WhisperGate: Silence-Aware Gating for Hallucination-Free Speech Recognition with Frozen Whisper

Paras Sharma
mail2paras.s@gmail.com

Abstract

Large-scale speech recognition models such as Whisper produce fluent but entirely fabricated transcriptions when presented with silence or non-speech audio—a phenomenon known as *hallucination*. We introduce WhisperGate, a lightweight silence-aware gating module that sits between Whisper's frozen encoder and decoder, learning to classify each encoder frame as speech or non-speech with a two-layer MLP bottleneck requiring only **12,353 trainable parameters** (<0.02% of Whisper-Tiny). Trained in a two-stage pipeline—first as a standalone binary classifier, then plugged into inference—WhisperGate **eliminates 100% of hallucinations** on pure silence and white noise inputs while preserving clean-speech word error rate (WER). We evaluate three gating strategies (soft multiplicative, hard binary, and cross-attention bias) and find that attention-bias gating achieves the best trade-off: **0% hallucination with only 0.03% clean WER overhead** on Whisper-Tiny, while reducing gap-30 WER by 40.8% relative to hard gating. The method generalizes to Whisper-Small with similar hallucination elimination. We further compare against energy-based VAD preprocessing and show that WhisperGate uniquely handles both silence and white noise, whereas energy VAD fails on noise inputs. All experiments use LibriSpeech test-clean (2,620 samples) with controlled silence-gap injection at 5%, 15%, 30%, and multi-gap levels.

1 Introduction

OpenAI's Whisper [Radford et al., 2023] achieves remarkable speech recognition accuracy through large-scale weak supervision on 680,000 hours of web audio. However, Whisper exhibits a critical failure mode: when presented with silence, noise, or non-speech audio, it generates fluent but entirely fabricated transcriptions—so-called *hallucinations*. Koenecke et al. [2024] document that approximately 1% of Whisper transcriptions contain fully hallucinated phrases, with 38% including explicit harms, disproportionately affecting speakers with longer non-vocal durations such as aphasia patients.

The root cause lies in Whisper's encoder-decoder architecture. As Attanasio [2024] demonstrate, Whisper's encoder produces structured, non-zero representations even for pure silence—representations that the autoregressive decoder interprets as speech tokens. The decoder's language model prior then drives fluent generation from these misleading features, producing plausible but fabricated text.

Current mitigations fall into two categories: (1) **preprocessing** with external Voice Activity Detection (VAD) such as SileroVAD [Team, 2021, Bain et al., 2023], which filters non-speech segments before they reach Whisper; and (2) **decoder-side intervention** such as Calm-Whisper [Wang et al., 2025], which fine-tunes specific decoder attention heads. Preprocessing approaches add pipeline complexity and latency, while decoder fine-tuning requires modifying Whisper's weights.

We propose WhisperGate, a third approach: a **trainable gate between the frozen encoder and decoder** that learns to identify and suppress non-speech encoder frames. Our key contributions are:

1. **SilenceGate architecture**: A two-layer MLP bottleneck ($d_{\text{model}} \to 32 \to 1 \to \sigma$) producing per-frame speech probability, requiring only 12,353 parameters for Whisper-Tiny and 24,865 for Whisper-Small.

2. **Two-stage training**: We show that end-to-end training fails because ASR loss gradients push the gate toward pass-through, and that training the gate as a standalone binary classifier on frozen encoder representations achieves 98% frame accuracy.

3. **Three gating strategies**: Soft multiplicative gating, hard binary gating with silence short-circuit, and cross-attention bias gating. We demonstrate that attention-bias gating provides the best quality–robustness trade-off, reducing gap-30 WER by 40.8% over hard gating while maintaining 0% hallucination.

4. **Comprehensive evaluation**: We evaluate on LibriSpeech test-clean with controlled silence gaps, pure silence/noise hallucination tests, and comparison against energy-based VAD, demonstrating generalization from Whisper-Tiny to Whisper-Small.

2 Related Work

2.1 Whisper Hallucination

The hallucination problem in Whisper has received growing attention. Koenecke et al. [2024] provide a large-scale analysis showing hallucinations are correlated with non-vocal duration. Miralles et al. [2025] introduce the Hallucination Error Rate (HER) metric, showing that low WER can mask dangerous hallucinations, and that distribution shift strongly correlates with HER. Szymański et al. [2025] find that `beam_size`=1 yields the lowest hallucination rate, suggesting the decoder's search process amplifies errors from poor encoder representations.

