MECHANISM-SPECIFIC SPEECH ENCODING FAILURES IN AUDITORY NEUROPATHY: A COMPUTATIONAL PHENOTYPING FRAMEWORK

Marta Campi[1,2,*], **Elie Partouche**[1], **Gregory Gerenton**[1], **Paul Avan**[1], **Clément Gaultier**[1,3]

[1]Université Paris Cité, Institut Pasteur, AP-HP, INSERM, CNRS, Fondation Pour l'Audition, Institut de l'Audition, IHU reConnect, F-75012, Paris, France

[2]University of Zurich, University Hospital Zurich, Dept. of Otorhinolaryngology, Switzerland

[3]MRC Cognition and Brain Sciences Unit, University of Cambridge, 15 Chaucer Road, Cambridge, UK, CB2 7EF

{mcampi, epartouche, ggerenton, pavan, cgaultier}@pasteur.fr

*Corresponding author: mcampi@pasteur.fr

March 8, 2026

ABSTRACT

Auditory Neuropathy Spectrum Disorders (ANSD) are best characterized by distorted patterns of auditory nerve activity despite preserved spectral analysis of sound in the cochlea. They should clinically translate into impaired speech recognition despite normal auditory sensitivity. While distinct pathophysiological mechanisms affecting auditory-nerve activity have been identified in animal models, current clinical speech tests cannot distinguish among them. Our working hypothesis is that current speech audiometry yields aggregate recognition scores that average across phoneme categories, obscuring mechanism-specific patterns instead of pinpointing distinctive signatures. Using computational modeling of auditory nerve responses, we tested four mechanism types, demonstrating that mechanism-specific encoding disruptions cascade into speech recognition failures. Brief consonants showed severe disruption while sustained vowels were preserved, with category-specific patterns differing across mechanisms. Models trained on ANSD-degraded signals developed compensation strategies that generalized to healthy signals, while the reverse failed completely: strategies exploiting slow temporal structure (sustained formants) generalize, while those requiring millisecond-scale timing do not. Noise training that benefited healthy models harmed ANSD models, explaining real-world listening difficulties. Phoneme-specific confusion patterns enable mechanism identification from behavioral testing alone, providing the missing diagnostic infrastructure for targeted intervention.

Keywords Auditory Neuropathy, Speech Processing, Formant Encoding, Neural Perturbations, Optimal Transport

1 Introduction

Hearing loss represents a global health challenge affecting 466 million people worldwide, with projections reaching 900 million by 2050. The associated economic burden exceeds €216 billion annually in Europe alone [1, 2]. Despite hearing aid development, approximately half of individuals who need amplification do not use it, partly due to poor performance in complex acoustic environments [3]. These limitations are particularly evident in Auditory Neuropathy Spectrum Disorders (ANSD), which affect an estimated 1–10% of those with hearing complaints [4, 5] and are associated with minimal benefit from conventional amplification despite preserved peripheral sensitivity.

Patients with ANSD exhibit preserved outer hair cell function, as probed by otoacoustic emissions and/or cochlear microphonic responses, yet show abnormal or absent auditory brainstem responses (ABRs) [6, 7], indicating distorted patterns of auditory nerve activity despite normal cochlear filtering. This neural disruption typically manifests as the characteristic syndrome of "hearing without understanding": patients detect sounds but struggle to comprehend speech, with audiometric thresholds that substantially underestimate their functional communication difficulties [8], though presentation varies widely and some cases may show minimal speech deficits despite measurable neural abnormalities. The finding that ABRs are disrupted suggests impairment of fine timing cues in auditory neuron activity, contrasting with typical cochlear hearing loss that reduces frequency selectivity and elevates detection thresholds. Reported patients show deficits in gap detection, amplitude modulation sensitivity, and speech envelope tracking not explained by their audiometric profiles, attributed to impaired auditory-neuron activity [9, 10]. Other patients with similar complaints of poor speech recognition and indirect evidence of impaired auditory-neuron activity may show no clear audiological pattern, a condition termed 'hidden hearing loss' [11, 12].

The clinical heterogeneity of ANSD underscores the complexity of its underlying pathophysiology. Patient presentations vary widely: some appear to struggle disproportionately with consonant discrimination, others with vowel recognition, and performance can fluctuate unpredictably across days or acoustic conditions [13]. Distinct pathophysiological mechanisms have been identified that operate at different sites along the auditory pathway. Temporal jitter from auditory nerve demyelination represents one major mechanism: myelin sheaths ensure rapid, synchronized action potential propagation, and demyelination introduces variable conduction delays that scramble precise timing relationships across frequency channels [14, 15]. Phase-locking and cross-frequency coincidence detection in the brainstem both rely on precisely aligned inputs [16, 17], and demyelination-induced jitter disrupts both. Synaptic loss represents a second major mechanism, particularly affecting low- and medium-spontaneous rate (SR) fibers that resist saturation and may be important for encoding speech in noise [18, 19]. Selective loss of these synapses reduces population coding strength while sparing the high-SR fibers that mediate quiet threshold detection, explaining why pure-tone audiograms can appear normal despite expected speech-in-noise deficits. Saturation deficits from reduced neural populations represent a third mechanism, compressing dynamic range and preventing proper formation of level-tolerant spectral representations [20, 21]. Additionally, conditions with sporadic loss of inner hair cells [22] or dead cochlear zones [23] may produce ANSD-like distortions in neural representations, a neurogram being a time chart of spiking activity across auditory neurons with different best frequencies, with little or no elevation in auditory thresholds, due to compensation by surviving hair cells or off-frequency listening. These distortions might make identification of some speech sounds difficult or impossible. We modeled these pathophysiological categories as four perturbation types: temporal jitter was implemented in two variants (uniform and scattered) to capture global versus patchy demyelination, while synaptic loss and saturation were modeled as selective fiber loss and response truncation (Fig. 1A–B).

Current diagnostic approaches face two fundamental limitations. First, while electrophysiological signatures (absent or abnormal ABRs with preserved OAEs) confirm the presence of ANSD, they cannot differentiate between specific pathophysiological mechanisms, and their absence does not preclude ANSD [7]. While recent work has demonstrated some success using compound action potential patterns to distinguish synaptopathy from myelinopathy [24], electrophysiological measures cannot reliably predict the specific phoneme confusions that characterize individual patients' functional deficits. Second, while speech recognition tests can reveal functional consequences of confirmed ANSD, standard speech audiometry provides aggregate intelligibility scores (percentage of words or sentences correctly identified) that collapse multidimensional confusion patterns into a single number. Phoneme-specific testing exists: the Nonsense Syllable Test [25], California Consonant Test [26], and similar materials yield detailed confusion matrices that could reveal category-specific deficits. Yet results are typically reported as aggregate scores for which normative data exist, rather than analyzed for category-specific patterns. If different ANSD mechanisms preferentially disrupt different phoneme categories, this aggregation systematically obscures mechanism-specific signatures.

Proving that a given group of patients corresponds to a single pathophysiological category, establishing that their phoneme confusions are mechanism-specific, would require histopathological data, which is rarely available. Conversely, computational models can emulate different ANSD mechanisms with known ground truth and establish whether they preferentially disrupt specific phoneme categories. This is the approach we take: numerical perturbations

