AERIN: ACOUSTIC EVIDENCE ROUTING AND INTEGRATION FOR COMPACT SPEECH REPRESENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Different speech tasks depend on different properties of the same signal, yet frozen self-supervised encoders leave the organization of those properties largely to pretraining. Downstream systems can select or combine internal representations, but only after those representations have already been formed. We ask whether a reusable speech representation can instead be structured around explicit acoustic evidence while still learning how that evidence should be used. We introduce AERIN (Acoustic Evidence Routing and Integration Network), a raw-waveform encoder that organizes speech into Boundary, Envelope, Phonation, and Texture evidence streams and learns their contribution through fine-grained routing and contextual integration. Thus, the sources of evidence are specified by design, while their encoding and combination remain learned from data. AERIN is distilled from WavLM-base without downstream task labels, frozen, and evaluated across 11 clinical, phonetic, and affective tasks. It remains competitive with substantially larger pretrained encoders and with its WavLM-base teacher despite using 25.66 M parameters. Routing and stream-ablation analyzes further show that the learned representation is not governed by one globally dominant stream or a fixed mixture: routing is selective at individual frame-feature positions, and downstream sensitivity to the streams varies across tasks. These results support evidence-organized representation learning as a broadly transferable alternative to relying only on emergent layer representations.

1 INTRODUCTION

A speech signal encodes multiple types of information, and each downstream task relies on a different subset of it. Phonetic and linguistic tasks rely strongly on what is spoken, whereas affective and clinical tasks can depend more on how speech is produced, including voice quality, prosody, timing, and other acoustic characteristics. Self-supervised speech learning nevertheless makes it possible to reuse a single pretrained encoder across such different tasks (Hsu et al., 2021; Chen et al., 2022; Yang et al., 2021). This is especially valuable when labeled data are limited because the encoder can remain frozen while only a lightweight downstream model is learned. The same strategy has therefore been applied to depression and Parkinson's disease assessment (Wu et al., 2023; La Quatra et al., 2024) and to affective speech, where models such as emotion2vec further specialize representation learning toward emotion-related information (Ma et al., 2024b). These applications expose a common challenge. A reusable encoder must support tasks that depend on different properties of the same signal and therefore make different forms of useful acoustic evidence available within a shared representation.

Where that evidence becomes available is largely determined during pretraining. Self-supervised speech encoders form a hierarchy of internal representations whose acoustic, phonetic, and linguistic content changes with depth and can differ across pretrained models (Pasad et al., 2021; 2023). Consequently, the representation most useful to a downstream task can depend on both the encoder and the task. Existing systems commonly address this by selecting one internal layer or by learning a combination of several layers (Yang et al., 2021; Shih & Harwath, 2024; Chiu et al., 2024). Such approaches can improve access to the information already present in the pretrained hierarchy, but they do not determine how that information was organized in the first place. The encoder decides what its internal states contain during pretraining, and the downstream task subsequently decides where in that hierarchy to read from.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This distinction becomes important when the same frozen encoder is expected to serve tasks that rely on different acoustic properties. Choosing a layer can change which learned state is exposed to the downstream model, but it does not make the underlying sources of acoustic evidence explicit. Conventional acoustic descriptors offer the complementary view. Feature sets such as eGeMAPS define in advance which frequency, energy, spectral, and temporal properties are measured (Eyben et al., 2016). Their acoustic content is therefore clear, but the resulting representation is much more constrained in how those measurements can be transformed and combined. Learned encoders and conventional descriptors thus offer different advantages. One provides flexible representation learning, while the other provides an explicit account of the acoustic evidence being measured.

This motivates the question at the center of our work. Can a speech representation retain explicit structure over the acoustic evidence from which it is built, while still learning how that evidence should be encoded and combined? Such a representation should not require a downstream task to infer all useful acoustic structure indirectly from an already learned hierarchy. At the same time, it should not reduce speech to a fixed descriptor vector or assume that the same acoustic evidence is equally useful everywhere. Instead, the model should know which forms of evidence are available and learn when and how strongly each should contribute.

We pursue this idea with AERIN (**Acoustic Evidence Routing and Integration Network**). AERIN organizes the waveform into four complementary sources of evidence: Boundary, Envelope, Phonation, and Texture. Boundary, Envelope, and Phonation retain waveform-derived measurements of distinct acoustic properties, while Texture contributes learned short-time waveform structure from a shared stem. The measurements determine what evidence enters the corresponding streams, but the representations formed from that evidence remain learned. AERIN then uses competitive routing at the frame–feature level so that the contribution of each stream can change locally throughout an utterance rather than being fixed in advance.

The model preserves these stream identities while progressively placing the evidence in context. Temporal interactions capture how each source evolves, cross-stream interactions allow complementary evidence to influence one another, and a relation-aware token stage integrates information over longer ranges while allowing measured signal relations to influence attention. The result is a small predefined hierarchy of downstream representations whose acoustic inputs are structured by the architecture rather than discovered only afterward by searching through a stack of hidden layers. In this way, AERIN fixes neither the representation nor the importance of any evidence source. It specifies what evidence is available and learns how that evidence should be represented, selected, and integrated.

We train AERIN without downstream task labels through layer-matched distillation from WavLM-base (Chen et al., 2022) and freeze the resulting encoder for all downstream evaluation. The same representation is then tested across clinical, phonetic, and affective problems, allowing us to examine whether evidence organized in this way remains useful when the downstream target changes. We compare AERIN with general-purpose self-supervised encoders, an affect-specialized encoder, and conventional acoustic feature sets. We additionally analyze the learned routing and individual evidence streams to test whether the proposed organization is actually used within the trained representation.

Our main contributions are:

- We introduce a raw-waveform speech encoder that constructs its representation from four complementary sources of acoustic evidence. The architecture makes the evidence entering the representation explicit while leaving the representations derived from that evidence learned from data.
- We introduce fine-grained competitive routing and structured integration of these evidence streams. Their contribution can vary across frames and feature dimensions, while their identities are preserved through temporal, cross-stream, and longer-range interactions.
- We evaluate the same frozen encoder across 11 clinical, phonetic, and affective evaluations against 11 public speech encoders and two conventional acoustic feature sets. We complement the transfer results with routing and stream-level analyses that examine whether and how the proposed evidence structure is used by the learned representation.

The main text focuses on the proposed representation, its evaluation, and the analyses needed to test its central claims. A broader discussion of related work is provided in Appendix A.

2 METHODOLOGY

2.1 PROBLEM FORMULATION

We consider a speech corpus $\mathcal{D} = \{(\mathcal{X}_i, y_i)\}_{i=1}^{N}$, where recording $\mathcal{X}_i$ is divided into C_i fixed-length waveform chunks and y_i is its recording-level label. The encoder f_θ is trained before any downstream task labels are introduced and is then frozen; for each corpus, only a lightweight probe p_ϕ is trained. In conventional pretrained speech encoders the downstream representation is typically chosen by selecting a layer from the model stack. AERIN instead exposes two architecturally defined read-outs that differ in the intermediate evidence they aggregate.

For a waveform chunk $x \in \mathbb{R}^L$ sampled at 16 kHz, the encoder follows

$$x \xmapsto{\Phi} [h_r]_{r\in\mathcal{R}} \xmapsto{\Gamma} [\tilde{h}_r]_{r\in\mathcal{R}} \xmapsto{\Psi} (M, U) \xmapsto{\Lambda} (Z_{\text{low}}, Z_{\text{high}}), \tag{1}$$

where Φ constructs and projects the evidence streams (Section 2.2), Γ routes them (Section 2.3), Ψ integrates them over local and longer temporal context (Section 2.4), and Λ forms the downstream read-outs (Section C.4). Fig. 1 summarizes the complete encoder.

The ordered stream set is $\mathcal{R} = (\text{Boundary}, \text{Envelope}, \text{Texture}, \text{Phonation})$. All stream outputs are aligned to a common $H = 96$-sample (6 ms) temporal grid, giving T frame positions. A validity mask $m \in \{0, 1\}^T$ removes padded positions from masked operations. After conditioning and projection, every stream has the common feature width $D = 384$.

The stream definitions determine which measurements are made available to each branch. They do not impose a semantic interpretation on the features learned from those measurements. The stream representations and their interactions are learned during distillation, while their usefulness for a downstream task is assessed only after the encoder has been frozen.

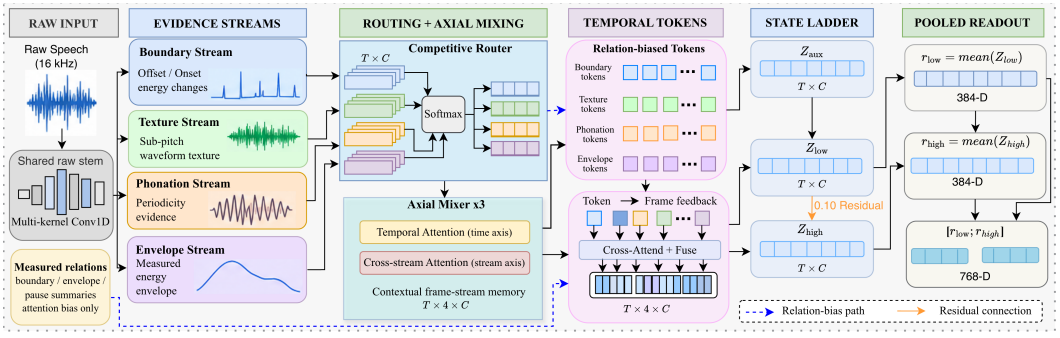


Figure 1: Overview of AERIN. Raw speech is first processed by a shared multi-kernel waveform stem and four complementary evidence streams capturing Boundary, Texture, Phonation, and Envelope information. The stream representations are combined by a competitive router and refined through repeated temporal and cross-stream axial mixer. Measured Boundary, Envelope, and pause relations provide an additional bias for the temporal-token stage, where stream-specific tokens are formed and fused back with the frame-level representation. The resulting states are organized into a hierarchical state ladder comprising Z_{aux}, Z_{low}, and Z_{high}, with a residual connection from Z_{low} to Z_{high}. Downstream evaluations use pooled read-outs from Z_{low}, Z_{high}, or their concatenation, while Z_{aux} is discarded after training.

2.2 STAGE I: EVIDENCE EXTRACTION

Stage I turns each waveform chunk into four descriptions of the signal. Boundary and Envelope streams begin from explicit amplitude and timing measurements, Phonation begins from a time-domain periodicity profile, and Texture learns short-timescale waveform structure from a shared

waveform stem. The four streams are subsequently aligned to the same temporal grid and projected to the same width before routing.

Let $q_n \in \{0, 1\}$ denote the waveform-validity mask for sample n. The amplitude-preserving waveform is

$$x_n^{\text{amp}} = q_n x_n. \tag{2}$$

It is used for measurements whose absolute amplitude should be retained. A second view is standardized over valid waveform samples,

$$\mu_x = \frac{\sum_n q_n x_n}{\sum_n q_n}, \qquad \sigma_x^2 = \frac{\sum_n q_n (x_n - \mu_x)^2}{\sum_n q_n}, \qquad x_n^{\text{norm}} = q_n \frac{x_n - \mu_x}{\sqrt{\sigma_x^2 + \varepsilon}}. \tag{3}$$

The shared stem and the Phonation stream read x^{norm}, whereas Boundary and Envelope read x^{amp}. Speech-pause measurements use the speech mask described in Appendix B.4.