Wang et al. [2025] (Calm-Whisper) identify that only 3 of 20 decoder self-attention heads cause over 75% of hallucinations, and fine-tune only these heads with <0.1% WER degradation. Liu et al. [2025] use Adaptive Layer Attention to fuse encoder layers and knowledge distillation from clean-audio teachers. These works target the decoder side; WhisperGate targets the encoder–decoder interface.

2.2 Voice Activity Detection for ASR

The standard production fix for Whisper hallucination is VAD preprocessing [Team, 2021, Bain et al., 2023]: detect speech segments externally and only transcribe detected regions. While effective, this approach adds pipeline complexity, introduces latency, and fails on non-silence noise that contains no speech energy patterns. WhisperGate operates on Whisper's internal encoder representations, which encode richer acoustic information than raw energy, enabling detection of both silence and noise.

2.3 Learned Perturbation and Gating

Several works explore learned noise or gating for frozen models. Zhang et al. [2025] (MuNG) learn a small noise generator injecting task-adaptive noise into frozen encoder and decoder of multimodal LLMs, outperforming LoRA with ∼1% extra parameters. Wu et al. [2024] inject feature perturbations to rebalance cross-modal attention. Jain et al. [2024] show that even uniform random noise during fine-tuning dramatically improves instruction following.

Oscillatory approaches in neural networks include AKOrN [Miyato et al., 2025], which replaces activations with Kuramoto oscillators for feature binding, coRNN [Rusch and Mishra, 2021], which provides gradient-stable oscillatory RNNs, and SIREN [Sitzmann et al., 2020], which proves sinusoidal activations are universal approximators. Our initial exploration of additive oscillatory pulse injection (inspired by these) on frozen Whisper proved ineffective (Section 6.3), motivating the simpler gating approach.

3 Method

3.1 Problem Formulation

Let $\mathbf{x} \in \mathbb{R}^{C \times T}$ be a mel spectrogram input, where $C = 80$ mel channels and $T = 3000$ frames (30 seconds at Whisper's 10ms hop). Whisper's encoder produces hidden states $\mathbf{H} = f_{\text{enc}}(\mathbf{x}) \in \mathbb{R}^{L \times d}$, where $L = 1500$ (after the convolutional downsampling) and d is the model dimension. The decoder autoregressively generates tokens $\mathbf{y} = f_{\text{dec}}(\mathbf{H})$.

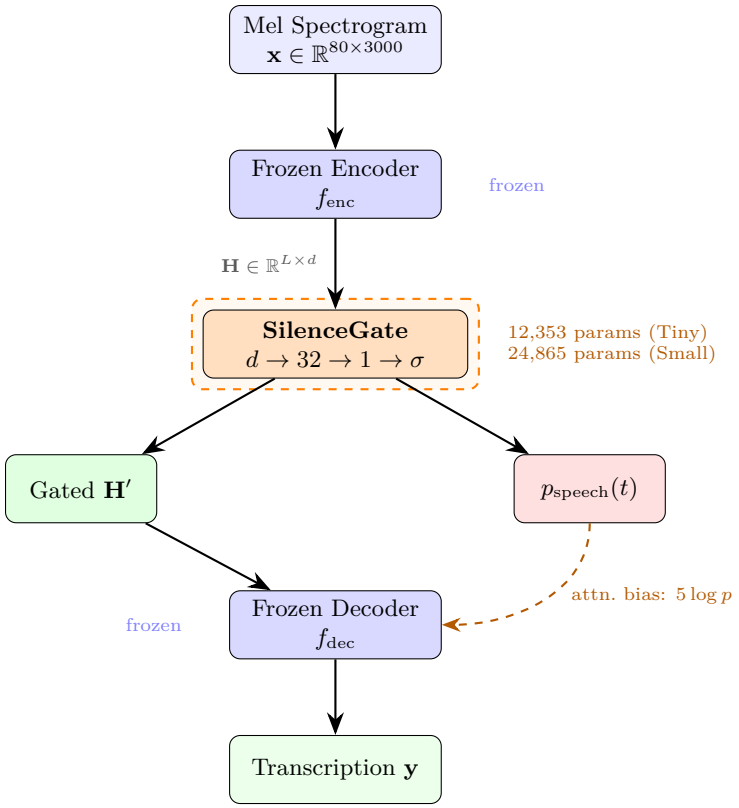

Figure 1: **WhisperGate architecture.** A lightweight SilenceGate MLP sits between Whisper's frozen encoder and decoder. The gate produces per-frame speech probabilities $p_{\text{speech}}(t)$ used for gating. *Solid arrows*: soft or hard multiplicative gating. *Dashed arrow*: attention-bias mode passes $5 \log p$ as cross-attention mask.