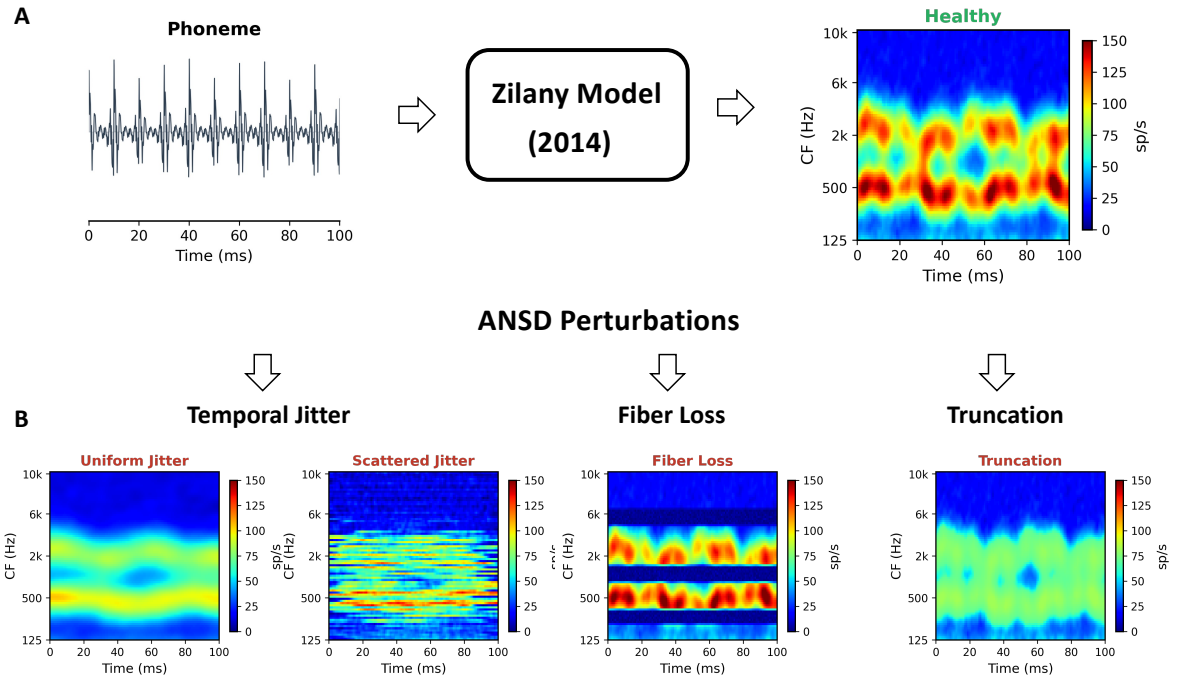

Figure 1: **Computational framework for ANSD neurogram perturbations. (A)** Neurogram generation: speech phonemes are processed through the Zilany auditory nerve model to produce healthy neurograms representing firing rates across characteristic frequency (CF) and time. **(B)** Four perturbation types modeling distinct ANSD mechanisms, each derived from the healthy neurogram. *Temporal jitter* models demyelination in two variants: *uniform jitter* applies identical temporal shifts across all channels, preserving cross-frequency timing; *scattered jitter* applies independent per-channel shifts, destroying cross-frequency synchrony. *Fiber loss* models synaptopathy through frequency-selective channel elimination, creating spectral gaps (dark bands). *Truncation* models neural hypoplasia through amplitude saturation, compressing dynamic range. All panels use identical colorscale (0–150 sp/s).

applied to neurograms feed a speech recognition network, enabling systematic exploration of all mechanism–phoneme combinations under controlled conditions. One goal was to test whether simulations might generate distinguishable patterns such that a patient scoring 60% on a speech recognition test could be failing on consonants while preserving vowels (suggesting disrupted timing cues), or failing uniformly across categories (suggesting spectral impairment), possibly depending on the presence of noise. This might help lift a barrier still preventing targeted intervention: without knowing which mechanism drives dysfunction in individual patients, clinicians cannot select appropriate rehabilitation strategies.

Speech sounds vary systematically in the temporal precision they demand from neural encoding. Vowels are produced with relatively stable vocal tract configurations, creating sustained formant patterns that persist for 50–200 ms and may survive temporal degradation through temporal integration. Consonants, by contrast, involve rapid articulatory transitions producing brief acoustic events: stop bursts lasting 5–40 ms, formant transitions over 20–50 ms, and flaps requiring only 20–30 ms of precise encoding. These properties predict that temporal perturbations should disproportionately affect brief consonants while sparing sustained vowels, and that frequency-selective fiber loss should affect formants encoded by the damaged region, whereas saturation should affect categories more uniformly. If these predictions hold, phoneme-specific confusion patterns could serve as diagnostic signatures for the underlying mechanism.

2 Results

If distinct neural pathologies disrupt different acoustic cues, they should produce distinct, phoneme-specific recognition failures. If so, category-specific confusion patterns would carry more diagnostic information than aggregate recognition scores, enabling mechanism identification from behavioral testing alone. We tested this hypothesis in three stages: first quantifying how each proposed ANSD mechanism alters neural encoding of speech (Experiment 1), then determining whether these encoding distortions produce phoneme-specific classification failures (Experiment 2),

and finally evaluating whether the resulting patterns persist under realistic noise conditions (Experiment 3). Testing this hypothesis required constructing three computational modules: a neurogram perturbation generator, a hierarchical speech classifier, and a noise robustness protocol, described below.

2.1 Model Construction

Each module described below supports one experiment. Starting from healthy neurograms generated from TIMIT speech stimuli [27] using a biophysically grounded auditory nerve model [28], we introduced four perturbation types simulating proposed ANSD mechanisms (Fig. 1): uniform temporal jitter (identical delays across frequency channels, preserving cross-frequency timing), scattered jitter (independent per-channel delays, disrupting cross-frequency synchrony), selective fiber loss (frequency-channel elimination), and truncation (amplitude saturation).

Module 1: Distortion generator. This module quantifies how each perturbation type alters neural encoding of speech using optimal transport theory [29]. The Gromov-Wasserstein distance measures whether phonemes that are acoustically similar remain neurally similar after perturbation. This approach echoes prior work modeling hearing damage effects on neural population responses [30, 31] but targets phoneme-specific disruption patterns rather than aggregate encoding quality. Encoding disruption does not guarantee perceptual failure: the auditory system might compensate downstream. Experiment 1 tests whether different mechanisms produce distinct category-specific encoding disruptions.

Module 2: Classifier. Building on [32], this module comprises a two-stage hierarchical speech recognition system: Stage 1 extracts acoustic-phonetic features from brief neurogram windows (100 ms), Stage 2 integrates predictions across longer temporal context (610 ms), roughly paralleling early auditory feature extraction and cortical integration. The two-stage design can reveal whether ANSD failures arise at the encoding level, the integration level, or from their interaction. The module also includes a misclassification analysis approach inspired by reverse correlation methods [33, 34, 35], averaging neurograms conditioned on classification outcome to reveal how perturbations physically transform the signal for specific phonemes. Experiment 2 tests whether encoding distortions translate to systematic classification failures.

Module 3: Noise robustness. This module evaluates robustness by testing individual perturbation types in isolation to establish mechanism-specific severity rankings, and by introducing environmental noise at varying signal-to-noise ratios to compare silence-trained and noise-trained models. Experiment 3 tests whether category-specific patterns persist under realistic listening conditions.

2.2 Experiment 1: Each Mechanism Distorts Specific Phoneme Categories

Optimal transport analysis quantified how ANSD perturbations alter the correspondence between acoustic formant trajectories and neural representations. The Gromov-Wasserstein (GW) distance measures whether this correspondence is preserved: higher distances indicate the neural code no longer maintains acoustic similarity structure; sounds that should be encoded similarly become neurally dissimilar. Concretely, GW distance quantifies whether phonemes that are acoustically similar (close formant trajectories) remain neurally similar after perturbation: a distance of zero indicates perfect preservation, while higher values indicate that perturbations have scrambled the relationship between acoustic and neural similarity.

Overall GW distances differed significantly across the four perturbation types (Kruskal-Wallis $H = 10.02$, $p = 0.018$; Fig. 2A). However, this between-mechanism difference is small compared to variation within mechanisms: different phoneme categories showed nearly 4-fold variation in GW distance under the same perturbation type. That is, phoneme-specific patterns carry greater diagnostic information than aggregate mechanism severity.

We analyzed GW distances both across phoneme categories and across individual formant trajectories (F1, F2, F3) to characterize category-specific and frequency-specific disruption patterns.

Category-specific patterns provide diagnostic signatures. Different ANSD mechanisms produced divergent phoneme profiles enabling mechanism identification (Table 1; Fig. 2C1–C2).

High-disruption categories. Fricatives, glottal stops, and flaps consistently showed highest GW distances. These categories are particularly affected by temporal disruption due to their brief durations, aperiodic components, and rapid spectral transitions. Flaps showed peak disruption with scattered jitter (0.070), consistent with their requirement for cross-frequency synchrony to detect the brief 20–30 ms ballistic gesture.

Low-disruption categories. Vowels, nasals, and liquids maintained low distances (0.018–0.037) across all perturbations. Their sustained, quasi-periodic structure can be extracted despite timing imprecision through temporal integration.

Perturbation-specific signatures. Uniform jitter disproportionately affected fricatives (0.052) and glottal stops (0.070), 58% higher fricative disruption than scattered jitter (0.033). This reflects uniform jitter's preservation of cross-frequency relationships: fricative identification depends on within-channel spectral shape, which uniform jitter degrades by smearing temporal structure. Scattered jitter produced the most severe flap disruption (0.070), the highest single category-perturbation combination, because flap detection requires cross-frequency coincidence that scattered jitter destroys. Loss perturbation showed highest impact on affricates (0.052), which require intact encoding across multiple frequency regions. Truncation showed most uniform effects across categories (SD = 0.015), consistent with dynamic range compression affecting all phonemes proportionally.

Category	Uniform	Scattered	Loss	Trunc
Vowel	0.028	0.023	0.033	0.032
Nasal	0.034	0.021	0.033	0.038
Liquid	0.037	0.033	0.033	0.032
Fricative	0.053	0.033	0.049	0.058
Stop	0.042	0.029	0.048	0.065
Affricate	0.051	0.037	0.052	0.047
Flap	0.083	0.070	0.071	0.068
Glottal	0.070	0.056	0.064	0.078

Table 1: Gromov-Wasserstein distances (dimensionless, relative scale) quantifying encoding disruption between acoustic formant trajectories and neural representations, by phoneme category and perturbation type. Higher values indicate greater distortion of acoustic-neural correspondence. Values range from 0.021 (Nasal, Scattered) to 0.083 (Flap, Uniform).

Formant-specific patterns. Higher formants showed greater disruption (Fig. 2B): F3 ($\sim$2500 Hz) exhibited GW distances of 0.046–0.058 across perturbations, while F1 ($\sim$500 Hz) showed lowest distances (0.035–0.039). This gradient reflects differential temporal demands: F1 tracks slow jaw movements (5–10 Hz modulations) resilient to timing imprecision; F3 encodes rapid articulatory gestures (15–25 Hz) where millisecond-scale jitter becomes comparable to modulation periods. Uniform jitter showed the steepest gradient (66% F1-to-F3 increase), while scattered jitter produced more uniform disruption across formants (24% gradient) by destroying both within-channel timing and cross-channel synchrony.

