SimuGAN-Whisper-ATC: Generative Noise Injection for Improved Automatic Speech Recognition in Air Traffic Control

Omar Garib*, Mohamed Ghanem†, Olivia J. Pinon Fischer‡, Dimitri N. Mavris§
Georgia Institute of Technology, Atlanta, GA 30332, USA

Air traffic control (ATC) operations are inherently complex and cognitively demanding, a challenge that continues to grow with increasing flight volumes and a more diverse set of airspace users. To reduce controller workload and minimize operational errors, the integration of automated support systems, particularly those powered by reliable Automatic Speech Recognition (ASR) models, has emerged as a compelling research avenue. This work explores enhancements to ASR performance using Whisper, a state-of-the-art, open-source model developed by OpenAI, known for its multilingual training and strong generalization across varied audio conditions. Despite these advantages, a significant obstacle remains: the limited availability of labeled datasets that accurately reflect real-world ATC environments. Prior efforts have fine-tuned Whisper on the simulation-based, clean (noise-free) ATCOSIM dataset (∼10 hours labeled) and the real-world but minimally labeled ATCO2 dataset (∼1 hour labeled). While this approach achieved excellent results on clean audio (Word Error Rate, WER: 1.19%), it showed a marked drop in performance on authentic ATC speech (WER: 14.66%). To address this gap, we introduce a novel pipeline, SimuGAN-Whisper-ATC, which leverages a Generative Adversarial Network (GAN) to learn realistic noise characteristics from real ATC corpora (ATCO2, TartanAviation) and injects them into clean ATCOSIM utterances. This generative augmentation approach effectively expands and diversifies the training data, enabling Whisper to better adapt to noisy, real-world conditions. When fine-tuned on this enriched dataset, Whisper achieves a significantly improved WER of 3.58% on real-world ATC speech, setting—to our knowledge—a new benchmark for this domain. We believe that this advancement represents a substantial step toward deployable, real-time ASR systems capable of proactively identifying communication issues and enhancing ATC safety. Further refinements and evaluations will be presented in the final version of this paper.

I. Introduction

Clear voice communication between air traffic controllers (ATCOs) and pilots is essential for ensuring aviation safety. Rigorous communication protocols have been established and continuously refined to meet evolving safety requirements and minimize operational risks. However, the rapid growth in air traffic volumes, coupled with increasingly complex airspace management scenarios, such as airports serving both commercial airlines and military operations, makes effective communication more challenging. At major high-traffic airports, controllers frequently handle multiple clearances simultaneously, relying heavily on memory or brief handwritten notes to track rapidly changing situations [1]. Similarly, airports with shared civil–military airspace, such as the Ronald Reagan Washington National Airport [2], require controllers to manage diverse aircraft types and operational missions, significantly elevating the complexity and cognitive demands. Although controllers are extensively trained and generally equipped to manage these demanding situations effectively, the potential for error persists due to intense workload and cognitive strain. Even minor miscommunications, such as a missed altitude clearance or an incorrectly acknowledged call sign, can escalate rapidly into catastrophic incidents, including runway incursions or loss-of-separation events.

Providing controllers with additional support or oversight mechanisms could substantially reduce the risk of miscommunication. The concept of an AI-based assistant capable of continuously monitoring communications, verifying pilot read-backs, and promptly flagging discrepancies presents a promising idea to explore for enhancing operational safety. However, developing such assistants first requires reliably accurate transcription of ATC voice

*Graduate Research Assistant, Aerospace Systems Design Laboratory
†Graduate Research Assistant, Georgia Institute of Technology
‡Principal Research Engineer, Aerospace Systems Design Laboratory, AIAA Associate Fellow
§Georgia Tech Distinguished Regents Professor and Director of ASDL, AIAA Fellow

exchanges. Real-time Automatic Speech Recognition (ASR) provides a foundational technology for capturing these spoken interactions, yet existing ASR models often struggle under realistic, noisy VHF (very high frequency) radio conditions typical of operational environments [3]. Enhancing ASR robustness to better handle these challenging acoustic scenarios thus remains an important research priority.

Significance of Reliable ATC-ASR Recent initiatives have aimed to integrate AI-based ASR into ATC operations to improve situational awareness and minimize errors [4]. However, building robust ASR solutions for aviation remains challenging due to the limited availability of large-scale, annotated ATC speech corpora; typical open-source real datasets contain only a few hours of labeled speech [5]. Gathering and transcribing these recordings is labor-intensive and constrained by privacy requirements. This challenge is further compounded by the specialized aviation jargon and audio distortions, including background noise and channel interference, present in ATC communications. As a result, even advanced ASR approaches often struggle to deliver consistent accuracy in live ATC environments [6].

Advancements in ASR Technology Automatic Speech Recognition (ASR) refers to the process of converting spoken language into written text. Recent advances in deep learning and large-scale data training have significantly improved ASR accuracy, with state-of-the-art systems approaching human-level performance on general speech benchmarks [7]. One notable example is Whisper, a transformer-based ASR model developed by OpenAI and pre-trained on 680,000 hours of weakly supervised audio across multiple languages and domains [8]. This extensive and diverse training enables Whisper to generalize well across various speech conditions, including different accents and background noise. When adapted to the air traffic control (ATC) domain, Whisper, fine-tuned as Whisper-ATC [6], achieved a Word Error Rate (WER) of approximately 14.66% on real-world ATC recordings and as low as 1.19% on clean, simulated data. These results demonstrate both the model's potential in specialized domains and the persistent challenges posed by the noisy, variable nature of operational ATC communications.

Generative Data Augmentation Data augmentation techniques, such as incorporating background noise or simulating reverberation—the persistence of sound after the original sound has stopped—have been shown to significantly enhance the robustness of ASR systems [9]. However, conventional noise-mixing methods often fall short in replicating the complexity of radio communications, such as transient channel interference, abrupt bursts of static from radio activation (squelch pops), and the characteristic hiss of VHF transmissions. Generative Adversarial Networks (GANs) [10] offer a promising alternative by learning and reproducing realistic noise patterns through modeling complex, domain-specific distributions. This capability is especially valuable in the ATC domain, where clean, labeled recordings (e.g., ATCOSIM) are readily available or can be effectively synthesized using modern generative approaches, but labeled data reflecting real-world radio conditions is severely limited. Crucially, large amounts of unlabeled real-world ATC audio are available, enabling GANs to effectively learn realistic acoustic characteristics directly from these unlabeled datasets. While GANs have previously been explored for text-to-speech synthesis [11–13], our work specifically focuses on generating realistic noise profiles to emulate VHF radio channel conditions. By injecting GAN-generated noise into clean, labeled ATC recordings, we create a more diverse and operationally realistic training set.

Domain Adaptation for ATC Large-scale pre-trained models such as Whisper provide a strong foundation but must be adapted to handle the specialized vocabulary and phraseology unique to ATC. Fine-tuning on available data can partially improve accuracy; however, the severe imbalance between simulation-based, clean recordings (e.g., approximately 10 hours from ATCOSIM data) and limited real-world labeled samples (less than 1 hour from ATCO2) significantly constrains noise modeling. Thus, an effective domain adaptation strategy needs to simultaneously address the distinctive linguistic features and the adverse radio conditions, such as noisy VHF channels.

Limitations of Existing Approaches Despite advancements in noise-robust training and domain adaptation, existing ATC-ASR models still exhibit substantial performance gaps when transitioning from simulated environments to operational conditions [6]. A reliance on limited labeled real-world recordings often leads to overfitting and domain mismatch, further hindering transcription accuracy. Current state-of-the-art systems typically achieve a WER around 15% on challenging ATC datasets, such as ATCO2, falling short of the accuracy requirements set by aviation authorities for safety-critical deployment. These operational standards demand near-perfect performance, emphasizing the necessity for further improvements in ASR robustness.

Research Contributions To address these challenges, this paper proposes *SimuGAN-Whisper-ATC*, a novel framework that integrates generative adversarial network (GAN)-based noise augmentation with domain-specific fine-tuning. Our approach synthesizes realistic VHF radio noise patterns learned directly from real ATC recordings, injecting them into clean simulation-based utterances to significantly expand and diversify the labeled training dataset. Fine-tuning Whisper on this enhanced corpus achieves substantial performance gains, reducing WER from 14.66% to 3.58%, representing, to our knowledge, the best-reported accuracy in open-source ATC-ASR benchmarks.

The remainder of this paper is structured as follows: Section II reviews related work in ASR and generative data augmentation; Section III details the datasets employed; Section IV outlines our methodology and overall approach; and Section V provides implementation specifics for SimuGAN training and Whisper fine-tuning. Section VI presents qualitative and quantitative evaluations of the SimuGAN-generated audio and resulting Whisper fine-tuning performance. Section VII outlines future work planned for inclusion in the final manuscript, and finally Section VIII summarizes the main findings and discusses practical considerations for deployment.

II. Background

Air Traffic Control (ATC) relies significantly on accurate and efficient voice communication to ensure safety and operational effectiveness. However, these communications often face real-time challenges, including radio congestion, linguistic variability, and environmental noise. This section reviews key aspects of ATC voice communication challenges, Automatic Speech Recognition (ASR) advancements, and the role of generative augmentation techniques to address existing limitations.

A. Real-Time Challenges in ATC Voice Communications

Air Traffic Control voice communications are subject to a range of real-time stressors, including high traffic density, congested radio frequencies, and linguistic variability. At busy terminals, controllers are tasked with managing complex air traffic flows and approach procedures as well as a mix of aircraft types and airspace users, all under time pressure and requiring high cognitive loads. For example, facilities such as San Diego's SoCal Terminal Radar Approach Control (TRACON) and Honolulu's joint civil–military airport must coordinate operations involving both commercial and military traffic, each with different flight profiles and priorities [14–16]. Traffic complexity is frequently compounded by difficult weather conditions, restricted areas, or operational constraints, which further increase cognitive fatigue and the risk of communication errors.

In such environments, the diversity in aircraft performance, particularly in mixed operations involving commercial jets, general aviation, and fast-moving military aircraft, adds further complexity. Research shows that controller performance degrades when traffic includes significant variations in climb rates or trajectories. Under such circumstances, immediate demands for conflict resolution may hinder the timely detection or correction of minor communication errors, potentially increasing the risk of escalation into critical incidents.

Compounding these issues are technical limitations of radio communication. Congested frequencies can delay the transmission of urgent messages, as pilots must wait for a break in ongoing communications. Simultaneous communications, commonly referred to as "stepped on transmission", can result in neither message being fully received [17]. Ground control frequencies at major airports often become extremely congested during peak periods, underscoring the operational challenge of frequency management. Additionally, audio clarity is frequently degraded by background cockpit noise, poorly tuned radios, or inadvertent microphone activation, all of which can result in incomplete or unintelligible transmissions [18]. Call sign confusion further complicates the communication environment. Similar-sounding identifiers, such as *Delta 1725* and *Delta 1735*, can lead to pilots acting on clearances intended for other aircraft. Since all transmissions on a frequency are publicly audible, a single misheard instruction or clearance can create confusion across multiple flight crews and disrupt traffic patterns.