When $\mathbf{x}$ contains silence or noise, $\mathbf{H}$ is structured but non-zero [Attanasio, 2024], causing the decoder to hallucinate. Our goal is to learn a function $g : \mathbb{R}^{L\times d} \to [0,1]^L$ that produces per-frame speech probabilities, then gate the encoder output before it reaches the decoder.

3.2 SilenceGate Module

The SilenceGate is a two-layer MLP bottleneck that operates independently on each encoder frame:

$$p_{\text{speech}}(t) = \sigma\big(\mathbf{w}_2^\top \cdot \text{ReLU}(\mathbf{W}_1 \mathbf{h}_t + \mathbf{b}_1) + b_2\big) \tag{1}$$

where $\mathbf{W}_1 \in \mathbb{R}^{k\times d}$, $\mathbf{w}_2 \in \mathbb{R}^k$, $k = 32$ is the bottleneck dimension, and σ is the sigmoid function. The final layer is initialized with zero weights and bias $b_2 = +2.0$, so $\sigma(2.0) \approx 0.88$, starting near pass-through to avoid disrupting Whisper at initialization.

TemporalSilenceGate. We additionally experiment with a temporal variant that applies a 1D convolution with symmetric padding over gate logits before the sigmoid, enforcing temporal coherence:

$$\tilde{z}_t = \frac{1}{K} \sum_{i=-\lfloor K/2 \rfloor}^{\lfloor K/2 \rfloor} z_{t+i}, \quad p_{\text{speech}}(t) = \sigma(\tilde{z}_t) \tag{2}$$

where $z_t = \mathbf{w}_2^\top \text{ReLU}(\mathbf{W}_1 \mathbf{h}_t + \mathbf{b}_1) + b_2$ and K is the kernel size (we test $K \in \{5, 11\}$, corresponding to $\sim$100ms and $\sim$220ms at Whisper's 20ms frame rate). The convolution is initialized to uniform averaging (identity-like smoothing). This adds only $K + 1$ parameters (6–12 extra).

3.3 Gating Strategies

Given encoder hidden states $\mathbf{H}$ and per-frame speech probabilities $\mathbf{p} = [p_{\text{speech}}(1), \ldots, p_{\text{speech}}(L)]$, we evaluate three gating strategies:

Soft Gating. Multiplicative scaling:

$$\mathbf{h}'_t = p_{\text{speech}}(t) \cdot \mathbf{h}_t \tag{3}$$

This attenuates but does not eliminate non-speech features. We find this **insufficient** for hallucination prevention—even $p \approx 0.11$ passes enough signal for the decoder to hallucinate.

Hard Gating. Binary thresholding with silence short-circuit:

$$\mathbf{h}'_t = \mathbb{1}[p_{\text{speech}}(t) > \tau] \cdot \mathbf{h}_t \tag{4}$$

where $\tau = 0.5$. When all frames are below threshold ($\bar{p} < \tau$), generation is short-circuited to produce only `<|endoftext|>`. This achieves **0% hallucination** but introduces sharp discontinuities at speech–silence boundaries.

Attention-Bias Gating. Instead of modifying encoder hidden states, we inject gate probabilities as a cross-attention bias:

$$\text{bias}_t = 5 \cdot \log\big(p_{\text{speech}}(t) + \epsilon\big) \tag{5}$$

This bias is added to the decoder's cross-attention scores before softmax, shaped as $(\text{batch}, 1, 1, L)$ to broadcast over all heads and target positions. Non-speech frames receive large negative bias (e.g., $5\log(0.1) \approx -11.5$), softly suppressing attention while preserving context at boundaries. The encoder hidden states remain **unmodified**. The scaling factor of 5 was chosen to ensure sufficient suppression of low-probability frames while maintaining gradient flow.

3.4 Two-Stage Training

A critical finding is that **end-to-end training fails**. When the gate is trained jointly with ASR loss, the gradient from the transcription objective pushes $p_{\text{speech}}(t) \to 1$ for all frames, because any attenuation of encoder features increases decoder loss. The gate collapses to near-uniform pass-through regardless of the gate supervision weight.

