

Can learned voice cues improve speech recognition in noise?

Spectrotemporal profiling and individualized auditory training

Amaç Erdem

Learning, perception and computational auditory research

Abstract

Familiar voices can improve speech recognition in noise, but this benefit does not identify which learned acoustic features are useful or how to train them. We propose a speech study in adults with sensorineural hearing loss. A measured bank of pre-generated utterances will support contrasts in spectral structure and segmental timing, with fundamental frequency constrained. Randomized familiarization, reciprocal component comparisons and separate-day replication will test learning-dependent cue use. A subsequent trial will compare profile-guided training with equally intensive control training. Its primary endpoint is recognition of new natural talkers four weeks after training. Clinical efficacy and the incremental predictive value of Musical Intelligence will be evaluated separately.

Keywords: speech-in-noise; talker learning; spectral envelope; temporal cue use; auditory rehabilitation

1. Question and prior evidence

Which spectral and temporal features of a learned talker improve recognition when carried into new speech? Do their contributions combine additively or interact? Can that profile identify training targets that produce retained gains for previously unheard talkers?

Familiar-talker benefits have been demonstrated in older listeners with hearing loss.¹ Laboratory familiarization can also improve intelligibility and reduce listening effort.² These findings motivate learning as an intervention target; they do not establish which acoustic components produce the benefit.

Speech-cue weighting varies across listeners with hearing loss,³ and changing masker modulation shifts sentence-level spectral weights.⁴ A listener's profile is therefore conditional on the task and acoustic scene. The objective is better recognition, not equal spectral and temporal weights.

2. From music to speech

Originating in musical source selection, the proposal asks whether learned acoustic structure helps listeners follow a target amid competition. Speech makes that benefit directly measurable as word recognition; transfer from music training is not assumed.

3. Testable hypotheses

H1 · Learned component use. Familiarization changes the contribution of at least one prespecified spectral or timing contrast beyond matched exposure to a control voice. Both adding a B-like component to A-like speech and replacing it in B-like speech are tested.

H2 · Joint dependence. Familiarization changes the spectral × timing interaction. Its sign is not prescribed; a null result is informative only if uncertainty excludes a prespecified meaningful interaction.

H3 · Retained transfer. A frozen profile-guided selection policy improves speech reception threshold for new natural talkers four weeks after training, relative to a dose-matched active policy.

4. Research approach

Spectral-weighting methods provide a direct measurement precedent.⁴ A rhythm/EEG preprint links target/background timing to recognition and neural tracking.⁵ The proposed bridge is to ask whether *learned* component use predicts trainability and retained transfer.

The project combines auditory assessment, learning experiments and computational modelling to relate individual cue use to speech recognition.

Table 1. Established starting points and the proposed next test.

Evidence	What it supports	What remains to be tested
Familiar-talker benefit ^{1,2}	Recognition depends on experience	Which learned components contribute?
Context-dependent cue weights ^{3,4}	Cue use varies with listener and scene	Is the learning-related profile repeatable?
Targeted auditory practice ⁶	Task-level gains are possible	Does profile-guided practice transfer?

5. MI resolves features and qualifies the stimulus contrasts

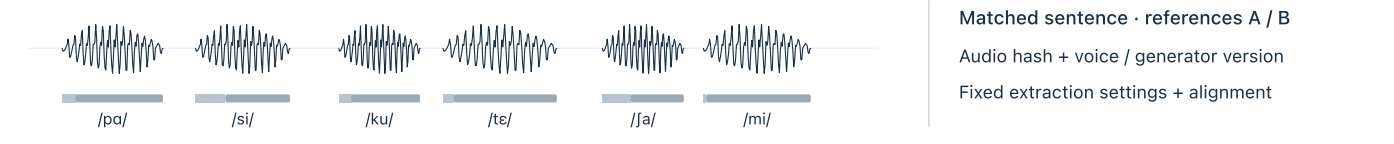
ElevenLabs will generate matched-text candidates. R^3/T^3 will measure the resulting audio, preserving the complete feature vectors and extraction settings. Two recordings with the same plotted distance may differ in their component measurements; X/Y proximity is therefore insufficient for experimental acceptance.

The stimulus pilot will select a small set of spectral and timing contrasts. Each comparison must achieve its intended feature change while bounding off-target changes. The proposed four-cell design proceeds only if enough recordings satisfy these rules across sentences and talker pairs.

