FH-RestoreASR: Frequency-Hopping Robust Air Traffic Control Speech Restoration and Recognition

Youngeun Kwon[1,2], Yeri Byun[1,2*], Hyunsung Cho[1,3*], and Jongwon Choi[1,4†]
[1] Defense AI Education College, Chung-Ang University, Seoul, Korea
[2] Yonsei University, Seoul, Korea
[3] Ajou University, Suwon, Korea
[4] Dept. of Advanced Imaging, GSAIM, Chung-Ang University, Seoul, Korea
E-mail: youngeunk@yonsei.ac.kr, yeari0206@yonsei.ac.kr, cho100c@ajou.ac.kr, choijw@cau.ac.kr

***Abstract*—We present FH-RestoreASR, a dropout-aware speech restoration and recognition framework designed for communication systems experiencing short-duration audio dropouts. Such dropouts are common in frequency-hopping systems, widely used for jamming resilience, but they also introduce brief interruptions. These interruptions pose significant challenges for Automatic Speech Recognition (ASR) systems, which typically assume temporally continuous input. FH-RestoreASR introduces a dropout-specific masking mechanism that enables targeted restoration by simulating realistic dropout patterns during training. Built on GAN based architecture, the system restores degraded speech and integrates with ASR models to improve transcription robustness. Experiments on simulated dropout scenarios demonstrate substantial improvements in Word Error Rate (WER) and perceptual quality compared to existing methods. These results highlight the effectiveness of FH-RestoreASR in achieving robust speech recognition under challenging and intermittently disrupted audio conditions.**

I. Introduction

Air Traffic Control (ATC) communications require high reliability and real-time performance to ensure airspace safety [1], but often operate under adverse wireless conditions with signal degradation or interference [2]. In particular, military ATC environments face more severe disruptions such as frequent short-duration audio dropouts, which complicate reliable voice-based operations [3].

In ATC communication, Frequency Hopping Spread Spectrum (FHSS) is widely used for jamming resilience, but its rapid frequency switching introduces short dropouts [4]. These interruptions degrade speech intelligibility and present significant challenges for Automatic Speech Recognition (ASR) systems, which are typically optimized for continuous and uninterrupted speech input.

Existing speech restoration methods mainly address stationary noise and uninterrupted speech, but are insufficient for structured, bursty dropouts common in spectrum-hopping scenarios [5]. These methods struggle under realistic dropout patterns, indicating the need for dropout-aware modeling that explicitly targets missing regions during training and inference.

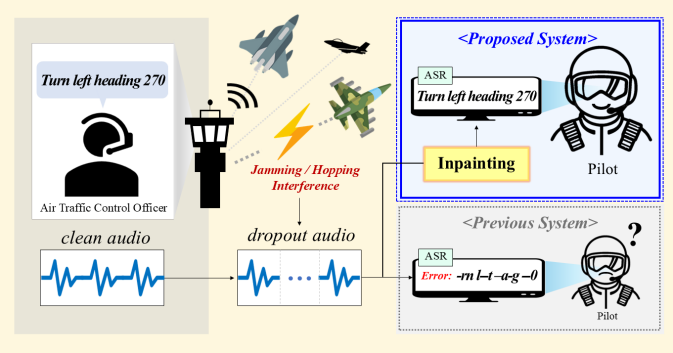


Fig. 1. System overview of FH-RestoreASR under frequency hopping environments. While previous systems produce error-prone transcripts under dropout conditions, our model reconstructs missing speech and integrates with ASR to deliver clear and reliable communication, reducing pilot confusion and enhancing situational awareness.

To tackle the limitations of conventional restoration methods under structured dropout conditions, we present FH-RestoreASR - a dropout-aware speech restoration framework designed to enhance downstream ASR performance. The proposed restoration network builds upon a generative architecture equipped with a dropout-specific masking mechanism, enabling the model to focus learning on realistically missing segments. The enhanced audio output is subsequently fed into ASR models (e.g., Wav2Vec2 [6], Whisper [7]) to yield readable transcripts. This joint design allows for end-to-end improvement in both perceptual speech quality and transcription accuracy under adverse communication scenarios.