Even when controllers maintain overall situational awareness, linguistic factors remain a persistent source of misunderstanding. While English is the mandated language of international aviation, accent differences, varying speech rates, and deviations from standardized phraseology frequently lead to miscommunication [19–21].

High-profile incidents, such as the 1977 Tenerife runway collision, have underscored the dangers of ambiguous language and missed read-back errors [22]. Despite efforts to enforce global standards, real-world communication often strays from prescribed phraseology during periods of high workload. Cultural norms may also discourage explicit requests for clarification. In the case of Avianca Flight 52, indirect phrasing such as "we're running out of fuel" failed to communicate an emergency situation [23]

Maintaining effective situational awareness thus requires addressing these intertwined challenges, particularly in high-stakes environments where miscommunication can lead to a loss of separation assurance, compromising the minimum safe distance required between aircraft. To address these risks, aviation authorities continue to strengthen controller and pilot training, promote awareness of linguistic diversity, and enforce the use of standardized phraseology and read-backs. Technological solutions, such as Controller–Pilot Data Link Communications (CPDLC), advanced noise-canceling headset systems, and improved radio-frequency management, further support these efforts by reducing audio distortion and communication delays.

Complementing these human and system-level improvements is the integration of automated tools, such as robust automatic speech recognition (ASR) systems. These systems hold significant promise for supporting real-time monitoring, transcription, and error detection in ATC environments. In this work, we investigate the use of generative data augmentation to enhance ASR robustness by simulating realistic radio noise conditions, thereby enabling more reliable and efficient communication management within complex operational environments.

B. Automatic Speech Recognition (ASR) Models and Fine-tuning

Automatic Speech Recognition (ASR) aims to transform spoken language into text, enabling hands-free interaction and machine-level processing of verbal inputs. Early efforts in the 1950s and 1960s relied on relatively simple, rule-based approaches, such as Bell Labs *Audrey* and IBM's *Shoebox*, which recognized small vocabularies using constrained acoustic matching techniques [24, 25]. Although these systems established fundamental principles, they required structured pauses and were limited in vocabulary and speaker independence.

The 1970s and 1980s saw the rise of statistical modeling, most notably through the adoption of Hidden Markov Models (HMMs). By integrating multiple sources of information, acoustic, phonetic, and linguistic, within a unified probabilistic framework, HMM-based systems achieved significant gains in both accuracy and scalability [25]. Projects like CMU's *Harpy*, which supported around a thousand words and enabled speaker-independent continuous recognition, and IBM's *Tangora*, which supported approximately twenty thousand words but remained speaker-dependent and limited to isolated-word recognition, significantly expanded vocabulary coverage and laid important groundwork for subsequent ASR advancements [24]. Despite refinements in acoustic feature extraction and language modeling, hybrid HMM-Gaussian Mixture Model (HMM-GMM) systems began to plateau in performance in the 1990s and early 2000s. This stagnation, combined with growing computational resources and the availability of larger annotated corpora, prompted a shift toward deep learning-based approaches, including early work on sequence modeling [26] and discriminative training [27].

A major breakthrough occurred between 2009 and 2012, when deep neural networks (DNNs) replaced the Gaussian Mixture Model (GMM) components within HMM systems. This change led to reductions in word error rates of up to 30% on standard benchmarks [27, 28]. The success of deep learning renewed interest in end-to-end architectures, including recurrent neural networks (RNNs) and convolutional neural networks (CNNs), which offered more efficient modeling of temporal and spectral patterns in speech [29]. This evolution gave rise to sequence-to-sequence models which eliminated the need for explicit alignment between acoustic input and text output. Among these, the Listen, Attend and Spell (LAS) model introduced attention mechanisms that dynamically aligned audio frames with the generated text, significantly improving recognition accuracy [30]. In parallel, recurrent neural network transducers (RNN-T) emerged as a key architecture for streaming ASR, enabling low-latency recognition by jointly modeling acoustic features and previously decoded tokens [31].

Transformer-based architectures emerged as a major advancement in speech recognition, replacing recurrent connections with a self-attention mechanism. Self-attention allows the model to assign dynamic importance weights across the input sequence, allowing it to capture long-range dependencies more effectively [32, 33]. The adoption of self-attention also enabled larger training batches and deeper model architectures, accelerating gains in performance.

Conformer models, which combine convolutional operations with Transformer blocks, further improved speech recognition by simultaneously learning local spectral features and global contextual information [34]. Alongside these architectural advances, self-supervised pre-training methods such as wav2vec 2.0 demonstrated that large volumes of unlabeled audio could effectively produce robust speech representations, significantly boosting ASR performance in low-resource scenarios [35].

OpenAI's Whisper represents one of the current frontiers of ASR, built on a Transformer-based encoder–decoder architecture and trained on a vast dataset comprising 680,000 hours of weakly supervised audio [8]. By training across multiple tasks, including transcription, translation, and timestamp prediction, across more than 90 languages, Whisper demonstrates strong robustness to noise, accents, and varied domains without specialized/task-specific fine-tuning.

While domain-specific models may outperform it on narrowly defined benchmarks, Whisper's generalization capabilities highlight the growing importance and relevance of large-scale, multitask frameworks. However, its baseline performance on real-world ATC audio underscores the persistent gap between general-purpose systems and the demands of highly specialized operational environments.

Fine-tuning is the final adaptation stage for large or pre-trained models, enabling them to perform effectively in specialized domains. In this process, model parameters are updated using a domain-specific dataset and a loss function such as cross-entropy. Through iterative weight adjustments via backpropagation, the model becomes increasingly attuned to the characteristics of the target environment, such as domain-specific noise, accents, or terminology. This refinement is particularly important in safety-critical applications like air traffic control, where high recognition accuracy is essential to maintaining operational safety.

C. Speech Enhancement and Noise Reduction

Speech enhancement techniques aim to improve the quality and clarity of speech signals corrupted by various types of noise. Effective speech enhancement is essential for reliable human and machine perception, particularly in challenging acoustic environments. Historically, classical methods such as spectral subtraction, which removes an estimated noise spectrum from the noisy signal, Wiener filtering, a linear filtering approach minimizing mean-square error based on estimated signal-to-noise ratios, and Minimum Mean Square Error (MMSE) estimators, statistical methods minimizing expected error under probabilistic noise assumptions, dominated the field due to their simplicity and computational efficiency [36]. However, these methods often introduced unnatural artifacts and struggled to adapt effectively to non-stationary or rapidly changing noise environments.

With the emergence of deep learning, speech enhancement techniques have made substantial progress. SEGAN (Speech Enhancement Generative Adversarial Network) was among the pioneering methods applying generative adversarial networks (GANs) to enhance speech directly in the waveform domain. SEGAN employs a convolutional encoder-decoder architecture with skip connections (similar to a U-Net, a widely-used encoder–decoder convolutional neural network architecture), trained *adversarially*, meaning a second network (the discriminator) evaluates whether outputs resemble real, clean speech, thus guiding the generator toward producing more realistic waveforms directly from noisy inputs [37]. Despite some mixed results on standard metrics like Perceptual Evaluation of Speech Quality (PESQ), SEGAN notably reduced speech distortion and suppressed background noise compared to traditional Wiener filtering [37].

Real-time applicability has become increasingly important, especially in communication scenarios. RNNoise addressed this by combining classical digital signal processing (DSP), techniques used to mathematically manipulate audio signals, with a lightweight recurrent neural network based on gated recurrent units (GRUs) to estimate spectral gains in real-time, achieving substantial improvements over classical approaches. This hybrid DSP and neural approach demonstrated exceptional performance in handling non-stationary noise conditions, making it practical for real-time applications such as teleconferencing and mobile communications [38].

Addressing the limitations of magnitude-only enhancement, DCCRN (Deep Complex Convolutional Recurrent Network) explicitly models phase information within the complex spectrogram domain [39]. Utilizing complex-valued convolutional and recurrent layers, DCCRN significantly improved speech clarity and perceptual quality, achieving top performance in the real-time track of the Interspeech 2020 Deep Noise Suppression Challenge [39]. Similarly, Defossez et al. introduced a waveform-domain speech enhancement model leveraging convolutional encoder-decoder architectures combined with multi-scale spectral losses, which are loss functions that simultaneously measure differences between predicted and actual speech across multiple time-frequency resolutions. Additionally, their approach employed extensive data augmentation techniques. This model achieved state-of-the-art results on standard benchmarks while maintaining computational efficiency suitable for real-time deployment [40].

These general advancements demonstrate the capability of deep neural architectures, particularly GAN-based and complex-domain models, to significantly outperform traditional methods, delivering high-quality, real-time speech enhancement across diverse applications. However, specialized contexts such as air traffic control introduce uniquely challenging noise conditions that often exceed the capabilities of general-purpose methods. In ATC environments, speech signals frequently suffer from intense, non-stationary noise sources, including cockpit noise, radio frequency interference, and overlapping communications, rendering conventional enhancement techniques insufficient. To address these specific challenges, recent research has increasingly focused on developing specialized deep-learning solutions explicitly tailored for aviation communication scenarios. Several notable studies exemplify this trend, leveraging advanced neural architectures to achieve significant improvements in enhancing ATC speech quality.

For instance, Chen et al. developed a GAN-based speech enhancement method tailored specifically for cockpit communications [41]. Their approach leveraged a Deep Convolutional GAN trained with a Wasserstein GAN (WGAN) objective to stabilize the adversarial training process. The model incorporated conditional inputs, spectral constraint layers, and a combination of adversarial and reconstruction losses, significantly outperforming conventional GANs by effectively reducing background noise and improving speech clarity in flight crew dialogues.

Building on this GAN-based direction, Liang et al. introduced DeCGAN, a model explicitly designed for ATC communications. DeCGAN leveraged a sophisticated DeConformer architecture that combined transformer-based self-attention with deformable convolutions to capture both global and local spectral details. Its dual-output structure, which predicts both complex spectrograms and time-frequency masks, enabled enhancement of both phase and magnitude components of ATC speech. This approach achieved state-of-the-art perceptual quality metrics, with PESQ scores up to 3.31 and Short-Time Objective Intelligibility (STOI) scores up to 0.96 on seen-noise test sets, and approximately 3.06 PESQ/0.94 STOI on the unseen-noise set [42].