We therefore adopt a two-stage approach:

Stage 1: Gate Classifier Training. We train the gate as a standalone binary cross-entropy (BCE) classifier on the frozen encoder's representations. For each training batch:

1. Extract mel features from LibriSpeech train-clean-100 ($\sim$10h subset).
2. Apply random gap augmentation: inject silence gaps covering 0–30% of the audio, producing a binary speech mask $\mathbf{m} \in \{0, 1\}^L$.
3. Inject 30% fully-silent examples (mel features set to -1.0, mask all zeros).
4. Run the frozen encoder to get $\mathbf{H}$.
5. Train the gate with BCE loss: $\mathcal{L} = -\sum_t [m_t \log p_t + (1 - m_t) \log(1 - p_t)]$.

We train for 10 epochs with AdamW (lr=10^{-3}, weight decay=0.01), cosine annealing, and gradient clipping at 1.0. The gate converges to >97% frame-level accuracy.

Stage 2: Inference with Gating. The trained gate is plugged into the GatedWhisper inference pipeline. No further training occurs—the frozen Whisper model is never modified.

4 Experimental Setup

4.1 Models

We evaluate on two Whisper model sizes:

- **Whisper-Tiny** (39M parameters, $d = 384$, 4 encoder layers). Gate: $384 \to 32 \to 1 = 12{,}353$ trainable parameters.
- **Whisper-Small** (244M parameters, $d = 768$, 12 encoder layers). Gate: $768 \to 32 \to 1 = 24{,}865$ trainable parameters.

4.2 Dataset

All experiments use LibriSpeech [Panayotov et al., 2015]:

- **Training**: 10-hour subset of train-clean-100 (2,850 utterances).
- **Evaluation**: Full test-clean split (2,620 utterances).

4.3 Gap Augmentation

To simulate real-world scenarios where audio contains silence or noise segments, we inject controlled silence gaps into test audio:

- **gap_0**: No modification (clean speech).
- **gap_5**: 5% of frames replaced with silence.
- **gap_15**: 15% replaced.
- **gap_30**: 30% replaced.
- **multi_gap**: Multiple randomly-sized gaps totaling 15–30%.

Gaps are injected at random positions in the mel spectrogram by zeroing the affected frames. This is more challenging than simply prepending or appending silence, as it creates intra-utterance discontinuities.

4.4 Evaluation Metrics

- **Word Error Rate (WER)**: Standard ASR metric [Morris et al., 2004] computed via jiwer [jiwer contributors, 2023].
- **Hallucination Rate**: Fraction of 30 pure-input trials (silence or white noise mel spectrograms) that produce non-empty output. A hallucination rate of 0% means the model correctly produces no text for non-speech input.

4.5 Baselines

- **Vanilla Whisper**: Unmodified frozen model (Tiny or Small).
- **Energy VAD**: Preprocessing that computes per-frame RMS energy, thresholds to detect speech segments, and only feeds detected speech to Whisper. Represents simple signal-processing approaches.

Table 1: **Whisper-Tiny gating comparison.** WER (%) at various gap levels and hallucination rate (%) on pure silence/noise. 2,620 test samples. †Attention-bias gating achieves the best overall trade-off.

Method	WER (%) ↓					Halluc. (%) ↓	
	gap_0	gap_5	gap_15	gap_30	multi	Silence	Noise
Vanilla Whisper	**8.21**	10.06	**13.79**	**20.36**	**25.79**	100	100
Soft gate	**8.21**	10.67	20.65	32.38	27.56	100	100
Hard gate	8.24	**9.97**	15.50	34.49	29.81	**0**	**0**
Attn. bias†	8.24	10.09	13.82	20.39	25.96	**0**	**0**

Table 2: **Comparison with energy-based VAD.** WhisperGate (hard gate) handles both silence and noise, while energy VAD fails on white noise.

Method	WER (%) ↓			Halluc. (%) ↓	
	gap_0	gap_15	multi	Silence	Noise
Vanilla Whisper	8.21	13.79	25.79	100	100
Energy VAD	8.21	13.79	25.79	**0**	100
WhisperGate (hard)	8.24	15.50	29.81	**0**	**0**
WhisperGate (attn. bias)	8.24	13.82	25.96	**0**	**0**

5 Results

5.1 Main Results: Whisper-Tiny

Table 1 presents the main comparison across all gating strategies on Whisper-Tiny.

