establish empirically grounded acceptance criteria — for instance, defining the joint $(\rho_{\text{spk}}, \text{WER}, D_{\text{mel}})$ regions under which a synthetic transmission would be indistinguishable from an authentic one under operational conditions. Formalizing these boundaries would support both the certification of synthetic speech for training and simulation use cases and the design of threshold-based detection systems for communication resiliency applications.

References

[1] International Civil Aviation Organization, "Procedures for Air Navigation Services: Air Traffic Management (Doc 4444)," ICAO, Tech. Rep., 2016.

[2] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021, available: https://arxiv.org/abs/2106.15561 (accessed 2026-03-02).

[3] H. Azzuni and A. El Saddik, "Voice cloning: Comprehensive survey," arXiv:2505.00579, 2025, available: https://arxiv.org/abs/2505.00579 (accessed 2026-03-02).

[4] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[5] DEF CON 33 Presentation, "Voice Cloning and ATC Spoofing Demonstration at DEF CON 33," 2025, gray literature. Available: https://www.youtube.com/watch?v=JKwxsGYcZq4 (accessed 2026-03-02).

[6] Resemble AI, "Chatterbox tts," https://github.com/resemble-ai/chatterbox, 2024, accessed 2026-03-02.

[7] Skybrary, "Unauthorised use of atc frequency," https://skybrary.aero/articles/unauthorised-use-atc-frequency, 2025, accessed 2026-03-02.

[8] IETF, "RFC 9372: L-Band Digital Aeronautical Communications (L-DAC) Framework and Considerations," https://datatracker.ietf.org/doc/rfc9372/, 2023, accessed 2026-03-02.

[9] M. Todisco, X. Wang, V. Vestman *et al.*, "ASVspoof 2019: Spoofing Countermeasures for the Detection of Synthesized Speech," 2019.

[10] J. van Doorn, "Whisper large-v2 atco2 asr model," https://huggingface.co/jlvdoorn/whisper-large-v2-atco2-asr-atcosim, 2023, accessed 2026-03-02.

[11] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed., 2026, draft of January 26, 2026. Available: https://web.stanford.edu/~jurafsky/slp3/ed3book_jan26.pdf (accessed 2026-03-02).

[12] A. Morris, V. Maier, and P. Green, "From WER and RIL to MER and WIL: Improved Evaluation Measures for Connected Speech Recognition," in *Proceedings of Interspeech*, 2004.

[13] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," *arXiv preprint arXiv:2005.07143*, 2020, available: https://arxiv.org/abs/2005.07143 (accessed 2026-03-02).

[14] M. Ravanelli, T. Parcollet, A. Rouhe, P. Plantinga, E. Rastorgueva, L. Lugosch, N. Dawalatabad, L. Ju-Chieh, A. Heba, F. Grondin *et al.*, "SpeechBrain: A General-Purpose Speech Toolkit," *arXiv preprint arXiv:2106.04624*, 2021, available: https://arxiv.org/abs/2106.04624 (accessed 2026-03-02).

[15] K. Hofbauer, S. Petrik, and H. Hering, "The ATCOSIM Corpus of Non-Prompted Clean Air Traffic Control Speech," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA), 2008. [Online]. Available: https://aclanthology.org/L08-1507/

[16] J. van Doorn, "ATCOSIM Dataset," https://huggingface.co/datasets/jlvdoorn/atcosim, 2023, huggingFace dataset mirror, revision b5839d9 (accessed 2026-03-02).