However, the 4-fold category variation in encoding distortion predicts that classification should fail differently across phoneme categories. Whether this prediction holds requires testing speech recognition directly.

2.3 Experiment 2: Encoding Distortions Produce Systematic Classification Failures

Experiment 1 established encoding distortions; Experiment 2 tests whether these translate to classification failures. If ANSD destroys temporal structure, integration that benefits healthy listeners may harm ANSD listeners by accumulating corrupted features rather than compensating for them. Cross-population transfer tests whether learned strategies remain adequate when the classifier receives neurograms from the population it was not trained on: if healthy-trained models fail on ANSD signals, they depend on features that ANSD destroys; if ANSD-trained models succeed on healthy signals, they have learned to exploit features that survive temporal degradation.

We trained four models to disentangle population and environmental effects. Two models were trained on healthy (unperturbed) neurograms: Healthy-Silence and Healthy-Noise. Two models were trained on ANSD-perturbed neurograms: ANSD-Silence and ANSD-Noise. ANSD training used mixed perturbations, each utterance received one randomly sampled perturbation type (Loss, Truncation, Uniform Jitter, or Scattered Jitter), to simulate clinical heterogeneity and ensure models learned general rather than mechanism-specific strategies. Noise-trained models incorporated additive environmental noise at SNRs uniformly sampled from 0–20 dB.

This design enables three key analyses: matched testing, cross-population transfer, and misclassification analysis. Robustness testing (Experiment 3) examines noise interactions and mechanism-specific effects. Here we report matched testing, cross-population transfer, and misclassification analysis; all results use TIMIT's held-out test set (168 speakers, no overlap with training). All category-specific accuracy differences were significant (McNemar's test, $p < 0.001$; bootstrap 95% CIs in SI Tables S6–S8).

Matched testing reveals hierarchical failure and category-specific patterns. Both populations showed hierarchical change, the difference between Stage 1 and Stage 2 accuracy, when tested on held-out speakers, but the magnitude differed significantly (Fig. 3A). Healthy-Silence models showed modest Stage 2 reduction (Stage 1: 67.0% to Stage

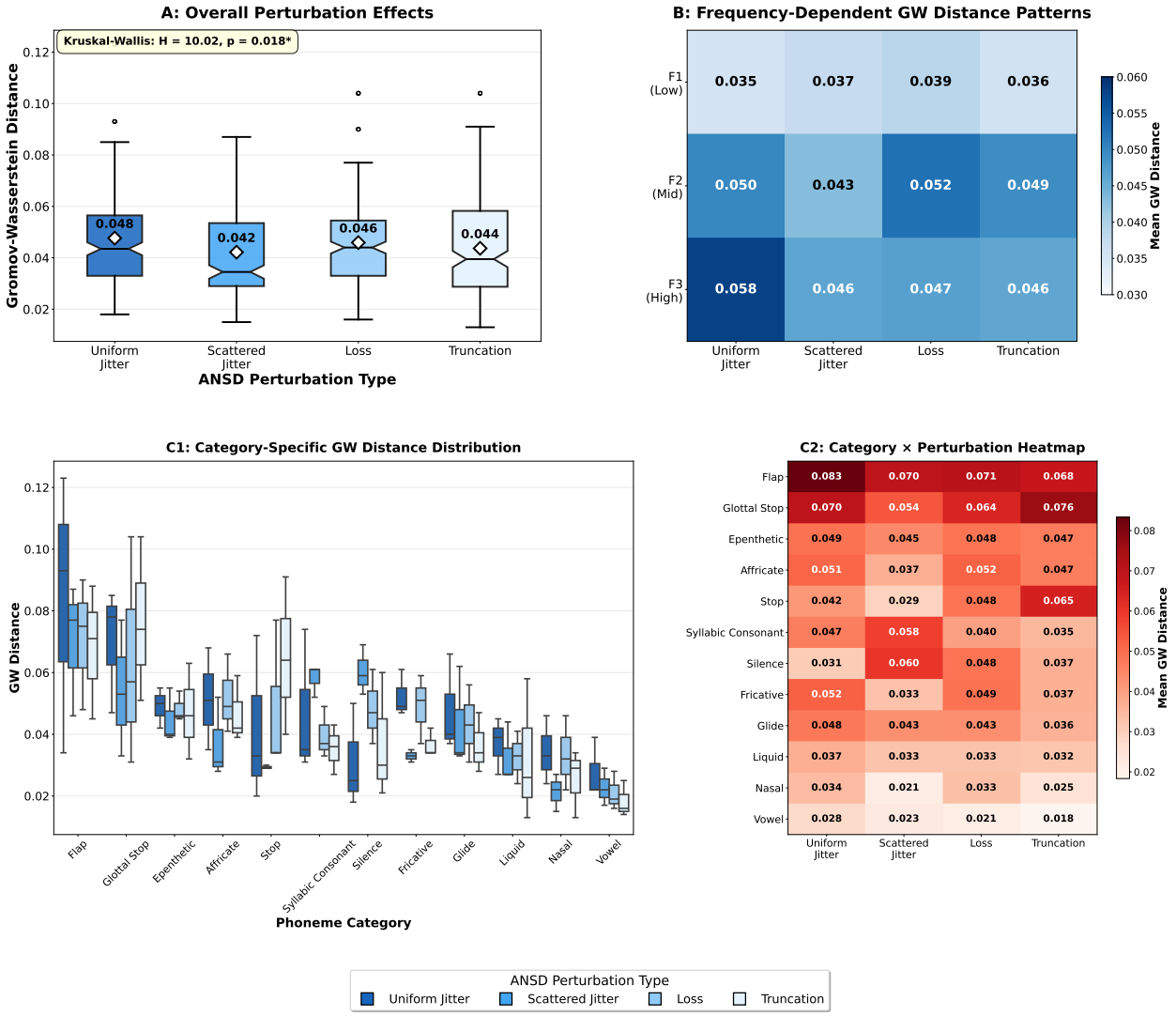

Figure 2: **Experiment 1: Optimal transport reveals phoneme-specific encoding disruptions.** **(A)** Overall perturbation effects show significant differences across ANSD mechanisms (Kruskal-Wallis $H = 10.02$, $p = 0.018$). White diamonds indicate means; boxes show interquartile range. Perturbation order: Uniform Jitter, Scattered Jitter, Loss, Truncation. **(B)** Frequency-dependent patterns: F3 shows greatest distances; uniform jitter shows steepest gradient (66% F1-to-F3 increase); loss and truncation peak at F2. **(C1)** Category-specific distribution showing diagnostic signatures across perturbation types. **(C2)** Category-by-perturbation heatmap: flaps most sensitive to uniform jitter (0.083); glottal stops to truncation (0.078); vowels and nasals preserved across all perturbations (0.021–0.033). Minor categories (epenthetic, syllabic consonant, silence, glide) shown for completeness but not discussed in text.

2: 61.0%). ANSD-Silence models showed severe degradation (48.0% to 25.5%), a 22.5-point loss, nearly four times the healthy reduction.

ANSD-Silence models achieved strong vowel accuracy (65.2%), approaching healthy performance (70.3%) and far exceeding other categories (Fig. 3C; SI Table S6 for all categories). Temporal consonants showed severe impairment that hierarchical processing could not rescue. Stops achieved only 18.2% (vs. 52.1% healthy), fricatives 20.1% (vs. 55.2%), and flaps 35.2% (vs. 80.2%). The 45-point flap deficit represents the largest category-specific gap, consistent with flaps' brief duration (20–30 ms) requiring precise cross-frequency synchrony.

Confusion matrices revealed systematic rather than random errors (Table 2). Flaps were misclassified as vowels 26.3% in ANSD versus 1.8% healthy (14.6-fold increase). Stops showed 22.1% vowel confusion (vs. 3.2%), fricatives 24.3% (vs. 2.1%; complete confusion matrices in SI Fig. S1).

Cross-population transfer reveals fundamental asymmetry. Cross-population testing revealed striking asymmetries (Fig. 3B).

Healthy models failed severely when tested on ANSD data: Stage 2 accuracy dropped from 61.0% (matched) to 28.7% (transfer) in silence, a 32-point loss. Remarkably, ANSD models *improved* when tested on healthy data, rising from 25.5% (matched) to 38.4% (transfer), a 13-point gain.

This asymmetry is unidirectional: strategies learned on degraded signals generalize to intact signals, but not the reverse.

Misclassification analysis reveals a stop-to-vowel transformation mechanism. To identify spectrotemporal features driving the consonant-to-vowel confusions documented above, we computed classification-conditioned templates: average neurograms for tokens grouped by predicted category and correctness. Because the classifier learns decision boundaries in neural space, averaging neurograms conditioned on classification outcome recovers the prototypical pattern underlying each decision. Comparing correctly and incorrectly classified templates therefore isolates the spectrotemporal changes that drive representations across category boundaries.

