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

January 24, 2026

ABSTRACT

Auditory Neuropathy Spectrum Disorder (ANSD) causes severe speech comprehension deficits despite preserved cochlear function. Patients vary widely in their impairments, likely reflecting distinct pathophysiological mechanisms that current diagnostics cannot distinguish. A fundamental barrier to diagnosis is that standard speech tests take a one-size-fits-all approach—yielding aggregate intelligibility scores that average across phoneme categories, obscuring mechanism-specific patterns. We hypothesized that different ANSD mechanisms produce distinct phoneme-specific signatures that targeted testing could reveal. Using computational modeling of auditory nerve responses, we tested four mechanism types through three sequential experiments demonstrating that mechanism-specific encoding disruptions cascade into speech recognition failures that remain robust under noise. Brief consonants (flaps, stops) showed severe disruption while sustained vowels were relatively preserved—and critically, these category-specific patterns differed across mechanisms. Speech recognition models trained on ANSD-degraded neural representations developed compensation strategies that generalize to healthy signals, while the reverse failed completely, establishing a fundamental asymmetry: spectral compensation generalizes while temporal integration requires intact timing. These patterns persist under noise, where the same training that benefits healthy listeners harms ANSD listeners, explaining why ANSD patients struggle disproportionately in real-world listening conditions. The results suggest that phoneme-targeted testing—examining specific categories rather than using phonetically balanced materials—may enable mechanism-specific diagnosis and personalized intervention.

***K*eywords** 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, encompassing reduced productivity, diminished quality of life, and healthcare costs [1, 2]. Despite decades of hearing aid development, approximately half of individuals who would benefit from amplification do not use it, largely due to poor performance in complex acoustic environments [3]. These limitations are particularly evident in Auditory Neuropathy Spectrum Disorder (ANSD), which affects an estimated 1–10% of those with hearing loss [4, 5] and is associated with minimal benefit from conventional amplification despite preserved peripheral sensitivity.

Patients with ANSD exhibit preserved outer hair cell function—demonstrated by present otoacoustic emissions and/or cochlear microphonic responses—yet show abnormal or absent auditory brainstem responses and severe speech comprehension deficits, particularly in noise [6, 7]. This creates 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]. Unlike typical cochlear hearing loss, which reduces frequency selectivity and elevates detection thresholds, ANSD preserves cochlear frequency tuning while profoundly impairing temporal processing. Patients show deficits in gap detection, amplitude modulation sensitivity, and speech envelope tracking that cannot be explained by their audiometric profiles [9, 10].

The clinical heterogeneity of ANSD underscores the complexity of its underlying pathophysiology. Patient presentations vary widely: some struggle disproportionately with consonant discrimination, others with vowel recognition, and performance can fluctuate unpredictably across days or acoustic conditions [11].

This heterogeneity likely reflects distinct pathophysiological mechanisms operating 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 [12]. Phase-locking and cross-frequency coincidence detection in the brainstem both rely on precisely aligned inputs [13, 14], and demyelination-induced jitter disrupts both mechanisms. Synaptic loss represents a second major mechanism, particularly affecting low- and medium-spontaneous rate (SR) fibers critical for encoding speech in noise [15, 16]. Selective loss of these synapses reduces population coding strength while sparing the high-SR fibers that mediate quiet threshold detection—explaining why audiograms can appear normal despite severe 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 [17, 18]. We modeled these three 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.

Current diagnostic approaches face two fundamental limitations. First, electrophysiological signatures—absent or abnormal ABR with preserved OAE/cochlear microphonic—define ANSD as a category but cannot differentiate between specific pathophysiological mechanisms [7]. While recent work has demonstrated some success using compound action potential patterns to distinguish synaptopathy from myelinopathy [19], electrophysiological measures cannot access speech perception or predict the specific phoneme confusions that define patients' functional deficits. Second, standard speech audiometry provides aggregate intelligibility scores—percentage of words or sentences correctly identified—that collapse multidimensional confusion patterns into a single number. If different ANSD mechanisms preferentially disrupt different phoneme categories, this averaging systematically obscures mechanism-specific signatures. A patient scoring 60% on a word recognition test could be failing on consonants while preserving vowels (suggesting temporal impairment) or failing uniformly across categories (suggesting spectral impairment)—the aggregate score cannot distinguish these patterns. This creates a fundamental barrier to targeted intervention: without knowing which mechanism drives dysfunction in individual patients, and without testing paradigms that reveal mechanism-specific patterns, clinicians cannot select appropriate rehabilitation strategies.

1.1 Speech acoustics and phoneme-specific sensitivity

We hypothesized that different ANSD mechanisms produce distinct phoneme-specific signatures because speech sounds vary systematically in their acoustic properties and neural encoding requirements. Understanding these predictions requires considering speech acoustics [20]. Vowels are produced with relatively stable vocal tract configurations, creating sustained formant patterns (resonance peaks) that persist for 50–200 ms. The first formant (F1, ∼300–800 Hz) correlates with jaw opening; the second formant (F2, ∼800–2500 Hz) reflects tongue position; the third formant (F3, ∼2000–3500 Hz) relates to lip rounding and front cavity length. Because vowels are spectrally defined and temporally sustained, they may survive temporal degradation through averaging—even with imprecise timing, formant structure can be extracted from extended signal segments.

