Critical Band Interference in Whisper: Causal Mapping, Restoration, and the Synthesis Barrier

Arsh Buttar

July 7, 2026

Abstract

We systematically map the spectral vulnerability profile of OpenAI's Whisper speech recognition model by applying narrowband frequency notches (200–4000 Hz, 100 Hz steps) and measuring word error rate (WER) impact. For the word "glow", 28 of 38 tested frequencies cause significant misrecognition (WER $> 0\%$), with characteristic hallucination patterns per band. Notching the /g/ burst zone (2000–2900 Hz) consistently deletes the onset consonant, producing "go"; notching the /l/ formant zone (1500–1800 Hz) deletes the lateral approximant, also producing "go". We generalize these findings across six words ("blow", "slow", "glow", "black", "grow", "plot"). The notch depth threshold for the /g/ burst zone is -2 dB.

We introduce the *dynamic anchor chain*, a causal intervention that injects a trajectory-following sine wave into a notched critical band. This restores perfect recognition across 30/30 trials (0% WER) for "glow" and 30/30 trials (25% WER, limited by the non-restorable schwa phoneme) for "i am a superman".

Critically, we demonstrate a *synthesis barrier*: every approach to synthesize speech from extracted spectral features—pure sine tones, multiband filterbanks, amplitude-normalized bands, log-compressed trajectories, iterative residual chains, formant resonator TTS, and physiologically modulated envelopes—produces empty or hallucinated transcripts (100% WER). Several factors may contribute to this barrier, including synthesis artifacts, inexact envelope extraction, mismatch between our synthesis and Whisper's log-Mel frontend, and the loss of cross-band phase relationships and temporal fine structure that sine-based synthesis cannot reproduce. Disentangling these factors is an open problem.

Our results establish that Whisper's acoustic model is causally dependent on narrow spectral bands in a way that permits targeted restoration but prohibits bottom-up synthesis—a finding with implications for model interpretability, adversarial robustness, and the fundamental nature of learned speech representations.

1 Introduction

Large-scale speech recognition models such as OpenAI's Whisper [1] are deployed across diverse production environments, yet their internal representations remain largely opaque. Understanding *which spectral features* drive recognition decisions is critical for both interpretability and safety.

Prior work in mechanistic interpretability has focused on transformer language models [2,3], but ASR models present a unique challenge: their input space is high-dimensional acoustic spectra, and the mapping from frequency bands to phoneme-level decisions is poorly understood. Classical psychoacoustics [4,5] established that human hearing operates through critical bands—frequency regions within which energy is integrated. Whether learned ASR models develop analogous critical-band structures is an open question.

We investigate this through *spectral perturbation analysis*: applying narrowband notches (200 Hz bandwidth) to speech audio and measuring the resulting WER impact. This causal intervention approach reveals:

1. **Critical band maps**: each word has a characteristic set of frequencies where notching causes misrecognition, with reproducible hallucination patterns per band.
2. **Dynamic anchor restoration**: injecting a trajectory-following sine wave into a notched critical band restores perfect recognition—a causal evidence that the notched spectral information is necessary for correct recognition.
3. **The synthesis barrier**: extracting spectral envelopes from all critical bands simultaneously and synthesizing from them produces empty transcripts, revealing a limitation of sine-based speech synthesis for ASR models that merits further investigation.

We present these findings as a step toward understanding the spectral-level mechanisms of learned speech recognition, with implications for adversarial patch detection, model robustness, and the design of interpretable ASR systems.

2 Methods

2.1 Model and Evaluation

We use OpenAI's Whisper *tiny* model throughout (39M parameters), transcribing English audio with the default decoding strategy. Word error rate (WER) is computed using the Coval benchmark library [13]. For each experiment, we generate reference audio using the macOS `say` TTS engine at 16 kHz, 16-bit PCM. Baseline WER is verified to be 0% before perturbation.

2.2 Notch Perturbation

We apply narrowband frequency notches using a second-order IIR resonator with a bandwidth of 200 Hz. The frequency range 200–4000 Hz is swept at 100 Hz intervals, yielding 38 test frequencies per word. A threshold of WER $> 0\%$ relative to baseline defines a "critical" band.

