Computational Approaches to Hearing Diagnostics

Cochlear Implants and Auditory Neuropathy

Dr. Marta Campi

Deep Hearing Lab · University Hospital Zurich

Computational Linguistics Group, UZH, 2026

Outline

Deep Hearing Lab

1. **Prediction of Speech Outcomes with Cochlear Implants**
   Postdoc, Deep Hearing Lab, USZ/UZH · with Prof. Alexander Huber & Prof. Tobias Goehring

2. **Auditory Neuropathy: What Phoneme Confusions Reveal**
   Postdoc, Hearing Institute – CERIAH, Institut Pasteur · with Prof. Paul Avan

Part 1

Prediction of Speech Outcomes with Cochlear Implants

with Prof. A. Huber & Prof. T. Goehring · Swiss National CI Database

The Swiss National CI Database

University Hospital Zurich · 2000–2024

Deep Hearing Lab

Sample

All patients	771
Adults (≥18)	538
With post-op Mono1	431
Pre + post Mono1	**266**
With trajectory data	156

Mean age 59 · 55% female · post-op $65 \pm 25\%$

Five speech tests

Test	What it measures	r_{Mono1}	Ceiling?
FM	Voice discrim. (M vs F)	0.48	Yes (68%)
Num	Number recognition	0.63	Yes (81%)
V08	Vowel discrimination	0.68	No
C12	Consonant discrimination	0.79	No
Mono1	Open-set words (WRS)	—	No (16%)

Pre-op: best aided (optimal hearing aid). Post-op: CI-only.

Pre-op variables (85 usable)

PTA (30 freq., CI + contra) · Demographics (age, sex) · Etiology & onset (duration, pre/post-lingual) · Communication (articulation, lip reading) · Socioeconomic · Surgical · QoL (SSQ12, THI)

Research Questions

Deep Hearing Lab

Can we predict individual speech outcomes with cochlear implants?

RQ1 **Can advanced AI models predict CI speech outcomes?**

8 algorithms incl. TabPFN · from 3 to 85 pre-op features · rigorous CV

RQ2 **Can we identify patients at risk of poor outcomes?**

Binary classification: post-op $< 40\%$ · risk factors · 6-month trajectory

RQ3 **If we profile patients by auditory dimensions, can we predict the profiles?**

Spectral index, FM–Num gap, top-down index · ongoing work

Previous studies report R^2 up to 0.27 — but with limited cross-validation and often no multi-centre validation.

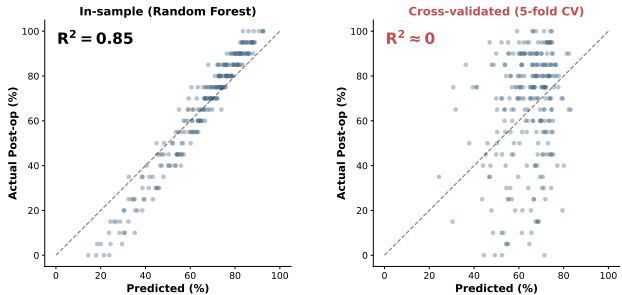
Pre-operative Prediction Fails

8 algorithms × 5 feature sets — all roads lead to the same ceiling

Model	Train	CV
Linear / Ridge	0.18	0.03
Random Forest	0.85	−0.03
XGBoost	0.55	−0.01
TabPFN	—	**0.10**

In-sample R^2 up to 0.85 — **all collapse** to $R^2 \leq 0.10$ under honest CV.
Published R^2 of 0.20–0.27 reflect methodological inflation (no CV).

N = 266 adults, Swiss National CI DB (Zurich). Mono1 (Freiburg monosyllables).

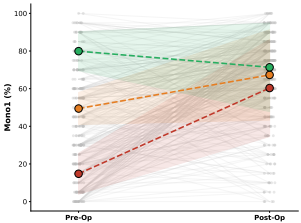


Left: model sees the patient — fits noise. **Right:** patient held out — predictions collapse to ∼65%.