Consonants, by contrast, involve rapid articulatory transitions producing brief acoustic events. Stop consonants (/p/, /t/, /k/, /b/, /d/, /g/) are characterized by complete vocal tract closure followed by release bursts lasting only 5–40 ms. Place of articulation is cued by formant transitions in the 20–50 ms following release, with front consonants (/t/, /d/) showing rising F2 and back consonants (/k/, /g/) showing falling F2. Fricatives (/s/, /f/, /sh/) involve sustained turbulent noise but require precise spectral analysis to distinguish sibilants from non-sibilants. Flaps (/dx/ as in "butter") represent an extreme case: single ballistic tongue gestures lasting only 20–30 ms, requiring exceptional temporal precision for accurate encoding.

These acoustic properties predict differential effects of ANSD mechanisms across phoneme categories. Temporal jitter should disproportionately affect brief consonants requiring precise timing, while sparing sustained vowels whose formant structure can be extracted despite timing imprecision. Frequency-selective fiber loss should affect formants encoded by the damaged frequency region. Saturation deficits should affect intensity-dependent contrasts across all phoneme categories more uniformly. If these predictions hold, phoneme-specific confusion patterns could serve as diagnostic signatures—but testing them directly in patients faces practical barriers: we cannot experimentally induce specific ANSD mechanisms, patients present with unknown mixtures of pathology, and systematic testing across all mechanism-phoneme combinations would require prohibitive clinical time. Computational modeling provides the necessary first step: establishing which phoneme-mechanism combinations produce diagnostic signatures before undertaking the technical challenge of translating these predictions into patient-testable stimuli.

1.2 Perturbation types modeling ANSD mechanisms

The Zilany-Bruce-Carney auditory nerve model [21] provides a biophysically detailed implementation incorporating inner hair cell transduction, synaptic vesicle dynamics, and auditory nerve spike generation. Traditional computational approaches to modeling auditory neuropathy operate at the level of underlying stochastic generative processes [22, 23]. We developed an alternative approach: neurogram-level perturbations that directly instantiate pathophysiological hypotheses at the neural population output level. Using the Zilany model to generate baseline healthy neurograms, we introduce controlled disruptions—temporal shifts, channel elimination, amplitude limiting—that reproduce hallmark ANSD features while isolating perceptual consequences.

These perturbation types correspond to distinct pathophysiological mechanisms (Fig. 1A–B):

Uniform jitter applies identical temporal delays across all frequency channels, modeling global demyelination that slows conduction uniformly. This preserves cross-frequency timing relationships while introducing overall temporal smearing—serving as a reference condition isolating within-channel temporal degradation.

Scattered jitter applies independent random delays to each frequency channel, modeling patchy demyelination affecting different cochlear regions differentially. This destroys cross-frequency synchrony essential for coincidence detection, representing severe temporal desynchronization.

Fiber loss eliminates randomly selected frequency channels, modeling cochlear synaptopathy or neural degeneration affecting specific tonotopic regions. This creates spectral gaps in the neural representation while preserving temporal precision in surviving channels.

Truncation limits response amplitudes at a fraction of maximum, modeling saturation from reduced neural populations. This compresses dynamic range uniformly across frequencies without affecting temporal structure.

1.3 Research questions and experimental structure

We structure our investigation as three sequential experiments, each building on the previous to establish a cascade from neural encoding to functional consequences to clinical robustness:

Experiment 1: Neural encoding distortions. Which phoneme categories show greatest sensitivity to specific ANSD mechanisms? Using optimal transport theory [24], which quantifies the preservation of similarity structure between representations, we measure how each perturbation distorts the correspondence between acoustic formant patterns and neural representations. This experiment alone—characterizing encoding disruptions—echoes prior computational work that models how hearing damage affects neural population responses [23, 22]. However, such approaches characterize encoding at the population level without testing phoneme-specific perceptual consequences. Encoding disruption does not guarantee perceptual failure: the auditory system might compensate downstream.

Experiment 2: Functional consequences. Do encoding distortions translate to classification failures? Building on [25], we use hierarchical automatic speech recognition to test whether ANSD-degraded representations support accurate phoneme identification—and whether the system can learn to compensate for distorted inputs. Cross-population transfer (training on healthy signals and testing on ANSD, or vice versa) tests whether recognition relies on signal