The contributions of this work are summarized as follows:

- A dropout-aware speech restoration framework enhancing ASR performance under lossy communication conditions.
- A masking strategy that simulates structured dropouts during training, enabling the models to focus learning on corrupted regions.
- Seamless integration with pretrained ASR models, demonstrating restoration–recognition synergy across diverse dropout scenarios.

* Equally contributed authors.
† Corresponding author.

979-8-3315-7206-8/25/$31.00 ©2025 IEEE

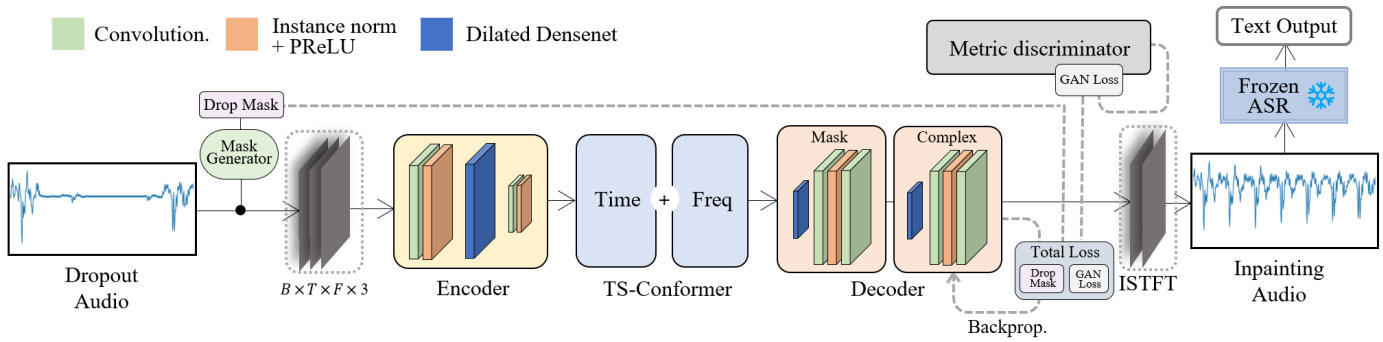


Fig. 2. Block diagram of the proposed FH-RestoreASR framework. Dropout audio is processed by a TS-Conformer-based inpainting architecture, utilizing dropout mask supervision and adversarial loss to reconstruct missing regions. The enhanced audio is then transcribed into text using a pre-trained ASR module.

II. Related Works

A. Speech Enhancement and Restoration

Speech enhancement aims to clarify noisy speech. Early methods like spectral subtraction showed limited robustness, while deep learning approaches achieved major advance [8], [9]. In particular, CMGAN integrates CNNs and transformers to capture both local and global dependencies [10].

Speech restoration focuses on reconstructing missing segments. Early interpolation estimated gaps from context, later improved by autoencoders [11]. GAN-based approaches enabled perceptually natural outputs, while transformer models captured long-range dependencies with attention.

However, most speech restoration methods have focused on continuous speech, which do not reflect frequent dropouts commonly encountered in ATC communication. While CMGAN achieves strong performance for generic denoising tasks, it is not explicitly optimized for bursty dropouts, which can result in residual artifacts or over-suppression. To address this limitation, we adopt a dropout-aware extension designed specifically for FHSS-induced dropouts.

B. Automatic Speech Recognition (ASR)

Early ASR systems relied on Connectionist Temporal Classification (CTC) for speech-text alignment, later improved by Seq2Seq [12] for better context and streaming. Recent models like Wav2Vec2 and Whisper advance ASR by learning from unlabeled data and supporting multilingual, multitask recognition, making them suitable for ATC communications [6], [7].