2.2.1 The Four Evidence Streams

The Boundary stream represents changes in energy together with the temporal organization of speech and silence. Boundary and Envelope use $W = 256$-sample (16 ms) windowed frames on the common 96-sample hop. Let $\mathcal{F}_t$ be the waveform samples in frame t and $N_t = |\mathcal{F}_t|$. Frame energy is

$$E_t = \frac{1}{N_t} \sum_{n \in \mathcal{F}_t} \left(x_n^{\text{amp}}\right)^2. \tag{4}$$

The Boundary input combines E_t with its first and second temporal differences and the speech-pause measurements s_t,

$$b_t = \left[E_t,\, \Delta E_t,\, \Delta^2 E_t,\, s_t\right], \qquad e_{\text{bound},t} = F_{\text{bound}}(b_t). \tag{5}$$

Thus, speech-boundary structure is made available explicitly rather than being left entirely for the network to recover from the raw waveform. The exact speech-pause construction and the fallback used when explicit voice-activity information is unavailable are given in Appendix B.4.

The Envelope stream follows continuous changes in waveform amplitude. Its basic descriptor is frame-level root-mean-square amplitude,

$$A_t = \sqrt{\frac{1}{N_t} \sum_{n \in \mathcal{F}_t} \left(x_n^{\text{amp}}\right)^2 + \varepsilon}. \tag{6}$$

The complete Envelope input is

$$a_t = \left[A_t,\, \Delta A_t,\, \Delta^2 A_t\right], \qquad e_{\text{env},t} = F_{\text{env}}(a_t). \tag{7}$$

Boundary and Envelope therefore begin from related amplitude information, but Boundary additionally uses speech-pause structure whereas Envelope follows the continuous amplitude trajectory.

Texture does not begin from a hand-computed scalar descriptor. Instead, a shared waveform stem applies a bank of learned one-dimensional convolutions directly to x^{norm}, producing $h^{\text{stem}} \in \mathbb{R}^{T \times d_s}$. The stem uses kernels of 5, 15, 31, and 63 waveform samples. The Texture stream is then derived from this shared representation,

$$e_{\text{tex},t} = F_{\text{tex}}\left(h_t^{\text{stem}}\right), \qquad e_{\text{tex},t} \in \mathbb{R}^{d_e}. \tag{8}$$

It therefore provides a learned description of local waveform shape rather than an explicit estimate of periodicity or pitch.

The Phonation stream measures periodic structure directly in the time domain. Its autocorrelation is evaluated with a 512-sample (32 ms) window on the same 96-sample hop and then aligned to the common frame grid. Let $\mathcal{L}$ denote the fixed set of waveform lags. For frame t and lag ℓ, normalized autocorrelation is

$$\rho_t(\ell) = \frac{\left\langle x_n^{\text{norm}} x_{n+\ell}^{\text{norm}} \right\rangle_t}{\sqrt{\left\langle (x_n^{\text{norm}})^2 \right\rangle_t} \sqrt{\left\langle (x_{n+\ell}^{\text{norm}})^2 \right\rangle_t} + \varepsilon}. \tag{9}$$

The lag profile and its temporal difference are

$$u_t = [\rho_t(\ell)]_{\ell \in \mathcal{L}}, \qquad v_t = [u_t, \Delta u_t], \qquad e_{\text{phon},t} = F_{\text{phon}}(v_t). \tag{10}$$

For this periodicity profile, the encoder compares the waveform with copies shifted by a fixed set of sample offsets,

$$\mathcal{L} = 40, 53, 67, 80, 100, 133, 160, 200, 267, 320. \tag{11}$$

Each value in $\mathcal{L}$ specifies the number of waveform samples by which one copy is shifted before computing the normalized autocorrelation. At 16 kHz, these offsets span periodicities corresponding approximately to 400–50 Hz. The complete profile across these offsets is retained; no single lag is selected by an $\arg\max$ and no explicit F_0 estimate is produced. Phonation therefore contributes periodicity evidence across a range of candidate repetition rates rather than reducing the signal to a pitch track. Further implementation details are given in Appendix C.1.3.

Before routing, each stream is conditioned on the shared waveform stem through a gated residual adapter,

$$e_r^{\text{cond}} = e_r + \alpha_s \, \gamma_r \odot \psi_r\big(e_r, h^{\text{stem}}\big), \tag{12}$$

where $\gamma_r \in [0,1]^{T \times d_e}$ is an element-wise conditioning gate. The conditioned stream is concatenated with the shared stem and projected to the common width,

$$h_r = P_r\big([e_r^{\text{cond}}, h^{\text{stem}}]\big) + \text{Emb}_{\text{stream}}(r), \qquad h_r \in \mathbb{R}^{T \times D}. \tag{13}$$

The shared stem therefore supplies common waveform context without replacing the stream-specific evidence. Appendix C.1.4 gives the adapter and projection details.

2.3 Stage II: Per-Element Routing

The relative usefulness of the four evidence streams can vary across both time and feature dimensions. AERIN therefore assigns routing weights at each frame–feature position rather than using a single weight for an entire stream or utterance.

To make each routing decision aware of the other evidence streams, the projected streams and the shared stem are first combined into a common routing context,

$$h^{\text{ctx}} = \text{MLP}\left([h_1, \ldots, h_{|\mathcal{R}|}, h^{\text{stem}}]\right), \qquad h^{\text{ctx}} \in \mathbb{R}^{T \times D}. \tag{14}$$

For each stream r, the gate network G_r receives its projected representation h_r, the common routing context h^{ctx}, and the shared waveform representation h^{stem}. Routing is competitive: the resulting logits are normalized across the four streams at every frame-feature position,

$$g_r = \frac{\exp\left(G_r\left([h_r, h^{\text{ctx}}, h^{\text{stem}}]\right)\right)}{\sum_{r' \in \mathcal{R}} \exp\left(G_{r'}\left([h_{r'}, h^{\text{ctx}}, h^{\text{stem}}]\right)\right)}, \qquad g_r \in [0,1]^{T \times D}. \tag{15}$$

The softmax is applied over the stream axis separately at each frame t and feature dimension d, such that

$$\sum_{r \in \mathcal{R}} g_{r,t,d} = 1. \tag{16}$$

Thus, $g_{r,t,d}$ describes the relative allocation of evidence to stream r at one position in the representation. It is not a global importance score for that stream over the complete utterance. Each gate then modulates a learned value update while preserving the original projected stream through a residual connection,

$$\tilde{h}_r = \text{LN}\left(h_r + \alpha g_r \odot V_r(h_r)\right). \tag{17}$$

Here V_r is the learned value transform for stream r, $\odot$ denotes element-wise multiplication, and α controls the magnitude of the routed update. The residual connection preserves the projected stream while the gated value path changes how strongly each element is updated. Invalid frames are removed with the mask m.

2.4 Stage III: Multi-Scale Contextualization

Routing determines the local allocation of evidence but does not by itself provide temporal context. Stage III therefore combines the routed streams at two temporal scales, using frame-level axial attention to model local context and a coarser token representation to capture structure across the complete chunk.

At the frame level, the model keeps the four evidence streams separate while allowing them to exchange information. Within each stream, self-attention is applied over nearby frames so that the representation at one time point can use information from surrounding frames. The model then applies attention across the four streams for the same frame, allowing Boundary, Texture, Phonation, and Envelope information to interact without merging them into a single stream.

After this step, the representation still has a separate sequence for each evidence stream,

$$M' \in \mathbb{R}^{T\times|\mathcal{R}|\times D}, \tag{18}$$

where M'_r denotes the contextualized frame sequence for stream r.

The frame-level representation is next summarized over larger temporal regions so that information can be related across the full chunk. For each stream, the valid frames of M'_r are divided into K contiguous, nearly equal segments, and the frames within each segment are averaged,

$$\bar{U}_r = \mathrm{SegMean}_K\left(M'_r, m\right), \qquad \bar{U} = \mathrm{Concat}_{r\in\mathcal{R}}\left(\bar{U}_r\right). \tag{19}$$

Here, $\bar{U}_r$ contains the ordered temporal tokens from stream r, while $\bar{U}$ combines the tokens from all four streams. The same segmentation procedure is applied to the corresponding teacher representations during distillation.

The token sequence $\bar{U}$ is then processed by a transformer that models interactions across temporal segments and evidence streams. Each token carries its stream identity and temporal-segment position. The attention scores also incorporate pairwise information about the temporal and stream relationship between tokens, together with differences in Boundary, Envelope, and speech-pause duration quantities computed earlier in the encoder (Section 2.2). These additional terms influence how tokens attend to one another without replacing their learned representations. Their exact construction is given in Appendix C.3.

The token transformer maps $\bar{U}$ to the contextualized sequence

$$U = \mathrm{TokenTransformer}\left(\bar{U}\right). \tag{20}$$

The longer-range context in U is then returned to the frame-level representation,

$$M = \mathrm{Fuse}\left(M', U, m\right), \qquad M \in \mathbb{R}^{T\times|\mathcal{R}|\times D}. \tag{21}$$

The resulting representation M combines the frame-level interactions with the longer-range context captured by the token transformer. Both M and U are then passed to the read-out stage.

2.5 Stage IV: Read-Out Construction

The final stage constructs two representations for downstream use from the contextualized frame representation M and the token sequence U. Unlike a conventional SSL encoder, where the downstream representation is selected from an internal layer, the information available to these two read-outs is fixed by the architecture. The lower read-out combines the contextualized evidence streams with their frame-to-frame changes and a learned summary of the token sequence. The higher read-out builds on this representation and additionally uses a separate token summary together with the mean over all valid tokens,

$$\begin{aligned} Z_{\mathrm{low}} &= F_{\mathrm{low}}\left(M,\, \Delta M,\, \mathrm{Pool}_{\mathrm{low}}(U)\right), \\ Z_{\mathrm{high}} &= F_{\mathrm{high}}\left(Z_{\mathrm{low}},\, M,\, \mathrm{Pool}_{\mathrm{high}}(U),\, \mathrm{Mean}(U)\right). \end{aligned} \tag{22}$$

Here ΔM contains the first temporal differences of the contextualized streams. $\mathrm{Pool}_{\mathrm{low}}$ and $\mathrm{Pool}_{\mathrm{high}}$ are separate learned pooling operations over U, allowing the two read-outs to obtain different summaries of the same token sequence.

The two representations therefore differ in the information made available to them by construction. Z_{low} uses the contextualized streams, their local temporal variation, and one token-level summary, whereas Z_{high} additionally receives Z_{low}, a separate summary of U, and the mean token representation over the chunk. The exact pooling, and residual operations are described in Appendix C.4.

2.6 Training Objective

Training uses three states at different points in the read-out construction. The auxiliary state Z_{aux} is aligned with WavLM layer 3, Z_{low} with layer 7, and Z_{high} with layer 11. The auxiliary state is used only during representation learning and is not exposed for downstream evaluation.