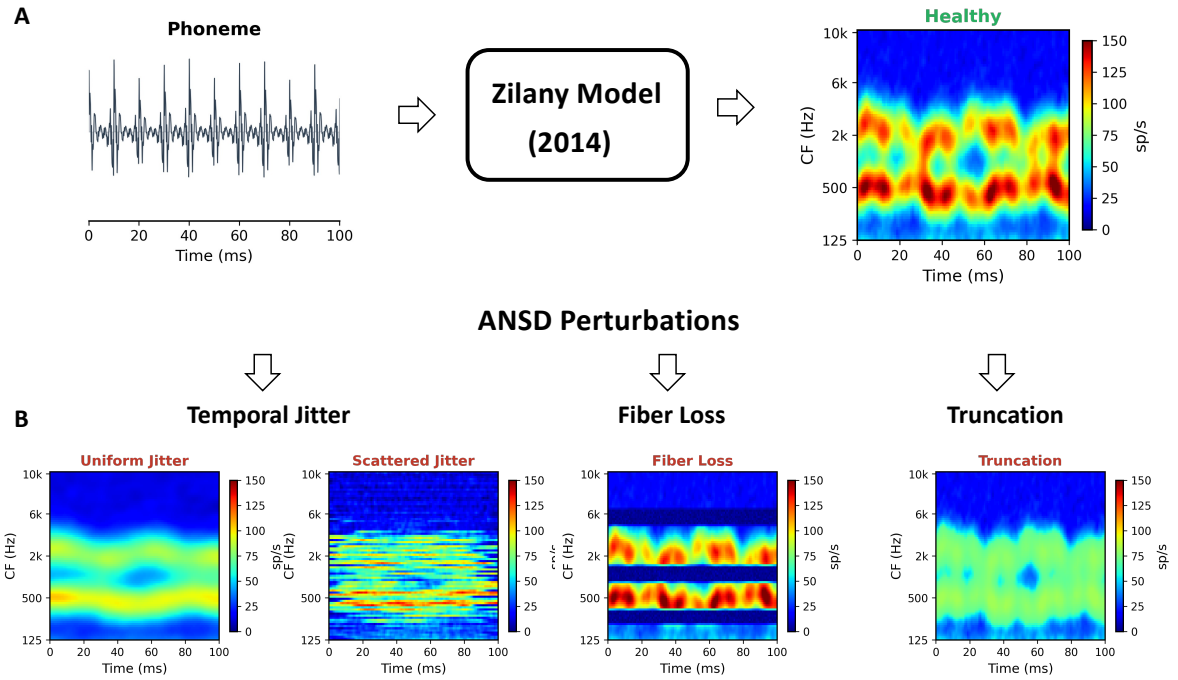

Figure 1: **Computational framework for ANSD neurogram perturbations. (A)** Neurogram generation: speech phonemes are processed through the Zilany auditory nerve model to produce healthy neurograms representing firing rates across characteristic frequency (CF) and time. **(B)** Four perturbation types 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).

properties degraded by ANSD: if healthy-trained models fail on ANSD signals, they depend on features unavailable to ANSD patients; if ANSD-trained models succeed on healthy signals, they have learned alternative cues that may support rehabilitation. To understand why models fail, template analysis—averaging neurograms by classification outcome [26, 27, 28]—identifies which spectrotemporal features drive misclassifications.

Experiment 3: Clinical robustness. Do these patterns persist under realistic listening conditions? Testing performance in noise evaluates whether ANSD-related failures reflect simple internal noise or a qualitative degradation of the auditory signal. Mechanism-specific testing isolates which individual perturbation types pose the greatest challenges. If ANSD simply added internal noise, increasing external noise should progressively equalize performance across populations, eliminating ANSD-specific deficits. In contrast, persistent performance differences under noise would indicate that ANSD alters the structure of the signal itself, rather than merely increasing variability.

The cascade from Experiment 1 through 3 is the contribution: each experiment closes an escape route. Encoding distortion (Exp 1) might be compensated downstream; classification failure (Exp 2) might occur only in artificial silence; noise testing (Exp 3) establishes that these signatures persist in conditions where patients actually struggle. Together, they establish that if you know which phonemes fail, you can infer the mechanism—suggesting that targeted, phoneme-specific testing (rather than phonetically balanced word lists) may enable mechanism-specific diagnosis.

2 Results

We implemented four perturbation types simulating proposed ANSD mechanisms (Fig. 1): starting from healthy neurograms generated from TIMIT speech stimuli [29] using the Zilany auditory nerve model [21], we applied uniform temporal jitter (identical delays across frequency channels, preserving cross-frequency timing), scattered jitter (independent per-channel delays, destroying cross-frequency synchrony), selective fiber loss (frequency-channel elim-

ination), and truncation (amplitude saturation). Results are organized as three sequential experiments establishing the cascade from encoding distortion to functional failure to clinical robustness.

2.1 Experiment 1: Optimal Transport Reveals Mechanism-Specific Encoding Disruptions

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 (0.049–0.083). 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 averaging.