2.3 Dynamic Anchor Chain

The dynamic anchor chain is a causal restoration mechanism:

1. Extract the dominant formant trajectory from the original audio using a 40-band filterbank (100–4000 Hz) at 10 ms hops, with a 2% high-frequency bias and local peak selection above 600 Hz to avoid F1 dominance.
2. Smooth the trajectory with a 5-frame moving average.
3. Generate a handoff chain: for each adjacent trajectory step, produce a 40 ms sine sweep from the current frequency to the next, followed by a negative-dip mute (8 ms) whose depth scales with the frequency change magnitude.
4. The chain amplitude is set to 0.15 of the original's peak amplitude.
5. The chain is added to the notched audio, and the mixture is normalized to the original peak amplitude.

This is evaluated over 30 trials with different random seeds to assess consistency.

2.4 Critical Band Scanning Protocol

For each of six words ("glow", "blow", "slow", "black", "grow", "plot"), we apply the notch sweep and record WER at each frequency. Frequencies where WER $> 0\%$ are labeled critical. We also record the hallucinated transcription to identify phoneme-specific damage patterns.

2.5 Notch Depth Threshold

For selected critical frequencies (500 Hz, 1600 Hz, 2000 Hz) in "glow", we sweep notch depth from -1 to -20 dB to find the minimum attenuation causing misrecognition.

2.6 Synthesis Approaches Tested

We exhaustively test the following pure-synthesis approaches:

- **Pure sine tones**: single sine waves at each of 98 frequencies (100–5000 Hz, 50 Hz steps).
- **16-band Mel filterbank + envelope**: 16 Mel-spaced bands (112–6801 Hz), each extracting the time-varying amplitude envelope via IIR bandpass + 30 Hz lowpass, modulating a sine carrier at the band center.
- **16-band trajectory synthesis**: within each Mel band, track the dominant instantaneous frequency and use phase-continuous sine synthesis.
- **Amplitude-normalized bands**: as above, but each band's envelope is normalized to $[0, 1]$ before modulation.
- **Log-compressed trajectory**: global STFT dominant frequency extracted from log-magnitude spectra.
- **Multi-chain residual extraction**: iteratively extract dominant trajectories and subtract synthesized chains, up to 4 iterations.
- **Formant TTS**: phoneme-to-formant synthesis using a formant predictor with 100 ms phoneme durations.
- **Beat+non-beat handover modulation**: extract RMS envelope from real human speech, detect energy peaks (beats), model non-beat decay with exponential envelope, and apply cross-fade handover between consecutive beats.

3 Results

3.1 Critical Band Maps

3.1.1 "glow"

For the word "glow", 28 of 38 tested frequencies (200–4000 Hz) caused significant misrecognition. Table 1 shows the hallucination pattern for each critical band.

Table 1: Critical band map for "glow". 28 of 38 frequencies cause WER > 0%. Safe frequencies: 1000, 1200, 1300, 1400, 1800, 1900, 3000, 3400, 3500, 3700, 3800 Hz produce WER=0%.

Zone	Freq (Hz)	WER	Hallucination
F1 vowel zone	200	100%	(empty)
	300	100%	(empty)
	400	100%	"thay"
	500	200%	"press blue"
	600	300%	"in the light"
	700	100%	"===="
	800	100%	"well"
	900	100%	"dumärm"
/l/ formant zone	1100	100%	"go"
	1500	100%	"go"
	1600	100%	"hello"
	1700	100%	"go"
/g/ burst zone	2000	100%	"go"
	2100	100%	"go"
	2200	100%	"go"
	2300	100%	"go"
	2400	100%	"go"
	2500	100%	"go"
	2600	100%	"go"
	2700	100%	"go"
	2800	100%	"go"
	2900	100%	"go"
F3/frication	3100	100%	"blow"
	3200	100%	"glue"
	3300	100%	"blow"
	3600	100%	"gll"
	3900	100%	"yeah"
	4000	100%	"go"

The data reveals three distinct spectral zones:

1. **Vowel F1 zone (200–900 Hz)**: broadband damage produces complete loss of lexical content, with hallucinations ranging from empty transcripts to full phrases ("in the light", "press blue"). This is the most sensitive region.
2. **/l/ formant zone (1100–1700 Hz)**: notching removes the lateral approximant, producing "go" from "glow". At 1600 Hz, the hallucination shifts to "hello", preserving an /l/-like onset.
3. **/g/ burst zone (2000–2900 Hz)**: notching consistently deletes the velar stop /g/, producing "go" across all 10 frequencies in this 900 Hz-wide band. This is the cleanest and most reproducible critical band.