We focused on stop /t/ (canonical high-frequency burst, prominent in confusions) and vowels /aa/, /ih/, /iy/ spanning distinct formant configurations (additional consonants in SI Fig. S5; perturbation-specific templates in SI Fig. S3).

When healthy models correctly classified /t/, templates showed intense, brief broadband bursts (0–25 ms, 2–10 kHz). When ANSD models misclassified /t/ as vowel (22.1% of tokens), templates showed reduced early burst energy (0–50 ms) accompanied by sustained activity extending 50–100 ms into formant-frequency regions (Fig. 4A).

This consonant-to-vowel bias was not limited to /t/ but generalized across manner classes (Table 2).

True → Predicted	**H-Sil**	**A-Sil**	**Ratio**
Flap → Vowel	1.8%	26.3%	14.6×
Stop → Vowel	3.2%	22.1%	6.9×
Fricative → Vowel	2.1%	24.3%	11.6×
Nasal → Vowel	4.5%	31.2%	6.9×
Vowel → Silence	12.3%	23.3%	1.9×

Table 2: Systematic confusion patterns showing consonant-to-vowel transformation in ANSD. H-Sil: Healthy-Silence; A-Sil: ANSD-Silence. Ratio indicates fold-increase in ANSD relative to healthy. Flaps show greatest transformation (14.6×), consistent with their brief duration and cross-frequency synchrony requirements.

The difference map (ANSD confused minus Healthy correct) revealed a redistribution of energy from early high-frequency transients (blue, 0–50 ms) to later low-frequency sustained structure (red, 50–100 ms, 500–3000 Hz), consistent with a shift from stop-like to vowel-like representation.

Importantly, healthy confused and ANSD confused templates were highly similar. ANSD does not generate a novel error pattern; rather, temporal perturbations systematically shift consonant representations into neural regions already associated with vowel decisions.

Vowel templates (/aa/, /ih/, /iy/) showed ANSD-induced distortions including high-frequency reduction and altered formant structure, yet preserved sufficient spectral organization for reliable classification (Fig. 4B).

2.4 Experiment 3: Mechanism-Specific Patterns Persist in Noise

Experiment 2 established category-specific failures using silent training conditions. Experiment 3 tested robustness: how noise training interacts with temporal disruption, whether deficits persist across SNR levels, and which individual perturbation mechanisms pose the greatest challenges. If ANSD simply adds internal noise, noise training should help both populations equally by teaching robustness to stochastic variability. Divergent responses would indicate that ANSD involves structural signal degradation, not just increased variability.

Noise-timing interaction. Noise training and temporal disruption interact: the same training regime that benefits healthy populations harms ANSD populations (Fig. 3A,D). Healthy models showed hierarchical benefit under noise (Stage 1: 58.5% to Stage 2: 63.2%), while ANSD models showed further degradation (41.3% to 17.1%). Noise training improved healthy vowel accuracy (from 70.3% to 75.1%) but degraded ANSD vowel accuracy (from 65.2% to 35.4%), a 30-point drop.

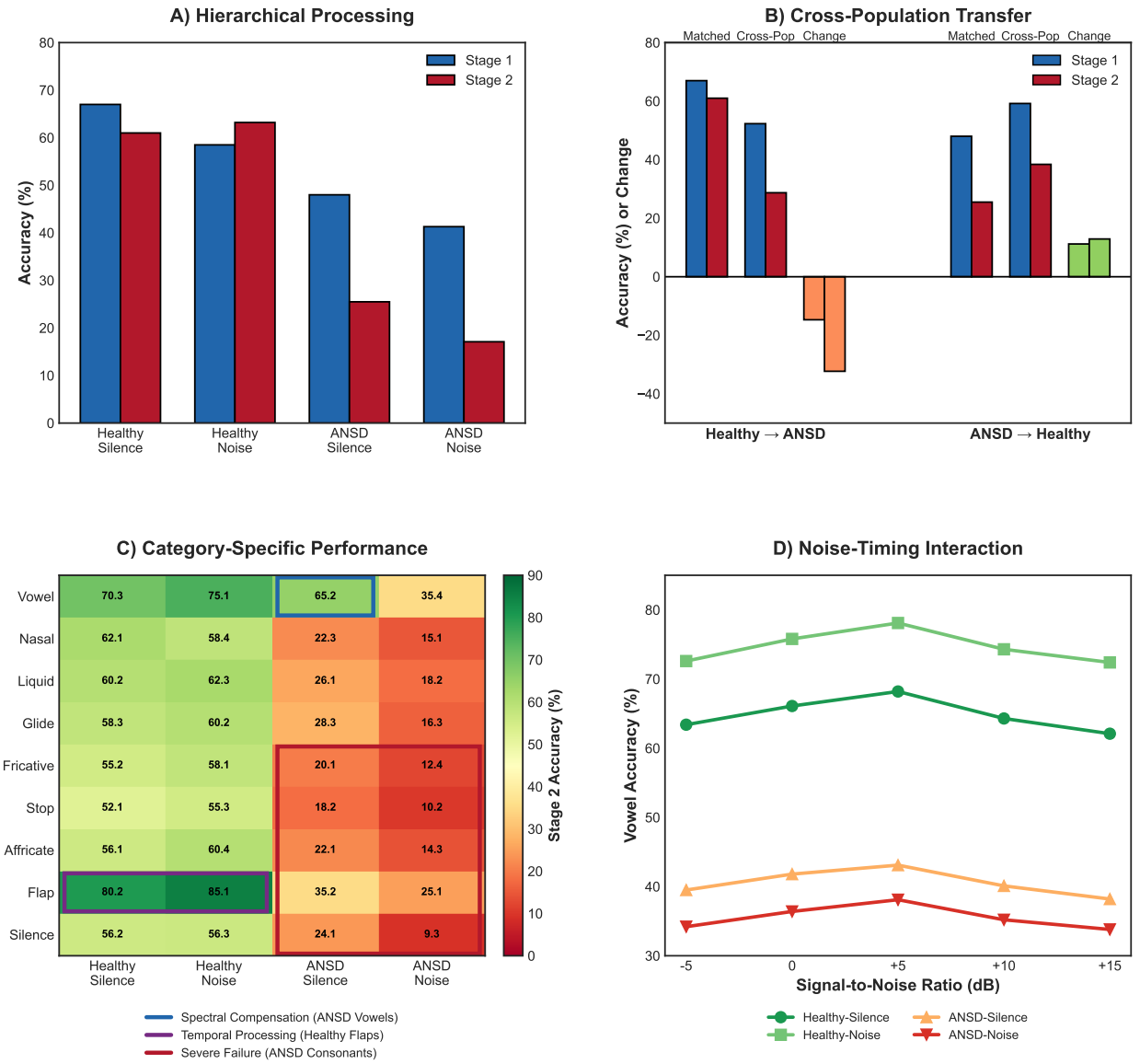

Figure 3: **Speech recognition reveals hierarchical failure, asymmetric transfer, and noise interactions (Experiments 2 and 3). (A)** Hierarchical processing on held-out speakers: Healthy-Silence models show modest reduction; ANSD-Silence models show severe degradation. Noise conditions discussed in Experiment 3. **(B)** Cross-population transfer: Healthy models fail on ANSD data; ANSD models improve on healthy data. **(C)** Category-specific patterns: ANSD achieves strong vowel discrimination (65.2%, blue box) but shows severe consonant failure (red box). Healthy flaps show intact temporal processing (purple box). **(D)** Noise-timing interaction (Experiment 3): Healthy models benefit from noise training; ANSD models are harmed.

This opposite effect demonstrates that ANSD does not simply add internal noise. Noise training reinforces temporal processing strategies that ANSD signals cannot support, which may explain why the performance gap widens under noise (35.5 points in silence to 46.1 points in noise).

The transfer asymmetry confirmed this pattern: healthy-trained models' failure on ANSD data grew from −32.3 points in silence to −43.8 under noise, while ANSD-trained models' success on healthy data remained stable (+12.9 to +13.5 points; SI Table S12).

Persistent deficits across SNR levels. Testing across signal-to-noise ratios (+15 to −5 dB) confirmed the Healthy-ANSD performance gap persists regardless of environmental noise level (Fig. 3D). The gap remains approximately

Population	Loss	Trunc	Uniform	Scattered
Stage 2 Accuracy (%)				
Healthy-Sil	73.5	62.7	59.8	57.5
Healthy-Noi	65.1	58.5	54.0	51.2
ANSD-Sil	54.1	26.1	22.7	19.3
ANSD-Noi	46.8	18.4	14.1	10.8
Hierarchical Δ (S2−S1, percentage points)				
Healthy-Sil	+2.3	−5.4	−5.5	−6.3
Healthy-Noi	+2.7	−1.3	−2.3	−2.9
ANSD-Sil	+1.9	−23.6	−24.1	−24.6
ANSD-Noi	+1.4	−24.4	−25.1	−25.9

Table 3: Mechanism-specific severity hierarchy. Models trained with mixed perturbations (ANSD) or healthy neurograms (Healthy) were tested on individual perturbation types applied in isolation. Stage 2 accuracy and hierarchical change (S2−S1) shown. Loss uniquely preserves hierarchical benefit across all populations; temporal perturbations produce failure, with scattered jitter most severe. Complete category-by-category accuracy breakdowns in SI Table S11.

constant across all SNRs. Noise training provided consistent benefit to healthy vowel accuracy (+10 points across SNRs) but consistently harmed ANSD vowel accuracy (−5 points; SI Table S13).