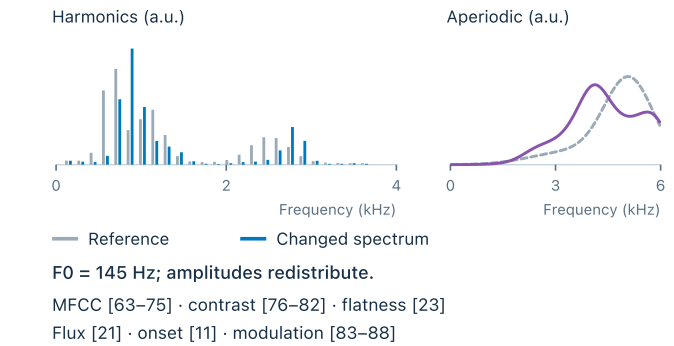
From waveform to qualified contrasts

MI ENGINE / FEATURE-RESOLVED MEASUREMENT AND STIMULUS ACCEPTANCE

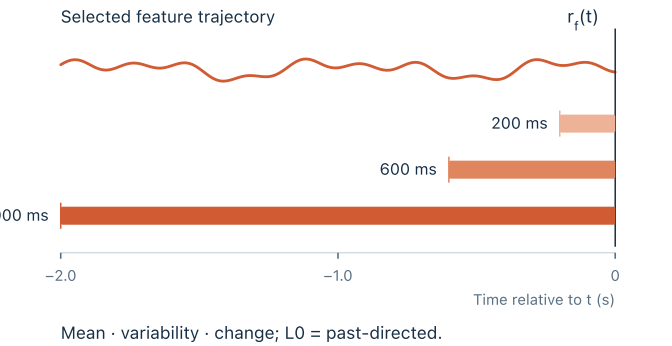
● a One signal, a traceable measurement record



● b R^3 · spectral components



● c T^3 · temporal organization

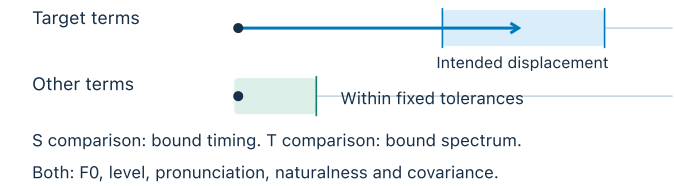


d Speech-specific integration

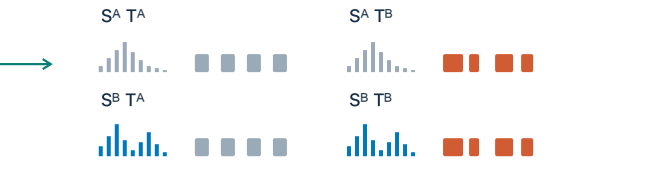
Checked phoneme timing · F0 · F1–F3 · harmonic / aperiodic measures · calibrated level

PROPOSED ADAPTER

● e Qualify the contrast



● f Require all four cells



FAIL → regenerate / resynthesize → re-measure

PASS → full vectors for analysis; X/Y for display

Procedural illustrations; no measured spectra, accepted research cells or participant results.

Figure 1. MI as the measurement and stimulus-audit layer. a–c, A schematic utterance, harmonic/aperiodic spectra and proportional window supports explain existing R^3/T^3 operators (zero-based indices). d–f, Proposed speech measurements feed target/off-target acceptance and four-cell qualification. The intervals express rules, not observed measurements. Analysis retains full vectors; X/Y provides a display. Stimulus selectivity and coverage remain untested.

5.1 Specify features before testing recognition

R^3 contributes envelope, contrast, flatness, flux and modulation channels. These are computational descriptors, not direct formant or phoneme measurements. A proposed speech layer adds validated F0, formant, aperiodic-energy and alignment measurements. Preserve durations on the original clock when aligning spectral comparisons. Preregister the target coordinates and nuisance set; if several terms remain coupled, interpret their joint contrast.