For each state, teacher-student alignment combines cosine distance, layer-normalized mean-squared error, and a relational term that preserves pairwise similarities between examples (Tung & Mori, 2019). The same teacher depths also supervise ordered temporal tokens, encouraging the student to preserve within-chunk temporal structure in addition to the state-level representation. We further apply VICReg-style variance and covariance regularization (Bardes et al., 2022), augmented with a cosine-spread term, and use a cross-state correlation penalty between Z_{low} and Z_{high} to discourage redundant feature directions (Zbontar et al., 2021). The complete objective is

$$\mathcal{L}_{\text{total}} = \lambda_{teach}\left(\mathcal{L}_{\text{align}} + \lambda_{temp}\,\mathcal{L}_{\text{temp}}\right) + \lambda_{geo}\,\mathcal{L}_{\text{geo}} + \lambda_{sep}\,\mathcal{L}_{\text{sep}}. \tag{23}$$

Here, $\mathcal{L}_{\text{align}}$ denote state-level teacher alignment, $\mathcal{L}_{\text{temp}}$ temporal-token alignment, $\mathcal{L}_{\text{geo}}$ within-state representation regularization, and $\mathcal{L}_{\text{sep}}$ cross-state decorrelation. λ_{teach}, λ_{temp}, λ_{geo}, λ_{sep} denote the respective weights of each loss term. Their exact definitions, state weights, projection heads, and optimization settings are given in Appendix D. After training, all distillation-specific components are discarded and the encoder is frozen for downstream evaluation.

3 Results

3.1 Frozen-Encoder Transfer Across Tasks

Table 1 evaluates AERIN across clinical, phonetic, and affective tasks. AERIN obtains the lowest mean rank in the panel, compared with WavLM-large and WavLM-base, the teacher used for distillation. This result is obtained with a 25.66 M-parameter encoder, about 3.7× smaller than WavLM-base and between 12× and 25× smaller than the large speech encoders in the comparison.

Table 1: Frozen-encoder transfer across eleven evaluations. All speech encoders are frozen, and only the downstream probe is fitted for each evaluation. Scores are F_1 for binary tasks and macro-F_1 for categorical tasks, except SAND‡, which is scored by the challenge server reports averaged-F_1. P_M denotes encoder parameters in millions, and the final column reports the mean rank across the eleven evaluations. For NeuroVoz, NV-H/P denotes HC/PD, NV-SV sustained vowels, NV-GM GANGA/MANGA, and NV-LR Listen-and-Repeat. RAV denotes RAVDESS-Speech. **Bold** marks the best score in each column. Evaluation protocols are given in Appendix E.

		Clinical					Phonetic			Affect			
System	P_M	DAIC	MODMA	IPVS	NV-H/P	SAND‡	NV-SV	NV-GM	NV-LR	TESS	RAV	ShEMO	**rank**
HuBERT-base	94.4	0.480	0.644	0.976	0.876	0.307	0.898	0.984	0.972	0.901	0.690	0.604	5.05
HuBERT-large	315.4	0.356	0.619	0.890	0.888	0.300	0.853	0.961	0.943	0.890	0.727	0.608	7.36
wav2vec2-large-robust	315.4	0.170	0.517	0.870	0.599	0.441	0.645	0.649	0.087	0.268	0.623	0.241	11.64
WavLM-base	94.4	0.524	0.626	**0.988**	0.858	0.313	0.853	0.980	0.974	**0.902**	0.691	0.608	5.00
WavLM-base+	94.4	0.369	0.676	0.976	0.814	0.425	0.893	0.978	0.973	0.896	0.719	0.626	5.32
data2vec-audio-large	313.3	0.000	0.697	0.937	0.664	0.548	0.453	0.902	0.952	0.772	0.642	0.440	9.05
XLS-R-300m	315.4	0.307	0.671	0.837	0.835	0.537	0.886	0.840	0.712	0.844	0.675	0.522	8.45
WavLM-large	315.5	0.352	0.696	0.848	0.842	0.446	0.885	**0.986**	0.976	0.897	0.747	0.680	4.64
w2v-BERT 2.0	580.5	0.042	0.484	0.648	0.797	0.334	0.604	0.940	0.802	0.878	0.729	0.492	10.09
Whisper-large-v3	637.0	0.044	0.604	0.831	0.843	0.417	0.929	0.981	**0.981**	0.893	0.747	0.760	5.82
emotion2vec+ large	164.0	**0.535**	0.455	0.804	0.662	0.239	0.712	0.687	0.224	0.891	**0.856**	**0.804**	8.64
ComParE-16	—	0.458	0.607	0.946	0.648	0.448	0.593	0.712	0.122	0.892	0.523	0.172	9.55
eGeMAPS	—	0.450	0.421	0.909	0.567	0.548	0.821	0.453	0.225	0.863	0.477	0.385	10.32
AERIN (ours)	**25.66**	0.361	**0.705**	0.982	**0.892**	**0.609**	**0.942**	0.980	0.972	0.893	0.685	0.661	**3.82**

The clinical results are mixed but span substantially different settings. AERIN gives the highest score on MODMA (Cai et al., 2022), NeuroVoz HC/PD (Mendes-Laureano et al., 2024), and SAND (Sannino et al., 2026), while reaching 0.982 on IPVS (Dimauro & Girardi, 2019), where several systems are already close to ceiling. DAIC-WOZ (Burdisso et al., 2024) is the clear exception, with a score of 0.361 below several baselines. The four NeuroVoz evaluations provide

a complementary test because the speaker set is shared while the downstream target or elicitation condition changes. They should not be counted as four independent corpus-level successes, but together show that the same representation remains useful across different uses of the same underlying speech collection.

The panel also tests transfer across language. The speech used to distill AERIN is English-language only, whereas MODMA, IPVS and SAND, NeuroVoz, and ShEMO (Mohamad Nezami et al., 2019) evaluate Chinese, Italian, Castilian Spanish, and Persian speech, respectively. These evaluations therefore test transfer beyond the language used to train the encoder. This is not unique to AERIN: HuBERT-base, HuBERT-large, and WavLM-base are also pretrained on English speech, while Whisper-large-v3 provides a multilingual baseline. We therefore interpret the panel as evidence that the compact representation remains useful across unseen languages.

The affective tasks show the clearest limit of the panel result. AERIN remains competitive on TESS but trails emotion2vec+ on RAVDESS-Speech (Livingstone & Russo, 2018) and ShEMO. This comparison also reflects different training objectives: emotion2vec+ is specialized for speech emotion recognition, whereas AERIN receives no emotion-specific supervision. The released emotion2vec+ training lineage begins with fine-tuning on EmoBox (Ma et al., 2024a) emotion data and continues with substantially larger emotion-focused training. EmoBox contains RAVDESS, TESS, and ShEMO. The affect results should consequently be interpreted in the context of much stronger emotion-specific training rather than as evidence that emotion2vec+ provides the stronger representation across the full panel. It is, however, less competitive on several clinical and phonetic corpora.

Mean rank is used as a compact description of this heterogeneous panel rather than as a statistical test. Several clinical evaluations contain small cohorts, the four NeuroVoz columns are not independent datasets, and the baseline training regimes differ substantially in data scale, language coverage, and task specialization. We therefore use Table 1 to support the conclusion that AERIN, a 25.66 M-parameter frozen encoder trained from English speech can remain competitive across heterogeneous tasks and languages, not that AERIN is decisively better than every baseline on every evaluation. Additional protocol limitations are discussed in Appendix E.7.

3.2 Evidence-Stream Contributions

Table 2 examines how the trained representation depends on the four evidence streams using two inference-time interventions. The first removes one stream while leaving the others unchanged, whereas the second retains only one stream. The intervention patterns vary substantially with the

Table 2: Evidence-stream interventions. Here s denotes the downstream evaluation score for a task, using the same metric as in Table 1. The left block reports $\Delta = s_{\text{intact}} - s_{\text{lesioned}}$ when one stream is removed. The right block reports $s_{\text{single}}/s_{\text{intact}}$ when only one stream is retained. The shared waveform stem remains active throughout. Here, C is the number of target classes and **base** is the score of the intact model.

			Remove one stream				**Retain one stream**			
Corpus	C	**base**	**bnd**	**env**	**pho**	**tex**	**bnd**	**env**	**pho**	**tex**
DAIC-WOZ	2	0.361	−0.056	−0.083	−0.056	0.000	1.00	1.38	1.38	1.15
MODMA	2	0.705	−0.023	−0.048	−0.017	−0.009	0.94	1.03	0.90	1.02
IPVS	2	0.982	+0.017	+0.031	+0.013	0.000	0.96	0.94	0.93	0.96
NV-H/P	2	0.892	−0.013	+0.002	+0.032	+0.091	0.93	0.86	0.88	0.93
NV-SV	5	0.942	+0.014	+0.014	+0.047	+0.061	0.95	0.91	0.90	0.95
NV-GM	2	0.980	0.000	−0.003	+0.003	+0.006	0.94	0.92	0.92	0.99
NV-LR	16	0.972	+0.004	+0.004	+0.015	+0.058	0.84	0.82	0.88	0.97
TESS	7	0.893	+0.003	0.000	+0.022	+0.007	0.96	0.98	0.99	0.98
RAVDESS-Speech	8	0.685	−0.007	+0.002	+0.111	+0.086	0.62	0.66	0.83	0.82
ShEMO	6	0.661	−0.002	+0.017	+0.087	+0.070	0.75	0.78	0.82	0.83

downstream target and do not reveal a fixed hierarchy among the four streams. Phonation and Texture show the clearest recurring effects, particularly on the affective evaluations and on several NeuroVoz conditions, whereas Boundary and Envelope have less consistent effects across the panel. The NeuroVoz evaluations make this target dependence especially visible: despite sharing the same

underlying speaker collection, their sensitivity to Texture changes markedly with the prediction target or elicitation condition.

The single-stream results reinforce the same interpretation. Some evaluations can be reproduced closely from one source of evidence, while RAVDESS-Speech and ShEMO retain a larger gap from the intact representation regardless of which individual stream is kept. The trained encoder therefore does not rely on a fixed source of evidence across tasks; the contribution of each stream changes with the downstream target.

These interventions should nevertheless be read as tests of dependence within the trained model. Some lesions or single-stream conditions improve performance on the smaller clinical evaluations, and IPVS is already close to ceiling. Moreover, the shared waveform stem remains available and the streams were learned jointly. The results therefore do not assign unique information to an individual stream. Additional remove-one, remove-pair, and single-stream analyses are reported in Appendix F.2.

3.3 Routing Behavior and Functional Effect

Appendix Table 11 shows that routing is relatively balanced after averaging over time and feature dimensions, but substantially more selective at individual frame-feature positions. Across the evaluation data, one stream receives more than 0.8 of the routing mass at 43.6% of valid positions, compared with 3.2% under a uniform-simplex reference. Routing also varies across corpora without a consistent ordering by task family. Thus, the aggregate stream weights obscure the local allocation decisions made by the router.

We further replace the learned gates by the uniform allocation $g_{r,t,d} = 1/R$ while leaving the rest of the trained encoder unchanged. The resulting representation has a cosine similarity of 0.744 to that obtained with learned routing, showing that the learned allocations materially change the representation. This intervention characterizes the functional effect of routing within the trained model. Appendix F.1 presents a broader view of routing.

4 Conclusion

We presented AERIN around a simple premise. A transferable speech representation need not treat the waveform as a single undifferentiated source of information. Instead, complementary acoustic evidence can be represented explicitly and combined according to what is present in the signal. AERIN realizes this idea through Boundary, Envelope, Phonation, and Texture evidence, together with competitive routing that determines their contribution locally, while hierarchical distillation provides representations at different levels of abstraction. Across diverse clinical, phonetic, and affective evaluations, the resulting encoder transfers without task-specific adaptation. More importantly, the routing and ablation results show that this transfer is not explained by one universally dominant stream or by a fixed mixture of evidence. Different streams become useful in different contexts and for different downstream problems. This supports a broader view of speech representation learning in which the objective is not only to learn a strong latent space, but also to preserve meaningful sources of acoustic evidence and learn how they should be combined.