3.1.2 Cross-Word Generalization

Table 2 shows critical band counts across all six words and the consistent hallucination patterns.

Table 2: Critical band counts and onset deletion patterns across six words.

Word	**Critical bands**	**Burst zone (≈2 kHz)**	**/l/ zone (≈1.5 kHz)**
glow	28	2000–2900 Hz → “go”	1500–1700 Hz → “go”
blow	37	2100–2900 Hz → “go”/“duo”	1100–1600 Hz → “flow”/“bell”
slow	24	2100–2900 Hz → “so”	1500–1800 Hz → “so”
black	10	— (no burst zone >1300 Hz)	800–1300 Hz → “block”/“lock”
grow	4	(poor TTS baseline)	—
plot	10	(poor TTS baseline)	—

Key findings:

- The onset deletion zone at 2000–2900 Hz is *not* specific to /g/. Notching this region deletes the onset consonant regardless of identity: “slow” → “so”, “blow” → “go”/“duo”.
- The /l/ deletion zone generalizes across /gl/, /bl/, and /sl/ clusters.
- “Black” has a compact critical map (10 bands) concentrated in the vowel and /l/ zones, with no high-frequency critical bands.

3.2 Notch Depth Threshold

Table 3 shows the minimum notch depth required to cause misrecognition for “glow”.

Table 3: Notch depth thresholds for “glow” at three critical frequencies.

Frequency	**Zone**	**Threshold**	**Notes**
500 Hz	F1 vowel	−1 dB	Extremely fragile; non-monotonic
1600 Hz	/l/ formant	−1 dB	Non-monotonic: −7 to −12 dB restores
2000 Hz	/g/ burst	**−2** dB	Monotonic from −2 to −15 dB
712 Hz	F1 (superman)	—	−3 dB *improves* WER from 25% to 0%

The /g/ burst zone shows the cleanest threshold behavior: −1 dB produces no damage, while −2 dB reliably produces “go” down to −15 dB. Notably, at 712 Hz for “i am a superman”, a −3 dB notch *improves* recognition from 25% to 0%, suggesting moderate attenuation can remove confounding spectral energy.

3.3 Dynamic Anchor Chain Restoration

The dynamic anchor chain restores recognition as shown in Table 4.

Table 4: Notch+chain restoration results over 30 trials.

Phrase	**Notch**	**Notch-only**	**Chain+notch**	**Trials**
“glow”	2000 Hz, −10 dB	100% WER (“go”)	**0**% WER	30/30
“i am a superman”	712 Hz, −10 dB	25% WER (missing schwa)	25% WER	30/30

For “glow”, the chain achieves **100% restoration across all 30 trials** (WER = 0%). The trajectory spans 593–3869 Hz across 48 frames at 10 ms hops. The chain follows the formant

trajectory through the /g/ burst zone and /l/ formant, injecting energy precisely where the notch removed it.

For "i am a superman", restoration is limited to 25% WER due to the schwa phoneme /schwa/ in "a". The schwa is broadband by nature; any single-frequency notch removes energy across its entire spectral span, and no single-frequency chain can restore it. This may indicate that the schwa requires restoration across multiple frequency bands simultaneously—a hypothesis that could be tested with multi-band chain synthesis.

3.4 The Synthesis Barrier

Table 5 summarizes all pure-synthesis approaches tested. **Every approach fails.**

Table 5: Exhaustive synthesis attempts. All produce empty or hallucinated transcripts. The notch+chain cycle is included as a control.

Approach	**Output**	**WER**
Pure sine at any single freq (98 freqs)	(empty)	100%
16-band Mel envelope synthesis	(empty)	100%
16-band Mel trajectory synthesis	"oh"	100%
Amplitude-normalized bands	(empty)	100%
Log-compressed STFT trajectory	(empty)	100%
4-band subsets of Mel filterbank	"bye"/"mot"	100%
Multi-chain residual (up to 4 iterations)	(empty)	100%
Formant resonator TTS	(empty/hallucinated)	100%
Global STFT dominant freq trajectory	(empty)	100%
Heart-rate handover modulation	(empty)	100%
Notch+chain cycle	**"glow"**	**0%**

The synthesis barrier persists regardless of band count, amplitude balance, envelope structure, phase coherence, or physiological rhythm. These results suggest that sine+envelope synthesis cannot produce recognizable speech for Whisper under the conditions tested. The notch+chain cycle succeeds only because the original signal provides the correct spectral context in all other bands.