5.2 Qualify comparisons, not just recordings

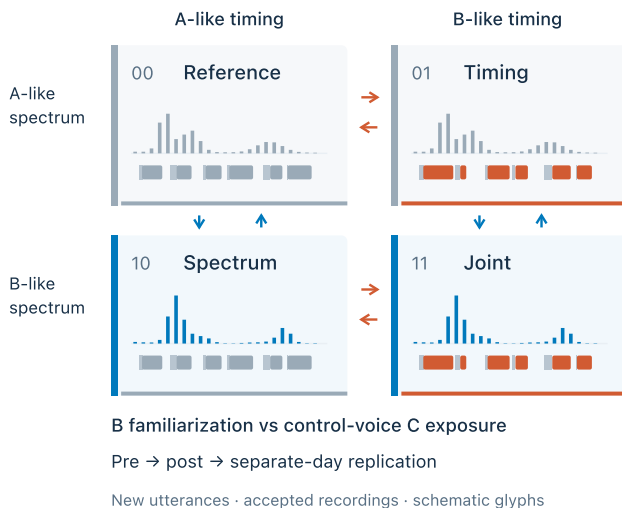
Retain per-feature target deviations, nuisance deviations and missing-value flags relative to both references. Fix scaling and acceptance tolerances in the independent pilot; audit residual covariance and all four cells across texts. Recheck after playback or hearing-aid processing.⁷ Failed candidates return to generation or resynthesis. Passing the audit supports a specified acoustic comparison; it does not establish perceptual independence.

6. Separate component learning from clinical transfer

Stage I will recruit English-speaking adults with bilateral sensorineural hearing loss and an age-matched normal-hearing comparison group. Audiometry, speech-in-noise performance, device use and language background will characterize the sample. Cochlear implant users will not be pooled with acoustic-hearing participants.

Talkers are counterbalanced across roles. Participants are randomized to familiarization with B or matched exposure to a third talker C, absent from the A/B test bank. A is an untrained reference, not a zero-feature voice. New sentences are used in each pre-training, post-training and replication session.

● a Reciprocal component comparisons



● b Independent transfer trial

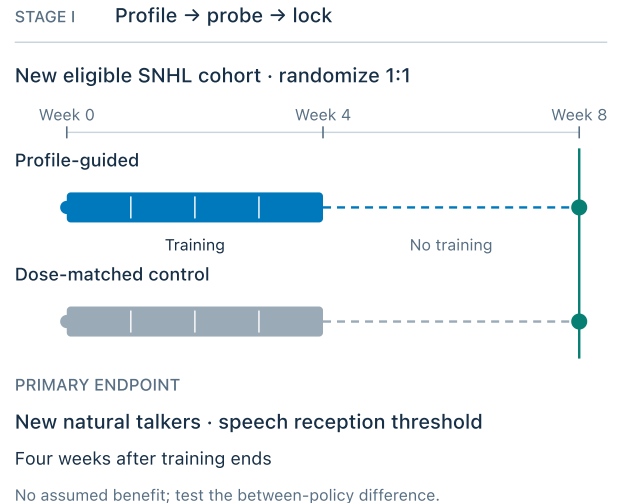


Figure 2. Learning and transfer require different controls. **a**, Spectral stems and timing strips depict the intended A-like/B-like component combinations; the glyphs are schematic. Reciprocal contrasts add or replace learned features in accepted recordings. **b**, Replicated, modifiable profiles determine a frozen selection policy. Its efficacy is tested in an independent cohort against active training, with the primary retention endpoint four weeks after training ends.

6.1 Mapping in competing sound

Primary mapping uses two competing talkers. A separately analysed speech-shaped-noise condition tests masker dependence. Set an individual challenge level before familiarization, then hold it fixed across that listener's matched cells. Target cues contain different words from the scored sentence. Counterbalance masker identities and trial order.

The listener types the target words. Primary scoring is open-set word accuracy; intrusions, voice identification and reported effort are secondary. Store raw responses and prespecify spelling normalization. Word errors are not automatically attributed to a phoneme or cue.

6.2 From a profile to a training target

Re-estimate spectral, temporal and joint effects on a separate day and new utterances. Then use independent, counterbalanced learning probes to test whether each candidate contribution can change. A small baseline effect is insufficient grounds for targeting it.

The policy uses repeatability, uncertainty and probe response. Thresholds and eligibility rules are fixed in Stage I; report the eligible proportion. Any benefit applies to this defined population of eligible profiles.

6.3 Independent policy trial

In a new SNHL cohort, eligible participants are randomized 1:1 to profile-guided or balanced control selection. Both arms receive the same task families, talker pool, feedback, duration and masker progression. The intended dose is 20 minutes/day, five days/week for four weeks, subject to feasibility confirmation.

Begin near the learned reference and introduce accepted recordings with reduced familiar support. Guidance may prioritize spectral, temporal or joint variation according to the locked rule. The control uses a prespecified balanced schedule; the manipulated factor is selection policy.

6.4 Outcome and retention

The primary endpoint is baseline-adjusted speech reception threshold (SRT) at week 8, four weeks after training ends. Test natural talkers and sentences never used in profiling or training; lower SRT is better. Assessors are blinded. Immediate post-training performance is secondary.