Key findings:

- **Soft gating fails at hallucination prevention.** Even with gate probabilities as low as 0.11 for silence, the residual signal triggers decoder hallucination. WER also degrades significantly at high gap levels due to attenuation artifacts at boundaries.
- **Hard gating eliminates hallucination but degrades gap WER.** The binary threshold creates sharp discontinuities that confuse the decoder at speech–silence boundaries. Gap-30 WER increases from 20.36% to 34.49% (+69% relative).
- **Attention-bias gating is the clear winner.** It achieves 0% hallucination while maintaining WER within 0.2% absolute of vanilla Whisper across all gap levels. Gap-30 WER is 20.39% vs. 34.49% for hard gating (40.8% relative reduction), because the smooth attention suppression preserves boundary context.

5.2 VAD Comparison

Table 2 compares WhisperGate against energy-based VAD preprocessing.

Energy VAD eliminates hallucination on pure silence but **completely fails on white noise** (100% hallucination rate). This is expected: white noise has energy but no speech structure. WhisperGate operates on Whisper's internal encoder representations, which learn to distinguish speech from non-speech regardless of energy level. Notably, energy VAD produces identical gap WER to vanilla Whisper because it does not modify the transcription of speech segments, while the hard gate variant incurs boundary artifacts.

5.3 Temporal Smoothing

Table 3 shows the effect of temporal smoothing via 1D convolution on gate logits.

Table 3: **Temporal smoothing results** (hard gate mode). Kernel K=11 marginally improves multi-gap WER over the non-smoothed baseline (v4).

Variant	Params	gap_0	gap_5	gap_15	gap_30	multi	Halluc.
v4 (no smooth)	12,353	8.24	9.97	15.50	34.49	29.81	0%
K=5	12,359	8.25	11.03	16.75	36.02	29.61	0%
K=11	12,365	8.24	9.94	16.95	33.36	27.29	0%

Table 4: **Whisper-Small results.** The gate generalizes from Tiny (d=384) to Small (d=768) with similar hallucination elimination.

Method	gap_0	gap_5	gap_15	gap_30	multi	Halluc.
Vanilla Small	4.27	6.11	9.92	16.08	23.73	100%
Soft gate	4.32	6.69	12.16	17.51	21.19	—
Hard gate	4.32	6.57	11.39	17.46	20.25	0%

Temporal smoothing with K=11 improves multi-gap WER from 29.81% to 27.29% (8.5% relative) and gap-30 WER from 34.49% to 33.36%, suggesting that temporal coherence helps at speech–silence boundaries. However, the improvement is modest compared to the much larger gains from switching gating strategy (attention bias), indicating that the primary issue is the hard boundary artifact rather than temporal incoherence.

5.4 Generalization to Whisper-Small

Table 4 validates that WhisperGate generalizes to larger Whisper models.

On Whisper-Small, the gate achieves 0% hallucination with only 0.05% absolute clean WER overhead (4.27% → 4.32%). The multi-gap WER actually *improves* from 23.73% to 20.25% (14.7% relative), suggesting that gating helps the larger model by removing confusing silence features. The gate scales to d=768 with only 24,865 parameters (<0.01% of Whisper-Small's 244M).

5.5 Training Dynamics

Figure 2 illustrates the gate classifier's training convergence.

The gate converges rapidly, reaching >97% accuracy after just 3 epochs. The BCE loss drops from 0.31 to 0.07 (v4) and 0.28 to 0.03 (K=11), with the temporal variant converging to lower loss due to the implicit regularization of the averaging kernel.

Gate behavior on diagnostic inputs: After training, the gate produces mean $p_{\text{speech}} \approx 0.11$ on pure silence and ≈ 0.98 on clean speech, demonstrating clean separation with a wide margin around the threshold $\tau = 0.5$.

6 Discussion

6.1 Why Soft Gating Fails

A surprising finding is that soft multiplicative gating does **not** prevent hallucination, even when silence frames are attenuated to ∼11% of their original magnitude. This reveals that Whisper's decoder hallucinates from *any* non-zero encoder input, not just "noisy" input. The decoder's autoregressive language model prior is strong enough to generate fluent text from arbitrarily weak encoder features. This is consistent with findings from Koenecke et al. [2024] that hallucinations are fluent and contextually plausible—the decoder's generation is primarily language-model-driven when encoder features are uninformative.