ATC-specific ASR research has explored contextual knowledge such as callsign identification and domain adaptation [13], [14]. Additionally, enhancing speech quality without modifying the underlying ASR model has been shown to improve recognition performance [15]. Nonetheless, most existing studies assume continuous speech and do not sufficiently address dropout conditions, where speech signals are intermittently disrupted due to factors such as frequency hopping or jamming. This study addresses this limitation by applying dropout-aware speech restoration as a front-end strategy to improve the robustness of automatic speech recognition under such conditions.

III. Methodology

Fig. 2 shows FH-RestoreASR, which is a dropout-aware speech restoration and recognition framework for short-duration audio dropouts in ATC. Built on CMGAN, it incorporates dropout mask generation and dropout-focused training [10]. Our model adopts a GAN structure, where the generator, which is called *TSCNet*, restores corrupted speech guided by dropout masks, while the discriminator evaluates the perceptual quality. The audio is then transcribed by pre-trained ASR models.

A. Dropout Mask Generation

Dropout masks are created by comparing clean and degraded spectrogram magnitudes. For each time-frequency bin (t, f), a bin is marked as dropped if the degraded magnitude $|\hat{X}_{t,f}|$ is lower than the clean magnitude $|X_{t,f}|$ by a threshold τ:

$$M_{t,f} = \begin{cases} 1, & \text{if } \frac{|\hat{X}_{t,f}|}{|X_{t,f}|+\epsilon} < \tau \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where $\tau = 0.6$ and $\epsilon = 10^{-8}$ ensures numerical stability.

To emphasize highly corrupted regions, mask values are averaged over time for each frequency bin f. A bin is marked as high-confidence if over 95% of its frames are dropped:

$$\tilde{M}_f = \begin{cases} 1, & \text{if } \frac{1}{T}\sum_{t=1}^{T} M_{t,f} > 0.95 \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where T is the number of time frames and F is the number of frequency bins.

B. Training Data Preparation

Pairs of clean and degraded waveforms were generated from the ATCOSIM dataset, each fixed to 2 seconds (32,000 samples). Spectrograms were obtained using Short-Time Fourier Transform (STFT), with real, imaginary, and magnitude components:

$$|\hat{X}_{t,f}| = \sqrt{\hat{R}_{t,f}^2 + \hat{I}_{t,f}^2}, \tag{3}$$

where $\hat{R}_{t,f}$ and $\hat{I}_{t,f}$ denote the real and imaginary parts. During training, dropout masks weight the loss to emphasize corrupted regions, while inference is performed without masks.

C. Network Architecture

FH-RestoreASR adopts a GAN architecture with a generator and discriminator. The generator is based on CMGAN's TSCNet, consisting of a DenseEncoder, four Time-Frequency Sequential Conformer Blocks (TSCBs), and two decoders [10]. The DenseEncoder uses dilated convolutions to capture long-range features, and TSCBs apply Conformer attention along time and frequency to model both local and global speech patterns. The Mask Decoder predicts a magnitude mask for spectrogram enhancement, while the Complex Decoder refines phase reconstruction by estimating residual real and imaginary components. Unlike the network-generated mask, our dropout mask is externally derived from simulated dropouts and used only as a loss weight. This dual-decoder GAN design with dropout-aware weighting effectively restores bursty dropouts, reducing artifacts common in generic denoising models. Enhanced waveforms are recovered via inverse STFT, while the discriminator enforces perceptual quality.

D. Training Strategy

To improve restoration in dropout-affected regions, we combine a dropout-aware loss with adversarial learning. The RI and magnitude losses adopt dropout-weighted mean squared error:

$$\mathcal{L}_{\text{RI}} = \frac{1}{TF}\sum_{t=1}^{T}\sum_{f=1}^{F}\big(1+\alpha\cdot M_{t,f}\big) \times \big(|\hat{R}_{t,f}-R_{t,f}|^2+|\hat{I}_{t,f}-I_{t,f}|^2\big), \tag{4}$$

where $t = 1,\ldots,T$ and $f = 1,\ldots,F$ index time-frequency bins; $\alpha = 0.5$ emphasizes corrupted regions.

$$\mathcal{L}_{\text{Mag}} = \frac{1}{TF}\sum_{t=1}^{T}\sum_{f=1}^{F}\big(1+\alpha\cdot M_{t,f}\big)|\hat{M}_{t,f}-M_{t,f}|^2. \tag{5}$$

Here, $M_{t,f}$ is a binary dropout mask generated from simulated dropouts and contains no learnable parameters. Power-law compression is applied to magnitude spectra to reflect perceptual scales.