Mechanism-specific severity hierarchy. The preceding analyses used mixed-perturbation training and testing. To isolate effects of individual ANSD mechanisms, the same four models were tested on single perturbation types applied in isolation (Table 3; see SI Table S11 for complete category-by-category accuracy breakdowns). Testing healthy models on perturbations provides a baseline; comparing to ANSD models reveals which mechanisms benefit most from adapted training.

This established a severity ranking (Table 3): Loss < Truncation < Uniform Jitter < Scattered Jitter. Loss perturbation uniquely showed hierarchical benefit across all four populations, the only perturbation where Stage 2 exceeded Stage 1. Temporal disruptions showed negative hierarchical Δ, with scattered jitter most severe (SI Table S11 for category breakdowns; statistical validation in SI Table S10).

3 Discussion

Three experiments establish that ANSD mechanisms produce distinct phoneme-specific signatures that cascade from encoding disruption (Experiment 1) through classification failure (Experiment 2) to persistent deficits under noise (Experiment 3). Critically, we discovered a fundamental asymmetry: strategies learned on degraded signals generalize to intact signals, but not the reverse (Fig. 3B; SI Table S12). This asymmetry arises from a single constraint: the features extracted at Stage 1 determine what Stage 2 can accomplish.

Mechanism: Stage 1 features determine Stage 2 outcomes. Models trained on healthy signals learn to extract timing-dependent features at Stage 1; models trained on ANSD signals learn to extract timing-independent features, because precise timing is unavailable. Timing-independent features generalize to intact signals (where timing is present but unnecessary), while timing-dependent features fail on degraded signals (where the timing they require has been destroyed). Stage 2 then integrates whatever Stage 1 provides; when those features are misaligned with the signal, integration amplifies errors rather than correcting them. Here, signals are made of phonemes that use different time scales. Fast transient cues (brief consonant bursts and rapid formant transitions distinguishing voiceless stops /t/, /k/, /p/) require 20–50 ms temporal precision; conversely, stable formant patterns with more slowly varying temporal structure persisting over tens of milliseconds survive temporal integration better than features that require millisecond-scale coordination. For most of their duration, vowels can be identified using purely spectral cues. It is thus possible to hypothesize that ANSD-trained models learn to exploit slow temporal structure because fast temporal cues are unavailable, whereas healthy-trained models efficiently exploit fast transient cues.

The mechanism-specific breakdown confirms this interpretation (Table 3). Among ANSD perturbations, only fiber loss preserves hierarchical benefit (+1.9 points); all three temporal perturbations produce severe hierarchical failure (−23.6 to −24.6 points). The key distinction is whether timing survives: fiber loss removes channels but preserves temporal precision within surviving channels, while jitter and truncation destroy the timing relationships that hierarchical integration requires.

This pattern has a structural parallel in predictive coding. Giraldi et al. [36] showed that phonologically plausible linguistic priors can override conflicting acoustic input, inducing listeners to perceive speech consistent with their expectations rather than the actual signal. Our Stage 1 features play an analogous role: they constitute a learned representation that determines what Stage 2 can accomplish. The key difference is directionality: Giraldi's effect is top-down (prior overrides sensory input), while our constraint is feedforward (Stage 1 output propagates to Stage 2 with no descending connections). But both demonstrate that processing outcomes depend critically on whether the operative representation matches the signal's actual structure.

Noise training confirms the asymmetry. The Stage 1 constraint predicts that noise training should have opposite effects in the two populations, and it does (Fig. 3A; Table 3). In healthy models, noise training reverses the direction of hierarchical change, from -6.0 percentage points in silence to +4.7 points under noise, indicating that hierarchical integration becomes beneficial only when the signal is degraded. In ANSD models, hierarchical processing is already strongly detrimental in silence and worsens further under noise (from -22.5 to -24.2 points). Statistical validation of all category-specific differences (McNemar's test with bootstrap 95% CIs) is provided in SI Tables S6–S10; significance thresholds vary by category sample size, from approximately 3 percentage points for vowels (N=3506) to 18 points for flaps (N=149).

This divergence confirms that the two populations rely on fundamentally different feature representations, as established in Experiment 3.

Phoneme-specific signatures enable mechanism-specific diagnosis. The Stage 1 constraint also explains why different perturbations produce distinct phoneme profiles. We predicted that phoneme categories differ systematically in their temporal encoding demands: vowels with sustained formant structure should survive temporal degradation through temporal integration, whereas brief consonants requiring precise timing should be most susceptible. At the encoding stage (Experiment 1), optimal transport analysis confirmed this prediction (Table 1; Fig. 2): phoneme categories relying on fast temporal coordination, particularly flaps and stops, were consistently the most disrupted, whereas vowels with slower spectrotemporal structure were relatively preserved across mechanisms. Crucially, variability across phoneme categories was nearly four-fold within a single perturbation type (e.g., uniform jitter: flaps 0.083 vs. nasals 0.021), far exceeding the 14% variation observed between mechanisms. This indicates that phoneme-level signatures carry more diagnostic information than aggregate performance measures.

The perturbation-specific patterns aligned with the slow/fast distinction. Uniform jitter disproportionately disrupted fricatives and flaps, but for different reasons. Fricatives are sustained sounds, yet distinguishing sibilants (/s/, /sh/) from non-sibilants (/f/, /th/) requires resolving spectral shape, i.e. the distribution of turbulent noise energy across frequency channels. Spectral shape is encoded by relative firing rates across channels, but accurate rate estimation depends on temporal precision *within* each channel: jitter smears the spike patterns from which channel-specific energy is estimated, degrading spectral resolution even for sustained signals. Uniform jitter specifically targets this mechanism because it preserves cross-frequency relationships (irrelevant for fricative identity) while degrading within-channel temporal structure (essential for spectral shape). Flaps failed for a different reason: their single ballistic tongue gesture lasts only 20–30 ms, requiring cross-frequency coincidence detection to capture the brief spectral transient. Scattered jitter, which destroys cross-frequency synchrony, accordingly produced the largest flap disruption (0.070), consistent with disruption of brainstem coincidence detection circuitry.

Loss perturbation most affected affricates, which combine a stop closure with subsequent frication and therefore require intact encoding across multiple frequency regions; loss uniquely preserved hierarchical benefit (Table 3), consistent with temporal structure remaining intact within surviving channels so that Stage 2 integration can interpolate across spectral gaps. Truncation produced the most uniform effects across categories, consistent with dynamic range compression affecting all phonemes proportionally rather than targeting specific temporal or spectral features.

These encoding distortions translated into classification failures (Experiment 2; Fig. 3B–C). Cross-population transfer quantified the asymmetry: models trained on healthy signals failed on ANSD data (-32.3 points), while models trained on ANSD signals improved on healthy data (+12.9 points), supporting the interpretation that timing-independent features generalize whereas timing-dependent features do not. Misclassification analysis revealed the mechanism (Fig. 4): temporal perturbations reduce the brief, broadband burst characterizing stops and replace it with sustained activity in formant regions, effectively shifting consonant representations toward vowel-like sustained structure. The resulting consonant-to-vowel confusions were systematic rather than random (Table 2).

Importantly, these patterns were stable under noise (Experiment 3; Fig. 3D). Performance gaps remained constant across SNR levels rather than converging, and the transfer asymmetry widened: healthy-trained models' failure on ANSD data grew from -32.3 points in silence to -43.8 under noise, while ANSD-trained models' success on healthy data remained stable (+12.9 to +13.5 points). This reinforces the conclusion that the underlying distortions are structural rather than stochastic.

Convergence with prior work. Our framework complements parameter-based auditory nerve models [24] by directly manipulating neurogram representations to instantiate hypothesized ANSD mechanisms. Several recent findings converge with ours: Garrett et al. [37] showed that synaptopathy selectively degrades temporal envelope coding at suprathreshold levels, paralleling our finding that fiber loss preserves hierarchical benefit while producing consonant deficits. Vasilkov et al. [38] demonstrated that spontaneous rate-specific fiber loss degrades word recognition using Zilany-based simulations, with constrained decoder architectures proving critical for revealing biologically meaningful deficits. Saddler & McDermott [39] showed that limiting phase-locking capacity alters speech recognition in task-dependent ways, aligning with our finding that temporal mechanisms dominate ANSD-related degradation. Carney's neural fluctuation contrast theory [40] provides a mechanistic basis for our frequency-dependent effects, emphasizing that high-frequency regions depend more critically on temporal envelope coding due to inherently weaker phase-locking. Viswanathan et al. [41] showed that confusion patterns distinguish processing strategies, supporting our claim that phoneme-specific patterns carry diagnostic information that aggregate scores obscure. Trinesh and Uppunda [42] showed that cochlear hearing loss and auditory neuropathy produce distinct cortical evoked potential signatures, with divergent activation patterns during early sound detection but convergent processing at later evaluation stages, providing electrophysiological evidence that different peripheral pathologies generate distinguishable central consequences. Giraud and Poeppel's [43] proposal that precise temporal scaffolding supports hierarchical linguistic computation connects directly to our finding that temporal perturbations selectively disrupt Stage 2 integration. More broadly, Ruggles et al. [44] demonstrated that clinically identical audiograms conceal large individual differences in suprathreshold temporal encoding, reinforcing the principle that aggregate measures systematically obscure encoding-level variation. Kaur et al. [45] provided direct histopathological confirmation: in a multivariable regression across multiple hearing loss etiologies, auditory nerve peripheral axon survival was the only significant histological predictor of word scores once audiometric thresholds were accounted for, demonstrating that neural degeneration invisible to the audiogram drives intelligibility deficits.