The goal is retained speech recognition in noise. Better recognition does not demonstrate recovery of cochlear sensitivity or normalization of a clinical audiogram.

7. Learning-specific contrasts

Fit word-correct responses with a hierarchical logistic model including component cell, exposure arm, session and hearing group. Model listener, sentence, talker-pair and masker variation. Let q_{st} denote estimated log odds for spectral state s and timing state t :

$$C_S = [(q_{10} - q_{00}) + (q_{11} - q_{01})] / 2$$
$$C_T = [(q_{01} - q_{00}) + (q_{11} - q_{10})] / 2$$
$$I = q_{11} - q_{10} - q_{01} + q_{00}$$

The learning-specific effect is the pre-to-post change in C_S , C_T or I after B training minus that after C training. Report probability-scale marginal effects as well. Retain the prespecified feature vectors in the model; report bundle effects where terms cannot be separated. These contrasts measure dependence on the accepted stimulus manipulation; they are not percentages of attention.

8. Clinical and computational tests

For H3, estimate the baseline-adjusted between-arm SRT difference at retention, using intention-to-treat analysis and prespecified missing-data sensitivity analyses. Improved trained-item performance alone is insufficient.

For MI, add T^3 summaries to a matched acoustic/exposure baseline. Candidate native spans of 200, 600 and 2,000 ms are modelling choices, not attention categories. L0 is past-directed; centred R^3 frames and clip normalization preclude online causality. Tune in development folds; test log loss and calibration on held-out listeners and talkers.

Table 2. Findings that distinguish the proposed claims.

Observed pattern	Permitted interpretation
Component learning; no retained transfer	Learning is task-specific; clinical hypothesis unsupported
Retained transfer; no predicted cue change	Policy may help; the proposed component account is weakened
Behavioral benefit; no MI predictive increment	Clinical result stands; incremental model claim fails
Confounded cells or imprecise effects	Mechanistic test is inconclusive; redesign or more precision is needed

References

1. Souza P et al. (2013). The advantage of knowing the talker. *J Am Acad Audiol*. [10.3766/jaaa.24.8.6](#)

2. Baxter F, Smith HJ, Holmes E (2025). Voice familiarization training improves speech intelligibility and reduces listening effort. *Trends Hear*. [10.1177/23312165251401318](#)

3. Souza P, Gallun F, Wright R (2020). Contributions to speech-cue weighting in older adults with impaired hearing. *JSLHR*. [10.1044/2019_JSLHR-19-00176](#)

4. Shen Y, Langley L (2023). Spectral weighting for sentence recognition in steady-state and amplitude-modulated noise. *JASA Express Lett*. [10.1121/10.0017934](#)

5. Schell E, Cheng TZ, Shen Y, Zhao TC (2026). Rhythm modulates perception and neural tracking of speech in a speech-in-noise task. *PsyArXiv preprint*, v2. [10.31234/osf.io/h7duq_v2](#)

6. Koprowska A et al. (2023). The effect of phoneme-based auditory training on speech intelligibility in hearing-aid users. *Int J Audiol*. [10.1080/14992027.2022.2135032](#)

7. Yun D, Lentz JJ, Shen Y (2024). The noise reduction algorithm may not compensate for the degradation in output SNR caused by wide dynamic range compression. *Am J Audiol*. [10.1044/2024_AJA-24-00011](#)

8. Yücel MA et al. (2021). Best practices for fNIRS publications. *Neurophotonics*. [10.1117/1.NPh.8.1.012101](#)

9. Optional neural extension

After behavioral feasibility, separate continuous-speech blocks may test target-versus-masker EEG tracking under acoustically counterbalanced conditions. The rhythm and neural-tracking preprint supplies a relevant precedent,⁵ not proof that improved tracking mediates rehabilitation. EEG remains secondary to recognition and transfer.

Available fNIRS could address slower prefrontal effort-related responses, with short-separation and physiological controls.⁸ It cannot identify a 150-ms attention window. Neither modality is needed for the primary policy trial.

10. Decisions before registration

An independent pilot will establish stimulus quality, profile precision and task burden. Simulation-based sample sizing must cover both the component interactions and the retained between-arm SRT contrast. Fix meaningful-effect bounds, stopping rules, exclusions, multiplicity control and missing-data handling before confirmatory enrollment.

Obtain ethics approval, voice-use permission and licensed assessment materials. The bank generator, feature definitions, model versions and selection policy must be versioned. The present interface is a demonstrator: no participant, generation or MI-extraction results are reported.