

Can learned voice cues improve speech recognition in noise?

Amaç Erdem | amace@bu.edu | ORCID 0009-0000-1711-0036

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.

Academic edition r4 · 12 October 2026 · Source: draft 1.1, 5 October 2026. Unregistered proposal; no participant results.

Spectrotemporal profiling and individualized auditory training

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

1. Question and prior evidence

We ask which spectral and temporal features of a familiar talker help listeners recognize new speech, whether these contributions interact, and whether the resulting profile can guide training that benefits recognition of 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 masker modulation changes sentence-level spectral weights [4]. We therefore expect a profile to depend on the task and acoustic scene. Training will aim to improve recognition; equal spectral and temporal weights are not a treatment goal.

2. From music to speech

The proposal grew from a question about following a musical source amid competing sounds. In speech, the contribution of learned acoustic structure can be tested directly through word recognition. The study does not assume that music training transfers to speech.

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 prespecified 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 measurement precedent [4]. A rhythm/EEG preprint also relates target/background timing to recognition and neural tracking [5]. This study extends those questions to learned component use, its change with training and its relationship to 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. Stimulus measurements and comparison criteria

ElevenLabs will generate matched-text candidates, which R^3/T^3 will measure. Complete feature vectors and extraction settings will be retained. Since recordings at the same plotted distance can differ in individual features, X/Y proximity alone will not determine 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.

5.1 Specify features before testing recognition

R^3 supplies envelope, contrast, flatness, flux and modulation descriptors. A proposed speech layer will add validated F0, formant, aperiodic-energy and alignment measurements; the existing channels do not directly measure formants or phonemes. Spectral comparisons will retain durations on the original clock. Target coordinates and nuisance features will be preregistered. Where several features remain coupled, their joint effect will be interpreted.

5.2 Check the acoustic contrasts

For each reference, records will include per-feature target and nuisance deviations and missing values. An independent pilot will set scaling and acceptance tolerances, assess residual covariance and check all four cells across texts. Comparisons will be reassessed after playback or hearing-aid processing [7]. Candidates that fail will be regenerated or resynthesized. These checks establish acoustic differences within specified tolerances; perceptual independence remains an empirical question.

6. Component learning and clinical transfer

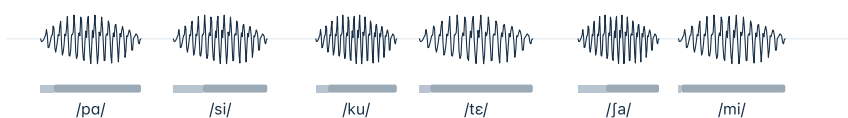
Stage I will recruit English-speaking adults with bilateral sensorineural hearing loss and an age-matched normal-hearing comparison group. Audiometry, speech-in-noise performance, device use and language background will characterize the sample. Cochlear implant users will not be pooled with acoustic-hearing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

From waveform to qualified contrasts

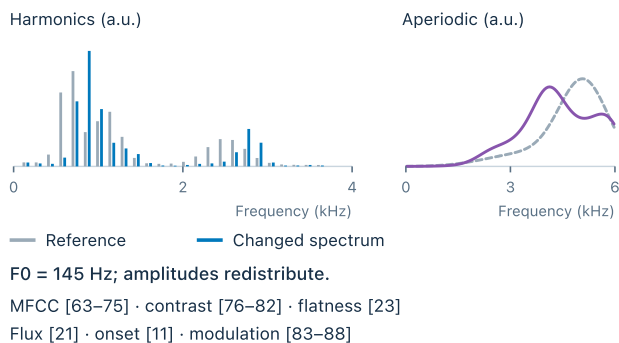
MI ENGINE / FEATURE-RESOLVED MEASUREMENT AND STIMULUS ACCEPTANCE

● a One signal, a traceable measurement record

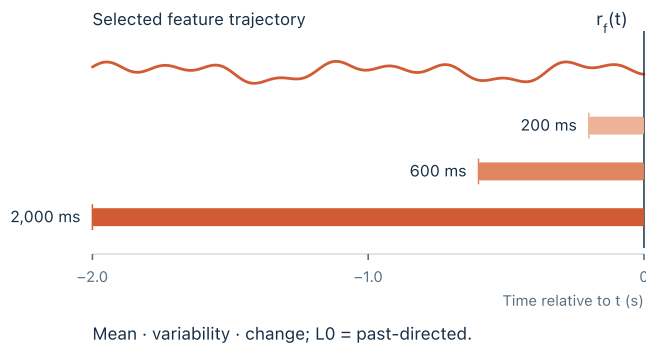


Matched sentence · references A / B
Audio hash + voice / generator version
Fixed extraction settings + alignment

● b R^3 · spectral components



● c T^3 · temporal organization

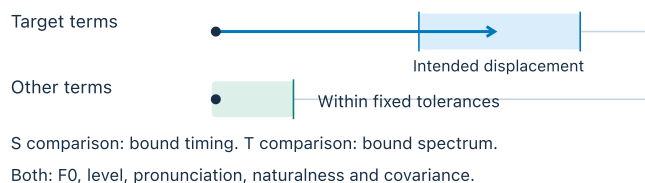


d Speech-specific integration

Checked phoneme timing · F_0 · F_1 – F_3 · 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.