The time-domain loss compares enhanced and clean waveforms:

$$\mathcal{L}_{\text{Time}} = \frac{1}{N}\sum_{i=1}^{N}(1+\beta\cdot\delta)|\hat{x}_i - x_i|, \tag{6}$$

where $i = 1,\ldots,N$ indexes time-domain samples; $\beta = 50.0$, and δ is a fixed scaling factor set to either 0.001 or 0.003 depending on dropout severity. In practice, the product $\beta\cdot\delta$ acts as a single weighting constant. The adversarial loss $\mathcal{L}_{\text{GAN}}$ employs a MetricGAN-based discriminator that predicts PESQ scores. The final loss combines all terms:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{RI}}\mathcal{L}_{\text{RI}} + \lambda_{\text{Mag}}\mathcal{L}_{\text{Mag}} + \lambda_{\text{Time}}\mathcal{L}_{\text{Time}} + \lambda_{\text{GAN}}\mathcal{L}_{\text{GAN}}. \tag{7}$$

Dropout masks are applied only during training to guide the model's learning, while inference proceeds without masks and relies on the model's generalization capability.

TABLE I
DROPOUT SIMULATION SCENARIOS TAILORED FOR MILITARY ATC, CATEGORIZED BY DOMAIN TYPE AND DIFFICULTY LEVEL (FHSET: S2+S3+S6+S7).

Domain Type	**Easy**		**Hard**	
	E1(Out)	**E2(In)**	**H1(In)**	**H2(Out)**
Dropout	**20%**	**30%**	**40%**	**50%**
3∼5 (ms)	S1	S2	S3	S4
8∼10 (ms)	S5	S6	S7	S8

E. Inference

During inference, the generator reconstructs waveforms from noisy spectrograms using inverse STFT, without dropout masks. These enhanced waveforms are transcribed by pre-trained ASR models such as Wav2Vec2 [6] and Whisper [7].

IV. EXPERIMENTS

A. Implementation Details

FH-RestoreASR was trained with the Adam optimizer (initial learning rate 5×10^{-3}, halved every 5 epochs) and a batch size of 2. Early stopping was applied after 10 epochs without validation improvement. Training ran up to 30 epochs with checkpoints saved each time; the 25th epoch gave the best validation performance and was used for reporting.

B. Dataset Simulation

We used the ATCOSIM dataset to simulate military ATC speech dropouts. Since it lacks frequency-hopping or jamming effects, short interruptions were introduced by muting waveform segments to emulate a 300 hops/s FHSS rate [16]. Additional bursty dropouts represent multipath fading or deliberate jamming. Although simplified, this setup provides a controlled testbed for structured speech loss. As shown in Table I, eight datasets were created by varying dropout duration and rate, categorized by domain type (in-/out-domain) and difficulty level (easy/hard).

C. Evaluation Setup

Word Error Rate (WER) was computed against ATCOSIM transcripts. We compared performance before and after restoration, analyzing WER distributions across models and dropout levels. Training used FHset for in-/out-domain evaluation and S8 for testing robustness under extreme dropout conditions.

D. Automatic Speech Recognition

FH-RestoreASR was evaluated on the ATCOSIM dataset using Wav2Vec2 and Whisper ASR models. WER was the primary metric, with tests covering both in-domain and out-domain scenarios under varying dropout severities. For comparison, baseline models including MetricGAN+ and CMGAN were used to validate the effectiveness of dropout-specific masking and domain-adapted training.