This formulation also leaves several directions open. The current evidence set can be expanded when other speech properties are important, without changing the underlying principle of structured evidence and learned combination. A particularly important next step is to test whether the same organization of evidence remains stable across languages, recording environments, speaker populations, and larger clinical cohorts. The framework could also be extended from fixed-duration recordings to long-form or streaming speech, where the relevance of different acoustic evidence may evolve over time. More broadly, future work can investigate whether structured, selectively combined evidence provides a useful foundation beyond the tasks considered here, particularly in settings where labelled data are limited and understanding which aspects of the speech signal support a prediction is as important as the prediction itself.

AI USE STATEMENT

Large language models (LLMs) were used exclusively to assist with language editing, stylistic polishing, and readability improvements during the preparation of this manuscript. All conceptualization, methodological design, data collection, analyses, and scientific interpretations were carried out solely by the authors. The authors have thoroughly reviewed and validated all AI-assisted text and take full responsibility for the integrity and final content of this manuscript.

ETHICS STATEMENT

Our research is based on publicly available datasets. Our proposed method advances research in speech-based clinical systems, including applications such as depression detection, stress detection, Parkinson's disease detection, and emotion recognition. While such systems hold significant promise for improving healthcare accessibility and early diagnosis, they also carry risks of misuse and harm, including privacy violations, misdiagnosis, discriminatory outcomes, and unauthorized surveillance, if deployed without appropriate care. We emphasize that these systems should be used only under proper clinical guidance and oversight, and should complement-not replace-professional medical judgment. We urge practitioners and downstream users to adhere to relevant ethical, legal, and regulatory standards, including informed consent and data privacy protections, when deploying speech-based clinical technologies.

REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility of our work. Implementation details, including model architecture, training hyperparameters, and dataset preprocessing steps, are provided in Section 2 and Appendices B, C, and D. Upon acceptance, we will release the trained model checkpoints and code, and provide dataset access paths (download URLs) to facilitate replication of our results. Additional experimental details, evaluation protocols, and metrics are provided in Appendix C.

REFERENCES

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Sergio Burdisso, Ernesto Reyes-Ramírez, Esaú Villatoro-tello, Fernando Sánchez-Vega, Adrian Lopez Monroy, and Petr Motlicek. DAIC-WOZ: On the validity of using the therapist's prompts in automatic depression detection from clinical interviews. In Tristan Naumann, Asma Ben Abacha, Steven Bethard, Kirk Roberts, and Danielle Bitterman (eds.), *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pp. 82–90, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.clinicalnlp-1.8. URL https://aclanthology.org/2024.clinicalnlp-1.8/.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

Carlos Busso, Reza Lotfian, Kusha Sridhar, Ali N Salman, Wei-Cheng Lin, Lucas Goncalves, Srinivas Parthasarathy, Abinay Reddy Naini, Seong-Gyun Leem, Luz Martinez-Lucas, et al. The MSP-Podcast corpus. *arXiv preprint arXiv:2509.09791*, 2025.

Hanshu Cai, Zhenqin Yuan, Yiwen Gao, Shuting Sun, Na Li, Fuze Tian, Han Xiao, Jianxiu Li, Zhengwu Yang, Xiaowei Li, et al. A multi-modal open dataset for mental-disorder analysis. *Scientific data*, 9(1):178, 2022.

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

Heng-Jui Chang, Shu-wen Yang, and Hung-yi Lee. DistilHuBERT: Speech representation learning by layer-wise distillation of hidden-unit BERT. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7087–7091, 2022. doi: 10.1109/ICASSP43922.2022.9747490.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Michael Zeng, and Furu Wei. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

Sheng-Chieh Chiu, Chia-Hua Wu, Jih-Kang Hsieh, Yu Tsao, and Hsin-Min Wang. Learnable layer selection and model fusion for speech self-supervised learning models. In *Interspeech 2024*, pp. 3914–3918, 2024. doi: 10.21437/Interspeech.2024-849.

Qingkun Deng, Saturnino Luz, and Sofia de la Fuente Garcia. An interpretable speech foundation model for depression detection by revealing prediction-relevant acoustic features from long speech. In *Interspeech 2025*, pp. 5248–5252, 2025. doi: 10.21437/Interspeech.2025-1968.

Daria Diatlova, Nikita Balagansky, Alexander Varlamov, and Egor Spirin. Varan: Variational inference for self-supervised speech models fine-tuning on downstream tasks. In *Interspeech*, 2025.

Giovanni Dimauro and Francesco Girardi. Italian parkinson's voice and speech, 2019. URL https://dx.doi.org/10.21227/aw6b-tg17.

Florian Eyben, Martin Wöllmer, and Björn Schuller. opensmile: The Munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 1459–1462, 2010.

Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, 2016.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Daniel Galvez, Greg Diamos, Juan Ciro, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter, Maximilian Lam, Mark Mazumder, and Vijay Janapa Reddi. The people's speech: A large-scale diverse english speech recognition dataset for commercial usage. *arXiv preprint arXiv:2111.09344*, 2021.

John J Godfrey, Edward C Holliman, and Jane McDaniel. Switchboard: Telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pp. 517–520. IEEE, 1992.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 2021.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141, 2018.

Wessel Kraaij, Thomas Hain, Mike Lincoln, and Wilfried Post. The ami meeting corpus. In *Proc. International Conference on Methods and Techniques in Behavioral Research*, pp. 1–4, 2005.

Moreno La Quatra, Maria Francesca Turco, Torbjørn Svendsen, Giampiero Salvi, Juan Rafael Orozco-Arroyave, and Sabato Marco Siniscalchi. Exploiting foundation models and speech enhancement for Parkinson's disease detection from speech in real-world operative conditions. In *Interspeech 2024*, pp. 1405–1409, 2024. doi: 10.21437/Interspeech.2024-522.

Steven R. Livingstone and Frank A. Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE*, 13:1–35, 05 2018. doi: 10.1371/journal.pone.0196391. URL https://doi.org/10.1371/journal.pone.0196391.

Ziyang Ma, Mingjie Chen, Hezhao Zhang, Zhisheng Zheng, Wenxi Chen, Xiquan Li, Jiaxin Ye, Xie Chen, and Thomas Hain. Emobox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark. In *Proc. INTERSPEECH*, 2024a.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024b. ACL Anthology 2024.findings-acl.931.

Bubai Maji, Rajlakshmi Guha, Aurobinda Routray, Shamila Nasreen, and Dipanwita Majumdar. Investigation of layer-wise speech representations in self-supervised learning models: A cross-lingual study in detecting depression. In *Interspeech*, pp. 3020–3024, 2024.

Janaína Mendes-Laureano, Jorge A Gómez-García, Alejandro Guerrero-López, Elisa Luque-Buzo, Julián D Arias-Londoño, Francisco J Grandas-Pérez, and Juan I Godino-Llorente. Neurovoz: a castillian spanish corpus of parkinsonian speech. *Scientific Data*, 11(1):1367, 2024.

Omid Mohamad Nezami, Paria Jamshid Lou, and Mansoureh Karami. Shemo: a large-scale validated database for persian speech emotion detection: O. mohamad nezami et al. *Language Resources and Evaluation*, 53(1):1–16, 2019.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 914–921. IEEE, 2021.

Ankita Pasad, Bowen Shi, and Karen Livescu. Comparative layer-wise analysis of self-supervised speech models. In *ICASSP*, pp. 1–5, 2023.

Yifan Peng, Yui Sudo, Shakeel Muhammad, and Shinji Watanabe. DPHuBERT: Joint distillation and pruning of self-supervised speech models. In *Interspeech*, 2023.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, 2018.

M. Kathleen Pichora-Fuller and Kate Dupuis. Toronto emotional speech set (TESS). Borealis, 2020. URL https://doi.org/10.5683/SP2/E8H2MF.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1050. URL https://aclanthology.org/P19-1050/.

Mirco Ravanelli and Yoshua Bengio. Speaker recognition from raw waveform with SincNet. In *IEEE Spoken Language Technology Workshop (SLT)*, pp. 1021–1028, 2018.

Anthony Rousseau, Paul Deléglise, and Yannick Esteve. Ted-lium: an automatic speech recognition dedicated corpus. In *Lrec*, pp. 125–129, 2012.

Giovanna Sannino, Ivanoe De Falco, Nadia Brancati, Laura Verde, Maria Frucci, Daniel Riccio, Vincenzo Bevilacqua, Antonio Di Marino, Lucia Aruta, Valentina Virginia Iuzzolino, et al. The speech analysis for neurodegenerative diseases challenge. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 21737–21739. IEEE, 2026.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

Yi-Jen Shih and David Harwath. Interface design for self-supervised speech models. In *Interspeech*, 2024.

Joel Shor and Subhashini Venugopalan. TRILLsson: Distilled universal paralinguistic speech representations. In *Interspeech*, pp. 356–360, 2022.

Joel Shor, Aren Jansen, Wei Han, Daniel Park, and Yu Zhang. Universal paralinguistic speech representations using self-supervised conformers. In *ICASSP*, pp. 3169–3173, 2022.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1365–1374, 2019. doi: 10.1109/ICCV.2019.00145.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 993–1003, 2021.

Rui Wang, Qibing Bai, Junyi Ao, Long Zhou, Zhixiang Xiong, Zhihua Wei, Yu Zhang, Tom Ko, and Haizhou Li. LightHuBERT: Lightweight and configurable speech representation learning with once-for-all hidden-unit BERT. In *Interspeech*, 2022.

Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition, 2018. URL https://arxiv.org/abs/1804.03209.

Wen Wu, Chao Zhang, and Philip C. Woodland. Self-supervised representations in speech-based depression detection. In *ICASSP*, pp. 1–5, 2023.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. SUPERB: Speech processing universal performance benchmark. In *Interspeech*, pp. 1194–1198, 2021.

Zhao You, Shulin Feng, Dan Su, and Dong Yu. SpeechMoE: Scaling to large acoustic models with dynamic routing mixture of experts. In *Interspeech*, 2021.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 2021.

Neil Zeghidour, Nicolas Usunier, Iasonas Kokkinos, Thomas Schatz, Gabriel Synnaeve, and Emmanuel Dupoux. Learning filterbanks from raw speech for phone recognition. In *ICASSP*, pp. 5509–5513, 2018.

Neil Zeghidour, Olivier Teboul, Félix de Chaumont Quitry, and Marco Tagliasacchi. LEAF: A learnable frontend for audio classification. In *International Conference on Learning Representations (ICLR)*, 2021.