Recognizing the benefit of closer integration between speech enhancement and automatic speech recognition, Yu et al. proposed ROSE, a recognition-oriented speech enhancement framework tailored specifically for the ATC domain. ROSE employed a multi-objective learning strategy to jointly optimize speech enhancement quality and ASR accuracy. The framework employed innovative attention-based modules within a convolutional U-Net architecture to effectively suppress radio echoes and reduce non-stationary noise, which are common issues in ATC audio. ROSE's modular design facilitated easy integration with existing ASR systems, leading to substantial improvements in speech clarity and transcription accuracy under realistic operational conditions [43].

Extending the integrated strategy even further, Jiang et al. proposed a noise-robust ASR architecture specifically for ATC applications based on the U2 framework, a unified architecture designed for both streaming and non-streaming ASR tasks. Their model incorporated a dedicated speech-enhancement front-end, jointly trained alongside the ASR components, achieving significant reductions in word error rates, especially under challenging noise scenarios typical of ATC environments. This joint enhancement-recognition model further demonstrated the value of explicitly optimizing enhancement objectives in tandem with recognition accuracy [44].

Collectively, recent research highlights the effectiveness of domain-specific speech enhancement methods for air traffic control. The progression from general-purpose GAN-based methods to specialized architectures and tightly integrated, recognition-oriented frameworks demonstrates substantial potential for improving communication clarity and ASR reliability within the demanding operational environments characteristic of air traffic control.

D. Data Augmentation for Robust Modeling

Data augmentation involves artificially expanding a training dataset by applying label-preserving transformations to existing samples. This approach helps mitigate overfitting, increases dataset diversity, and improves generalization by exposing models to a wider range of input variations, without the need for additional manual labeling [45]. Below we examine how augmentation strategies have evolved in speech/audio applications.

In speech and audio, augmentation techniques aim to enhance robustness by simulating realistic acoustic variability. Common methods include injecting background noise, altering speech speed or pitch, and convolving clean audio with room impulse responses (RIRs) to mimic reverberant environments. RIRs capture how sound reflects off surfaces such as walls and ceilings, creating overlapping echoes that simulate real-world acoustics. Ko et al. showed that speed perturbation and simulated reverberation can significantly improve recognition accuracy in diverse acoustic conditions [9].

While traditional augmentation methods are valued for their simplicity, computational efficiency, and reliable label preservation, they often fall short in specialized domains like air traffic control. In particular, classical techniques struggle to model complex, non-stationary artifacts characteristic of radio communication, such as VHF radio hiss or transient bursts from squelch activation. Simply overlaying Gaussian noise or generic background audio fails to replicate these nuanced signal characteristics.

To address these limitations, researchers have explored more sophisticated generative methods beyond simple additive noise. Time-frequency masking approaches, notably SpecAugment [46], introduce synthetic, non-stationary distortions by randomly masking or warping spectral regions. Other methods employ large, carefully curated noise libraries like MUSAN [47], providing extensive collections of diverse acoustic conditions to augment speech training data. Variational autoencoder (VAE)-based augmentation has also demonstrated success, leveraging learned transformations from clean to noisy speech to improve robustness [48]. More recently, diffusion-based generative approaches have been investigated, synthesizing realistic noisy speech by iteratively transforming clean signals through learned probabilistic

diffusion processes [49].

Alongside these promising non-adversarial approaches, generative adversarial networks (GANs) have further advanced noise modeling capabilities. GAN-based noise injection methods have shown substantial improvements in ASR robustness, often outperforming conventional augmentation techniques in challenging noise conditions [50, 51].

E. Generative Adversarial Networks

Generative Adversarial Networks (GANs), introduced by Goodfellow et al. [10], are central to our SimuGAN-Whisper-ATC augmentation pipeline, offering an effective way to model realistic noise profiles in ATC speech data. A GAN framework consists of two competing neural networks: a generator (G), which synthesizes data samples, and a discriminator (D), which attempts to differentiate these synthetic samples from genuine ones. Formally, GAN training involves solving the following two-player optimization problem:

$$\min_G \max_D V(G, D) \;=\; \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[\log D(\mathbf{x})] \;+\; \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}}[\log(1 - D(G(\mathbf{z})))]\,, \tag{1}$$

where $\mathbf{x}$ represents real data samples drawn from the true distribution p_{data}, and $\mathbf{z}$ denotes random noise vectors sampled from a latent distribution $p_{\mathbf{z}}$ (typically Gaussian). The generator's goal is to produce samples $G(\mathbf{z})$ that closely resemble real data, whereas the discriminator aims to correctly classify each sample as either real or generated. Training proceeds iteratively, adjusting D to enhance its discriminative accuracy and updating G to generate increasingly realistic samples. Ideally, equilibrium occurs when D can no longer distinguish real from synthetic samples [10].

However, achieving this ideal equilibrium can be challenging. In practice, GAN training often encounters stability issues, including mode collapse, where the generator repeatedly outputs overly similar examples (such as producing only static noise), and vanishing gradients, arising when the discriminator becomes excessively accurate too early. To overcome these difficulties, improved architectures like Deep Convolutional GANs (DCGAN) [52] and modified loss functions, such as the Wasserstein GAN (WGAN) [53] and WGAN with Gradient Penalty (WGAN-GP) [54], have stabilized training. These developments have broadened GAN applicability across computer vision, natural language processing, and, more recently, speech tasks.

When applying GANs specifically to speech data, the nature of available datasets, whether parallel or unparalleled, becomes especially important. Parallel datasets include explicit pairs of noisy and corresponding clean recordings, enabling direct mappings between domains. Conversely, real-world datasets, especially in specialized fields such as aviation, are typically unparalleled (or unpaired), comprising separate collections of clean and noisy audio samples without direct alignment. SimuGAN [51] addresses the challenge of realistic noise synthesis by injecting plausible noise into clean speech without requiring parallel (or paired) clean–noisy samples. It employs a discriminator to ensure that the generated noisy outputs closely resemble real-world conditions.

Building on this approach, recent studies have shown that GAN-based noise injection can substantially improve ASR performance, especially in low-resource settings. For example, Chen et al. trained SimuGAN using just ten minutes of noisy audio alongside a larger set of clean speech, successfully capturing the structure of complex noise patterns found in real environments. By generating a diverse range of "clean-plus-noise" training examples, SimuGAN improved ASR robustness under realistic acoustic conditions. Related GAN-based methods have achieved similar results by emphasizing critical time–frequency regions within the audio signal [55]. These generative approaches directly support our goal of bridging clean simulation datasets (e.g., ATCOSIM) with the acoustic realities of air traffic control communications. Unlike conventional augmentation techniques, GANs are particularly well-suited to modeling the non-stationary and domain-specific noise characteristics of aviation radio transmissions, such as VHF radio hiss and squelch pops. As discussed in more detail in Section IV, SimuGAN's ability to learn from unpaired data makes it a compelling choice for integration into our SimuGAN-Whisper-ATC pipeline. Its demonstrated performance advantages over traditional augmentation methods were a key factor in its selection for this work.

III. Datasets

Open-source labeled corpora of air traffic control speech remain scarce, which limits the training and validation of domain-specific ASR models. In this work, we leverage three complementary open-source datasets: *ATCO2*, *ATCOSIM*, and *TartanAviation*. Each dataset fulfills a distinct role: *ATCO2*, with limited but accurately transcribed real-world recordings, provides a realistic benchmark for evaluating ASR models under authentic operational noise conditions. *ATCOSIM*, featuring fully labeled but noise-free simulated ATC communications, provides domain-specific phraseology and clean reference data suitable for noise injection. Lastly, the extensive but unlabeled *TartanAviation* dataset serves

as the source for realistic noise profiles used in training our GAN-based noise augmentation pipeline. By integrating these datasets strategically, we construct an augmented training corpus that reflects authentic ATC acoustic conditions, significantly enhancing model robustness and reliability.

A. ATCO2

ATCO2 is a large-scale corpus of real ATC voice communications collected at more than ten airports worldwide, including European locations such as Prague and Zurich, and others in Australia [5]. The total dataset exceeds 5,000 hours, although only about four hours are fully transcribed by human annotators. An approximately one-hour publicly released portion forms the open-source test set. The remainder of the data is unlabeled, although pseudo-transcripts, produced by an in-domain ASR system, are included as weak labels. These recordings capture genuine VHF transmissions, complete with static, overlapping calls, and variable signal-to-noise ratios, making the ATCO2 test subset an ideal benchmark for real ATC speech. In our experiments, we utilize approximately 80% of the accurately transcribed speech for training purposes, reserving the remaining 20% as a validation set to evaluate model performance under realistic conditions.

B. ATCOSIM

ATCOSIM was developed at the EUROCONTROL Experimental Centre in France under simulation conditions [56]. Professional controllers communicated with human pseudo-pilots in a realistic but noise-free environment, where only the controller's side of each exchange was recorded. ATCOSIM contains around 10 hours of speech, fully segmented and transcribed at the utterance level. While the audio is clean (no radio artifacts) and free from overlapping speech, its content closely matches typical en-route or approach controller phraseology. Despite its limited duration, ATCOSIM's high-quality labels and domain fidelity make it a valuable resource for fine-tuning. In our approach, we inject real noise characteristics into this simulation-based dataset to expand its realism and ultimately enhance model robustness.

C. TartanAviation

Developed by Carnegie Mellon University's Robotics Institute, TartanAviation is a multi-modal repository collected at two regional airports in Pennsylvania, USA [57]. It comprises over 3,300 hours of raw VHF radio audio, accompanied by extensive video and ADS-B track data. Unlike ATCOSIM and ATCO2, these recordings have no human transcripts and incorporate both towered and non-towered communications, leading to a wide range of operational styles and accent variability. Although the speech is unlabeled, it provides an extensive pool of authentic ATC noise conditions, including radio interference, background chatter, and inconsistent signal quality. We leverage these realistic noise patterns, in combination with additional noise samples derived from the labeled portion of ATCO2, to accurately simulate operational radio distortions and inject them into ATCOSIM's clean utterances.

Using this combination of data, we train a Generative Adversarial Network (GAN) to learn realistic noise characteristics from TartanAviation, as well as from the labeled portion (approximately 80% of the available 1-hour dataset) of ATCO2. By training the generative pipeline directly on authentic radio conditions, the GAN synthesizes realistic noise features and injects them into the clean ATCOSIM recordings, thereby significantly expanding the size and acoustic diversity of our labeled dataset. We then fine-tune Whisper on this augmented corpus and evaluate real-world performance using the held-out validation portion (approximately 20%) of the ATCO2 test set. Leveraging these three datasets collectively addresses the gap between simulation-based training and real ATC operational environments, reducing the mismatch that frequently hinders ASR performance due to current limitations in open-source datasets.

IV. Methodology

This section describes our overall methodology for improving ATC-focused ASR models through generative data augmentation and domain-specific fine-tuning. First, we summarize the conceptual approach to dataset preparation and noise injection, adopting prior work's splits to ensure comparability. Next, we introduce the high-level SimuGAN-based framework for generating realistic noisy speech from clean data, and outline our Whisper adaptation strategy. Finally, we describe why a customized domain-specific word error rate (WER) calculation is necessary to evaluate performance under the unique linguistic demands of air-traffic control (ATC).