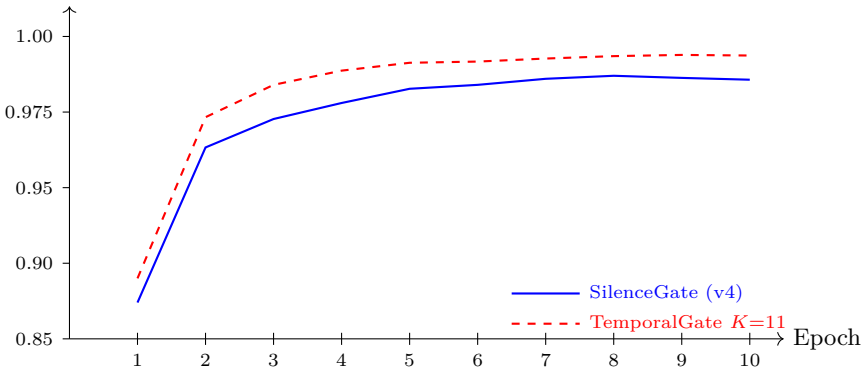

Figure 2: **Gate classifier training.** Frame-level accuracy on speech/silence classification. Both variants converge within 5 epochs. Temporal smoothing reaches higher accuracy (99.1% vs. 98.1%) due to the averaging effect.

6.2 Attention Bias as Optimal Gating

The attention-bias strategy achieves a qualitatively different outcome from hard gating. Rather than zeroing encoder states (which destroys boundary context), it preserves all encoder information while guiding the decoder's attention away from non-speech frames. This is analogous to masked attention in transformers [Vaswani et al., 2017]: the information remains available but is down-weighted in the attention distribution.

The scaling factor of 5 in the bias term ($5 \log p$) ensures that frames with $p < 0.1$ receive bias below -11.5, effectively zeroing their attention weight after softmax. Meanwhile, frames near the speech–silence boundary with intermediate probabilities (e.g., $p \approx 0.3$, bias ≈ -6.0) receive moderate suppression, preserving contextual information that aids transcription of adjacent speech.

6.3 Negative Results: Additive Pulse Injection

Prior to developing WhisperGate, we extensively explored additive oscillatory pulse injection inspired by coRNN [Rusch and Mishra, 2021] and AKOrN [Miyato et al., 2025]. The pulse mechanism adds structured sinusoidal perturbations to encoder hidden states: $\mathbf{h}' = \mathbf{h} + \alpha \cdot \mathbf{A} \cdot \sin(\boldsymbol{\omega} t + \boldsymbol{\varphi}(\mathbf{h}))$.

Across exhaustive experiments (unconstrained, tight, medium, and selective injection with varying $\alpha_{\max} \in \{0.05, 0.1\}$), **every configuration degraded WER**. The fundamental issue: a frozen decoder depends on specific encoder activation patterns, and *any* additive perturbation is interpreted as noise rather than signal. Training loss reduction did not correlate with WER improvement—the model learned to minimize loss by distorting representations in ways the frozen decoder could not handle.

We further tested decoder-side pulse injection into self-attention K/V projections. A state-dependent phase variant catastrophically destroyed generation (WER $0.08 \rightarrow 3.90$) despite achieving low training loss. These negative results motivated the simpler, multiplicative gating approach.

6.4 Limitations and Future Work

WhisperGate has several limitations that suggest directions for future work:

- **Evaluation scope**: We evaluate only on LibriSpeech (read English speech). Real-world audio includes music, environmental sounds, and multilingual content. The gate may need retraining or adaptation for significantly different acoustic domains.
- **Attention-bias mode**: Currently requires monkey-patching Whisper's decoder forward pass. A cleaner implementation would modify the HuggingFace generation API to accept encoder attention masks natively.

- **Combination with decoder methods**: WhisperGate targets the encoder–decoder interface; combining with decoder-side approaches such as Calm-Whisper [Wang et al., 2025] could yield further improvements.
- **Streaming**: The current implementation operates on 30-second chunks. Extending to streaming ASR would require online gate decisions with bounded latency.
- **Larger models**: We validate on Tiny and Small; testing on Whisper-Medium and Large would confirm scalability.