APPENDIX

Contents

A Related Work **14**
- A.1 Frozen self-supervised speech encoders . . . 14
- A.2 Reading a frozen encoder: layers, weights, and interfaces . . . 15
- A.3 Distillation and compact speech encoders . . . 15
- A.4 Structured acoustic evidence and learned combination . . . 16
- A.5 Positioning . . . 16

B Corpora **17**
- B.1 The distillation mixture . . . 17
- B.2 The evaluation panel . . . 18
- B.3 Diagnostic outside the mean rank . . . 19
- B.4 Preprocessing . . . 19
- B.5 Separation between distillation and evaluation . . . 19
- B.6 What the panel does and does not span . . . 20
- B.7 Access, licensing, and consent . . . 21

C Methodological Details **21**
- C.1 Evidence front ends and stream construction . . . 21
- C.2 Routing implementation . . . 24
- C.3 Contextualization details . . . 24
- C.4 Read-out construction details . . . 26

D Training and optimization details **27**
- D.1 Full distillation objective . . . 27
- D.2 Checkpointing and training diagnostics . . . 29
- D.3 Model, training, and distillation configuration . . . 29
- D.4 Knowledge-Distillation and Routing Diagnostics . . . 31

E Panel evaluation protocol **31**
- E.1 Evaluation order . . . 32
- E.2 Probe architecture . . . 34
- E.3 Probe optimization and model selection . . . 34
- E.4 Split construction and early stopping . . . 35
- E.5 Cross-validation protocol . . . 36
- E.6 Read-out search budget . . . 36
- E.7 Protocol limitations . . . 36

F Extended results and analysis **37**
- F.1 Routing analysis . . . 37
- F.2 Stream ablation and sufficiency analysis . . . 38
- F.3 Read-out ladder analysis . . . 40
- F.4 Metric and coverage sensitivity . . . 41

A Related Work

This section reviews the literature most closely related to our setting and organizes it around four themes: (a) frozen self-supervised speech encoders for clinical and paralinguistic analysis, (b) methods for extracting and selecting downstream representations from pretrained speech encoders, (c) compact speech encoders obtained through knowledge distillation and model compression, and (d) structured acoustic front ends and mechanisms for combining multiple sources of speech evidence. Across these areas, we focus on the aspects most relevant to AERIN and clarify how the proposed encoder differs from existing approaches. Table 3 summarizes these distinctions.

A.1 Frozen self-supervised speech encoders

Self-supervised speech encoders such as wav2vec 2.0, HuBERT, and WavLM learn representations from large unlabeled corpora and transfer them to downstream tasks with little or no encoder adap-

tation (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022). This frozen-encoder setting, formalized broadly by SUPERB (Yang et al., 2021), is particularly useful when labeled data are limited because task-specific learning can be restricted to a lightweight probe. It has consequently been adopted for clinical speech problems including depression and Parkinson's disease assessment (Wu et al., 2023; La Quatra et al., 2024), as well as for paralinguistic tasks such as emotion recognition. Other encoders specialize the pretraining distribution itself; emotion2vec, for example, learns specifically from emotional speech and transfers strongly to affective tasks (Ma et al., 2024b).

Our work adopts this frozen-transfer setting rather than proposing an alternative to it. The difference lies in what the frozen encoder makes available to the downstream task. Standard self-supervised encoders learn their internal organization during pretraining, whereas AERIN explicitly structures the input representation around complementary acoustic evidence before learning how that evidence should be encoded and combined. This makes the question of *which representation a downstream task reads* especially relevant.

A.2 Reading a frozen encoder: layers, weights, and interfaces

A frozen self-supervised encoder does not provide a single canonical representation. Layer-wise analyses show that acoustic, phonetic, and linguistic information changes with depth and that the resulting hierarchy depends on the pretraining objective (Pasad et al., 2021; 2023). Consequently, downstream systems often select one internal layer or learn how several layers should be combined (Yang et al., 2021). These choices improve access to the representations already present in the pretrained model, but they operate after the contents of those representations have been established.

Recent work has made this interface increasingly adaptive. (Shih & Harwath, 2024) show that a scalar weighted sum can lose information when features from different layers are combined dimension-wise and propose a hierarchical interface instead. (Chiu et al., 2024) learn discrete layer selection, including a dimension-wise variant, while (Diatlova et al., 2025) make layer aggregation input-dependent. Related clinical studies examine which pretrained layers are most useful for depression and cross-lingual assessment (Wu et al., 2023; Maji et al., 2024), and (Deng et al., 2025) recover acoustic explanations from a frozen foundation model after training.

These methods are closely related to our problem, but they intervene at a different point. Layer selection and aggregation decide how to read an already-formed hierarchy; they do not determine the acoustic sources from which that hierarchy was constructed. AERIN instead introduces structure before this read-out decision by separating complementary acoustic evidence and learning how strongly that evidence should contribute locally. The dimension-wise layer selection of (Chiu et al., 2024) is therefore a particularly close conceptual comparison: both approaches allow fine-grained selection, but one selects among pretrained depths whereas the other routes among explicitly defined evidence streams. Likewise, post-hoc interpretation methods such as (Deng et al., 2025) characterize acoustic information after a representation has been learned, whereas the evidence categories in AERIN are present in the computational graph by construction. This does not imply that the contribution of each measurement is known in advance; only the sources of evidence made available to the learned representation are specified.

A.3 Distillation and compact speech encoders

Knowledge distillation has been widely used to transfer representations from large self-supervised speech models into smaller students. DistilHuBERT, for example, supervises a shallow student from several HuBERT layers, while later methods such as LightHuBERT and DPHuBERT combine distillation with architecture search or structured pruning (Chang et al., 2022; Wang et al., 2022; Peng et al., 2023). For paralinguistic speech, TRILLsson similarly distils a larger representation model into compact encoders intended for transfer (Shor et al., 2022; Shor & Venugopalan, 2022).

AERIN also uses distillation, but compression is not the organizing principle of the method. WavLM-base provides supervision for a student whose representation is structured differently from the teacher, around explicit acoustic evidence and a small hierarchy of downstream states. Distillation is therefore used to transfer broad speech knowledge into this representation rather than to preserve the teacher architecture in smaller form. The compact parameter count is a consequence of this design rather than the primary objective.

A.4 Structured acoustic evidence and learned combination

Conventional acoustic representations make the contents of the representation explicit by defining the measurements in advance. GeMAPS and its extended variant eGeMAPS encode hand-specified frequency, energy, spectral, and temporal properties motivated by voice and affective analysis (Eyben et al., 2016), with openSMILE providing a widely used extraction toolchain (Eyben et al., 2010). This gives such representations a clear acoustic interpretation, but the measurements and their subsequent organization are largely fixed before learning.

Learnable front ends relax this constraint while retaining structure in the waveform transformation. SincNet parameterizes its first-layer filters as learnable band-pass filters (Ravanelli & Bengio, 2018), time-domain filterbanks learn waveform filters initialized from conventional spectral representations (Zeghidour et al., 2018), and LEAF replaces several operations of a mel-style front end with learnable counterparts (Zeghidour et al., 2021). These approaches learn how a waveform should be filtered into a representation, but they still produce a common learned front end rather than maintaining several semantically distinct sources of acoustic evidence. LEAF, for example, is explicitly designed as a learnable alternative to mel-filterbank front ends (Zeghidour et al., 2021).

A separate line of work concerns learned routing. Mixture-of-experts models use a router to choose among expert subnetworks (Shazeer et al., 2017; Fedus et al., 2022), including speech-specific variants such as SpeechMoE (You et al., 2021). In such models, routing is primarily a mechanism for dynamically allocating computation among learned experts; SpeechMoE, for instance, routes speech representations to expert subnetworks using learned routing signals. Feature-wise modulation mechanisms such as squeeze-and-excitation and FiLM likewise show that learned interactions can vary across channels or feature dimensions (Hu et al., 2018; Perez et al., 2018).

AERIN combines ideas from these directions but uses them for a different representation goal. Its streams are heterogeneous by construction: Boundary, Envelope, and Phonation retain different waveform-derived measurements, while Texture provides learned short-time waveform evidence. Routing therefore does not choose among interchangeable expert networks, but among explicitly different sources of acoustic evidence, and it does so at the frame-feature level while keeping the streams separately represented for subsequent contextualization. The novelty is consequently not gating itself, but the use of fine-grained learned routing to preserve and integrate a predefined organization of acoustic evidence within a transferable frozen encoder.

A.5 Positioning

The closest lines of work differ primarily in where they introduce structure into the speech representation. Self-supervised encoders learn a hierarchy of representations during pretraining, while layer-selection and interface methods improve how that already-formed hierarchy is accessed (Yang et al., 2021; Shih & Harwath, 2024; Chiu et al., 2024; Diatlova et al., 2025). Conventional acoustic descriptors take the opposite approach by specifying the measured signal properties in advance, but provide less flexibility in how those measurements are subsequently represented and combined (Eyben et al., 2016). Learnable front ends relax the measurement stage, and mixture-of-experts methods introduce adaptive routing, but neither is designed to maintain an explicit organization over complementary acoustic evidence throughout a reusable frozen representation.

AERIN occupies the intermediate design point developed in this work. The sources of acoustic evidence are structured explicitly as Boundary, Envelope, Phonation, and Texture, while their representations, local selection, and subsequent interactions remain learned. Fine-grained routing determines how strongly these sources contribute at individual frame-feature positions, and the resulting evidence is contextualized into a small predefined hierarchy of downstream states. Thus, AERIN does not replace learned representation with handcrafted features, nor does it merely introduce a new mechanism for reading an existing pretrained layer stack. It changes what is explicitly organized before the downstream read-out is formed. Distillation from WavLM supplies the training signal for this architecture; because we do not separately isolate the effect of distillation from the architecture, all empirical comparisons are reported as system-level results.

Table 3: Where AERIN sits among the seven lines of prior work reviewed in this appendix. *Specified* asks whether what the front end measures is settled by construction or is an outcome of pretraining. The last column asks what a downstream user chooses between when drawing a representation; for AERIN that evidence set is fixed by the computational graph rather than by training. One representative citation is given per line; the others are named in the text above.

Line	Representative work	Specified?	What a read-out selects over
Large SSL encoders	(Chen et al., 2022)	no	a depth index
Layer interfaces	(Shih & Harwath, 2024)	no	learned depth weights
Dimension-wise layers	(Chiu et al., 2024)	no	a depth index per dimension
Compressed SSL	(Peng et al., 2023)	no	a depth index
Paralinguistic KD	(Shor & Venugopalan, 2022)	no	nothing; single output
Descriptor sets	(Eyben et al., 2016)	yes	nothing; single output
Learnable front ends	(Zeghidour et al., 2021)	partly	nothing; single output
AERIN	Proposed approach	**yes**	**an evidence set**

B CORPORA

This paper uses two disjoint collections of datasets. The first is the unlabeled mixture from which AERIN is distilled. The second is the evaluation panel, which AERIN is never trained on and which supplies every number reported in this paper. No corpus appears in both, so AERIN encounters no evaluation corpus until that corpus's probe is fitted. Appendix B.5 states this separation in full, including the part of it we cannot control.

B.1 THE DISTILLATION MIXTURE

AERIN is distilled from WavLM-base on 539.1 hours of unlabeled speech assembled from nine public corpora and packed as 467,008 five-second chunks at 16 kHz. No clinical, emotion, or speaker label is used at any point in distillation: the mixture supplies waveforms and the teacher supplies targets. Table 4 gives the composition.

The mixture was chosen for its variety of speaking conditions rather than its size. Roughly two-thirds of it by duration is conversational (meetings, telephone calls, podcasts, and acted dialogue), and the remainder is lecture, parliamentary, or read speech. This matters for a front end built from timing, energy, and periodicity. Trained on read speech alone, the Boundary and Envelope streams would see almost no pausing behavior; trained on telephone speech alone, the Phonation stream would see only one narrow band. Recording conditions vary in the same deliberate way, from AMI's (Kraaij et al., 2005) far-field meeting microphones to Switchboard's (Godfrey et al., 1992) telephone channel of 8 kHz origin.

Proportions are set per corpus by a target number of hours except for IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019), and CREMA-D (Cao et al., 2014) rather than by taking each corpus whole, so some values in Table 4 are not the release size. A corpus larger than its target is subsampled without replacement, and one smaller than its target is oversampled with replacement. This prevents the small acted sets from being drowned out by the large conversational ones. Per-corpus counts, the sampling seed, and the teacher layers used are fixed in the mixture manifest, so the exact mixture can be reconstructed without redistributing any audio.