Table II shows that FH-RestoreASR consistently outperformed baselines in both easy and hard conditions. Particularly with Whisper, it achieved the lowest WERs even under severe

TABLE II
ASR PERFORMANCE ON IN-DOMAIN AND OUT-DOMAIN TEST SETS

Model	Trained	Test	WER (%) ▼	
			Wav2Vec2	Whisper
In-domain				
MetricGAN+	VCTK	E2	88.61	77.87
CMGAN	VCTK		87.22	68.73
CMGAN	FHset		76.89	57.88
Ours	**FHset**		**75.57**	**54.62**
MetricGAN+	VCTK	H2	91.54	83.04
CMGAN	VCTK		90.79	74.04
CMGAN	FHset		78.74	62.28
Ours	**FHset**		**77.19**	**57.71**
Out-domain				
MetricGAN+	VCTK	E1	84.75	71.32
CMGAN	VCTK		82.30	61.65
CMGAN	FHset		75.79	56.05
Ours	**FHset**		**74.15**	**52.68**
MetricGAN+	VCTK	H1	93.81	86.84
CMGAN	VCTK		92.98	82.16
CMGAN	FHset		81.24	63.32
Ours	**FHset**		**79.75**	**60.87**

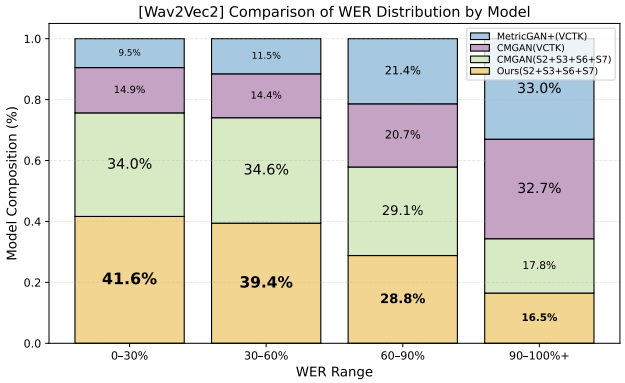
TABLE III
ASR PERFORMANCE ON S1, S5, AND S8 TEST SETS WITH S8-TRAINED MODELS(EXTREME DROPOUT TRAINING)

ASR	Model	Trained	WER (%) ▼		
			S1	S5	S8
Wav2Vec2	Raw	–	80.25	94.12	99.99
	MetricGAN+	VCTK	82.60	88.62	98.80
	CMGAN	VCTK	79.58	87.43	99.19
	CMGAN	S8	78.94	81.90	94.49
	Ours	**S8**	**74.42**	**78.20**	**87.33**
Whisper	Raw	–	60.10	65.33	97.41
	MetricGAN+	VCTK	68.93	75.64	97.97
	CMGAN	VCTK	61.43	67.90	97.00
	CMGAN	S8	66.31	60.62	70.91
	Ours	**S8**	**52.49**	**56.43**	**66.36**

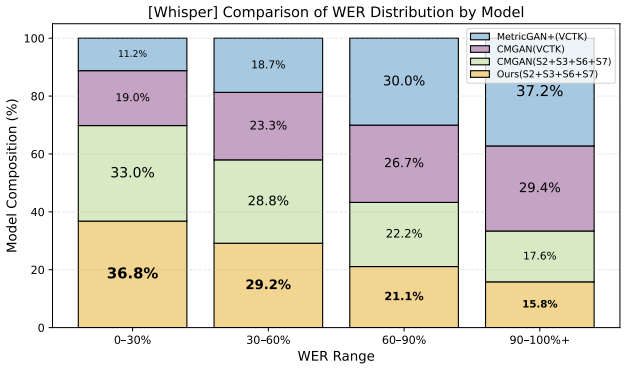
dropout, demonstrating its robustness. Wav2Vec2 also showed improvements over baselines, though overall WER remained higher than Whisper. Whisper's robustness likely stems from training on large, diverse datasets. These results confirm the benefit of combining ATC-specific training with dropout-aware restoration.