A. Data Preprocessing (Conceptual Overview)

To enable direct comparison with previous work, we adopt the same dataset splits for *ATCO2* as defined by van Doorn et al. [6], including their balanced speaker assignments across training and validation subsets. Specifically, we use approximately 80% of the available one-hour labeled ATCO2 data for training and reserve the remaining 20% for validation. For *ATCOSIM*, we utilize the full 10 hours of clean, fully labeled data for training, following the same partitioning scheme as in the referenced study.

In contrast, the *TartanAviation* dataset presented unique challenges due to its real-world collection methodology; microphones were left open continuously for extended periods, resulting in recordings that included silence and irrelevant content. While this raw data offers rich diversity in ATC transmissions, extensive cleaning was required to isolate the speech-bearing segments required for accurate GAN-based noise modeling. After preprocessing, these three datasets collectively formed the foundation of our SimuGAN augmentation pipeline and the subsequent fine-tuning of Whisper.

B. SimuGAN Noise-Injection (Conceptual Overview)

To learn realistic noise patterns in the absence of strictly parallel clean–noisy pairs, we adopt a SimuGAN-style unpaired training approach, as illustrated in Fig. 1. The pipeline begins by converting clean ATCOSIM speech into magnitude spectrograms using Short-Time Fourier Transforms (STFT). The generator G, structured as a U-Net architecture, then synthesizes realistic VHF noise artifacts, producing artificial noisy spectrograms. In parallel, the discriminator D, implemented as a PatchGAN classifier, learns to differentiate authentic noisy spectrograms (sampled from real ATC recordings) from the synthetic spectrograms generated by G. Through iterative adversarial training, the generator progressively learns to reproduce ATC-specific radio distortions, such as narrow-band spikes and squelch pops, which simpler additive-noise methods often fail to replicate accurately.

The generator architecture is specifically designed to preserve the overall structure of the input speech while injecting realistic spectral distortions. To further stabilize the training process and enhance noise realism, we incorporate multiple regularization objectives, detailed in Section V. Finally, the augmented spectrogram is recombined with the original audio phase to produce a time-domain "clean+noise" waveform, enriching the training dataset used for fine-tuning Whisper.

Section V provides comprehensive details regarding the data preprocessing steps, SimuGAN network architectures, specific loss function formulations, and training hyper-parameters necessary to implement this conceptual pipeline.

C. Whisper Adaptation Strategy (Conceptual Overview)

While Whisper [8] is pre-trained on a vast and diverse corpus, prior work reports zero-shot Word Error Rates (WER) as high as 29% on ATCO2 data [6], highlighting the significant performance gap within the ATC domain. To address this, we fine-tune Whisper on a combination of real ATCO2 recordings and synthetic ATCOSIM samples enhanced with GAN-generated noise. This approach aims to bridge the gap between clean, simulated data and the complexity of real VHF transmissions. During training, we maintain a 2:1 ratio of real to augmented data within each batch, ensuring the model is consistently exposed to authentic noisy transmissions while also benefiting from the controlled diversity introduced through simulated distortions.

D. ATC-Specific Word Error Rate

Standard Word Error Rate often treats minor variations, such as differences in call sign formatting (e.g., "A K L M" vs. "Alpha Kilo Lima Mike"), as equally significant as critical command terms like "descend". To better reflect the linguistic nuances of air traffic communication, we apply a domain-specific normalization procedure, largely adapted from Van Doorn et al. [6]. This includes handling NATO phonetic alphabet expansions, splitting numeric digits, and recognition of common synonyms used in ATC phraseology.

By accounting for these factors, our metric provides a more meaningful assessment of transcription accuracy in operational contexts. For example, correct expansions such as converting "AFR123" to "air france one two three" are no longer penalized as token mismatches. This adapted ATC-WER thus offers a more relevant benchmark for evaluating Whisper's performance in realistic air traffic control scenarios.

V. Implementation

This section discusses every design choice and numeric detail necessary to replicate our results, from data-cleaning thresholds to hyperparameter settings for SimuGAN and Whisper. We also describe the domain-aware normalization

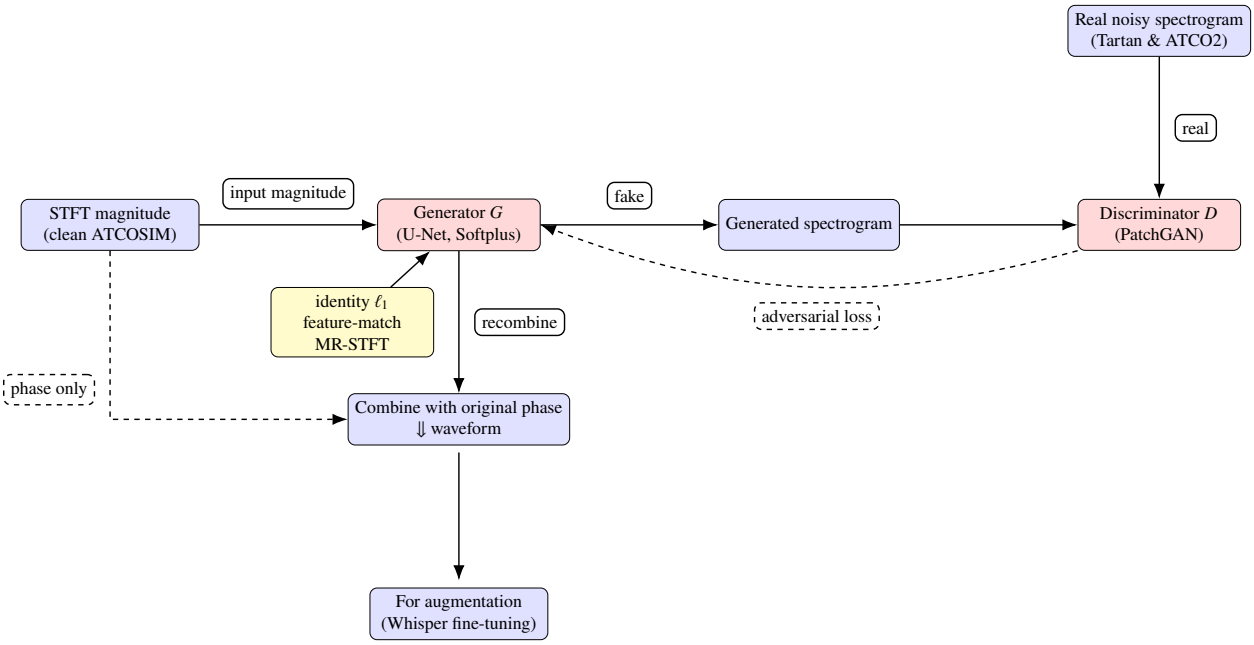

Fig. 1 Data flow in the SimuGAN noise-injection pipeline. Clean STFT magnitudes from the ATCOSIM dataset are input to the generator G, which produces *synthetic* (fake) noisy spectrograms. The discriminator D compares these with *real* noisy spectrograms sampled from TartanAviation or ATCO2 and outputs an adversarial loss (dashed arrow). The generator G is trained using a combination of this adversarial loss and additional terms: identity loss, feature-matching loss, and multi-resolution STFT loss (yellow block). The generated magnitudes are then combined with the original phase using an *inverse STFT*, resulting in "clean + noise" waveforms that augment ATCOSIM for Whisper fine-tuning.

rules and how we integrate them during evaluation.

A. Data Preprocessing (Detailed Steps)

Cleaning Criteria for TartanAviation Unlike the smaller, well-defined subsets of ATCOSIM and ATCO2, the TartanAviation dataset consists of continuous multi-month recordings, necessitating extensive preprocessing. Our filtering process included:

- Discarding segments shorter than one second.
- Removing segments with an estimated signal-to-noise ratio (SNR) below 3 dB.
- Splitting excessively long clips (some exceeding 10 minutes) into smaller segments using WebRTC-based voice activity detection (VAD) [58].

These steps reduced the initial 40 hours of raw TartanAviation data (as selected during preliminary experiments) to approximately 4.5 hours of usable speech. For training the GAN pipeline, we utilized the training subset of the labeled ATCO2 data (80% of one hour) combined with the 4.5 hours of processed TartanAviation data. Subsequently, for Whisper fine-tuning, we employed the entire 10-hour ATCOSIM dataset along with the same ATCO2 training subset.

Summary of ATC Speech Corpora Used Table 1 summarizes the datasets used in this study, distinguishing between the total publicly available hours and the specific subsets employed in our experiments. ATCO2 provides real, labeled but noisy speech samples for evaluation, TartanAviation supplies unlabeled VHF radio traffic for noise modeling, and ATCOSIM serves as our primary clean corpus for supervised training and GAN-based augmentation.

Table 1 Summary of ATC speech corpora used in this study. "Total" refers to the full publicly available dataset, while "Used" indicates the subset used in our experiments.

Dataset	Total (h)	Labeled (h)	Used (h)	Notes
TartanAviation	~ 3,374	0	~ 4.5	Untranscribed VHF traffic from two regional U.S. airports; realistic noise for GAN.
ATCO2	~ 5,285	~ 1	~ 1	Real-world VHF speech from >10 airports; 80% used for training, 20% for validation.
ATCOSIM	~ 10	~ 10	~ 10	Clean, simulated controller audio (EUROCONTROL). All used.

B. GAN Implementation (Technical Details)

This section outlines the technical foundations of our SimuGAN framework for injecting realistic radio noise into clean ATC speech. The goal is to bridge the gap between simulated and real-world conditions by learning to generate domain-specific distortions directly in the spectral domain. We describe the preprocessing pipeline, model architectures, loss functions, and training strategy in detail below.

STFT and Data-Flow Mechanics Clean ATCOSIM audio is first transformed into magnitude spectrograms using the Short-Time Fourier Transform (STFT), while the corresponding phase is retained for later reconstruction. To encourage coverage diversity during training, we apply random cropping to the spectrograms. The time dimension is zero-padded to the nearest multiple of eight to match U-Net pooling requirements. After the generator synthesizes a "fake" noisy spectrogram, it is recombined with the original phase and converted back into the time-domain waveform using inverse STFT. This waveform carries the learned radio distortions and is used to augment the training data for Whisper fine-tuning.

Generator and Discriminator Architecture The generator is a U-Net consisting of encoder (downsampling) and decoder (upsampling) blocks with skip connections. A final `Softplus` activation ensures strictly positive output magnitudes, eliminating the need for explicit clamping. The discriminator follows a PatchGAN architecture with spectral normalization applied to all convolutional layers, promoting stability during adversarial training on non-parallel data.