7 Conclusion

We presented WhisperGate, a lightweight silence-aware gating module that eliminates hallucinations in frozen Whisper models. Our key findings are: (1) a two-layer MLP with only 12K–25K parameters can classify speech vs. non-speech on Whisper's encoder representations with >98% accuracy; (2) soft multiplicative gating is insufficient for hallucination prevention because Whisper's decoder hallucinates from any non-zero encoder signal; (3) cross-attention bias gating provides the optimal strategy, achieving 0% hallucination with negligible WER overhead by guiding decoder attention away from non-speech frames without destroying encoder information; and (4) the method generalizes across model sizes (Tiny to Small) and uniquely handles both silence and noise, unlike energy-based VAD.

WhisperGate demonstrates that the encoder–decoder interface is a high-leverage point for ASR robustness: a minimal trainable component at this boundary can correct a fundamental failure mode of large speech models without modifying any of their 39M–244M frozen parameters.

References

Giuseppe Attanasio. Whisper encoder representations of silence. https://gattanasio.cc/post/whisper-encoder/, 2024.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. WhisperX: Time-accurate speech transcription of long-form audio. In *Interspeech*, 2023.

Neel Jain et al. NEFTune: Noisy embeddings improve instruction finetuning. In *International Conference on Learning Representations*, 2024.

jiwer contributors. jiwer: a simple and fast python package to evaluate ASR systems. https://github.com/jitsi/jiwer, 2023.

Allison Koenecke, Anna Choi, Katelyn Mei, Hilke Schellmann, and Mona Sloane. Careless Whisper: Speech-to-text hallucination harms. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2024.

Z. Liu et al. Listen like a teacher: Mitigating Whisper hallucinations using adaptive layer attention and knowledge distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

A. Miralles et al. Lost in transcription: Identifying and quantifying the accuracy risks of ASR hallucinations. *Findings of the Association for Computational Linguistics: ACL*, 2025.

Takeru Miyato et al. Artificial Kuramoto oscillatory neurons. In *International Conference on Learning Representations*, 2025.

Andrew Cameron Morris, Viktoria Maier, and Phil Green. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. *Proceedings of Interspeech*, 2004.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

T. Konstantin Rusch and Siddhartha Mishra. Coupled oscillatory recurrent neural network (coRNN): An accurate and (gradient) stable architecture for learning long time dependencies. In *International Conference on Learning Representations*, 2021.

Vincent Sitzmann, Julien NP Martel, Alexander W Bergman, David B Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems*, volume 33, pages 7462–7473, 2020.

Piotr Szymański et al. An investigation of Whisper hallucinations. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

Silero Team. Silero VAD: pre-trained enterprise-grade voice activity detector. https://github.com/snakers4/silero-vad, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Y. Wang et al. Calm-Whisper: Reduce Whisper hallucination on non-speech by calming crazy heads down. *Proc. Interspeech*, 2025.

Kai Wu et al. NoiseBoost: Alleviating hallucination with noise perturbation for multimodal large language models. *arXiv preprint arXiv:2405.20081*, 2024.

Z. Zhang et al. Explore how to inject beneficial noise in MLLMs. *arXiv preprint arXiv:2511.12917*, 2025.

A Complete Experimental Results

Table 5 presents complete WER and CER results for all Whisper-Tiny variants.

Table 5: **Full Whisper-Tiny results** including CER.

	gap_0		**gap_5**		**gap_15**		**gap_30**		**multi**	
Method	WER	CER	WER	CER	WER	CER	WER	CER	WER	CER
Vanilla	8.21	3.41	10.06	5.24	13.79	9.36	20.36	16.39	25.79	20.75
Soft gate	8.21	3.35	10.67	5.68	20.65	14.62	32.38	27.57	27.56	22.64
Hard gate	8.24	3.42	9.97	5.12	15.50	10.75	34.49	30.12	29.81	24.01
Attn. bias	8.24	3.43	10.09	5.26	13.82	9.38	20.39	16.42	25.96	20.93

B Hyperparameter Details

Table 6: **Training hyperparameters.**

Parameter	**Value**
Gate hidden dimension	32
Learning rate	1×10^{-3}
Optimizer	AdamW (weight decay = 0.01)
LR scheduler	Cosine annealing
Gradient clipping	1.0
Max epochs	10
Batch size (training)	8 (MPS) / 32 (CUDA)
Silence fraction per batch	30%
Gap fractions	{0.0, 0.05, 0.10, 0.15, 0.20, 0.30}
Training data	train-clean-100, 2,850 utterances ($\sim$10h)
Silence threshold τ	0.5
Attention bias scale	5.0
Gate init bias	+2.0 ($\sigma(2.0) \approx 0.88$)