For detailed evaluation, three representative test sets (S1, S5, S8) were selected to reflect different dropout lengths and severities as defined in Table I. As shown in Table III, FH-RestoreASR consistently outperformed all baselines, with the strongest improvements under extreme dropout (S8). MetricGAN+ even degraded ASR by introducing artifacts, while CMGAN trained on VCTK showed limited gains, particularly with Whisper. Notably, in S1 with Whisper, only FH-RestoreASR improved over raw input, indicating that targeted restoration preserves intact speech and synergizes with robust ASR models like Whisper.

Fig. 3 illustrates WER distribution across models and ASR backends. FH-RestoreASR showed a higher proportion of utterances in lower WER ranges (0–30%) and fewer in higher WER ranges (90–100%+) compared to baselines, with both



(a) WER distribution with Wav2Vec2



(b) WER distribution with Whisper

Fig. 3. Distribution of WER ranges across different speech restoration models using (a) Wav2Vec2 and (b) Whisper ASR backends. Each bar represents the model composition percentage within specified WER intervals (0–30%, 30–60%, 60–90%, and 90–100%+). Ours consistently shows higher proportions in the lower WER ranges and reduced proportions in the higher WER ranges compared to baseline models.

Wav2Vec2 and Whisper. This confirms the effectiveness of dropout-specific restoration in improving ASR accuracy and reliability. Unlike MetricGAN+ and CMGAN, which showed balanced or skewed distributions toward higher WERs, FH-RestoreASR shifted results toward lower error regions, demonstrating robustness in handling dropout-affected speech.

E. Speech Restoration

The performance of FH-RestoreASR was evaluated using six objective metrics: PESQ, STOI, CSIG, CBAK, COVL, and SSNR [17]–[19], which assess perceptual quality, intelligibility, and signal restoration. Evaluations included in-/out-domain conditions and S1, S5, and S8 test sets with extreme dropouts.

Tables IV–V show that FH-RestoreASR consistently outperformed baselines, confirming superiority in perceptual quality and intelligibility and leading to fewer ASR errors across mild and severe dropout conditions. While SSNR scores on S1 and S5 were slightly lower, this mainly reflected signal energy similarities rather than perceptual loss. High PESQ and STOI scores further emphasize the model's fidelity to perceptual quality, validating its robustness for downstream ASR under challenging dropout scenarios.

TABLE IV
SPEECH RESTORATION PERFORMANCE ON IN-DOMAIN AND OUT-DOMAIN TEST SETS

Domain	Model	Trained	Test	PESQ ▲	STOI ▲	CSIG ▲	CBAK ▲	COVL ▲	SSNR ▲
In-domain	MetricGAN+	VCTK	E2	1.16	0.74	2.90	2.18	2.00	3.61
	CMGAN	VCTK		1.18	0.79	3.39	2.51	2.31	6.20
	CMGAN	FHset		2.31	0.91	4.28	2.98	3.33	4.90
	Ours	**FHset**		**3.05**	**0.94**	**4.76**	**3.89**	**4.02**	**12.84**
	MetricGAN+	VCTK	H2	1.12	0.70	2.82	2.08	1.93	2.90
	CMGAN	VCTK		1.14	0.75	3.32	2.38	2.25	4.74
	CMGAN	FHset		2.05	0.90	4.10	2.82	3.11	4.36
	Ours	**FHset**		**2.79**	**0.93**	**4.61**	**3.64**	**3.79**	**10.80**
Out-domain	MetricGAN+	VCTK	E1	1.23	0.78	3.00	2.33	2.10	4.62
	CMGAN	VCTK		1.25	0.83	3.47	2.72	2.40	8.47
	CMGAN	FHset		2.59	0.92	4.48	3.17	3.58	5.55
	Ours	**FHset**		**3.34**	**0.96**	**4.90**	**4.21**	**4.26**	**15.53**
	MetricGAN+	VCTK	H1	1.10	0.66	2.76	2.00	1.87	2.39
	CMGAN	VCTK		1.11	0.70	3.26	2.28	2.20	3.73
	CMGAN	FHset		1.84	0.88	3.95	2.68	2.92	3.86
	Ours	**FHset**		**2.54**	**0.91**	**4.46**	**3.42**	**3.57**	**9.24**