Why Prediction Fails: Three Structural Reasons

The ceiling is not methodological — it is a property of how CI transforms the outcome distribution

1. CI Equalizer Effect



All patient profiles converge to same post-op mean (∼65%).
$\sigma_{\text{pre}} > \sigma_{\text{post}}$: between-group variance erased, within-group variance stays.

$E[Y|X] \approx \text{const} \Rightarrow R^2 \approx 0$.

2. Simpson's Paradox

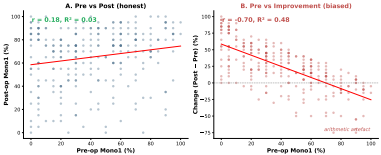
$$r_{\text{overall}} = 0.18$$

$$r_{\text{within strata}} \approx 0$$

The marginal pre–post correlation is entirely ecological — driven by between-group differences in hearing-loss severity.

Pre-op score tells you the **group**, not the individual landing point.

3. Mathematical Coupling



Same data, different formula. $\Delta = Y - X$:

$$\text{Cov}(X, \Delta) = \text{Cov}(X, Y) - \text{Var}(X)$$

Unless $\text{Cov}(X, Y) > \text{Var}(X)$, coupling is guaranteed negative. Studies reporting "pre-op predicts improvement" measure arithmetic, not biology.

What CAN We Predict?

Reframe: from regression to risk classification, from pre-op to post-op monitoring

Deep Hearing Lab

RQ2: Risk of poor outcome

Question	Result
Predict exact score?	$R^2 \approx 0.10$
Poor outcome <40%?	**AUC 0.67**
From 6-month data?	$R^2 = 0.48$

Risk factors: pre-lingual onset (OR = 5.5), articulation (OR = 3.4). Confirmed by TabPFN permutation importance.

Clinical translation

Before surgery: identify patients at elevated risk for counselling. AUC = 0.67 is modest but clinically useful.

After surgery: clinically actionable information arrives at **6 months** post-CI, not before.

Ongoing: perceptual profiling — can we predict *which* dimensions will be strong or weak? (RQ3)

Pre-op data can't predict *how well* — but can identify *who is at risk*.

Part 1: Take-Home Messages

Deep Hearing Lab

What we found

- $R^2 \leq 0.10$ across *all* models under honest CV — not a modelling failure, an **information limit**
- The CI Equalizer Effect compresses outcomes: all profiles → same post-op mean
- Researchers must account for Simpson's Paradox and Mathematical Coupling when evaluating prediction claims

What to do instead

- **Reframe:** regression → binary risk classification (poor outcome AUC = 0.67)
- **Shift timing:** 6-month post-op score predicts final outcome ($R^2 = 0.48$, 5× better than any pre-op variable)
- **Risk factors:** odds ratios and permutation tests reveal factors consistent with the literature

Part 2

Auditory Neuropathy: What Phoneme Confusions Reveal

with Prof. P. Avan, E. Partouche, C. Gaultier, G. Gerenton

bioRxiv: 2026/701491

Auditory Neuropathy Spectrum Disorder

The cochlea works fine — the neural signal is disrupted. Patients hear sounds but cannot understand speech.

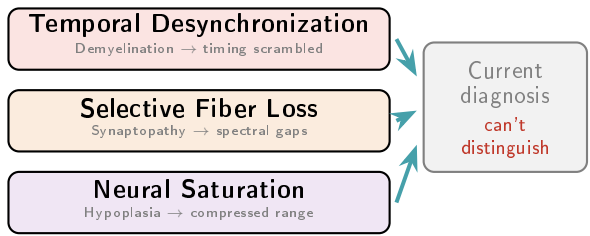
"Hearing but not understanding"

Diagnosis:

- OAEs present
  (cochlea works)
- ABR absent/abnormal
  (nerve signal disrupted)
- Speech
  disproportionately poor
  (worse than thresholds predict)

Candidate Pathophysiology

Anatomical cause → neural coding disruption → speech perception failure



ANSD affects 1–10% of hearing-impaired patients.