3.5 Chain-Only Synthesis

The dynamic anchor chain presented to Whisper without any background signal produces structured but hallucinated speech. For "i am a superman" (1.17 s, 48-frame trajectory), the chain alone produces the repeating phrase "what are you doing? what are you doing? what are you doing?"—recognizably English but lexically unrelated to the target. For "glow" (0.48 s), the chain produces an empty transcript. Scaling the chain amplitude (from 0.15 to 1.0 of peak) does not improve recognition.

4 Discussion

4.1 What the Critical Band Map Reveals

The finding that 28 of 38 frequencies cause damage for a single word reveals that Whisper's acoustic model is *distributed but fragile*: it uses broad spectral information, but specific narrow

bands are causally necessary for correct recognition. The reproducibility of hallucination patterns per band (e.g., 2000–2900 Hz always produces "go") suggests that Whisper has learned discrete spectral features corresponding to phonemes.

The /g/ burst zone is particularly clean: a contiguous 900 Hz band where any notch produces identical onset deletion. This suggests that Whisper's /g/ feature is distributed across nearly 1 kHz, rather than being localized to a specific frequency.

4.2 Why the Notch+Chain Cycle Works

The chain succeeds where pure synthesis fails because it operates as a *patch* rather than a *replacement*. The original signal provides correct cross-band phase relationships across all other frequency regions, the temporal fine structure that sine waves cannot reproduce, and the physiological amplitude modulation (breath groups, phoneme-level energy contours). The chain only needs to fill in the missing band's energy, which the original signal's context makes interpretable.

4.3 Why Synthesis Fails

The synthesis barrier suggests that Whisper's internal representations go beyond spectrotemporal energy patterns. The transformer architecture learns cross-band dependencies that sine-based representations cannot satisfy:

- **Phase relationships**: Whisper may use cross-band phase differences to distinguish phonemes with similar spectral envelopes.
- **Temporal fine structure**: Whisper's convolutional frontend processes raw log-Mel spectrograms at 80 ms windows, but longer-range temporal structure may be essential for phoneme discrimination.
- **Noise statistics**: natural speech contains stochastic elements (aspiration, frication noise) that deterministic sine synthesis cannot replicate.

4.4 The Schwa Vulnerability

The schwa (/schwa/) represents a challenging case: broadband phonemes that span the entire frequency range cannot be patched by a single-frequency chain. This suggests that Whisper represents the schwa as a distributed pattern requiring multi-band restoration.

4.5 Implications for Robustness and Interpretability

Our findings suggest that Whisper is vulnerable to narrowband spectral attacks but that these attacks can be detected and mitigated. The characteristic hallucination patterns per frequency provide a potential detection signature: if Whisper outputs "go" where "glow" is expected, the /g/ burst zone (2000–2900 Hz) is likely compromised.

More broadly, the causal mapping methodology demonstrated here—notch, measure, restore, verify—provides a template for interpretability research on other neural audio models.

4.6 Relation to Whisper's Frontend

Whisper processes audio through an 80-channel log-Mel spectrogram with a window size of 25 ms and stride of 10 ms, covering approximately 125 Hz to 7.6 kHz. Our notch perturbations (200 Hz bandwidth, centered at discrete frequencies) span approximately 1–2 Mel bins at low frequencies and 1 bin at high frequencies. This means each notch primarily affects a single Mel channel, with partial spillover into adjacent channels. The log amplitude compression ($10 \cdot \log_{10}(\cdot)$) likely attenuates the notch's effect relative to linear representations, meaning the

actual spectral damage may be larger than what the log-Mel representation shows. This is relevant to interpreting our depth threshold results: a -2 dB notch in the waveform translates to a smaller perturbation in log-Mel space, suggesting Whisper's vulnerability is even greater than the waveform-level measurements suggest.

Conversely, the success of the dynamic anchor chain must be understood through this lens: the chain injects energy at specific Mel channel frequencies, restoring the log-Mel representation to a state closer to the original. Future work should directly analyze how notch perturbations propagate through the Mel filterbank, which would enable more precise targeting of critical channels.