TABLE V
SPEECH RESTORATION PERFORMANCE ON S1, S5, AND S8 TEST SETS (EXTREME DROPOUT TRAINING)

Model	Trained	Test	PESQ ▲	STOI ▲	CSIG ▲	CBAK ▲	COVL ▲	SSNR ▲
MetricGAN+	VCTK	S1	1.29	0.82	3.05	2.42	2.17	5.23
CMGAN	VCTK		1.30	0.88	3.48	2.82	2.44	**9.54**
CMGAN	S8		1.37	0.85	3.49	2.43	2.46	3.04
Ours	**S8**		**2.38**	**0.92**	**4.39**	**3.06**	**3.43**	4.85
MetricGAN+	VCTK	S5	1.16	0.74	2.95	2.24	2.04	4.01
CMGAN	VCTK		1.20	0.79	3.46	2.61	2.36	**7.39**
CMGAN	S8		1.57	0.88	3.70	2.61	2.67	4.36
Ours	**S8**		**2.38**	**0.91**	**4.42**	**3.19**	**3.45**	6.95
MetricGAN+	VCTK	S8	1.05	0.58	2.72	1.92	1.82	1.77
CMGAN	VCTK		1.06	0.60	3.21	2.17	2.14	2.68
CMGAN	S8		1.27	0.77	3.53	2.31	2.42	2.70
Ours	**S8**		**1.68**	**0.82**	**3.93**	**2.74**	**2.85**	**5.28**

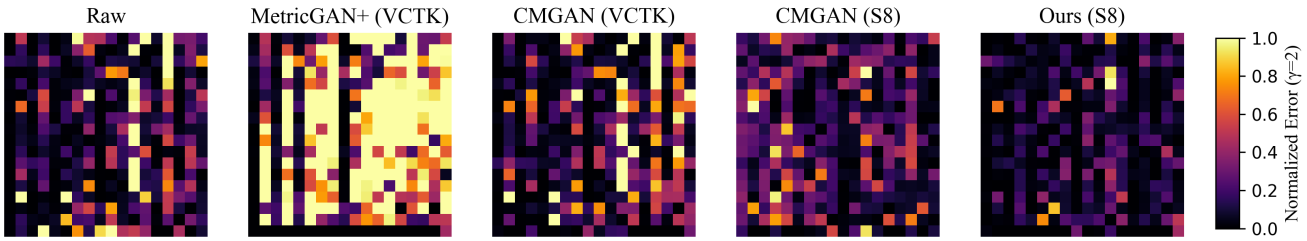


Fig. 4. Error heatmap comparison of various speech restoration models under a challenging dropout condition (S8). Bright regions indicate greater deviation from the original signal, while dark regions reflect better reconstruction. The proposed model shows the least residual error.

Fig. 4 shows spectrogram-based error heatmaps comparing original and restored audio under the most challenging condition (S8). Raw and baseline models displayed bright vertical streaks, indicating dropout-induced reconstruction errors. In contrast, FH-RestoreASR effectively suppressed these artifacts, producing the darkest heatmap with minimal residual error and demonstrating superior restoration of missing regions.

TABLE VI
TRAINING STABILITY RESULTS ON FHSET

Epoch	PESQ▲	STOI▲	CSIG▲	CBAK▲	COVL▲	SSNR▲
20	2.86	0.94	4.66	3.73	3.85	11.73
22	2.92	0.94	4.69	3.77	3.91	11.80
25	**2.93**	**0.94**	**4.69**	**3.77**	**3.91**	**11.83**
29	2.89	0.94	4.68	3.75	3.88	11.83

F. Training Stability

Model robustness was evaluated on FHset checkpoints at epochs 20, 22, 25, and 29. As shown in Table VI, performance was stable, with PESQ (2.86–2.93) and STOI (0.94) showing minimal variance. Despite the instability common in GANs, our training strategy achieved consistent convergence, and the 25th-epoch checkpoint was chosen as the final model.