Linking speech perception to neural coding disruption could inform **diagnosis** and enable mechanism-specific **intervention** (electrical stimulation for CI, speech enhancement for HA).

Deep Hearing Lab

Computational Framework

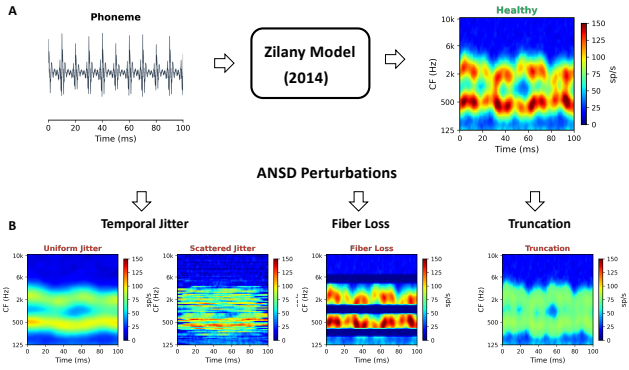
We simulate each mechanism as a **perturbation** to auditory nerve responses.

Model: Zilany et al. (2014)

Stimuli: TIMIT corpus, split into phonemes

Each produces a distinct neurogram:

- Jitter → timing smeared
- Fiber loss → spectral gaps
- Truncation → range compressed



A: Phoneme → Zilany model → healthy neurogram

B: Four perturbation types (jitter variants, fiber loss, truncation)

Three Questions

Deep Hearing Lab

Q1 **Does each mechanism distort specific phoneme categories?**

Optimal Transport: measuring distances between acoustic and neural representations

Q2 **Do these distortions produce systematic classification failures?**

OT shows encoding distortion — distortion $\neq$ perceptual failure · DNN tests if failures cascade

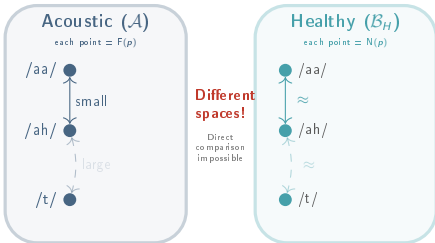
Q3 **Do patterns persist in noise?**

Noise robustness: same training that helps healthy listeners harms ANSD

Optimal Transport: Comparing Acoustic and Neural Spaces

How do we compare representations that live in completely different spaces?

Acoustic: $\mathsf{F}(p) = [F_1(p), F_2(p), F_3(p)] \in \mathbb{R}^{T_p \times 3}$ **Neural:** $\mathsf{N}(p) \in \mathbb{R}^{T_p \times 150}$

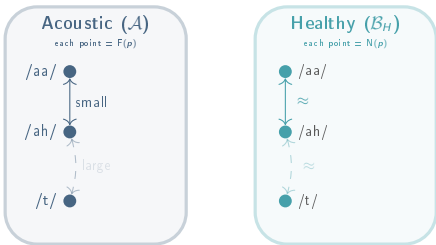


Optimal Transport: Comparing Acoustic and Neural Spaces

How do we compare representations that live in completely different spaces?