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

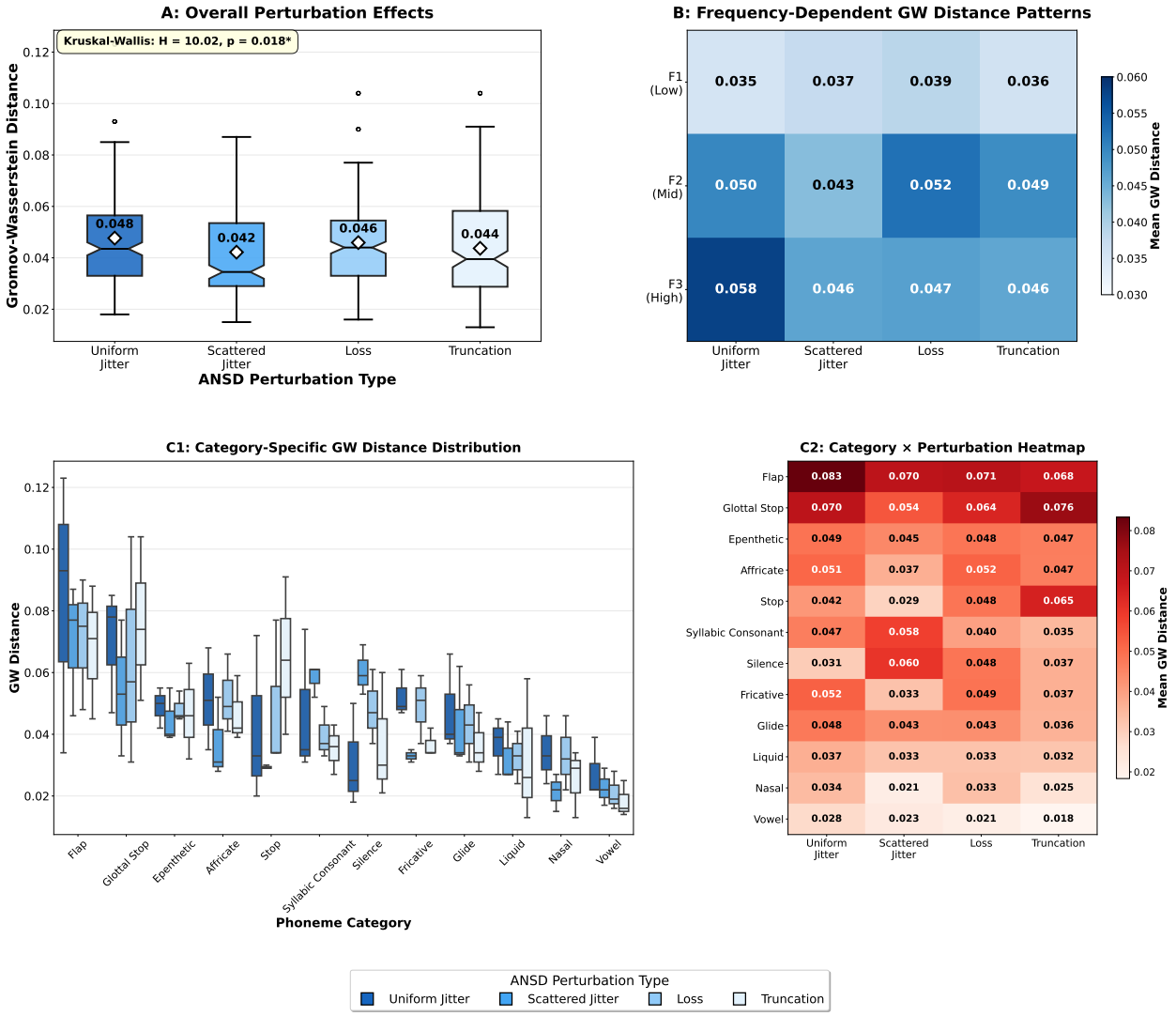

Figure 2: **Experiment 1: Optimal transport reveals phoneme-specific encoding disruptions.** **(A)** Overall perturbation effects show significant differences across ANSD mechanisms (Kruskal-Wallis $H = 10.02$, $p = 0.018$). White diamonds indicate means; boxes show interquartile range. Perturbation 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.

jitter produced more uniform disruption across formants (24% gradient) by destroying both within-channel timing and cross-channel synchrony.

However, Experiment 1 alone cannot establish functional consequences: the auditory system compensates for acoustic noise through downstream processing. Whether these encoding distortions translate to perceptual failures requires testing speech recognition.

2.2 Experiment 2: Speech Recognition Reveals Functional Consequences

Experiment 1 established encoding distortions; Experiment 2 tests whether these translate to classification failures. We used a two-stage hierarchical automatic speaker recognition (ASR) architecture. Stage 1 extracts acoustic-phonetic

features from brief neurogram windows (100 ms); Stage 2 integrates predictions across longer temporal context (610 ms), enabling coarticulatory processing.

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 comparisons: (1) matched testing reveals each population's processing capacity on its own signal type; (2) cross-population transfer tests whether learned strategies require specific signal properties; (3) template analysis identifies spectrotemporal features driving systematic misclassifications. Robustness testing (Experiment 3) examines noise interactions and identifies which individual perturbation types pose the greatest challenges. Here we report matched testing, cross-population transfer, and template analysis; all results use TIMIT's held-out test set (168 speakers, no overlap with training).

Hierarchical processing failure in ANSD. 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 models showed modest Stage 2 reduction in silence (Stage 1: 67.0% to Stage 2: 61.0%) but hierarchical benefit under noise (58.5% to 63.2%). ANSD models showed severe degradation across both conditions (Silence: 48.0% to 25.5%; Noise: 41.3% to 17.1%), a 22.5-point loss in silence alone, nearly four times the healthy reduction.