Loss Function and Optimization Strategy The generator is trained using a combination of adversarial and auxiliary objectives to enforce both realism and content preservation. The full loss function is defined as:

$$\mathcal{L}_G \;=\; \underbrace{\mathcal{L}_{\text{adv}}(G,D)}_{\text{adversarial}} + \lambda_1\,\mathcal{L}_{\text{id}} \;+\; \lambda_2\,\mathcal{L}_{\text{fm}} \;+\; \lambda_3\,\mathcal{L}_{\text{mr-stft}}. \tag{2}$$

where

- **Identity loss** (ℓ_1) penalizes deviations from the clean input.
- **Feature-matching loss** aligns internal feature maps in the discriminator between real and fake inputs.
- **Multi-resolution STFT (MR-STFT) loss** enforces consistency across multiple time–frequency scales.

Both generator (G) and discriminator (D) are optimized using the Adam optimizer with $\beta = (0.5, 0.999)$. We set the learning rates $lr_D = 2 \times 10^{-4}$ and $lr_G = 1 \times 10^{-4}$, with label smoothing applied to real samples ($y_{real} = 0.9$) to improve training stability. The key hyperparameters are summarized in Table 2.

Training Procedure Each training iteration pairs a clean ATCOSIM spectrogram with a randomly selected noisy reference from either TartanAviation or ATCO2. These real spectrograms serve as positive samples for the discriminator, while the generator learns to produce spectrally realistic alternatives. Over successive epochs, the generator increasingly captures salient distortion patterns typical of VHF radio communications.

Table 2 Key hyper-parameters for SimuGAN training.

Symbol / Item	Value	Applies to	Purpose
λ_1 (identity)	4	G loss	stabilize content
λ_2 (feat-match)	10	G loss	feature coherence
λ_3 (MR-STFT)	1	G loss	multi-scale detail
lr_G	1×10^{-4}	Adam	slower drift
lr_D	2×10^{-4}	Adam	stronger critic
Label smooth y_{real}	0.9	D targets	avoid confidence spikes
Adam β	$(0.5, 0.999)$	both	standard GAN default

C. Whisper Fine-tuning (Technical Details)

Fine-tuning large-scale speech models such as Whisper for domain-specific applications like air traffic control communications requires careful data curation and hyperparameter tuning. This section details our approach for adapting Whisper to noisy, accented, and specialized ATC speech using both real and artificially augmented datasets.

Dataset Mixture and Hyperparameters Although Whisper was originally trained on approximately 680,000 hours of general-purpose audio [8], van Doorn *et al.* [6] reported a substantial performance gap (29.05% zero-shot WER) when applied directly to ATCO2. To address this, we fine-tuned the Whisper *large-v2* checkpoint on a combined dataset composed of approximately 1 hour of authentic ATCO2 recordings and 10 hours of SimuGAN-augmented ATCOSIM data. Despite this raw duration ratio of 1:10, we employed an oversampling strategy, ensuring each mini-batch contained approximately twice as many real ATCO2 samples as synthetic ones, thus establishing a 2:1 sampling ratio. This approach prioritizes the model's exposure to genuine radio transmissions while maintaining broad coverage of realistic noise conditions provided by the augmented data.

Fine-tuning was conducted for 10 epochs using the AdamW optimizer with a cosine-decay learning rate schedule [59]. Additional details about batch size, learning rate specifics, hardware, and decoding settings are summarized in Table 3.

Table 3 Key hyperparameters and settings for Whisper fine-tuning on mixed ATC data.

Item	Value / Description
Base Whisper model	large-v2 (OpenAI)
GPU hardware	NVIDIA L40 (PACE Phoenix cluster)
Batch size	8
Epochs	10
Learning rate	1×10^{-5} (cosine decay)
Optimizer	AdamW, $\beta = (0.9, 0.999)$
Data ratio	2:1 sampling per batch (real ATCO2 : augmented ATCOSIM)
Logging & tracking	Weights & Biases (wandb)
Checkpoint selection	min WER on dev set
Decoding strategy	Greedy decoding (baseline)
Total training time	$\sim$ 8 hours

Logging, Dev-Set Monitoring, and Decoding We continuously evaluate on a held-out ATCO2 dev set after each epoch, using Whisper's built-in WER to select checkpoints. We log training curves, WER, and hyperparameters via Weights & Biases [60]. Although beam search or advanced decoding heuristics could further reduce WER, we currently focus on greedy decoding in this study. Future work will include investigating beam widths, real-time latency constraints, and domain-specific expansions of the lexical search.

D. Scoring & Evaluation (ATC-Specific WER Metrics)

Evaluating automatic speech recognition systems in the air traffic control domain requires domain-sensitive scoring strategies. WER metrics often fail to account for ATC-specific terminology, call sign conventions, and spoken formatting. To address this, we employ a customized normalization and evaluation pipeline, largely adapted from Van Doorn et al. [6], aligning WER computation with operational realities.

Normalization and Token Mapping A specialized text-normalization process is applied to both reference transcripts and hypotheses to account for ATC-specific phraseology and formatting. This includes expanding single-letter tokens into their NATO phonetic equivalents (e.g., "B" → "bravo"), splitting numeric sequences into individual digits, and unifying common abbreviations or synonyms (e.g., "FL" → "flight level"). Airline call signs are normalized by isolating numeric suffixes, thereby avoiding mismatches caused by variability in naming conventions. Table 4 illustrates an example of such transformations.

Table 4 Transcript-normalization example demonstrating ATC-specific transformations.

Stage	Example transcript
Before normalization	*Hotel November X-ray ninety five crossing is approved QNH one zero two two*
After normalization	*hotel november xray 9 5 crossing is approved qnh 1 0 2 2*

WER Computation and Evaluation Strategy Following normalization, WER is calculated using the standard formula:

$$\text{WER} = \frac{S + D + I}{N}, \tag{3}$$

where S denotes substitutions, D deletions, I insertions, and N denotes the total number of reference tokens. This ATC-aware approach ensures that semantically equivalent expansions are not unfairly penalized. While this specialized WER computation provides a more accurate measure of ASR performance in operational contexts, it is not used during training due to its computational overhead. Instead, it is applied post-training to evaluate model accuracy under realistic ATC communication conditions.

VI. Evaluation and Results

In this section, we evaluate the effectiveness of our SimuGAN augmentation pipeline through spectrogram-based qualitative analyses and objective acoustic metrics. We subsequently quantify the impact of this augmentation on Whisper ASR performance.

A. Spectrogram Analysis of SimuGAN-Synthesized ATC Noise

A spectrogram visually represents audio signals as frequency versus time, with amplitude shown as color intensity. These visualizations clearly highlight acoustic patterns, such as vowel resonances, harmonic structures, sudden short-lived noise bursts, and persistent background noise, that are challenging to discern from raw waveforms alone. In air-traffic control, spectrogram analysis is particularly valuable for examining distinctive VHF transmission distortions, including continuous hiss, narrow-band interference spikes, and abrupt signal dropouts, all of which significantly degrade speech recognition performance.

Figure 2 compares three spectrograms: *(i)* the clean ATCOSIM reference, *(ii)* the same audio after our SimuGAN generator has added "synthesized noise" artifacts, and *(iii)* a real–noisy clip from the ATCO2 corpus. Based on the current implementation of SimuGAN, several observations can be highlighted:

- **Preserved speech structure:** Horizontal bands corresponding to primary vowel resonances (formants) remain clearly visible in the GAN spectrogram, closely matching those in the original clean ATCOSIM recording. This demonstrates that, despite the significant noise introduced, critical harmonic features essential for distinguishing speech sounds remain intact. Maintaining these resonance patterns is vital to ensure the synthetic noise supports training robust ASR models without masking essential linguistic information.
- **Modeling Transient Interference Artifacts:** The real ATCO2 spectrogram exhibits thin vertical streaks arising from transient radio phenomena such as squelch noise, transmitter activation bursts, or carrier frequency

interference. Although less pronounced, the GAN-generated spectrogram similarly captures these brief, high-energy disturbances, indicating that the model effectively reproduces transient noise patterns rather than introducing generic stationary noise. Accurately modeling these short-duration interference events is key for realistic domain adaptation in ASR training.

- **Partial Representation of Transmission Dropouts:** Real-world ATC transmissions occasionally experience brief interruptions, visible in spectrograms as faint vertical bands, typically caused by weakened VHF signals or activation of receiver noise suppression thresholds. The GAN-generated spectrogram replicates these interruptions, although in shorter durations and with reduced intensity compared to real ATCO2 examples. Nevertheless, the presence of these partial signal interruptions enhances the realism of the augmented data, aligning it more closely with authentic operational conditions. Future work will focus on refining the GAN-generated dropout patterns to better match the severity and duration observed in real-world ATC communications.
- **Realistic High-Frequency Attenuation:** Both the GAN-synthesized and real ATCO2 spectrograms exhibit a clear upper-frequency boundary around 3–4 kHz, unlike the clean ATCOSIM reference, which extends energy beyond 6 kHz due to laboratory recording conditions. This phenomenon accurately reflects the bandwidth constraints imposed by standard 8.33 kHz VHF communication equipment, which inherently filters higher frequencies. The GAN's replication of this frequency limitation demonstrates its ability to realistically incorporate actual hardware-related characteristics, rather than simply adding generic noise artifacts. Future work will further refine the GAN-generated frequency attenuation to better emulate real-world operational conditions.