Deep Hearing Lab

Acoustic: $\mathrm{F}(p) = [F_1(p), F_2(p), F_3(p)] \in \mathbb{R}^{T_p \times 3}$ **Neural:** $\mathrm{N}(p) \in \mathbb{R}^{T_p \times 150}$



1. Within $\mathcal{A}$: compute $D^{\mathcal{A}}_{ij} = \|\mathrm{F}(p_i) - \mathrm{F}(p_j)\|$ for all phoneme pairs → distance matrix
2. Within $\mathcal{B}$: compute $D^{\mathcal{B}}_{ij} = \|\mathrm{N}(p_i) - \mathrm{N}(p_j)\|$ for all phoneme pairs → distance matrix
3. Compare matrices: $d_{\mathrm{GW}} = \min_\pi \sum |D^{\mathcal{A}}_{ij} - D^{\mathcal{B}}_{kl}|^2 \pi_{ik} \pi_{jl}$

Intuition

In acoustic space:

/aa/ ≈ /ah/ (both vowels, close)
/aa/ ≠ /t/ (vowel vs stop, far)

Healthy neural code:
same distance pattern
$D^{\mathcal{B}} \approx D^{\mathcal{A}}$

$\Rightarrow d_{GW} \approx 0$

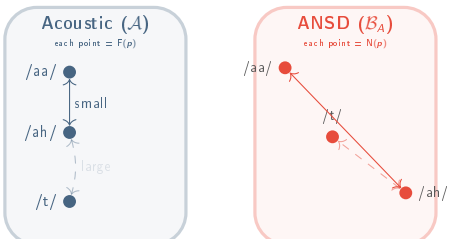
GW compares distance matrices, not points → works across spaces of different dimensions.

Optimal Transport: Comparing Acoustic and Neural Spaces

How do we compare representations that live in completely different spaces?

Deep Hearing Lab

Acoustic: $\mathrm{F}(p) = [F_1(p), F_2(p), F_3(p)] \in \mathbb{R}^{T_p \times 3}$ **Neural:** $\mathrm{N}(p) \in \mathbb{R}^{T_p \times 150}$



1. Within $\mathcal{A}$: compute $D^{\mathcal{A}}_{ij} = \|\mathrm{F}(p_i) - \mathrm{F}(p_j)\|$ for all phoneme pairs → distance matrix
2. Within $\mathcal{B}$: compute $D^{\mathcal{B}}_{ij} = \|\mathrm{N}(p_i) - \mathrm{N}(p_j)\|$ for all phoneme pairs → distance matrix
3. Compare matrices: $d_{\mathrm{GW}} = \min_\pi \sum |D^{\mathcal{A}}_{ij} - D^{\mathcal{B}}_{kl}|^2 \pi_{ik} \pi_{jl}$

Intuition

In acoustic space:

/aa/ ≈ /ah/ (both vowels, close)
/aa/ ≠ /t/ (vowel vs stop, far)

Healthy neural code:

$D^{\mathcal{B}_H} \approx D^{\mathcal{A}} \Rightarrow d_{GW} \approx 0$

ANSD neural code:

distances distorted
/aa/ no longer ≈ /ah/

$D^{\mathcal{B}_A} \neq D^{\mathcal{A}} \Rightarrow d_{GW} \gg 0$

GW compares distance matrices, not points → works across spaces of different dimensions.

† DTW (Dynamic Time Warping): acoustic space · wPEW (weighted Persistent Empirical Wiener): neural space

Q1: Each Mechanism Distorts Specific Phonemes

GW distance between formant trajectories and neurograms — 4-fold variation within mechanisms

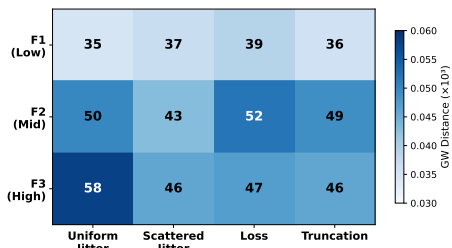
Deep Hearing Lab

By phoneme category

Category	Healthy	Uni.	Scat.	Loss	Trunc
Vowel	5	28	23	33	32
Nasal	7	34	21	33	38
Liquid	6	37	33	33	32
Fricative	8	**53**	33	49	58
Stop	7	42	29	48	**65**
Affricate	6	51	37	**52**	47
Flap	9	**83**	**70**	71	68
Glottal	8	70	56	64	**78**

GW distance ($\times 10^3$). Higher = more distortion. Bold = peak per mechanism.