Category-specific patterns. ANSD-Silence models achieved strong vowel accuracy (65.2%), approaching healthy performance (70.3%) and far exceeding other categories (Fig. 3C).

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.

Systematic consonant-to-vowel transformation. 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%).

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.

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.

Template analysis reveals transformation mechanism. To identify spectrotemporal features driving the consonant-to-vowel confusions documented above, we computed classification-conditioned templates: average neurograms for tokens classified as each category, stratified by correctness. We focused on stop /t/ (high-frequency, canonical burst signature, prominent in confusions) and vowels /aa/, /ih/, /iy/ (spanning formant configurations).

When healthy models correctly classified /t/, templates showed canonical signatures: intense, brief bursts (0–25 ms) with broadband activity (2–10 kHz). When ANSD models misclassified /t/ as vowel (22.1% of tokens), templates

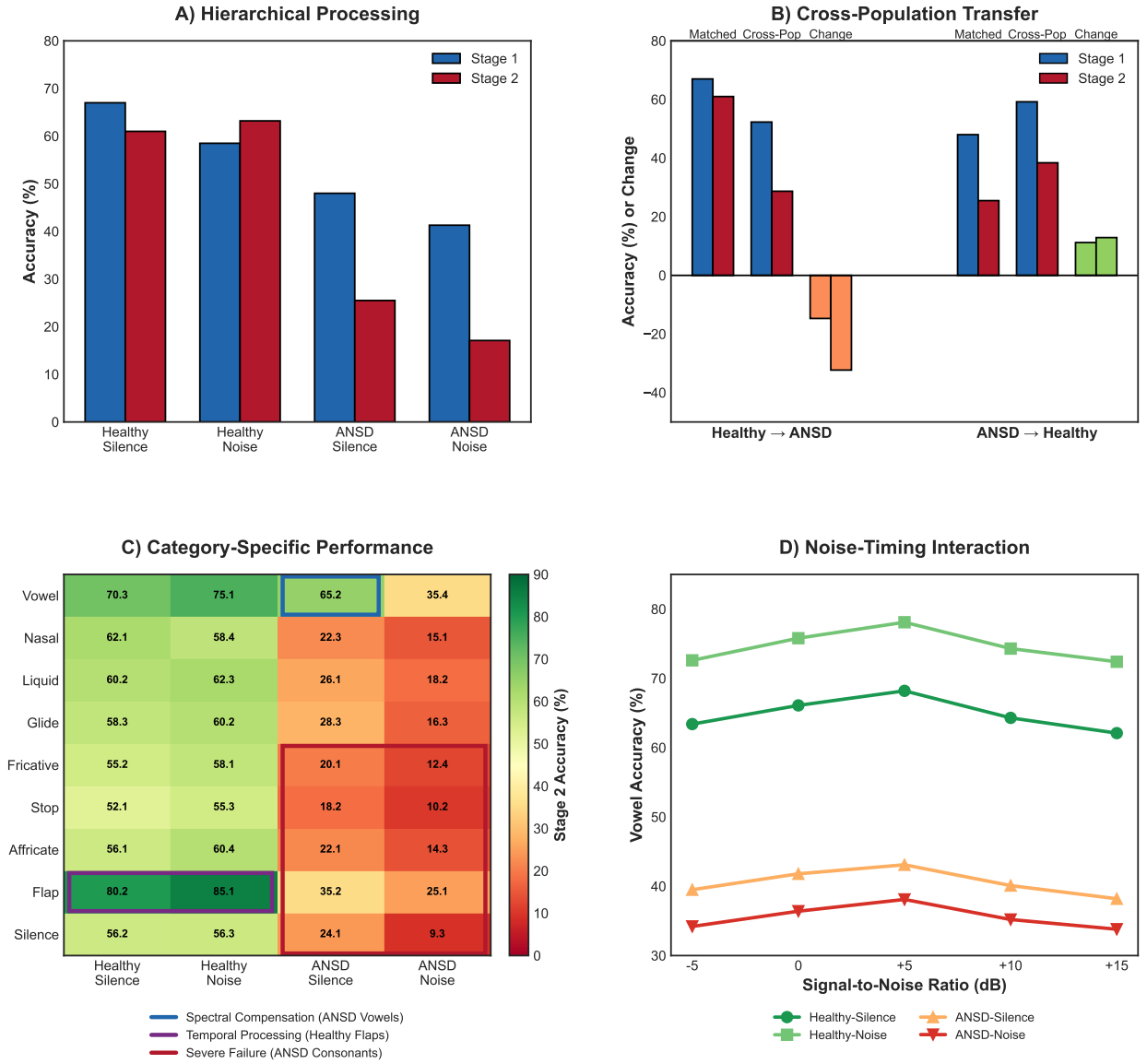

Figure 3: **Experiment 2: Speech recognition reveals hierarchical failure and asymmetric transfer.** **(A)** Hierarchical processing on held-out speakers: Healthy models show modest reduction in silence but benefit under noise; ANSD models show severe degradation across both conditions. **(B)** Cross-population 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 (discussed in Experiment 3): Healthy models benefit from noise training; ANSD models are harmed.

showed reduced burst intensity (0–50 ms) combined with sustained activity extending 50–100 ms beyond the initial transient (Fig. 4A).