Clinical implications. Current phonetically balanced speech tests represent the best available clinical method: normative data enable identification of patients performing outside normal ranges, and phonetic balancing maximizes sensitivity to any defect. However, this sensitivity comes at the cost of dilution by correct answers to phonemes unaffected by the specific mechanism, explaining why aggregate scores obscure the category-specific patterns our results reveal (within-perturbation variation nearly 4-fold, versus 14% between-perturbation variation). Building mechanism-specific test batteries faces a circularity: validating a jitter-specific list requires identified jitter patients, which presupposes the diagnostic capacity that does not yet exist. Our computational framework addresses this bottleneck by providing the mechanistic mapping needed to interpret existing behavioral data. This has two immediate applications: first, understanding what ANSD patients naturally compensate and what they cannot is essential before planning any targeted rehabilitation—our results predict that temporal training could degrade performance by 30 percentage points in jitter patients, making mechanism identification a prerequisite for safe intervention. Second, signal processing algorithms can be systematically evaluated against specific mechanisms: for instance, temporal enhancement may benefit fiber loss patients but harm those with jitter, enabling algorithm selection matched to pathophysiology [46].

Limitations. This study modeled high-SR fibers exclusively; incorporating mixed-SR populations would better assess distributed synaptic loss. The neurogram perturbation approach directly manipulates neural representations without modeling the biophysical mechanisms that produce them; bridging to parameter-based models [24, 28] would enable predictions from histological data. The ASR classifier does not capture linguistic context, top-down predictions, or perceptual learning that human listeners exploit, though it serves as an ideal observer revealing which acoustic-phonetic information survives neural degradation. Our misclassification analysis reveals how ANSD transforms the signal but does not directly probe the classifier's decision strategy; reverse correlation methods [47] could provide independent confirmation. Furthermore, our ASR architecture is purely feedforward; computational frameworks involving top-down predictions [36] capture mechanisms our model cannot.

Validating these predictions admits two paths. The first requires collecting detailed confusion matrices from ANSD patients with characterized pathophysiology: our framework predicts distinct patterns (scattered jitter produces disproportionate flap confusions; fiber loss preserves hierarchical benefit; uniform jitter shows stop-to-vowel transformations), and testing requires clinical partnerships combining electrophysiological characterization with phoneme-specific behavioral assessment, which is technically feasible but time-intensive. A complementary approach bypasses the validation bottleneck entirely: rather than testing existing predictions against patient data, the computational model can be used prospectively to design new speech tests. By identifying stimulus contrasts that maximally discriminate between perturbation types (for instance, minimal pairs where jitter predicts vowel confusion but loss predicts correct identification), one could generate mechanism-specific test batteries optimised for diagnostic contrast, then validate these targeted materials clinically. Critically, this path requires no neural recording from patients, only detailed scoring of existing speech tests. A fundamental challenge for clinical translation is the inversion problem: while our framework predicts which confusion patterns each perturbation type produces, clinical diagnosis requires inferring mechanisms

from observed confusions, a harder task given that multiple mechanisms may produce overlapping patterns and real patients likely present with mixed pathology.

Conclusions. This study establishes that the same intervention, noise training, reveals opposite hierarchical regimes in healthy versus ANSD listeners (Fig. 3A). The fundamental constraint: compensation based on slow spectrotemporal structure generalizes, while strategies requiring fast temporal coordination fail when timing is degraded. This predicts why patients may develop functional strategies in quiet but cannot overcome deficits under noise. Phoneme-specific confusion patterns, varying nearly 4-fold within perturbation types (Table 1), can reveal underlying pathophysiology that aggregate testing obscures. This framework enables the translational pathway: mechanism identification from behavioral testing unlocks mechanism-specific intervention research, patient stratification, and targeted rehabilitation, none of which is possible without first diagnosing which patients have which mechanism.

4 Materials and Methods

Auditory nerve modeling. Auditory nerve responses were generated using the Zilany et al. model [28]: 150 logarithmically-spaced characteristic frequencies (125 Hz–10 kHz), high spontaneous rate fibers (100 spikes/s), 65 dB SPL stimulus level. Temporal resolution: 2 ms (500 Hz sampling). Speech stimuli drawn from the TIMIT corpus using standardized train/test splits; phonemes grouped into 9 categories (Vowel, Nasal, Liquid, Glide, Fricative, Stop, Affricate, Flap, Silence).

Speech stimuli. Speech materials were drawn from the TIMIT corpus [27], using the standard train/test partition: 462 training speakers, 168 test speakers, no overlap between sets. Phonemes were grouped into 9 categories (Vowel, Nasal, Liquid, Glide, Fricative, Stop, Affricate, Flap, Silence). For noise-trained models, additive noise from the WHAM! corpus [48] was mixed with speech at SNRs uniformly sampled from 0–20 dB.

Neural perturbations. Four types modeled ANSD mechanisms: *Uniform jitter*: delay $\delta_t \sim \mathcal{U}(3, 10)$ ms applied identically across all frequency channels (preserving cross-frequency timing as reference condition). *Scattered jitter*: independent delays $\delta_{t,f} \sim \mathcal{U}(3, 10)$ ms per channel (destroying cross-frequency synchrony). *Fiber loss*: 1–4 randomly selected channels zeroed. *Truncation*: amplitude limited at ratio $\alpha \sim \mathcal{U}(0.3, 0.7)$ of maximum. See SI Appendix for mathematical definitions and physiological rationale.

Optimal transport analysis. We asked whether acoustic similarity structure is preserved in neural representations: if two phonemes sound similar, do they produce similar neural responses? The challenge is that acoustic features (formant frequencies in Hz) and neural responses (firing rates in spikes/s) live in different spaces and cannot be directly compared. Gromov-Wasserstein (GW) distance [49, 29] solves this by comparing *relationships* rather than values. For example, if /aa/ and /ah/ are acoustically close while /aa/ and /iy/ are far apart, GW asks whether this same pattern holds in the neural code. Unlike correlation, GW is not bounded within a specific range; it is a distance where zero means perfect preservation and larger values indicate greater distortion. To illustrate: a GW distance of 0.02 for vowels versus 0.07 for flaps tells us that flap encoding is more severely disrupted, even though neither number has meaning in isolation.

For each phoneme category and perturbation type, we computed GW distance between two similarity matrices: one capturing pairwise acoustic distances (from formant trajectories), the other capturing pairwise neural distances (from neurograms). This yields a single number quantifying how much each perturbation disrupts the acoustic-neural correspondence for each phoneme category. Computational details are provided in SI Appendix.

Speech recognition architecture. Following [32], we used a two-stage hierarchical automatic speech recognition system: Stage 1 extracts acoustic-phonetic features from brief windows; Stage 2 integrates predictions across longer temporal context. Stage 1 (456K parameters): three-layer unidirectional GRU (128 units), multi-head attention (4 heads), dense layers with dropout (0.3); input: 50-frame windows (100 ms). Stage 2 (406K parameters): three-layer bidirectional GRU (192, 192, 128 units), multi-head attention (6 heads), dense layers with dropout (0.4); input: 305-frame sequences (610 ms context). Output: 9 phoneme categories. Training: Adam optimizer, focal loss ($\gamma = 2$), early stopping. See SI Appendix for full training procedures and hyperparameters.

Misclassification analysis. The ASR results showed that ANSD perturbations cause systematic consonant-to-vowel confusions, but they do not reveal which neural features cause these errors. To investigate the patterns in the neural representation that drive these misclassifications, we adapted a reverse-correlation approach [34]. Reverse correlation uncovers the input features that trigger specific outputs: by averaging the neurograms that lead the ASR system to predict a particular output category, one can recover the characteristic neural pattern, or "template", associated with that prediction.

Because our ASR outputs a probability for each category rather than a single choice, we computed a weighted average of neurograms for each true–predicted phoneme pairing. Tokens that the ASR classifies with high confidence contribute more, while ambiguous cases contribute less, emphasizing the patterns that actually drive the model's decisions. Formally, the template for true phoneme ϕ classified as category c is

$$\mathbf{T}_{c,\phi} = \frac{\sum_i P(c|\mathbf{X}_i)\mathbf{X}_i}{\sum_i P(c|\mathbf{X}_i)},$$

where the sum runs over all tokens of phoneme ϕ, and $P(c|\mathbf{X}_i)$ is the ASR's probability for category c given neurogram $\mathbf{X}_i$.