By formant frequency



F3 (high) most disrupted.

Uniform jitter: steepest gradient (66% F1→F3).

Overall variation **4-fold** (flaps 83 vs nasals 21); between-mechanism only ∼14%.

Two-Stage DNN Classifier

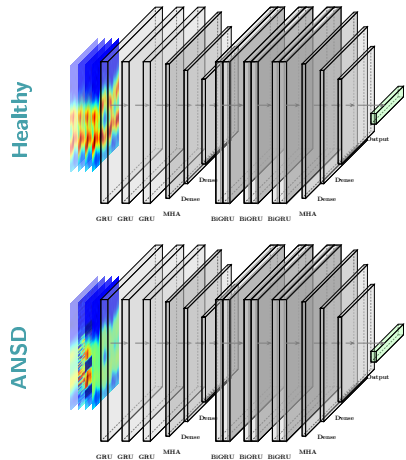
Extends Brochier et al. (2022): 9 categories not 39 phonemes · attention added · 4 populations trained separately

	Brochier	Ours
GRU layers	2	3
Units (S1)	64	128
Units (S2)	128	192/128
Attention	–	MHA
Output	40 phonemes	9 categories
Populations	1	4
Loss	Cross-entropy	Focal
Parameters	∼100K	862K

Stage 1: 100 ms windows → acoustic features

Stage 2: 610 ms context → temporal integration

862K parameters total · TIMIT corpus (462 train / 168 test speakers).



Experimental Design: Four Models, Three Tests

Disentangling population effects (healthy vs ANSD) from environmental effects (silence vs noise)

Deep Hearing Lab

Training: 4 models

	Silence	Noise
Healthy	H-Sil	H-Noi
ANSD	A-Sil	A-Noi

ANSD training: each utterance receives one *random* perturbation (jitter/loss/truncation) — simulates clinical heterogeneity.

Noise: additive WHAM! at 0–20 dB SNR.

Testing: 3 protocols

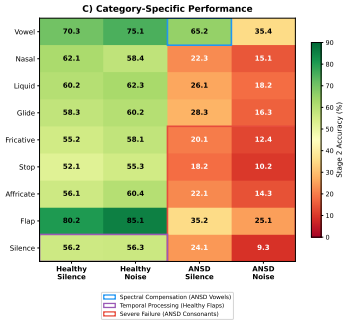
1. **Matched:** train & test on same population
   Does ANSD degrade classification?
2. **Cross-population:**
   Train Healthy → Test ANSD
   Train ANSD → Test Healthy
   Do learned strategies transfer?
3. **Mechanism-specific:**
   Each perturbation tested in isolation → severity ranking

Key question: If ANSD simply adds internal noise, noise training should help both populations equally. Divergent responses → ANSD involves **structural** signal degradation, not just increased variability.

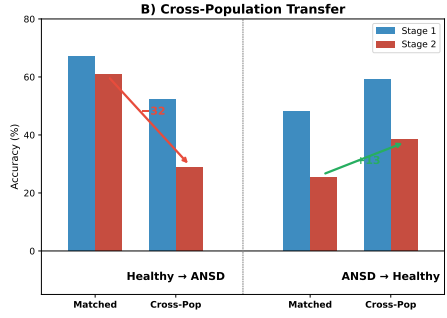
Classification Failures and Noise Robustness

Encoding distortions translate to systematic failures — and noise makes it worse for ANSD

Category-specific failures



Cross-population transfer (silence)



Vowels spared (65%). **Consonants** collapse (18%). **Flaps**: 80% → 35% (as OT predicted).

Transfer: Healthy-trained fail on ANSD (−32 pts) · ANSD-trained generalise to healthy (+13 pts).

ANSD models learn spectral strategies that generalise · healthy models learn timing strategies that don't.