Fig. 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.

6.1 Mapping in competing sound

Primary mapping uses two competing talkers. A separately analysed speech-shaped-noise condition tests dependence on the masker. An individual challenge level will be set before familiarization and held constant across matched cells. Target cues use words absent from the scored sentence. Masker identities and trial order will be counterbalanced.

Listeners type the target words. Open-set word accuracy is the primary outcome; intrusions, voice identification and reported effort are secondary. Raw responses will be stored, with spelling normalization specified in advance. Word errors will not be assigned to particular phonemes or cues without further evidence.

6.2 From a profile to a training target

Spectral, temporal and joint effects will be estimated again on a separate day with new utterances. Independent, counterbalanced learning probes will then test whether each contribution can change. A small baseline effect alone will not justify selecting it as a training target.

The policy will use repeatability, uncertainty and probe response, with thresholds and eligibility rules set in Stage I. We

will report the proportion meeting these rules; estimates of benefit will apply to that eligible population.

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.

Training begins near the learned reference and introduces accepted recordings with progressively less familiar support. The prespecified guidance rule may emphasize spectral, temporal or joint variation. Control training uses a balanced schedule, allowing the trial to isolate the effect of selection policy.

6.4 Outcome and retention

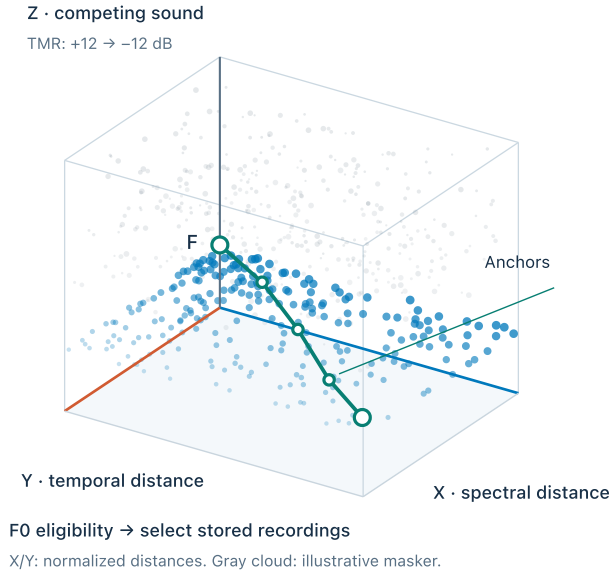
The primary endpoint is baseline-adjusted speech reception threshold (SRT) at week 8, four weeks after training ends. Blinded assessors will test natural talkers and sentences absent from profiling and training; lower SRT indicates better recognition. Immediate post-training performance is secondary.

This endpoint tests retained speech recognition in noise. It does not measure recovery of cochlear sensitivity or normalization of the clinical audiogram.

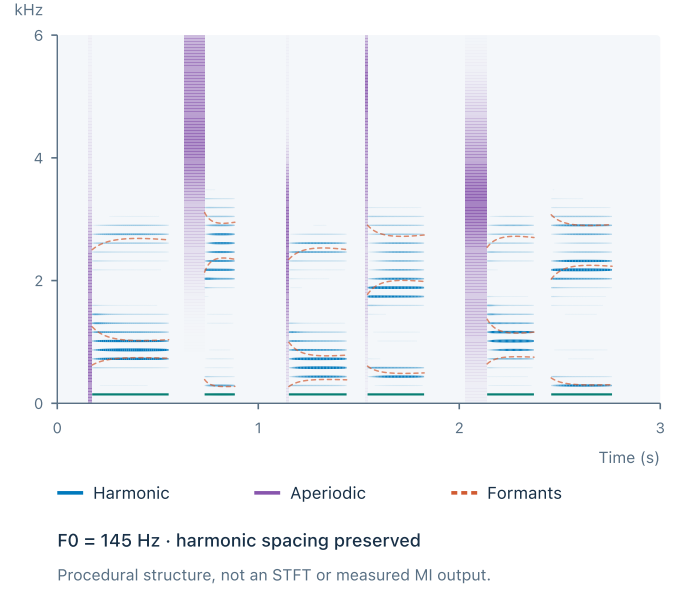
Explore the stimulus space

INTERACTIVE 3D COMPANION · Original static figure below.

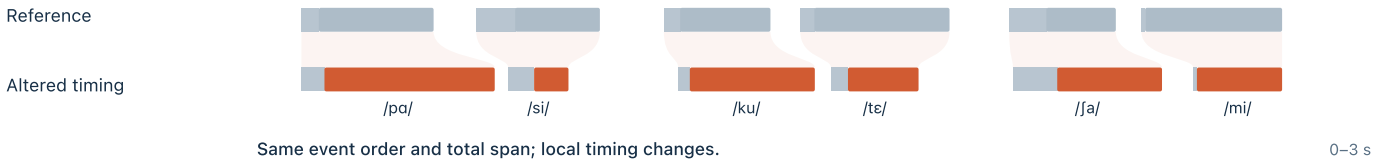
● a A discrete, reference-relative bank



● b Speech structure at fixed F0



● c Timing redistributes segments and pauses



Interactive companion. The stimulus map and its acoustic interpretation. **a**, X/Y select an F0-compatible bank anchor; Z specifies target-to-masker ratio after target selection. **b**, The procedural speech model shows harmonics, formant emphasis and aperiodic consonant energy at fixed F0. **c**, Timing redistributes segment durations and pauses at fixed utterance span. All points, curves and amplitudes are schematic. The companion web article animates the same controls; it contains no generated research audio or MI measurements.