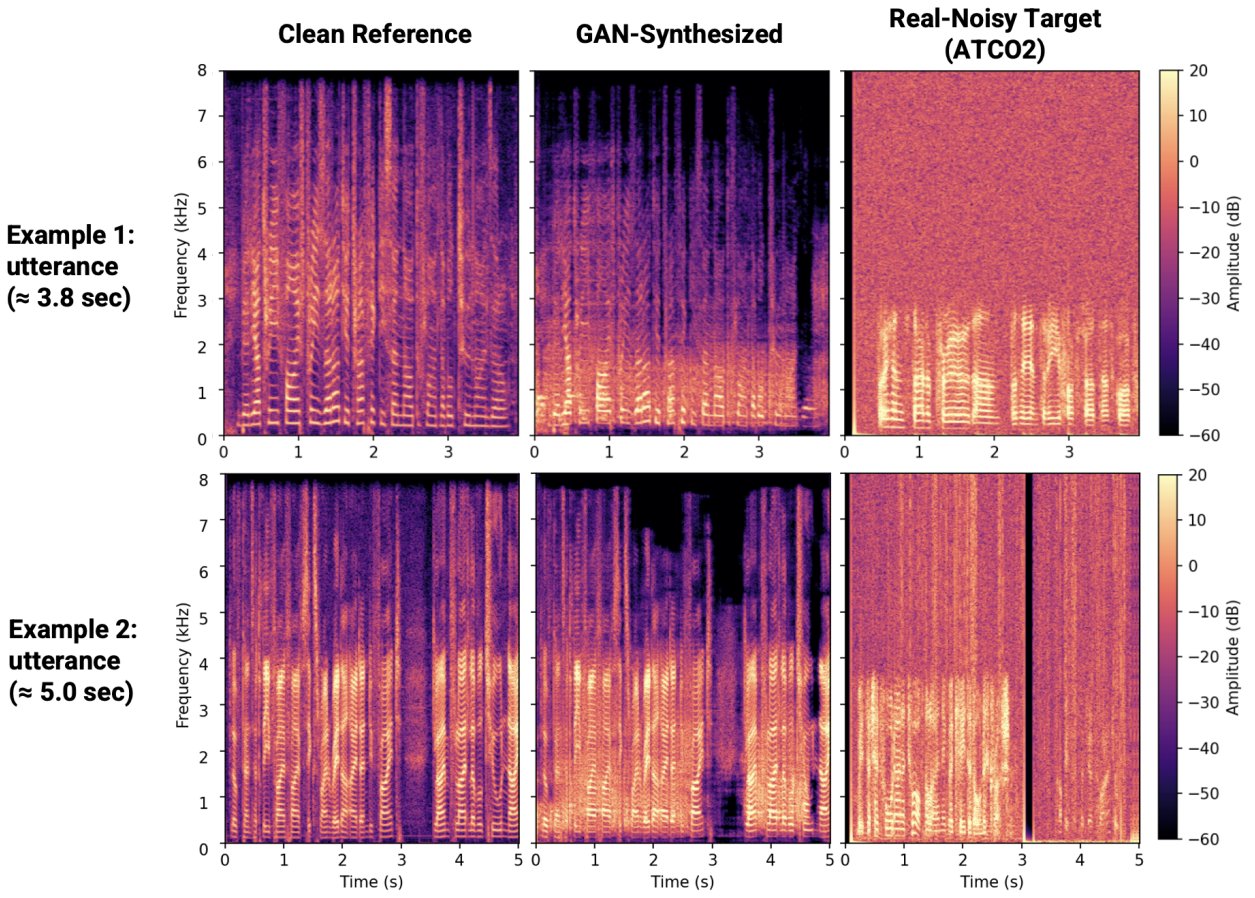

Fig. 2 Composite log–magnitude spectrograms produced by our SimuGAN evaluation. Each row displays a different utterance (top: ≈ 3.8 s, bottom: ≈ 5.0 s), comparing three variants: *clean reference*, *GAN-synthesized (fake-noisy)*, and *real-noisy* target. The generator convincingly recreates the elevated noise floor, narrow-band spikes, and frequency-selective dropouts characteristic of real VHF ATC radio transmissions, while preserving the underlying speech spectral structure. All spectrograms share the same amplitude scale (+20 to –60 dB) to facilitate direct comparison.

Having visually assessed these GAN outputs, we next evaluate objective acoustic metrics on the (clean, fake, real) triplets. This analysis aims to confirm that the synthetic noisy signals align not only with subjective impressions but also with statistical and perceptual measures of realism. The following subsection details these metric definitions and summarizes our findings.

B. Quantitative Evaluation of GAN-Synthesized ATC Noise (Objective Metrics)

To complement our spectrogram-based qualitative analysis, we quantitatively evaluated the realism and perceptual fidelity of our GAN-synthesized noise. We randomly selected 500 triplets (clean ATCOSIM reference, GAN-synthesized noisy, and real noisy from ATCO2/TartanAviation). Table 5 summarizes the mean, standard deviation, and two-sided 95% confidence intervals (CI) for each metric. Future work will involve assessing potential sample bias by employing sampling techniques that ensure representation across different speaker groups, noise levels, or operational scenarios, while also ensuring balanced duration distributions.

We evaluated several metrics grouped into three categories: spectral metrics, perceptual metrics and speaker and noise-dynamics metrics. They are defined in more detail below.

Spectral Metrics:

- **Log-Spectral Distance (LSD, dB)**: Measures frequency-domain differences between synthetic and real spectrograms. Lower values indicate closer spectral similarity.
- **Multi-resolution STFT loss (MR-STFT) [61]**: Assesses reconstruction fidelity across multiple STFT resolutions. Lower values indicate better multi-scale spectral realism.
- **Scale-Invariant Signal-to-Distortion Ratio (SI-SDR, dB)**: Quantifies the clarity of original speech despite added noise, with higher values reflecting better preservation of speech content.

Perceptual Metrics:

- **Embedding Distance (EmbedDist, dimensionless)**: Measures cosine distance between wav2vec2 embeddings of synthetic versus real signals, reflecting phonetic similarity important for ASR; lower values indicate greater perceptual realism.
- **Short-Time Objective Intelligibility (STOI, dimensionless [0–1])**: Predicts human intelligibility at the word level; higher scores indicate less intelligibility degradation by noise.
- **Fréchet Audio Distance (FAD, dimensionless) [62]**: Compares feature distributions of real and synthetic audio globally, capturing broader perceptual quality. Lower values indicate better global realism. (Single-pass computation, hence no Std./CI.)

Speaker and Noise-Dynamics Metrics:

- **Speaker-Embedding Distance (SpkDist, dimensionless)**: Cosine distance between speaker embeddings of clean versus noisy signals, indicating preservation of speaker identity; lower values imply better preservation.
- **SNR Histogram KL-Divergence ($\mathbf{SNR_{KL}}$, dimensionless)**: Measures how closely synthetic signals' SNR distributions match real ones. We use KL divergence instead of simpler metrics (like RMSE) due to its sensitivity in capturing temporal dynamics and variability inherent to ATC noise characteristics; lower values indicate better dynamic realism.

Table 5 Aggregate GAN evaluation across 500 randomly selected utterances. Lower values are preferable for Log–Spectral Distance (LSD, dB), multi-resolution STFT loss (MR-STFT), Embedding Distance (EmbedDist), SNR Histogram KL-Divergence (SNR_{KL}), Speaker-Embedding Distance (SpkDist), and Fréchet Audio Distance (FAD). Higher values are preferable for Scale-Invariant Signal-to-Distortion Ratio (SI-SDR, dB) and Short-Time Objective Intelligibility (STOI).

Metric	Mean	Std.	95% CI
LSD (dB)	10.43	1.08	±0.09
MR-STFT	0.068	0.021	±0.002
SI-SDR (dB)	3.96	2.81	±0.25
EmbedDist	0.113	0.067	±0.006
STOI	0.833	0.040	±0.004
FAD	35.73	—	—
SpkDist	0.113	0.067	±0.006
SNR_{KL}	15.31	1.29	±0.11

Overall, these quantitative evaluations support our visual assessment, confirming that our SimuGAN pipeline achieves a suitable balance between acoustic realism and intelligibility:

- The GAN-synthesized noise demonstrates strong spectral similarity to real ATC noise, with an LSD (10.43 dB vs. baseline real pairs of ~ 9.8 dB) and an MR-STFT of 0.068, both close to authentic samples.
- A high SI-SDR of ~ 4 dB confirms that speech remains distinguishable after noise injection, which is key for subsequent ASR training.
- Low embedding and speaker distances (~ 0.11) show effective preservation of both perceptual characteristics and speaker identities, highlighting the GAN's ability to augment data without compromising critical attributes.
- The STOI score of 0.83 indicates minimal intelligibility degradation, suggesting that noise augmentation does not unduly mask essential speech cues.
- The SNR-histogram KL divergence of 15.31 captures temporal variability realistically, though future work aims to refine this aspect further.

These acoustic evaluations collectively anticipate the substantial improvements in word-error rate demonstrated in Section VI.C. Future efforts will continue optimizing the GAN pipeline, refining the selection process of test samples, and validating generalization across broader and more diverse ATC datasets.

C. Whisper Fine-Tuning Results on ATCO2 Dev Set

Beyond the low-level spectral evaluations in Section VI.A, our ultimate goal is to improve ASR performance on authentic ATC audio. To quantify this improvement, we measured word error rate (WER) on the ATCO2 validation set previously used by van Doorn et al. [6], thereby ensuring direct comparability. Table 6 summarizes our key findings.

- **Baseline (Whisper-ATC, no additional augmentation):** Prior work [6] fine-tuned Whisper *large-v2* solely on noise-free ATCOSIM data with limited real ATCO2 examples, without incorporating synthesized noise augmentation. Although their approach reduced WER to approximately 14.66 %, a substantial performance gap remained due to the absence of realistic noise conditions in their training set.
- **SimuGAN–Whisper–ATC (ours):** Incorporating SimuGAN-generated realistic noise into ATCOSIM significantly expands the variability of labeled "clean + noise" examples, helping Whisper bridge the gap between simulated conditions and real VHF transmissions. Consequently, our fine-tuned Whisper achieves a WER of **3.58%**, representing a relative improvement of **75.6%** compared to the 14.66% baseline. These results were obtained using greedy decoding. We anticipate further gains from beam-search decoding, which we will explore and report in the final manuscript.
- **Towards deployable ASR systems:** Reducing WER into the low single-digit range positions our system closer to operational benchmarks set by aviation authorities, which typically seek WER around 5 % or lower for safety-critical ASR applications. Achieving such accuracy enables the detection of miscommunications or pilot–controller misunderstandings more reliably. Systems trained via SimuGAN–Whisper–ATC thus hold significant promise for integration with large language models, real-time monitoring tools, and future safety-oriented systems in aviation.

- **Future directions:** Our ongoing research agenda includes evaluating alternative GAN architectures, diversifying training datasets, and applying advanced augmentation methods. Planned efforts also encompass real-time latency tests and integration with safety-monitor prototypes, aiming to refine GAN-generated noise realism further and drive resulting WERs lower for practical ATC deployments.

Table 6 ASR performance on the ATCO2 dev set.

Model / Approach	WER (%)
Zero-shot Whisper (no fine-tune)	29.05
Whisper-ATC baseline [6]	14.66
SimuGAN–Whisper–ATC (Ours)	**3.58**

Taken together, these results demonstrate the effectiveness of incorporating GAN-generated realistic noise into a domain-focused fine-tuning pipeline. By narrowing the performance gap between simulations and realistic VHF-radio conditions, the SimuGAN–Whisper–ATC framework demonstrates significant potential in developing robust, operationally relevant ASR systems in aviation.