Several columns of Table 4 require explanation, because their meaning is not constant across the nine rows. *Task* records what each corpus was collected and labeled for, so that the composition of the mixture is legible; none of those labels is used here. *Hours* are summed from the per-chunk counts of valid-sample at 16 kHz, so they measure retained audio rather than nominal chunk duration. Chunks are ragged at utterance boundaries, and mean chunk length ranges from 2.54 s on CREMA-D through 3.24 s on AMI to 4.98 s on People's Speech, averaging 4.16 s over the mixture. Percentages are of total hours and are rounded independently, so that column sums to 99.8 rather than 100.0. *Ids* is the grouping identifier carried by the pack, and its granularity is the one quantity in the table that cannot be read uniformly. An ID denotes a speaker or speaker-side for AMI, Switchboard, IEMOCAP, VoxPopuli, and MSP-Podcast; a talk for TED-LIUM; a segment for Peo-

Table 4: Composition of the distillation mixture after chunking. *Hours* is the summed valid duration of the packed training chunks and does not denote the published size or unique audio duration of the source corpus. Counts likewise refer to the packed mixture rather than release sizes. Appendix B.1 gives the source and packing conventions.

Corpus	Task	Condition	Chunks	Hours	%	Ids
MSP-Podcast (Busso et al., 2025)	emotion	podcast	114,021	150.0	27.8	1,385
Switchboard (Godfrey et al., 1992)	ASR	telephone	91,317	100.0	18.5	4,867
AMI (Kraaij et al., 2005)	ASR	meetings	111,225	100.0	18.5	155
TED-LIUM (Rousseau et al., 2012)	ASR	lectures	61,978	80.0	14.8	271
VoxPopuli (Wang et al., 2021)	ASR	parliament	29,327	40.0	7.4	490
People's Speech (Galvez et al., 2021)	ASR	read, found	28,897	40.0	7.4	6,066
IEMOCAP (Busso et al., 2008)	emotion	acted dyads	12,712	14.2	2.6	8‡
MELD (Poria et al., 2019)	emotion	TV dialogue	12,322	11.2	2.1	9,706†
CREMA-D (Cao et al., 2014)	emotion	acted	5,209	3.7	0.7	5,209†
Total			**467,008**	**539.1**	**100.0**	

†utterance-level identifiers, not speakers. ‡speaker-sides, sessions 2-5 only.

ple's Speech; and one utterance for MELD and CREMA-D, which average 1.27 and 1.00 chunks per ID, respectively. The ID counts for these last two corpora are therefore not speaker counts, and CREMA-D's 91 actors cannot be recovered from this pack at all. IEMOCAP's IDs are genuine speaker-sides, but only Sessions 2-5 are packed, so its 8 IDs correspond to 8 of the release's 10 speakers rather than to a relabeling.

The four emotion corpora are included for their speaking style, not their labels. No label of any kind reaches the loss. The distillation objective defined in Section 2.6 consists solely of teacher alignment and the geometry floors.

B.2 The evaluation panel

The main comparison contains eleven evaluations drawn from eight source corpora. Five are clinical evaluations, three are phonetic/content evaluations constructed from NeuroVoz (Mendes-Laureano et al., 2024), and three are affective evaluations. This distinction between evaluations and corpora matters because NeuroVoz contributes four columns from the same or largely overlapping speaker cohort. The mean rank in Table 1 is therefore an evaluation-level summary rather than a corpus-weighted statistic. Speech Commands is reported separately as a lexical diagnostic and does not enter the mean rank (Appendix B.3).

Clinical evaluations. *DAIC-WOZ* (Burdisso et al., 2024) contains English clinical-interview speech for depression assessment. We follow its official split, with 106 participants in training and 35 in the reported development set; the small reported cohort is one reason this dataset is treated cautiously in Appendix E.7. *MODMA* (Cai et al., 2022) contains Mandarin depression speech and contributes 52 recordings to the evaluation. *IPVS* (Dimauro & Girardi, 2019) contains Italian Parkinson's speech and is close to ceiling for several encoders, so it provides limited separation between systems. *NeuroVoz* provides the HC/PD evaluation (NV-H/P) in Castilian Spanish. The corpus contains 108 speakers, with 55 healthy controls and 53 participants with Parkinson's disease. *SAND* (Sannino et al., 2026) evaluates Italian speech from participants with amyotrophic lateral sclerosis and healthy controls. We use the challenge's five-class dysarthria severity classification task and report the averaged-F_1 score evaluated on the official challenge server.

NeuroVoz phonetic/content evaluations. Three additional evaluations are constructed from the NeuroVoz recordings so that the downstream target changes while the underlying speaker population is largely fixed. NV-SV is a five-way task over sustained-vowel recordings, NV-GM is a binary GANGA/MANGA discrimination task, and NV-LR is a 16-way task over the Listen-and-Repeat recordings. These columns are useful precisely because they change the prediction target or

elicitation condition without introducing an independent source corpus. They should therefore be interpreted as different views of NeuroVoz rather than as three additional corpus-level replications.

Affective evaluations. *TESS* (Pichora-Fuller & Dupuis, 2020) and *RAVDESS-Speech* (Livingstone & Russo, 2018) contain acted English emotional speech, while *ShEMO* (Mohamad Nezami et al., 2019) contains Persian emotional speech. TESS contains only two actors, so a speaker-disjoint evaluation is not possible; we instead use a carrier-word-disjoint split, which tests generalization across lexical material while retaining the same voices. RAVDESS-Speech provides an eight-class evaluation over 24 actors. ShEMO provides a six-class Persian evaluation and is the non-English affective dataset in the ranked panel.

Language coverage and dependence between evaluations. The ranked panel spans five languages: English, Mandarin Chinese, Italian, Castilian Spanish, and Persian. Eight of the eleven evaluation columns are non-English, although four of those columns come from the same NeuroVoz source; at the source-corpus level, five of the eight corpora are non-English. Both the WavLM-base teacher and the speech used to distill AERIN are English-language data, so the non-English columns test transfer beyond the language used to train the encoder. They are not a controlled language-generalization benchmark, however, because language changes together with corpus, task, speakers, and recording conditions. We therefore use them as evidence of cross-lingual transfer within the broader evaluation panel rather than to isolate an effect of language alone.

B.3 Diagnostic outside the mean rank

Speech Commands (Warden, 2018) is reported as an additional lexical diagnostic rather than as part of the main transfer ranking. Our formulation is a 36-way keyword-identification task consisting of the 35 keywords in the standard release plus an explicit *silence* class. It measures lexical content rather than the clinical, phonetic, or affective targets that define Table 1, so its score is reported separately and does not enter the mean rank. Results on this dataset are reported in Appendix F.4.2.

B.4 Preprocessing

Every corpus in the distillation mixture and evaluation panel passes through the same audio-loading and chunking pipeline. Audio is decoded from the source container, converted to 16 kHz mono float32, and segmented into 5 s chunks with a 2.5 s hop. Each chunk carries a validity mask and a speech mask, and chunks containing less than 0.25 s of speech are removed. Chunks are ragged at utterance boundaries, so the packed duration in Table 4 is obtained by summing valid audio samples rather than from the nominal number of chunks.

Speech masks correspond to voice-activity information where the source corpus provides it. Where no such annotation is available, the mask represents valid audio rather than speech activity. The Boundary stream therefore uses the mask when genuine voice-activity information is present and otherwise falls back to its relative-energy threshold, so that no stream is constant on a corpus lacking voice-activity annotation.

Speaker identifiers are prefixed by corpus, so that each chunk's identity remains unique across the mixture and no two corpora can collide on a speaker key. The same prefix is also the key on which the Affect column of Table 4 is computed.

B.5 Separation between distillation and evaluation

The distillation data and downstream evaluation data are separated at the source-corpus level. None of the nine corpora in Table 4 appears among the eight source corpora that form the ranked evaluation panel, and Speech Commands is also absent from the distillation mixture. The reported encoder therefore receives no waveform from any downstream evaluation corpus during distillation.

The same separation holds for downstream labels. Distillation uses only speech and cached WavLM targets; clinical, phonetic, affective, and lexical labels from the source corpora are not used by the objective in Section 2.6. This distinction is important for the emotion corpora in the mixture. IEMOCAP, MELD, CREMA-D, and MSP-Podcast contain affective speech, but their emotion annotations

are not used. Thus, the encoder is exposed to emotional speech during representation learning without being trained to predict emotion.

Task families can consequently overlap even though corpora and labels do not. The distillation mixture contains conversational, read, acted, and affective speech, while the evaluation panel asks clinical, phonetic, and affective questions of entirely different recordings. Our separation claim is therefore that AERIN does not train on the evaluation recordings or their labels, not that its distillation data exclude every acoustic condition relevant to the downstream tasks.

This separation applies to AERIN itself and should not be extended to all public baselines. Their pretraining regimes differ substantially in data scale, language coverage, and task specialization. WavLM-base, the teacher used here, is an English-speech model, while other systems in Table 1 include multilingual and task-specialized encoders. We therefore treat the benchmark as a comparison of frozen representations under a common downstream evaluation protocol, rather than as a controlled comparison of pretraining data exposure.

B.6 What the panel does and does not span

The ranked comparison contains eleven evaluations drawn from eight source corpora. Its breadth is useful for testing whether one frozen representation remains usable as the downstream target changes, but it is not a collection of eleven independent datasets.

Language coverage. The panel spans English, Mandarin Chinese, Italian, Castilian Spanish, and Persian. Eight of the eleven evaluation columns are non-English: MODMA, IPVS, SAND, the four NeuroVoz evaluations, and ShEMO. At the source-corpus level, five of the eight ranked corpora are non-English. Both the WavLM-base teacher and the speech used to distill AERIN are English-language data, so these columns evaluate transfer beyond the language used for representation learning. They do not, however, isolate language as an experimental variable. Language changes together with corpus, speakers, recording conditions, and in many cases the downstream task. The results therefore support cross-lingual transfer, but not a causal claim about language invariance.

Dependence between evaluations. Four columns come from NeuroVoz and share the same or largely overlapping speaker population. NV-H/P evaluates Parkinson's status, whereas NV-SV, NV-GM, and NV-LR change the prediction target or elicitation condition within the same source collection. This makes NeuroVoz useful for studying target dependence without changing the corpus, but the four scores should not be interpreted as four independent corpus-level replications. TESS provides a different form of dependence, its recordings come from only two actors, so its evaluation measures generalization across held-out lexical material rather than across unseen speakers.

Cohort size and score resolution. Several clinical datasets contain relatively small participant cohorts, and their scores can consequently be sensitive to the particular split and probe training run. DAIC-WOZ, for example, is reported on its 35-participant development set. IPVS presents a different limitation because several encoders already operate close to ceiling, leaving little room to separate systems. For these reasons, individual clinical columns are more informative as part of the broader transfer pattern than as precise estimates of small performance differences.

Task and class-count coverage. Task family, corpus, language, elicitation condition, and number of target classes are not independently controlled across the full panel. Most clinical evaluations are binary, whereas SAND uses five severity classes and the affective evaluations use six to eight classes. The NeuroVoz tasks span binary, five-way, and sixteen-way prediction within one source corpus, which provides some variation in target structure without changing the underlying collection, but it does not remove the other differences between tasks. We therefore avoid attributing differences between corpus-level scores to any single factor such as class count, language, or task family.