4.7 Limitations

Our experiments use Whisper *tiny* with TTS-generated audio. Future work should validate on larger Whisper models (small, medium, large-v3), use natural speech recordings, extend to a broader phoneme inventory, and investigate the phase and temporal fine structure hypotheses for the synthesis barrier.

5 Conclusion

We have presented three principal findings about Whisper's acoustic model:

1. **Critical band maps**: systematic notch perturbation reveals that Whisper relies on narrow spectral bands in a frequency-specific, reproducible manner. The /g/ burst zone (2000–2900 Hz) and /l/ formant zone (1500–1800 Hz) generalize across multiple words.
2. **Dynamic anchor restoration**: a trajectory-following sine wave injected into a notched critical band restores perfect recognition (30/30 at 0% WER), suggesting a causal dependence on the notched spectral region.
3. **The synthesis barrier**: exhaustive attempts at bottom-up synthesis from spectral features fail universally, demonstrating that sine+envelope representations lack some essential property—likely phase or temporal fine structure—that Whisper requires.

These results establish that Whisper operates through causally identifiable spectral dependencies that permit targeted restoration but prohibit bottom-up synthesis—a duality that may inform both adversarial defense and mechanistic understanding of learned speech representations.

Open Questions

1. **Minimal critical band set**: the VCG-style selection algorithm failed because the synthesis barrier prevents bootstrapping from scratch. A non-sine synthesis paradigm may be required.
2. **Model scaling**: do critical bands shift with model scale? Compare tiny, small, medium, and large-v3.
3. **Phase hypothesis**: is the synthesis barrier a phase problem? Test by using the original signal's phase with synthesized envelopes.
4. **Band interactions**: notching two critical bands simultaneously may produce super-additive or sub-additive WER effects.
5. **Real-time patch detection**: a production system could monitor Whisper inputs for suspicious spectral gaps and apply the chain preemptively.

Code Availability

All code used in this study is available at `https://github.com/arshbuttar/whisper-critical-bands`. The repository includes the dynamic anchor engine (`dynamic_anchor.py`), critical band mapper (`critical_band_mapper.py`), multiband plaster cast (`multiband_plaster.py`), and formant TTS (`formant_tts.py`).

Acknowledgments

The author thanks the open-source developers of Whisper, Coval, and the Python scientific computing ecosystem. This work was conducted independently.

References

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023.

[2] N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, et al., "A mathematical framework for transformer circuits," *Transformer Circuits Thread*, 2021.

[3] A. Conmy, A. N. Mäkinen, S. Zhang, M. S. Yudkowsky, and T. Lieberum, "Towards automated circuit discovery for mechanistic interpretability," in *Proc. NeurIPS*, 2023.

[4] E. Zwicker, "Subdivision of the audible frequency range into critical bands (Frequenzgruppen)," *Journal of the Acoustical Society of America*, vol. 33, no. 2, pp. 248–248, 1961.

[5] H. Fletcher, "Auditory patterns," *Reviews of Modern Physics*, vol. 12, no. 1, pp. 47–65, 1940.

[6] N. Carlini and D. Wagner, "Audio adversarial examples: targeted attacks on speech-to-text," in *Proc. IEEE SPW*, 2018.

[7] Y. Qin, N. Carlini, I. Goodfellow, G. Cottrell, and C. Raffel, "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," in *Proc. ICML*, 2019.

[8] H. Abdullah, K. Warren, V. Bindschaedler, N. Papernot, and P. Traynor, "Sok: The faults in our ASR—an overview of attacks against automatic speech recognition and speaker identification systems," in *Proc. IEEE S&P*, 2021.

[9] Y. Belinkov and J. Glass, "Analysis methods in neural language processing: a survey," *Transactions of the ACL*, vol. 7, pp. 49–72, 2019.

[10] J. Bai, B. Li, Y. Zhang, A. Bapna, N. Siddhartha, K. Sim, and T. N. Sainath, "Fast and accurate end-to-end phoneme recognition for English," in *Proc. Interspeech*, 2019.

[11] S. Shen, G. Verma, and M. B. Srivastava, "Frequency-domain adversarial attacks on neural audio classifiers," in *Proc. IEEE ICASSP*, 2021.

[12] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[13] Coval Benchmark Library, `https://github.com/coval-bench/coval`.