Intuitively, $\mathbf{T}_{c,\phi}$ represents the *typical neural pattern of phoneme ϕ that causes the ASR to predict category c.* For example, the template for "/t/ misclassified as vowel" highlights which features of /t/ make it appear vowel-like to the classifier. By computing difference maps between templates derived from distinct population and classification-outcome conditions, we quantified how ANSD perturbations redistribute energy across time and frequency, revealing the neural transformations responsible for these confusions. Analysis focused on stop /t/ and vowels /aa/, /ih/, /iy/; additional consonants are reported in the SI Appendix.

Statistical testing. Significance of category-specific accuracy differences was assessed using McNemar's test on paired token-level predictions within each phoneme category. For each comparison (e.g., Healthy-Silence vs ANSD-Silence on stop consonants), a 2×2 contingency table categorised tokens by whether each model classified them correctly. Bootstrap 95% confidence intervals were computed using 10,000 resamples with replacement (percentile method), preserving token pairing to ensure the same tokens appear in both model predictions within each resample. All reported comparisons were significant at $p < 0.001$ unless otherwise noted. Complete statistical results are provided in SI Tables S6–S10.

Data Availability

TIMIT corpus available from Linguistic Data Consortium (LDC93S1). WHAM! noise corpus at `http://wham.whisper.ai/`. Analysis code and trained models available at `https://github.com/mcampi111/SimsACues_ASR`.

Contributions

M.C.: Conceptualization, data curation, formal analysis, investigation, methodology, software, writing—original draft, writing—review & editing.
E.P.: Conceptualization, formal analysis, investigation, methodology, writing—original draft, writing—review & editing.
G.G.: Formal analysis, investigation, writing—review & editing.
P.A.: Conceptualization, investigation, supervision, funding acquisition, writing—review & editing.
C.G.: Conceptualization, formal analysis, investigation, methodology, writing—original draft, writing—review & editing.

Acknowledgments

This work was supported by a grant from Fondation Pour l'Audition (FPA) to the Institut de l'Audition (FPAIDA09 to Hung Thai-Van and FPAIDA10 to Paul Avan and the Ceriah) and a French government grant managed by the Agence Nationale de la Recherche under the France 2030 program, reference ANR-23-IAHU-0003. CG was supported by the Pasteur-Roux-Cantarini Fellowship from Institut Pasteur.

MC thanks Prof. Birger Kollmeier for guidance, Dr. Léo Varnet and Dr. Emmanuel Ponsot for template analysis feedback, and Prof. Romain Serizel for ASR suggestions.

References

[1] World Health Organization et al. *World report on hearing*. World Health Organization, 2021.

[2] Bridget Shield. Evaluation of the social and economic costs of hearing impairment. *Hear-it AISBL*, pages 1–202, 2006.

[3] Wade Chien and Frank R Lin. Prevalence of hearing aid use among older adults in the united states. *Archives of internal medicine*, 172(3):292–293, 2012.

[4] Astrid Foerst, Dirk Beutner, Ruth Lang-Roth, Karl-Bernd Huttenbrink, Hasso von Wedel, and Martin Walger. Prevalence of auditory neuropathy/synaptopathy in a population of children with profound hearing loss. *International journal of pediatric otorhinolaryngology*, 70(8):1415–1422, 2006.

[5] Charles I Berlin, Linda J Hood, Thierry Morlet, Diane Wilensky, Li Li, Kelly Rose Mattingly, Jennifer Taylor-Jeanfreau, Bronya JB Keats, Patti St John, Elizabeth Montgomery, et al. Multi-site diagnosis and management of 260 patients with auditory neuropathy/dys-synchrony (auditory neuropathy spectrum disorder*). *International journal of audiology*, 49(1):30–43, 2010.

[6] Arnold Starr, Terence W Picton, Yvonnc Sininger, Linda J Hood, and Charles I Berlin. Auditory neuropathy. *Brain*, 119(3):741–753, 1996.

[7] Rosamaria Santarelli, Pietro Scimemi, Chiara La Morgia, Elona Cama, Ignacio Del Castillo, and Valerio Carelli. Electrocochleography in auditory neuropathy related to mutations in the otof or opa1 gene. *Audiology Research*, 11(4):639–652, 2021.

[8] Romolo Daniele De Siati, Flora Rosenzweig, Guillaume Gersdorff, Anaïs Gregoire, Philippe Rombaux, and Naïma Deggouj. Auditory neuropathy spectrum disorders: from diagnosis to treatment: literature review and case reports. *Journal of clinical medicine*, 9(4):1074, 2020.

[9] Fan-Gang Zeng, Sandy Oba, Smita Garde, Yvonne Sininger, and Arnold Starr. Temporal and speech processing deficits in auditory neuropathy. *Neuroreport*, 10(16):3429–3435, 1999.

[10] Fan-Gang Zeng, Ying-Yee Kong, Henry J Michalewski, and Arnold Starr. Perceptual consequences of disrupted auditory nerve activity. *Journal of neurophysiology*, 93(6):3050–3063, 2005.

[11] Roland Schaette and David McAlpine. Tinnitus with a normal audiogram: physiological evidence for hidden hearing loss and computational model. *Journal of Neuroscience*, 31(38):13452–13457, 2011.

[12] Hari M Bharadwaj, Salwa Masud, Golbarg Mehraei, Sarah Verhulst, and Barbara G Shinn-Cunningham. Individual differences reveal correlates of hidden hearing deficits. *Journal of Neuroscience*, 35(5):2161–2172, 2015.

[13] Kelsie J Grant, Aravindakshan Parthasarathy, Viacheslav Vasilkov, Benjamin Caswell-Midwinter, Maria E Freitas, Victor de Gruttola, Daniel B Polley, M Charles Liberman, and Stéphane F Maison. Predicting neural deficits in sensorineural hearing loss from word recognition scores. *Scientific reports*, 12(1):8929, 2022.

[14] Rosamaria Santarelli. Information from cochlear potentials and genetic mutations helps localize the lesion site in auditory neuropathy. *Genome Medicine*, 2(12):91, 2010.

[15] Guoqiang Wan and Gabriel Corfas. Transient auditory nerve demyelination as a new mechanism for hidden hearing loss. *Nature Communications*, 8:14487, 2017.

[16] Jerzy E Rose, John F Brugge, David J Anderson, and Joseph E Hind. Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey. *Journal of neurophysiology*, 30(4):769–793, 1967.

[17] Philip X Joris, Christoph E Schreiner, and Adrian Rees. Neural processing of amplitude-modulated sounds. *Physiological reviews*, 84(2):541–577, 2004.

[18] Sharon G Kujawa and M Charles Liberman. Adding insult to injury: cochlear nerve degeneration after "temporary" noise-induced hearing loss. *Journal of Neuroscience*, 29(45):14077–14085, 2009.

[19] M Charles Liberman and Sharon G Kujawa. Cochlear synaptopathy in acquired sensorineural hearing loss: Manifestations and mechanisms. *Hearing research*, 349:138–147, 2017.

[20] Craig A Buchman, Patricia A Roush, Holly FB Teagle, Carolyn J Brown, Carlton J Zdanski, and John H Grose. Auditory neuropathy characteristics in children with cochlear nerve deficiency. *Ear and hearing*, 27(4):399–408, 2006.

[21] Joaquin T Valderrama, Angel De la Torre, and David McAlpine. The hunt for hidden hearing loss in humans: From preclinical studies to effective interventions. *Frontiers in neuroscience*, 16:1000304, 2022.

[22] Richard Salvi, Wei Sun, Dalian Ding, Guang-Di Chen, Edward Lobarinas, Jian Wang, Kelly Radziwon, and Benjamin D Auerbach. Inner hair cell loss disrupts hearing and cochlear function leading to sensory deprivation and enhanced central auditory gain. *Frontiers in Neuroscience*, 10:621, 2017.

[23] Brian C. J. Moore. Dead regions in the cochlea: Diagnosis, perceptual consequences, and implications for the fitting of hearing aids. *Trends in Amplification*, 5(1):1–34, 2001.

[24] Maral Budak, Karl Grosh, Aritra Sasmal, Gabriel Corfas, Michal Zochowski, and Victoria Booth. Contrasting mechanisms for hidden hearing loss: Synaptopathy vs myelin defects. *PLoS computational biology*, 17(1):e1008499, 2021.

[25] Judy R. Dubno, Donald D. Dirks, and L. R. Langhofer. Evaluation of hearing-impaired listeners using a nonsense-syllable test. II. Syllable recognition and consonant confusion patterns. *Journal of Speech and Hearing Research*, 25(1):141–148, 1982.

[26] Elmer Owens and Earl D. Schubert. Development of the California Consonant Test. *Journal of Speech and Hearing Research*, 20(3):463–474, 1977.

[27] John S Garofolo, Lori F Lamel, William M Fisher, David S Pallett, Nancy L Dahlgren, Victor Zue, and Jonathan G Fiscus. TIMIT acoustic-phonetic continuous speech corpus. Linguistic Data Consortium, Philadelphia, 1993. LDC93S1.

[28] Muhammad SA Zilany, Ian C Bruce, and Laurel H Carney. Updated parameters and expanded simulation options for a model of the auditory periphery. *The Journal of the Acoustical Society of America*, 135(1):283–286, 2014.