Baseline training regimes. The baseline models in Table 1 also differ substantially before the downstream probe is fitted. Some are general-purpose English speech encoders, some are multilingual, and emotion2vec+ is explicitly specialized for speech emotion recognition. These differences are part of the practical comparison but prevent the table from isolating architecture alone. In particular, the strong affective performance of emotion2vec+ should be interpreted together with its

substantially greater emotion-specific training, as discussed in Section 3.1. The main claim of the panel is therefore narrower: a compact frozen representation can remain competitive across a heterogeneous set of downstream tasks and languages.

B.7 Access, licensing, and consent

All speech used in this study comes from existing research datasets or public corpus releases; we collected no new recordings and recruited no participants. We use each corpus under the access and licensing conditions specified by its publisher and rely on the original dataset documentation for its collection, consent, and participant-protection procedures.

We do not redistribute source audio. Reproducibility artifacts instead contain the preprocessing code and, where permitted by the corresponding licenses, the metadata needed to reconstruct the distillation and evaluation splits from the original releases. Corpora that require registration, agreement to dataset-specific terms, or challenge access must still be obtained from their original distributors.

Clinical recordings require particular care because they may contain sensitive health-related speech. Our processing does not introduce new participant labels beyond those supplied by the original datasets, and downstream labels are never used during distillation. For SAND, the metric reported in Table 1 is obtained through the challenge's official scoring procedure; its held-out labels are not required locally.

C Methodological Details

C.1 Evidence front ends and stream construction

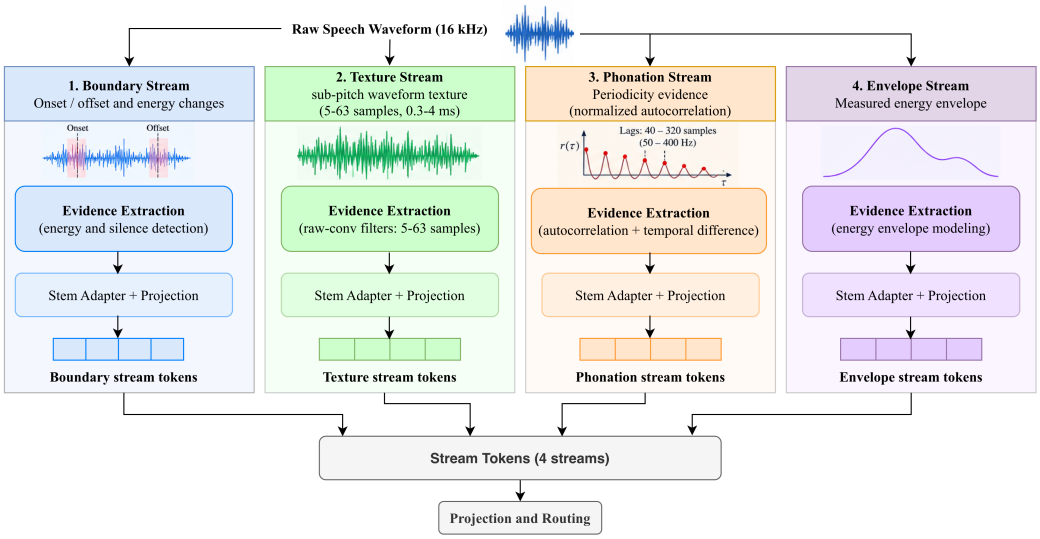


Figure 2: Four complementary acoustic evidence streams extracted directly from the waveform. Boundary, Envelope, and Phonation retain explicit waveform-derived measurements, while Texture is constructed from the learned waveform stem. All streams are aligned to the same temporal grid before routing.

The evidence quantities are introduced in Section 2.2. Here, we specify their implementation, temporal alignment, and conversion into the frame-level representations consumed by the routing stages. All streams are aligned to the common encoder grid with hop size $H = 96$ samples.

C.1.1 SHARED WAVEFORM STEM

The first learned description of the waveform is a compact convolutional stem. It operates directly on the normalized waveform x^{norm} using four parallel one-dimensional convolutional branches:

$$c^{(k)} = \mathrm{Conv1D}_k(x^{\mathrm{norm}}), \qquad k \in \{5, 15, 31, 63\}, \tag{24}$$

followed by

$$h^{\mathrm{stem}} = P_{\mathrm{stem}}\Big([c^{(5)}, c^{(15)}, c^{(31)}, c^{(63)}]\Big), \qquad h^{\mathrm{stem}} \in \mathbb{R}^{T\times 320}. \tag{25}$$

Each branch contains 48 channels and advances with the common 96-sample hop. A depthwise convolution with kernel size 3 refines each branch before the branch outputs are temporally aligned, concatenated, normalized, and mapped by P_{stem}.

At $16\,\mathrm{kHz}$, the four kernels span approximately 0.31, 0.94, 1.94, and $3.94\,\mathrm{ms}$ of waveform. Equivalently, their reciprocal time scales range from approximately $3.2\,\mathrm{kHz}$ to $254\,\mathrm{Hz}$. The filters themselves are learned, so we do not assign a fixed acoustic interpretation to any individual branch. The architectural range instead determines the short-waveform structures that are available to the stem. In particular, even the longest kernel remains shorter than a typical pitch period, separating the short-timescale waveform structure available to the Texture stream from the longer periodic structure measured explicitly by the Phonation stream in Appendix C.1.3.

The stem is shared and does not constitute a fifth evidence stream. Texture reads the stem directly, whereas the remaining streams receive it through the conditioning operation described in Appendix C.1.4. Consequently, each stream preserves its defining private measurement while still having access to common waveform context. The complete stem operates in the time domain and does not use STFT, mel-filterbanks, MFCCs, or another fixed time–frequency transform.

C.1.2 BOUNDARY, ENVELOPE, AND TEXTURE CONSTRUCTION

The Boundary and Envelope measurements are those defined in Section 2.2; their mathematical definitions are not repeated here. Their frame-level measurement sequences are aligned to the common T-frame encoder grid and passed through separate learned front ends. Each front end maps its private measurement vector to the evidence-stream width $d_e = 384$:

$$e_r = F_r(m_r), \qquad r \in \{\mathrm{Boundary}, \mathrm{Envelope}\}, \qquad e_r \in \mathbb{R}^{T\times 384}, \tag{26}$$

where m_r denotes the corresponding measured frame-level sequence and F_r is stream specific.

The original Boundary and Envelope measurements are retained alongside their learned representations. This distinction is required because Stage III uses temporally segmented summaries of the measured quantities when constructing the measured attention prior. The Stage III prior therefore receives the explicit measurements rather than an intermediate hidden state. The learned representations e_r, by contrast, are passed through the stem-conditioning and common-projection path in Appendix C.1.4 before routing.

Texture has no additional hand-computed descriptor. Its private input is obtained directly from the shared waveform stem:

$$e_{\mathrm{Texture}} = F_{\mathrm{Texture}}\big(h^{\mathrm{stem}}\big), \qquad e_{\mathrm{Texture}} \in \mathbb{R}^{T\times 384}, \tag{27}$$

where F_{Texture} is a learned projection. Texture therefore shares the same downstream stream interface as Boundary and Envelope, but it does not carry a separate measured sequence into Stage III.

All three pathways use the same frame-validity mask. Frames introduced only by convolutional or batch padding are excluded before routing. The same validity information is retained when the Boundary and Envelope measurements are temporally segmented for the measured attention prior.

C.1.3 PERIODICITY EVIDENCE

The Phonation stream measures periodic structure directly from the normalized waveform x^{norm}. For each frame position t, we use an autocorrelation window of $Q = 512$ waveform samples ($32\,\mathrm{ms}$), advanced with the same hop $H = 96$ samples used by the encoder. The periodicity measurements therefore remain aligned with the common frame axis.

Let $\mathcal{W}_t$ denote the waveform-sample indices contained in the periodicity-analysis window at frame position t. For each lag $\ell \in \mathcal{L}$, the normalized autocorrelation is

$$\rho_t(\ell) = \frac{\frac{1}{Q}\sum_{n\in\mathcal{W}_t} x_n^{\mathrm{norm}} x_{n+\ell}^{\mathrm{norm}}}{\sqrt{\frac{1}{Q}\sum_{n\in\mathcal{W}_t}\left(x_n^{\mathrm{norm}}\right)^2}\sqrt{\frac{1}{Q}\sum_{n\in\mathcal{W}_t}\left(x_{n+\ell}^{\mathrm{norm}}\right)^2}+\varepsilon}, \qquad \rho_t(\ell)\in[-1,1]. \tag{28}$$

The numerator and the two energy terms in the denominator are evaluated by average pooling. For every lag, the lagged waveform product is right-padded before pooling so that all lag sequences are produced on the same temporal grid. The resulting coefficients are clipped to $[-1, 1]$ and then aligned to the encoder frame count T.

A large positive $\rho_t(\ell)$ indicates that similar waveform structure recurs after delay ℓ, whereas a value close to zero provides little periodicity evidence at that lag. Energy normalization removes dependence on the overall waveform amplitude. Evaluating the full lag set requires $O(L|\mathcal{L}|)$ operations for waveform length L and is implemented entirely in the time domain.

The reported model uses ten lags. Stacking their coefficients produces

$$\rho_t = [\rho_t(\ell_1), \ldots, \rho_t(\ell_{10})] \in \mathbb{R}^{10}. \tag{29}$$

We additionally compute the first temporal difference of the lag profile and concatenate it with the profile itself:

$$p_t = [\rho_t,\, \Delta\rho_t] \in \mathbb{R}^{20}. \tag{30}$$

For the first valid frame, the temporal-difference term is initialized so that no information is introduced from a preceding padded position. The resulting 20-dimensional vector is mapped by the Phonation front-end MLP to

$$e_{\mathrm{Phonation}} = F_{\mathrm{Phonation}}(p), \qquad e_{\mathrm{Phonation}} \in \mathbb{R}^{T\times 384}. \tag{31}$$

The maximum autocorrelation value across lags is logged only as a diagnostic. It is not concatenated with the Phonation representation, does not control the router, and is not used to select a lag. No explicit F_0 trajectory is decoded, and no frame-dependent lag selection is performed.

C.1.4 Stem conditioning and projection

The four streams begin from different private evidence representations, but each is conditioned on the shared waveform stem before routing. For stream r, let

$$e_r \in \mathbb{R}^{T\times d_e}, \qquad d_e = 384,$$

denote the private representation produced by the corresponding front end. The gated conditioning update is

$$e_r^{\mathrm{cond}} = e_r + \alpha_s\, \gamma_r \odot \psi_r\big(e_r, h^{\mathrm{stem}}\big), \tag{32}$$

where ψ_r is the learned stream-specific conditioning transform and $\gamma_r \in [0,1]^{T\times d_e}$ is an element-wise gate. The gate controls the amount of shared-stem information added at each frame and feature dimension, while the residual path preserves the private stream representation. The reported encoder uses $\alpha_s = 0.25$, followed by layer normalization.

The conditioned representation is then concatenated with the shared stem and mapped to the common routing dimension:

$$h_r = P_r\big([e_r^{\mathrm{cond}},\, h^{\mathrm{stem}}]\big) + \mathrm{Emb}_{\mathrm{stream}}(r), \qquad h_r \in \mathbb{R}^{T\times D}. \tag{33}$$

Here P_r is a learned stream-specific projection and $\mathrm{Emb}_{\mathrm{stream}}(r) \in \mathbb{R}^D$ is a learned stream embedding that is added at every valid frame. Stacking the four projected streams in their canonical order gives

$$h = [h_{\mathrm{Boundary}}, h_{\mathrm{Envelope}}, h_{\mathrm{Texture}}, h_{\mathrm{Phonation}}] \in \mathbb{R}^{T\times|\mathcal{R}|\times D}, \tag{34}$$

which is passed to the routing stage.