VII. Future Work

The results presented in this paper demonstrate promising directions for enhancing ATC-focused ASR systems; however, several opportunities remain to be explored. First, we plan to investigate additional GAN architectures, specifically AttentionGAN [63] and CycleGAN [64], due to their demonstrated effectiveness in capturing complex transformations in unpaired data. AttentionGAN incorporates an attention mechanism that allows the model to dynamically focus on crucial regions within spectrograms, potentially improving the capture of transient, non-stationary noise artifacts prevalent in ATC communications. CycleGAN, on the other hand, employs cycle consistency to ensure that learned mappings between clean and noisy audio remain faithful, thus potentially enhancing the realism and stability of synthesized noise. We intend to systematically compare these methods against SimuGAN to determine their relative strengths and suitability for modeling ATC-specific radio noise characteristics. Second, to further improve transcription accuracy, we plan to explore alternative decoding strategies beyond greedy decoding, systematically varying beam search parameters, and examining trade-offs between accuracy and inference latency.

We will also validate our findings across additional ATC speech datasets beyond ATCO2, to better understand the generalizability of our approach under diverse operational and acoustic conditions. Furthermore, since the current study used a fixed mixture ratio of ATCO2 and TartanAviation samples for GAN-based noise synthesis, we intend to evaluate various mixture ratios systematically, determining optimal balance points for synthesizing maximally realistic noise artifacts. Additional hyperparameter tuning for both Whisper fine-tuning (e.g., learning rate schedules, optimizer variants) and SimuGAN training (e.g., λ weights in Eq. 2) will also be explored comprehensively.

Finally, recognizing the significant potential of generative methods in creating realistic labeled speech data, we plan to investigate further artificial expansion of ATC datasets. By combining GAN-generated acoustic noise with synthetic linguistic content (possibly leveraging text-to-speech synthesis), future research could substantially alleviate the labeled data scarcity problem, paving the way for more robust and accurate ASR systems in air traffic control.

VIII. Conclusion

This paper introduced *SimuGAN–Whisper–ATC*, a novel pipeline designed to significantly enhance Automatic Speech Recognition performance for air-traffic control communications. By leveraging a Generative Adversarial Network to inject realistic acoustic noise derived from real-world recordings (ATCO2 and TartanAviation datasets) into the clean, simulated ATCOSIM corpus, we effectively expanded the labeled training data to better reflect operational VHF communication conditions. Fine-tuning Whisper under these enriched acoustic conditions yielded a substantial reduction in word error rate, achieving a relative improvement of approximately 75.6% compared to previous state-of-the-art baselines.

Through comprehensive qualitative and quantitative analyses, including spectrogram inspections and quantification across diverse acoustic metrics, we confirmed that our GAN-generated audio not only captures important noise characteristics (transient interferences, signal dropouts, high-frequency attenuation) but also maintains speech intelligibility

and speaker identity, which are essential for reliable ASR training. The resulting ASR performance of 3.58% WER on real-world ATCO2 data represents a substantial step toward practical deployment, aligning closely with operational benchmarks and enabling future integration into real-time monitoring and safety-support systems.

Future work will focus on exploring beam search decoding strategies to further improve transcription accuracy, expanding dataset diversity, refining GAN architectures, and assessing real-time inference capabilities. Additionally, we will pursue extensive experimentation and validation to confirm the robustness and stability of our reported WER improvements. Collectively, these enhancements will continue to push the boundaries of what is achievable with ASR technologies in complex and safety-critical aviation environments.

IX. Acknowledgments

The authors would like to thank Pietro Barbera and Luca Gianantonia for their contributions during the early phases of this project. Additionally, the authors are grateful for the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology for providing computational resources, specifically the NVIDIA L40 GPUs on the Phoenix cluster, which were instrumental for the experiments presented in this paper.

References

[1] Cardosi, K., Falzarano, P., and Han, S., "Pilot–Controller Communication Errors: An Analysis of Aviation Safety Reporting System (ASRS) Reports," Tech. Rep. DOT/FAA/AR-98/17, Federal Aviation Administration, Office of Aviation Research, 1998. URL https://rosap.ntl.bts.gov/view/dot/8490.

[2] "Aircraft Noise: Better Information Sharing Could Improve Responses to Washington, D.C. Area Helicopter Noise Concerns," Tech. Rep. GAO-21-200, U.S. Government Accountability Office, Jan. 2021. URL https://www.gao.gov/assets/gao-21-200.pdf, quantifies civil–military traffic mix near Reagan National Airport.

[3] Zuluaga-Gomez, J. P., Motlicek, P., Zhan, Q., Veselý, K., and Braun, R., "Automatic Speech Recognition Benchmark for Air-Traffic Communications," *Proceedings of INTERSPEECH 2020*, 2020, pp. 2297–2301. https://doi.org/10.21437/Interspeech.2020-2173, URL https://www.isca-archive.org/interspeech_2020/zuluagagomez20_interspeech.pdf.

[4] Helmke, H., Ohneiser, O., Buxbaum, J., and Kern, C., "Increasing ATM Efficiency with Assistant-Based Speech Recognition," *12th USA/Europe Air Traffic Management Research & Development Seminar (ATM 2017)*, Seattle, WA, 2017. URL https://www.malorca-project.de/wp/wp-content/uploads/HelmkeOhneiserBuxbaumKern_ATMS2017-eingereicht.pdf.

[5] Zuluaga-Gomez, J. P., Veselý, K., Szöke, I., Blatt, A., Motlicek, P., Kocour, M., Rigault, M., Choukri, K., Prasad, A., Sarfjoo, S. S., Nigmatulina, I., Cevenini, C., Kolčárek, P., Tart, A., Černocký, J., and Klakow, D., "ATCO2 corpus: A Large-Scale Dataset for Research on Automatic Speech Recognition and Natural Language Understanding of Air Traffic Control Communications," *arXiv preprint*, Vol. arXiv:2211.04054, 2023. URL https://arxiv.org/abs/2211.04054, 5,281 h raw audio; 4 h manually transcribed; 1 h open-source test set.

[6] van Doorn, R., Sun, J., Hoekstra, J. M., Jonk, P., and de Vries, V., "Whisper-ATC: Open Models for Air Traffic Control Automatic Speech Recognition with Accuracy," *Proceedings of the International Conference on Research in Air Transportation (ICRAT)*, 2024, pp. 1–8. URL https://pure.tudelft.nl/ws/portalfiles/portal/218298256/ICRAT2024_paper_83.pdf.

[7] Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., and Zweig, G., "Achieving Human Parity in Conversational Speech Recognition," arXiv:1610.05256, 2017. URL https://arxiv.org/abs/1610.05256.

[8] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I., "Robust Speech Recognition via Large-Scale Weak Supervision," Tech. rep., OpenAI, 2022. URL https://arxiv.org/abs/2212.04356.

[9] Ko, T., Peddinti, V., Povey, D., Seltzer, M. L., and Khudanpur, S., "Audio Augmentation for Speech Recognition," *Proc. Interspeech*, 2015, pp. 3586–3589. https://doi.org/10.21437/Interspeech.2015-711.

[10] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y., "Generative Adversarial Networks," , 2014. URL https://arxiv.org/abs/1406.2661.

[11] Bińkowski, M., Donahue, J., Dieleman, S., Clark, A., Elsen, E., Casagrande, N., Cobo, L. C., and Simonyan, K., "High Fidelity Speech Synthesis with Adversarial Networks," *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020. https://doi.org/10.48550/arXiv.1909.11646, URL https://arxiv.org/abs/1909.11646, originally posted as arXiv:1909.11646 [v1] on 25 Sep 2019; oral presentation.

[12] Kumar, K., Kumar, R., de Boer, J., Sarma, L., Catre, V., Subramanian, T., Ronanki, S., and Gamper, H., "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis," arXiv:1910.06711, 2019. https://doi.org/10.48550/arXiv.1910.06711, URL https://arxiv.org/abs/1910.06711.

[13] Kong, J., Kim, J., and Bae, J., "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," arXiv:2010.05646, 2020. https://doi.org/10.48550/arXiv.2010.05646, URL https://arxiv.org/abs/2010.05646.

[14] Hilburn, B., "Cognitive Complexity in Air Traffic Control – A Literature Review," Tech. Rep. EEC Note 04/04, EUROCONTROL Experimental Centre, 2004. Identifies factors contributing to controller workload and complexity, e.g. traffic mix, military operations, radio congestion.

[15] Federal Aviation Administration, "Southern California TRACON (SCT) Facility Overview," https://www.faa.gov/about/office_org/headquarters_offices/ato/service_units/air_traffic_services/tracon/sct, 2024. Lists 2.2 million annual operations across 9,000 sq mi, making SCT one of the busiest approach-control facilities worldwide.

[16] United States Marine Corps, "MCAS Miramar Air Installations Compatible Use Zone (AICUZ) Study Update," Tech. rep., Department of the Navy, 2020. URL https://www.miramar.marines.mil/Portals/164/2020%20MCAS_Miramar_AICUZ_final%20version%201.pdf, section 3.2 highlights San Diego's "very complex airspace environment" arising from mixed civil–military operations.

[17] EUROCONTROL, "Blocked Transmissions / Undetected Simultaneous Transmissions (USiT)," https://skybrary.aero/articles/blocked-transmissions-undetected-simultaneous-transmissions-usit, 2023. Explains how simultaneous transmissions on congested ATC frequencies block or garble messages and raise safety risk.

[18] Federal Aviation Administration, *Aeronautical Information Manual: Official Guide to Basic Flight Information and ATC Procedures*, Federal Aviation Administration, 2024. URL https://www.faa.gov/Air_traffic/publications/media/AIM-Basic-w-Chg1-and-Chg2-dtd-3-21-24.pdf, see Ch. 4, §4-2-3 "Radio Technique" and §4-2-12 "Stuck Microphone" for guidance on minimising cockpit noise and avoiding blocked transmissions.

[19] Tiewtrakul, T., and Fletcher, S. R., "The Challenge of Regional Accents for Aviation English Language Proficiency Standards: A Study of Difficulties in Understanding in Air Traffic Control–Pilot Communications," *Ergonomics*, Vol. 53, No. 2, 2010, pp. 229–239. https://doi.org/10.1080/00140130903470033.

[20] Tajima, A., "Fatal Miscommunication: English in Aviation Safety," *World Englishes*, Vol. 23, No. 3, 2004, pp. 451–470. https://doi.org/10.1111/j.0883-2919.2004.00368.x.

[21] Molesworth, B. R. C., and Estival, D., "Miscommunication in General Aviation: The Influence of External Factors on Communication Errors," *Safety Science*, Vol. 73, 2015, pp. 73–79. https://doi.org/10.1016/j.ssci.2014.11.004.

[22] Dirección General de Aviación Civil (Spain), "Aircraft Accident Digest Circular 153-AN/56: Boeing 747 PH-BUF and Boeing 747 N736PA Collision at Tenerife Airport, Spain, 27 March 1977," Tech. rep., International Civil Aviation Organization, 1978. URL https://www.faa.gov/sites/faa.gov/files/2022-11/Spanish_Findings_0.pdf, official Annex-13 report; see pp. 22–68 for phraseology and read-back findings.