Mechanism-Specific Severity

Hierarchical delta (S2−S1) reveals which mechanisms allow integration and which destroy it

Population	Loss	Truncation	Uniform Jitter	Scattered Jitter
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

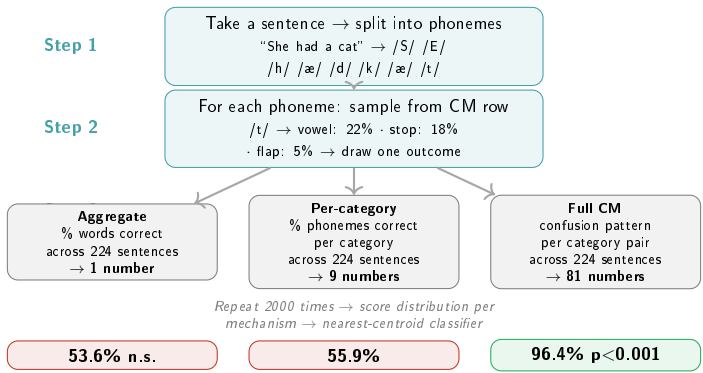
Loss only: Stage 2 integration **helps** across all populations.
Temporal structure preserved within surviving channels.

Temporal perturbations: Stage 2 integration **catastrophically fails** (~ -25 pp).
Timing destroyed $\rightarrow$ integration amplifies errors.

Same aggregate score, completely different integration regimes $\rightarrow$ aggregate testing cannot distinguish them.

How We Simulate the Speech Test

From confusion matrices to mechanism identification



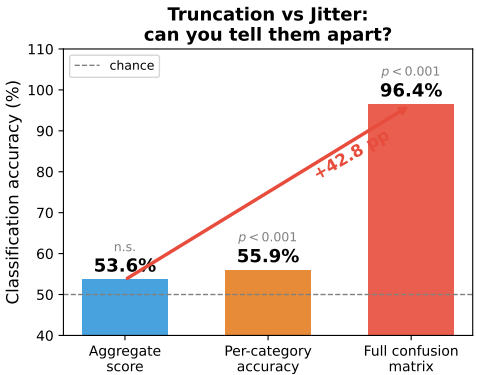
Towards Clinical Validation

Aggregate scoring destroys 92% of mechanism-discriminative information

Deep Hearing Lab

Simulated clinical test

We simulate a speech test from the DNN confusion matrices. Can the score tell Truncation from Jitter?



What this means clinically

Aggregate score: % correct — one number. Loses all pattern information. Cannot distinguish mechanisms.

Per-category accuracy: diagonal of the 9×9 matrix — how often each category is correct, but not *what it gets confused with*.

Full 9-category confusion matrix: off-diagonal entries reveal which categories are confused with which — the mechanism signature lives here. Same test, scored differently.

Part 2: Take-Home Messages

Deep Hearing Lab

What we showed

- Different ANSD mechanisms produce **distinct phoneme-specific** encoding failures (4-fold variation within mechanisms)
- Brief consonants severely disrupted; sustained vowels preserved — as predicted by temporal demands
- Cross-population transfer is **asymmetric**: ANSD strategies generalise, healthy strategies fail

Why it matters

- **Phoneme confusion patterns** carry more diagnostic information than aggregate scores
- **Mismatched intervention can harm:** temporal training may reinforce cues that jitter patients cannot encode
- Framework enables **computational phenotyping** — mechanism identification from behavioural testing alone

Bottom line: Same % correct, completely different confusion patterns. The diagnostic information is in the pattern, not the score.

Thank you

Questions?

Dr. Marta Campi

Deep Hearing Lab · University Hospital Zurich

marta.campi@usz.ch

CI outcome prediction: M. Campi, A. Huber, T. Goehring · Swiss National CI Database

ANSD: M. Campi, E. Partouche, C. Gaultier, G. Gerenton, P. Avan · bioRxiv 2026/701491