The difference map (ANSD confused minus Healthy correct) showed predominantly negative values (blue) in early periods (0–50 ms) and predominantly positive values (red) in later periods (50–100 ms) concentrated in formant regions (500–3000 Hz). Healthy confused and ANSD confused templates appeared similar, indicating ANSD-perturbed stops resemble naturally confusable cases in healthy hearing.

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

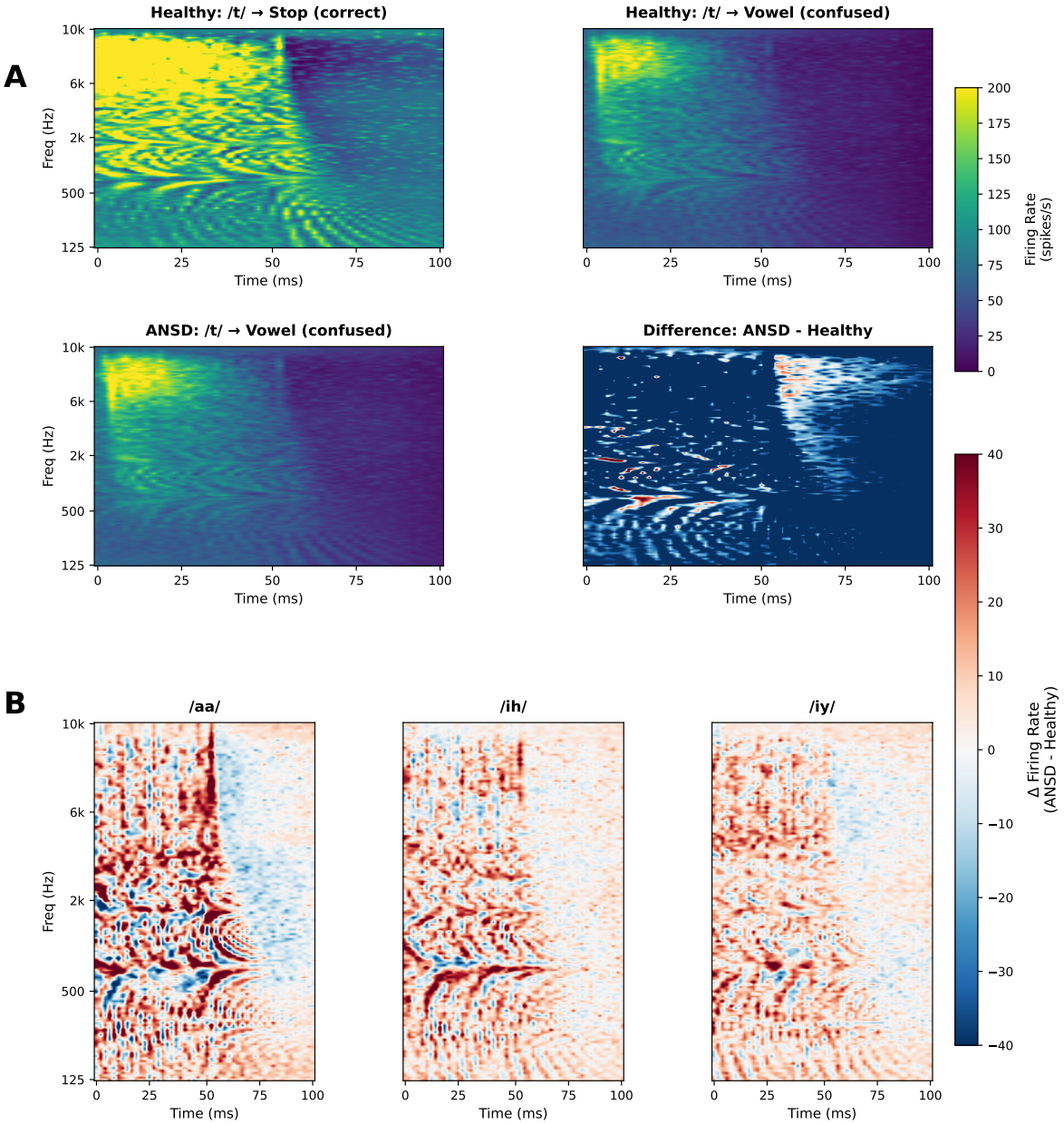

Figure 4: **Template analysis reveals stop-to-vowel transformation. (A)** Stop /t/ templates: Healthy correct shows canonical brief burst (0–25 ms); ANSD confused shows reduced burst and increased sustained activity. Difference map (ANSD−Healthy) shows reduced early activity (blue, 0–50 ms) and increased late activity (red, 50–100 ms). **(B)** Vowel encoding (/aa/, /ih/, /iy/) shows ANSD-induced distortions but preserves spectral structure.