Open the interactive 3D figure: mymusicalmind.io/speech#space. Drag to rotate and move X, Y and Z in the web view. The PDF preserves the original static companion.

7. Learning-specific contrasts

Word-correct responses will be fitted with a hierarchical logistic model including component cell, exposure arm, session and hearing group, with variation among listeners, sentences, talker pairs and maskers. 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 the change after C training. Probability-scale marginal effects will also be reported. The model retains the prespecified feature vectors; inseparable terms will be reported as joint effects. These contrasts describe sensitivity to the accepted stimulus manipulation, rather than percentages of attention.

8. Clinical and computational tests

H3 will be tested through the baseline-adjusted between-arm SRT difference at retention, with intention-to-treat analysis and prespecified missing-data sensitivity analyses. Improvement restricted to trained items would not support retained transfer.

To test MI's contribution, T^3 summaries will be added to a matched acoustic/exposure baseline. Candidate native spans of 200, 600 and 2,000 ms are modelling choices, not attention categories. Although L0 uses past information, centred R^3 frames

and clip normalization prevent an online causal interpretation of the full calculation. Tuning will use development folds; log loss and calibration will be tested on held-out listeners and talkers.

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.

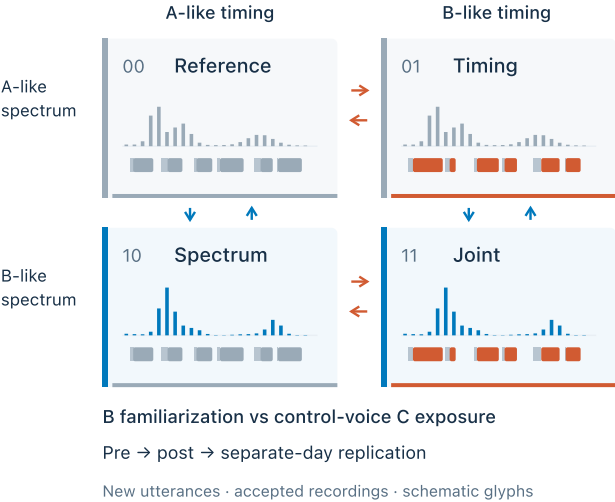
Available fNIRS could address slower prefrontal effort-related responses, with short-separation and physiological controls [8]. It cannot identify a 150-ms attention window. Neither modality is needed for the primary policy trial.

10. Decisions before registration

An independent pilot will assess stimulus quality, profile precision and task burden. Sample-size simulations will cover the component interactions and retained between-arm SRT contrast. Meaningful-effect bounds, stopping rules, exclusions, multiplicity control and missing-data procedures will be specified before confirmatory enrollment.

Recruitment will require ethics approval, voice-use permission and licensed assessment materials. The bank generator, feature definitions, model versions and selection policy will be archived by version. The current interface demonstrates the design; no participant, generation or MI-extraction results are reported.

● a Reciprocal component comparisons



● b Independent transfer trial

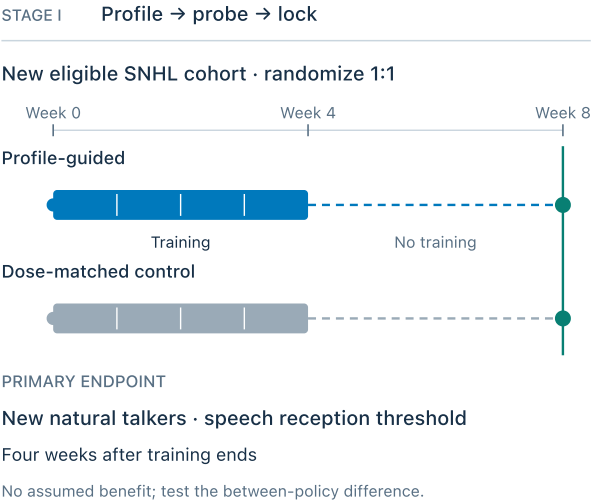


Fig. 2. Learning and transfer require different controls. **a**, Spectral stems and timing strips depict the intended A-like/B-like component combinations; the glyphs are schematic. Reciprocal contrasts add or replace learned features in accepted recordings. **b**, Replicated, modifiable profiles determine 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.

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

Protocol and materials

Source proposal: draft 1.1, 5 October 2026. mymusicalmind.io/speech. This corrected edition restores the original stimulus-space companion alongside the two numbered manuscript figures. All three are preserved as original vector PDF objects with their original content and captions. The live 3D workspace is linked from the companion in the main text; the PDF contains its original static illustration.