[29] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[30] Sarah Verhulst, Alessandro Altoè, and Viacheslav Vasilkov. Computational modeling of the human auditory periphery: Auditory-nerve responses, evoked potentials and hearing loss. *Hearing Research*, 360:55–75, 2018.

[31] Ian C Bruce, Yousof Erfani, and Muhammad SA Zilany. A phenomenological model of the synapse between the inner hair cell and auditory nerve: Implications of limited neurotransmitter release sites. *Hearing research*, 360:40–54, 2018.

[32] Tim Brochier, Josef Schlittenlacher, Iwan Roberts, Tobias Goehring, Chen Jiang, Deborah Vickers, and Manohar Bance. From microphone to phoneme: an end-to-end computational neural model for predicting speech perception with cochlear implants. *IEEE Transactions on Biomedical Engineering*, 69(11):3300–3312, 2022.

[33] Léo Varnet, Kenneth Knoblauch, Fanny Meunier, and Michel Hoen. Using auditory classification images for the identification of fine acoustic cues used in speech perception. *Frontiers in Human Neuroscience*, 7:865, 2013.

[34] Albert J Ahumada Jr. Classification image weights and internal noise level estimation. *Journal of Vision*, 2(1):8, 2002.

[35] Richard F Murray. Classification images: A review. *Journal of Vision*, 11(5):2, 2011.

[36] T. Giraldi et al. Cross-linguistic auditory illusions reveal phonological priors in speech perception. 2025. In press.

[37] Markus Garrett, Viacheslav Vasilkov, Manfred Mauermann, Pauline Devolder, John L Wilson, Leslie Gonzales, Kenneth S Henry, and Sarah Verhulst. Deciphering compromised speech-in-noise intelligibility in older listeners: the role of cochlear synaptopathy. *eneuro*, 12(2), 2025.

[38] Viacheslav Vasilkov, M. Charles Liberman, and Stéphane F. Maison. Modeling the impact of cochlear nerve degeneration on speech recognition performance. *PLoS ONE*, 20(11):e0336299, 2025.

[39] Mark R Saddler and Josh H McDermott. Models optimized for real-world tasks reveal the task-dependent necessity of precise temporal coding in hearing. *Nature Communications*, 15:10750, 2024.

[40] Laurel H Carney. Neural fluctuation contrast as a code for complex sounds: The role and control of peripheral nonlinearities. *Hearing Research*, 443:108966, 2024.

[41] Vibha Viswanathan, Barbara G Shinn-Cunningham, and Michael G Heinz. Speech categorization reveals the role of early-stage temporal-coherence processing in auditory scene analysis. *Journal of Neuroscience*, 42(2):240–254, 2022.

[42] Rakesh Trinesh and Ajith Kumar Uppunda. Cortical processing of speech sounds in individuals with cochlear hearing loss and auditory neuropathy. *Hearing Research*, 473:109573, 2026.

[43] Anne-Lise Giraud and David Poeppel. Cortical oscillations and speech processing: emerging computational principles and operations. *Nature neuroscience*, 15(4):511–517, 2012.

[44] Dorea Ruggles, Hari Bharadwaj, and Barbara G Shinn-Cunningham. Normal hearing is not enough to guarantee robust encoding of suprathreshold features important in everyday communication. *Proceedings of the National Academy of Sciences*, 108(37):15516–15521, 2011.

[45] Charanjeet Kaur, Pei-Zhe Wu, Jennifer T. O'Malley, and M. Charles Liberman. Cochlear neural degeneration is key to the variation in word scores among people with similar thresholds. *Hearing Research*, 473:109574, 2026.

[46] Kai Siedenburg, Matthieu Kowalski, and Monika Dörfler. Audio declipping with social sparsity. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1577–1581. IEEE, 2014.

[47] Etienne Thoret, Thomas Andrillon, Damien Léger, and Daniel Pressnitzer. Probing machine-learning classifiers using noise, bubbles, and reverse correlation. *Journal of Neuroscience Methods*, 362:109297, 2021.

[48] Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux. Wham!: Extending speech separation to noisy environments. *arXiv preprint arXiv:1907.01160*, 2019.

[49] Facundo Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11(4):417–487, 2011.

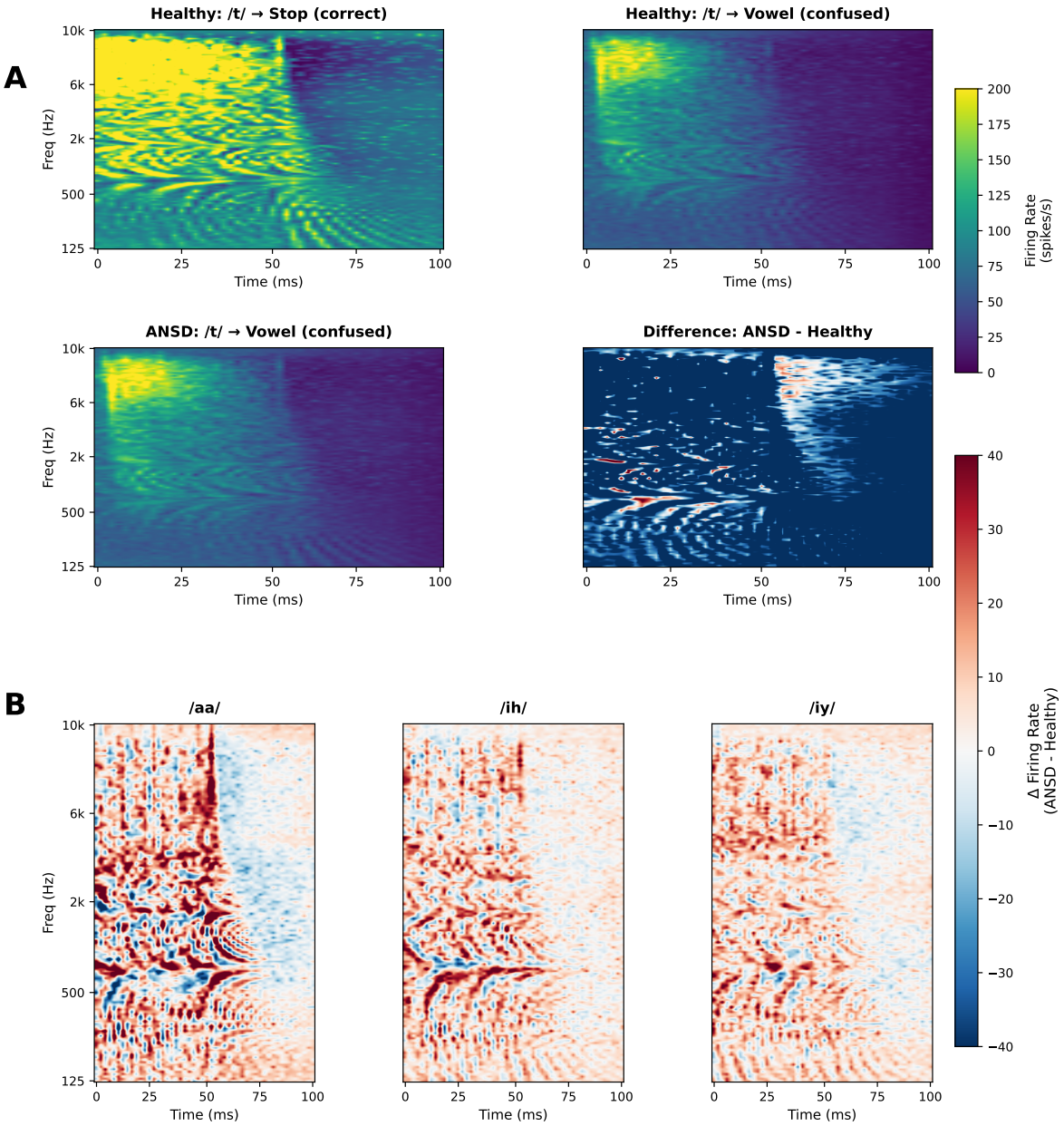

Figure 4: **Misclassification analysis reveals a stop-to-vowel shift in neural representation.** **(A)** Stop /t/ templates separated by population and classification outcome. Healthy correct classifications (top left) show the canonical brief high-frequency burst (0–25 ms, 6–10 kHz). When classified as vowels, healthy tokens (top right) exhibit attenuated burst structure. In ANSD, tokens confused as vowels (bottom left) show marked reduction of early burst energy accompanied by increased sustained low-frequency activity (50–100 ms, 500–3000 Hz). ANSD confused tokens closely resemble healthy confused tokens, indicating that temporal perturbations shift consonant representations toward neural regions associated with vowel decisions, rather than generating qualitatively new error patterns. The difference map (bottom right; ANSD confused minus Healthy correct) quantifies this redistribution: blue regions (0–50 ms) indicate diminished transient burst energy, whereas red regions (50–100 ms, formant frequencies) indicate enhanced sustained activity. **(B)** Vowel difference maps (/aa/, /ih/, /iy/) show ANSD-induced distortions (red: low-frequency excess below 2 kHz; blue: mid-frequency reduction) but preservation of overall spectral organization, consistent with retained vowel classification.