V. Conclusion

We proposed FH-RestoreASR, a dropout-aware speech restoration and recognition framework for ATC scenarios with frequent dropouts. By combining dropout-specific masking with a CMGAN backbone, the model restores missing speech and integrates with ASR systems such as Wav2Vec2 and Whisper. Experiments showed consistent gains in WER and perceptual quality, confirming its effectiveness in improving communication reliability and situational awareness. Training with simulated dropout data further suggests applicability to other specialized conditions.

Future Work

As no publicly available FHSS ATC dataset exists, our evaluation relied on simulated dropouts. Future studies could incorporate real FHSS or jamming-affected recordings to strengthen practical validation. The proposed framework may also extend to other domains with bursty dropouts, including maritime VHF communications, tactical military links, and real-time teleconferencing.

Acknowledgment

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MND) (RS-2022-II220601, Military-Specialized AI Curriculum Establishment and Operation (Military AI Development and Management Program)).

References

[1] J. Zuluaga-Gomez, I. Nigmatulina, A. Prasad, *et al.*, "Lessons learned in atco2: 5000 hours of air traffic control communications for robust automatic speech recognition and understanding," *Aerospace*, vol. 10, no. 10, p. 898, 2023.

[2] S. Özmen, R. Hamzaoui, and F. Chen, "Survey of ip-based air-to-ground data link communication technologies," *Journal of Air Transport Management*, vol. 116, p. 102 579, 2024, ISSN: 0969-6997.

[3] R. Palacios and R. J. Hansman, "Short-term consequences of radio communications blackout on the u.s. national airspace system," *Aerospace Science and Technology*, vol. 29, no. 1, pp. 426–433, 2013.

[4] D. Torrieri, *Principles of Spread-Spectrum Communication Systems*, 3rd. New York, NY, USA: Springer, 2011.

[5] J. Choi, S. Lee, J. Choi, and H. Cho, "Speech inpainting based on multi-layer long short-term memory networks," *Future Internet*, vol. 16, no. 2, p. 63, 2024.

[6] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[7] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, 2023, pp. 28 492–28 518.

[8] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979. DOI: 10.1109/TASSP.1979.1163209.

[9] S.-W. Fu, C. Yu, T.-A. Hsieh, *et al.*, "Metricgan+: An improved version of metricgan for speech enhancement," in *Proceedings of Interspeech 2021*, 2021, pp. 3296–3300.

[10] R. Cao, S. Abdulatif, and B. Yang, "Cmgan: Conformer-based metric gan for speech enhancement," in *Proc. Interspeech*, 2022, pp. 936–940.

[11] W. Etter, "Restoration of a discrete-time signal segment by interpolation based on the left-sided and right-sided autoregressive parameters," *IEEE Transactions on Signal Processing*, vol. 44, no. 5, pp. 1124–1135, 1996.

[12] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. NIPS*, 2014.

[13] J. Zuluaga-Gomez, I. Nigmatulina, A. Prasad, *et al.*, "Contextual semi-supervised learning: An approach to leverage air-surveillance and untranscribed ATC data in ASR systems," in *Proceedings of Interspeech 2021*, 2021, pp. 3296–3300.

[14] J. Zuluaga-Gomez, A. Prasad, I. Nigmatulina, *et al.*, "How does pre-trained wav2vec 2.0 perform on domain-shifted asr? an extensive benchmark on air traffic control communications," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 205–212.

[15] X. Yu, D. Guo, J. Zhang, and Y. Lin, "Rose: A recognition-oriented speech enhancement framework in air traffic control using multi-objective learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3365–3378, 2024.

[16] B. Kaiser and J. Droll, "Saturn: Comparison of saturn and havequick," Collins Aerospace, Cedar Rapids, IA, USA, Tech. Rep., Mar. 2019, White Paper.

[17] R. ITU-T, "Perceptual evaluation of speech quality (pesq) : An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P.862*, 2001.

[18] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[19] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.