2.3 Experiment 3: Noise Testing Establishes Clinical Robustness

Experiment 2 established category-specific failures using fixed 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.

Noise-timing interaction. Noise training and temporal disruption interact: the same training regime that benefits healthy populations harms ANSD populations (Fig. 3C–D). 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.

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. See SI for phoneme category breakdowns.

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 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).

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 for phoneme category 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: 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.

3 Discussion

Three sequential experiments establish mechanism-specific speech encoding impairments in ANSD. Experiment 1 revealed category-specific disruption patterns varying nearly 4-fold within perturbation types (uniform jitter: flaps 0.083 vs. nasals 0.021)—substantially exceeding 14% between-perturbation variation—suggesting phoneme-specific profiles may provide greater diagnostic power than aggregate mechanism identification. Experiment 2 demonstrated severe hierarchical processing failure despite comparable initial encoding, with asymmetric cross-population transfer: healthy models failed on ANSD data (−32.3 points) while ANSD models improved on healthy data (+12.9 points). We interpret this asymmetry as reflecting spectral versus temporal processing strategies: ANSD models exploit sustained formant structure that transfers when spectral information is cleaner, while healthy models rely on timing information that ANSD perturbations destroy—an interpretation consistent with the observed vowel preservation versus consonant failure in ANSD models. Experiment 3 confirmed these patterns persist under noise: the same training that benefits healthy listeners harms ANSD listeners. If ANSD were simply internal noise, external noise exceeding this level would eliminate the performance gap; instead, the gap remains constant across SNRs, suggesting ANSD involves temporal structure destruction rather than additive noise. Template analysis of stop consonant /t/ revealed a candidate mechanism: temporal perturbations transform brief consonant bursts into sustained vowel-like patterns, consistent with the observed consonant-to-vowel confusions (additional stops /k/, /s/, /d/ in SI).

Mechanisms of encoding failure. Each ANSD mechanism produces distinct failures through different computational disruptions:

Uniform jitter causes temporal smearing that blurs rapid spectrotemporal transitions while preserving cross-frequency relationships. Template analysis revealed transformation of brief stop bursts into sustained vowel-like patterns through temporal redistribution. The asymmetry between phoneme durations explains unidirectional confusions: a 100 ms vowel smeared by 10 ms jitter retains 90 ms of recoverable formant information, while a 25 ms stop burst loses its defining signature entirely.

Scattered jitter destroys cross-frequency synchrony essential for coincidence detection, consistent with its selective devastation of flaps (GW: 0.070)—brief consonants requiring precise cross-channel coordination. Scattered jitter also produced the most severe hierarchical failure (-24.6 points), consistent with hierarchical processing depending on cross-frequency integration.

Loss perturbation creates spectral gaps that, unlike temporal distortions, may be partially bridged through interpolation from neighboring frequencies—a potential explanation for the unique preservation of hierarchical benefit (+1.9 points versus -24 points for jitter perturbations).

Truncation compresses dynamic range through amplitude saturation, producing the most uniform effects across categories (SD = 0.015). Unlike jitter perturbations, truncation preserves temporal relationships between channels; unlike loss, it preserves spectral completeness. The uniform category profile suggests truncation affects overall signal fidelity rather than selectively disrupting specific phonetic features.

Convergence with prior work. Our framework complements parameter-based auditory nerve models [19] by directly manipulating neurogram representations to instantiate hypothesized ANSD mechanisms. Convergence with recent empirical and modeling studies strengthens the validity of this approach. Garrett et al. [30] show that synaptopathy selectively degrades temporal envelope coding at suprathreshold levels using electrophysiological markers validated in animal models, paralleling our finding that synaptic-loss perturbations produce consonant deficits while preserving hierarchical benefit—both studies converge on the insight that fiber-specific loss creates frequency-dependent vulnerabilities that audiometry fails to capture. Vasilkov et al. [31] demonstrate that spontaneous rate-specific fiber loss degrades word recognition using Zilany-based simulations, with constrained decoder architectures proving critical for revealing biologically meaningful deficits—paralleling our finding that decoder architecture affects sensitivity to encoding degradation. Saddler & McDermott [32] show that limiting phase-locking capacity above 320 Hz alters speech and voice recognition in task-dependent ways, aligning with our finding that temporal mechanisms dominate ANSD-related degradation. Carney's neural fluctuation contrast theory [33] 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.

Our use of consonant confusion matrices to reveal processing mechanisms parallels recent work using similar approaches to study temporal coherence in auditory scene analysis [34]. Viswanathan and colleagues demonstrated that consonant confusions can distinguish within-channel from across-channel processing strategies, analogous to our finding that category-specific confusions distinguish temporal from spectral mechanisms. The hierarchical failure connects to broader theories of temporal processing, suggesting temporal degradation compounds across processing stages, consistent with models proposing precise timing scaffolds higher-level linguistic computations [35].

Clinical implications for diagnosis. The cross-population transfer asymmetry has potential diagnostic implications. Current speech tests provide scalar intelligibility scores that collapse multidimensional confusion patterns into aggregate metrics, potentially obscuring mechanism-specific information.