[23] National Transportation Safety Board, "Aircraft Accident Report: Avianca, The Airline of Colombia, Flight 052, Boeing 707-321B, Fuel Exhaustion, Cove Neck, New York, 25 January 1990," Tech. Rep. NTSB/AAR-91/04, National Transportation Safety Board, 1991. URL https://www.ntsb.gov/investigations/AccidentReports/Reports/AAR9104.pdf, chapter 2.2 discusses indirect fuel-emergency phrasing and cultural factors.

[24] Huang, X., Baker, J. K., and Reddy, R., "A Historical Perspective of Speech Recognition," *Communications of the ACM*, Vol. 57, No. 1, 2014, pp. 94–103. https://doi.org/10.1145/2500887.

[25] Rabiner, L. R., "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, Vol. 77, No. 2, 1989, pp. 257–286. https://doi.org/10.1109/5.18626.

[26] Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J., "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," *Proceedings of the 23rd International Conference on Machine Learning (ICML 2006)*, 2006, pp. 369–376. https://doi.org/10.1145/1143844.1143891.

[27] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., and Kingsbury, B., "Deep Neural Networks for Acoustic Modeling in Speech Recognition: Shared Views of Four Research Groups," *IEEE Signal Processing Magazine*, Vol. 29, No. 6, 2012, pp. 82–97. https://doi.org/10.1109/MSP.2012.2205597.

[28] Mohamed, A.-r., Dahl, G. E., and Hinton, G., "Acoustic Modeling Using Deep Belief Networks," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 20, No. 1, 2012, pp. 14–22. https://doi.org/10.1109/TASL.2011.2109382, URL https://ieeexplore.ieee.org/document/6120347.

[29] Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., and *et al.*, "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin," *Proceedings of the 33rd International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 48, 2016, pp. 173–182. URL https://proceedings.mlr.press/v48/amodei16.html.

[30] Chan, W., Jaitly, N., Le, Q. V., and Vinyals, O., "Listen, Attend and Spell," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960–4964. https://doi.org/10.1109/ICASSP.2016.7472621, URL https://ieeexplore.ieee.org/document/7472621.

[31] Graves, A., "Sequence Transduction with Recurrent Neural Networks," ICML Workshop on Representation Learning, 2012. URL https://arxiv.org/abs/1211.3711.

[32] Bahdanau, D., Cho, K., and Bengio, Y., "Neural Machine Translation by Jointly Learning to Align and Translate," *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015. https://doi.org/10.48550/arXiv.1409.0473, URL https://arxiv.org/abs/1409.0473, originally posted on arXiv in 2014.

[33] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I., "Attention Is All You Need," *Advances in Neural Information Processing Systems (NeurIPS) 30*, 2017, pp. 5998–6008. https://doi.org/10.48550/arXiv.1706.03762, URL https://arxiv.org/abs/1706.03762.

[34] Gulati, A., Qin, J., Chiu, C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., and Pang, R., "Conformer: Convolution-augmented Transformer for Speech Recognition," *Proceedings of INTERSPEECH 2020*, 2020, pp. 5036–5040. https://doi.org/10.21437/Interspeech.2020-3015, URL https://www.isca-archive.org/interspeech_2020/gulati20_interspeech.html.

[35] Baevski, A., Zhou, H., Mohamed, A., and Auli, M., "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," *Advances in Neural Information Processing Systems (NeurIPS) 33*, 2020, pp. 12449–12460. https://doi.org/10.48550/arXiv.2006.11477, URL https://arxiv.org/abs/2006.11477.

[36] Loizou, P. C., *Speech Enhancement: Theory and Practice*, 2nd ed., CRC Press, Boca Raton, FL, 2013.

[37] Pascual, S., Bonafonte, A., and Serrà, J., "SEGAN: Speech Enhancement Generative Adversarial Network," *Proceedings of INTERSPEECH 2017*, 2017, pp. 3642–3646. https://doi.org/10.21437/Interspeech.2017-1428, URL https://www.isca-archive.org/interspeech_2017/pascual17_interspeech.pdf.

[38] Valin, J., "A Hybrid DSP/Deep Learning Approach to Real-Time Full-Band Speech Enhancement," *IEEE International Workshop on Multimedia Signal Processing (MMSP 2018)*, 2018, pp. 1–6. https://doi.org/https://doi.org/10.48550/arXiv.1709.08243, URL https://arxiv.org/abs/1709.08243.

[39] Hu, Y., Liu, Y., Lv, S., Xing, J., Qian, Y., Meng, G., Xu, B., and Gong, Y., "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," *Proceedings of INTERSPEECH 2020*, 2020, pp. 2472–2476. URL https://www.isca-archive.org/interspeech_2020/hu20g_interspeech.pdf.

[40] Défossez, A., Lamothe, G., Sun, W., and Zhang, R., "Real-Time Speech Enhancement in the Waveform Domain," arXiv:2006.12847, 2020. URL https://arxiv.org/abs/2006.12847.

[41] Chen, Z., Wang, W., Zhang, Y., and Pan, H., "Improved Generative Adversarial Network Method for Flight Crew Dialog Speech Enhancement," *Journal of Aerospace Information Systems*, Vol. 20, No. 9, 2023, pp. 645–655. https://doi.org/10.2514/1.I011168.

[42] Liang, H., Zhao, M., Yang, X., and Liu, F., "DeCGAN: Speech Enhancement Algorithm for Air Traffic Control," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2025)*, 2025, pp. XXX–XXX. https://doi.org/10.3390/a18050245, URL https://www.mdpi.com/1999-4893/18/5/245, to appear.

[43] Yu, M., Li, X., Zhang, R., and Gao, P., "ROSE: A Recognition-Oriented Speech Enhancement Framework in Air Traffic Control Using Multi-Objective Learning," arXiv:2312.06118, 2024. URL https://arxiv.org/abs/2312.06118.

[44] Jiang, T., Xu, C., Peng, J., and Li, H., "A Noise-Robust U2 Scheme for Automatic Speech Recognition of Air Traffic Voice Communication," *Integrated Communications, Navigation and Surveillance (ICNS) Conference 2025*, 2025, pp. XXX–XXX. URL https://www.researchgate.net/publication/391497241.

[45] Shorten, C., and Khoshgoftaar, T. M., "A Survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, Vol. 6, No. 1, 2019, p. 60. https://doi.org/10.1186/s40537-019-0197-0, URL https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0197-0, article 60.

[46] Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V., "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," *Proc. Interspeech*, 2019, pp. 2613–2617. https://doi.org/https://doi.org/10.48550/arXiv.1904.08779.

[47] Snyder, D., Chen, G., and Povey, D., "MUSAN: A Music, Speech, and Noise Corpus," *Proc. Interspeech*, 2015, pp. 3278–3282. https://doi.org/10.48550/arXiv.1510.08484.

[48] Hsu, W.-N., Zhang, Y., and Glass, J., "Unsupervised Domain Adaptation for Robust Speech Recognition via Variational Autoencoder-Based Data Augmentation," , 2017. https://doi.org/https://doi.org/10.48550/arXiv.1707.06265, URL https://arxiv.org/abs/1707.06265.

[49] Lemercier, J.-M., Richter, J., Welker, S., and Gerkmann, T., "Analysing Diffusion-based Generative Approaches versus Discriminative Approaches for Speech Restoration," , 2023. https://doi.org/10.48550/arXiv.2211.02397, URL https://arxiv.org/abs/2211.02397.

[50] Hu, H., Tan, T., and Qian, Y., "Generative Adversarial Networks Based Data Augmentation for Noise-Robust Speech Recognition," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 5044–5048. https://doi.org/10.1109/ICASSP.2018.8462624.

[51] Chen, C., Hou, N., Hu, Y., Shirol, S., and Chng, E. S., "Noise-Robust Speech Recognition With 10 Minutes Unparalleled In-Domain Data," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 4298–4302. https://doi.org/10.1109/ICASSP43922.2022.9747755.

[52] Radford, A., Metz, L., and Chintala, S., "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," arXiv preprint arXiv:1511.06434, 2015. URL https://arxiv.org/abs/1511.06434.

[53] Arjovsky, M., Chintala, S., and Bottou, L., "Wasserstein GAN," *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017, pp. 214–223. URL https://proceedings.mlr.press/v70/arjovsky17a.html.

[54] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A., "Improved Training of Wasserstein GANs," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5767–5777. URL https://arxiv.org/abs/1704.00028.

[55] Shah, N., Patil, H. A., and Soni, M. H., "Time-Frequency Mask-based Speech Enhancement using Convolutional Generative Adversarial Network," *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 1246–1251. https://doi.org/10.23919/APSIPA.2018.8659692.

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

[57] Patrikar, J., Dantas, J., Moon, B., Hamidi, M., Ghosh, S., Keetha, N., Higgins, I., Chandak, A., Yoneyama, T., and Scherer, S., "TartanAviation: Image, Speech, and ADS-B Trajectory Datasets for Terminal Airspace Operations," *Scientific Data*, Vol. 12, 2025, p. 468. https://doi.org/10.1038/s41597-025-04775-6, URL https://www.nature.com/articles/s41597-025-04775-6, see also arXiv:2403.03372 for preprint.

[58] Wiseman, J., "py–webrtcvad: Python Interface to the WebRTC Voice Activity Detector," https://github.com/wiseman/py-webrtcvad, 2020. Accessed 02 May 2025.

[59] Loshchilov, I., and Hutter, F., "SGDR: Stochastic Gradient Descent with Warm Restarts," *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017. https://doi.org/10.48550/arXiv.1608.03983, URL https://arxiv.org/abs/1608.03983, arXiv:1608.03983.

[60] Biewald, L., "Experiment Tracking with Weights and Biases," , 2020. URL https://www.wandb.com/, software available from wandb.com.

[61] Yamamoto, R., Song, E., and Kim, J.-M., "Parallel WaveGAN: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6199–6203. https://doi.org/10.1109/ICASSP40776.2020.9053795.

[62] Kilgour, K., Zuluaga, M., Roblek, D., and Sharifi, M., "Fréchet Audio Distance: A Reference-free Metric for Evaluating Music Enhancement Algorithms," *Proceedings of INTERSPEECH 2019*, 2019, pp. 2350–2354. https://doi.org/10.21437/Interspeech.2019-2219, URL https://www.isca-archive.org/interspeech_2019/kilgour19_interspeech.pdf.

[63] Tang, H., Liu, H., Xu, D., Torr, P. H. S., and Sebe, N., "AttentionGAN: Unpaired Image-to-Image Translation using Attention-Guided Generative Adversarial Networks," , 2021. https://doi.org/https://doi.org/10.48550/arXiv.1911.11897, URL https://arxiv.org/abs/1911.11897.

[64] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A., "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251. https://doi.org/10.1109/ICCV.2017.244.