Our results generate specific predictions for clinical testing. Scattered jitter produced the most severe flap deficits (GW: 0.070) among all perturbation-category combinations, predicting that patients with patchy demyelination may show disproportionate flap confusions. Loss perturbation uniquely preserved hierarchical benefit (+1.9 points), predicting that patients with frequency-selective synaptopathy may show less severe hierarchical degradation than patients with temporal deficits. These predictions require validation against confusion matrices from patients with characterized pathophysiology.

If validated, implementation would require speech materials isolating specific encoding dimensions. Stop consonants varying in place (/pa/-/ta/-/ka/) could probe high-frequency encoding through F2-F3 transitions. Minimal pairs contrasting manner (/ta/-/na/) could probe temporal processing through stop-nasal discrimination. Critically, such testing should be phoneme-targeted rather than phonetically balanced: our finding that within-perturbation variation (nearly 4-fold) exceeds between-perturbation variation (14%) suggests that category-specific patterns carry more diagnostic information than aggregate scores.

Clinical implications for rehabilitation. Template findings suggest potential rehabilitation targets, though these predictions require clinical validation.

The stop-to-vowel transformation observed in our template analysis suggests that temporal enhancement strategies should focus on sharpening onset transients rather than amplifying sustained activity. Signal processing algorithms emphasizing brief transients while attenuating prolonged responses might partially restore consonant discrimination [36], though this remains untested in ANSD populations.

The noise-timing interaction generates predictions about noise reduction strategies. Our finding that noise training harms ANSD performance suggests that conventional noise reduction algorithms exploiting temporal modulation

structure may actually degrade perception for ANSD patients by further smearing temporal features. We predict that ANSD-appropriate noise reduction should operate in spectral domains, preserving whatever temporal structure survives peripheral degradation.

The transfer asymmetry predicts realistic rehabilitation expectations. ANSD models developed spectral compensation strategies that transferred to intact signals, but could not develop temporal strategies that transferred to degraded signals. If this pattern holds in patients, rehabilitation should emphasize environmental modification (noise reduction, visual speech cues) and communication strategies rather than auditory training targeting temporal skills that may not be acquirable.

Limitations. This study modeled high-SR fibers exclusively; incorporating mixed-SR populations would better assess distributed synaptic loss and may reveal greater deficits in challenging conditions where low- and medium-SR fibers contribute disproportionately to robust encoding. The neurogram perturbation approach directly manipulates neural representations without modeling the biophysical mechanisms that produce them; bridging to parameter-based models [19, 21] would enable predictions from histological data.

The ASR classifier does not capture linguistic context, top-down predictions, or perceptual learning that human listeners exploit. However, the ASR serves as an ideal observer revealing which acoustic-phonetic information survives neural degradation—a constraint that applies to any decoder, biological or artificial. The template analysis strengthens this interpretation: the stop-to-vowel transformation reflects physical destruction of temporal structure, not an artifact of model learning.

Validating these predictions requires collecting detailed confusion matrices from ANSD patients with characterized pathophysiology. Validation studies should include comprehensive electrophysiological assessment alongside speech testing with confusion analysis, ideally in patients with known etiologies providing ground truth for mechanism identification. A fundamental challenge for clinical translation is the inversion problem: while our framework predicts which confusion patterns each mechanism 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 a computational framework for mechanism-specific ANSD phenotyping through three sequential experiments. The cascade from encoding distortion (Exp 1) through classification failure (Exp 2) to noise robustness (Exp 3) closes successive escape routes, establishing that phoneme-specific confusion patterns reveal underlying mechanisms. The fundamental constraint we identify—spectral compensation generalizes while temporal integration requires intact timing—predicts why patients may develop functional strategies in quiet but cannot overcome deficits under noise. Category-specific patterns with nearly 4-fold within-perturbation variation demonstrate that phonetically balanced assessment obscures mechanism-specific signatures essential for targeted intervention. The translational pathway forward: validate computational predictions against patient confusion matrices, develop mechanism-specific diagnostic protocols using targeted (not balanced) phoneme testing, and implement rehabilitation strategies matched to underlying neural dysfunction.

4 Materials and Methods

Auditory nerve modeling. Auditory nerve responses were generated using the Zilany et al. model [21]: 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 [29], 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 [37] 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—we cannot directly compare them. Gromov-Wasserstein (GW) distance [38, 24] 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 [25], we used a two-stage hierarchy automatic speaker 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.

Template 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 [27]. 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 comparing templates between ANSD and healthy neurograms (difference maps), we can see how ANSD perturbations alter specific time–frequency components, 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.

Code available at `https://github.com/mcampi111/SimsACues_ASR`.

Data Availability

TIMIT corpus available from Linguistic Data Consortium (LDC93S1). WHAM! noise corpus at `http://wham.whisper.ai/`. Analysis code and trained models released on GitHub upon publication.

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.

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] 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.

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

[13] 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.

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

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[20] Kenneth N Stevens. *Acoustic phonetics*, volume 30. MIT press, 2000.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

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

[34] 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.

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

[36] 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.

[37] 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.

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