This construction keeps each stream's private evidence distinct from the shared waveform context. In particular, ablating a stream removes its private representation e_r while leaving h^{stem} available to the remaining streams. The common validity mask is applied after projection so that padded frames do not contribute to routing or subsequent temporal aggregation.

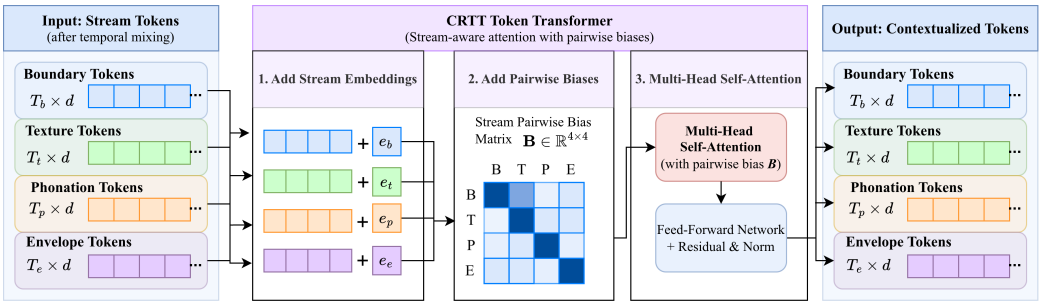


Figure 3: Token-transformer block. The temporally mixed Boundary, Texture, Phonation, and Envelope tokens are first augmented with learned stream embeddings so that their stream identity is retained after concatenation. A learned stream-pair bias matrix $\mathbf{B} \in \mathbb{R}^{4\times4}$ is then added to the attention scores, allowing the interaction between any two tokens to depend on the streams from which they originate. The resulting sequence is processed by multi-head self-attention followed by the feed-forward, residual, and normalization operations of the token transformer. The output is finally separated back into the four stream-specific token sequences, preserving the corresponding token counts T_b, T_t, T_p, and T_e and token width d.

C.2 Routing implementation

Section 2.3 defines the routing operator. In the implementation, the common routing context maps the concatenation of the four projected streams and the 320-dimensional stem to width $D = 384$. Each stream has its own gate network,

$$\mathrm{LN} \to \mathrm{Linear}(384) \to \mathrm{GELU} \to \mathrm{Dropout} \to \mathrm{Linear}(384),$$

applied to $[h_r, h^{\mathrm{ctx}}, h^{\mathrm{stem}}]$. The four resulting logits are stacked and normalized by a softmax over the stream axis separately for every frame and feature coordinate. Each stream also has a separate $384 \to 384 \to 384$ value MLP. The routed update uses the residual scale $\alpha = 0.25$ followed by layer normalization, as in Eq. 17. Routing returns four separate $T \times 384$ streams; the stream axis is not collapsed by a weighted sum.

The interventions in Section 3.2 are applied before stem conditioning. For a lesion set $\mathcal{S} \subseteq \mathcal{R}$, the private evidence tensor e_r is replaced by zero for each $r \in \mathcal{S}$, while h^{stem} and all learned parameters remain unchanged. The stem adapters, competitive router, contextualization modules, and read-out construction, therefore, continue to operate on the intervened representation. A Boundary lesion also zeros the measured Boundary summary and the speech/pause-duration quantity supplied to the Stage III attention prior. An Envelope lesion likewise zeros the measured Envelope quantity used by that prior. Texture and Phonation have no additional measured-prior input to remove. This convention is why the shared waveform stem remains available in both the remove-one and retain-one-stream analyses.

C.3 Contextualization details

This section gives the implementation details of the two contextualization stages summarized in Section 2.4. The first stage operates directly on the frame-level evidence streams, while the second stage operates on temporally pooled tokens spanning the complete chunk.

The routed representation retains one feature sequence for each evidence stream. The axial mixer first applies temporal self-attention independently within each stream and then applies cross-stream attention among the evidence streams at the same frame position. This preserves the stream axis while allowing each stream to use nearby temporal context and information from the other streams. The reported encoder uses three axial layers with four attention heads, a temporal-window parameter of 64 frames, an FFN expansion factor of 2, and attention dropout of 0.10. Cross-stream attention is enabled. Learned stream embeddings identify the evidence streams, and a learned stream-pair bias is used during cross-stream attention. The frame-validity mask m excludes invalid positions from the attention operations. The output of the axial mixer is

$$M' \in \mathbb{R}^{T\times|\mathcal{R}|\times D}. \tag{35}$$

The same temporal-token operator is used for the student representations and the corresponding teacher representations. For a frame sequence $V \in \mathbb{R}^{T\times D}$ with validity mask m, the valid frames are ordered by time and divided into K contiguous, nearly equal groups. Each group is then mean-pooled to form one token. We denote this operation by

$$\mathrm{SegMean}_K(V, m). \tag{36}$$

Applying Eq. 36 independently to each stream gives

$$\bar{U}_r = \mathrm{SegMean}_K\left(M'_r, m\right), \qquad \bar{U} = \mathrm{Concat}_{r\in\mathcal{R}}\left(\bar{U}_r\right). \tag{37}$$

The four streams therefore provide $4\times K$ token positions in total. Only valid frames contribute to the segment means. If a segment contains no valid frames, it is filled with the global mean over the valid frames and marked invalid in the token mask. If a chunk contains no valid frames, all tokens are set to zero and all entries in the token mask are false. The same segmentation rule is used for the teacher representations so that student and teacher tokens correspond to the same relative regions of the chunk.

Before token-level self-attention, the concatenated sequence $\bar{U}$ is projected from the common stream width $D = 384$ to an internal width of 576. A learned stream embedding and a learned segment-position embedding are then added to each token,

$$X^{(0)} = W_{\mathrm{in}}\bar{U} + E_{\mathrm{stream}} + E_{\mathrm{pos}}. \tag{38}$$

The token transformer contains two layers with four attention heads and an FFN width of 1152. Its attention dropout is 0.10. Invalid token positions are removed through the token-validity mask, and an output projection maps the transformer representation back to the common width D.

In addition to the usual query–key similarity, each attention head receives a structural term and an acoustic term,

$$\mathrm{Attn}^{(j)} = \mathrm{softmax}\left(\frac{Q^{(j)}{K^{(j)}}^{\top}}{\sqrt{d_{\mathrm{h}}}} + B^{(j)}_{\mathrm{struct}} + B^{(j)}_{\mathrm{meas}}\right). \tag{39}$$

The structural contribution is

$$B^{(j)}_{\mathrm{struct}} = B^{(j)}_{\mathrm{rel}} + B^{(j)}_{\mathrm{stream}} + B^{(j)}_{\mathrm{loc}}. \tag{40}$$

$B^{(j)}_{\mathrm{rel}}$ is a learned bias indexed by the relative positions of two temporal segments, while $B^{(j)}_{\mathrm{stream}}$ is a learned bias indexed by the ordered pair of evidence streams from which the two tokens originate. The local component restricts the first attention head to token pairs whose temporal-segment positions differ by at most two,

$$B^{(1)}_{\mathrm{loc},ii'} = \begin{cases} 0, & |k_i - k_{i'}| \le 2, \\ -10^4, & |k_i - k_{i'}| > 2, \end{cases} \tag{41}$$

where k_i is the temporal-segment position associated with token i. The corresponding local term is zero for the remaining attention heads, allowing them to retain non-local interactions.

The acoustic term uses three quantities computed earlier in the encoder: Boundary, Envelope, and speech/pause duration. We denote this set by

$$\mathcal{P} = \{\mathrm{bound}, \mathrm{env}, \mathrm{dur}\}. \tag{42}$$

Each quantity is divided using the same K temporal segments used for the learned stream representations, producing an ordered sequence τ_p. These sequences are repeated across the four stream groups so that they align with the complete token-attention matrix. For attention head j, the acoustic contribution between token positions i and i' is

$$B^{(j)}_{\mathrm{meas},ii'} = \sum_{p\in\mathcal{P}} \beta^{(j)}_p \left\|\tau_{p,i} - \tau_{p,i'}\right\|_1. \tag{43}$$

The L_1 term measures how different the two temporal positions are according to acoustic quantity p. Each quantity has a separately learned coefficient $\beta^{(j)}_p$ for every attention head, and these coefficients are not constrained in sign. The model can therefore learn whether a larger difference in a particular

acoustic quantity should increase or decrease the corresponding attention score. This term modifies only the attention logits; the token representations remain learned from $\bar{U}$.

After the token-transformer layers, the output projection produces the contextualized sequence

$$U \in \mathbb{R}^{K|\mathcal{R}|\times D}. \tag{44}$$

This longer-range context is then returned to the frame-level representation through the frame–token fusion module,

$$M = \mathrm{Fuse}\left(M', U, m\right), \qquad M \in \mathbb{R}^{T\times|\mathcal{R}|\times D}. \tag{45}$$

The fusion module receives the frame representation, the contextualized token sequence, the token-validity mask, and the frame-validity mask. Its output preserves the frame and stream axes and is passed directly to the read-out ladder in Section 2.5. The token sequence U is also retained because the read-out stage uses token-level summaries directly.

C.4 Read-out construction details

Section 2.5 introduces the two representations exposed to downstream probes. This appendix gives their exact construction and the auxiliary state used during representation learning.

Let

$$M \in \mathbb{R}^{T\times|\mathcal{R}|\times D}$$

denote the contextualized frame-level stream representation produced by Stage III, and let

$$U \in \mathbb{R}^{K|\mathcal{R}|\times D}$$

denote the contextualized token sequence. The three states use separate learned token summaries,

$$c_{\mathrm{aux}} = \mathrm{Pool}_{\mathrm{aux}}(U), \qquad c_{\mathrm{low}} = \mathrm{Pool}_{\mathrm{low}}(U), \qquad c_{\mathrm{high}} = \mathrm{Pool}_{\mathrm{high}}(U). \tag{46}$$

Each pooling operation produces a D-dimensional summary of the valid tokens, which is expanded across the T frame positions before being combined with the frame-level stream representations.

The auxiliary state is constructed from the non-Boundary evidence streams and its token summary,

$$z_{\mathrm{aux}} = P_{\mathrm{aux}}\left(\left[\{M_r\}_{r\in\mathcal{R}\setminus\{\mathrm{Boundary}\}},\, c_{\mathrm{aux}}\right]\right). \tag{47}$$

This state is used during representation learning and distillation but is not exposed as an independent downstream read-out.

The lower read-out receives all contextualized evidence streams, their first temporal differences, and a separate token summary,

$$Z_{\mathrm{low}} = P_{\mathrm{low}}\left(\left[M_1, \ldots, M_{|\mathcal{R}|},\, \Delta M_1, \ldots, \Delta M_{|\mathcal{R}|},\, c_{\mathrm{low}}\right]\right), \tag{48}$$

where ΔM_r denotes the first temporal difference of stream r. Thus, Z_{low} has direct access to all four contextualized evidence streams, their frame-to-frame variation, and token-level context.

The higher read-out uses a third token summary together with the mean over all valid tokens,

$$u_{\mathrm{global}} = \mathrm{Mean}(U). \tag{49}$$

An intermediate higher representation is formed as

$$\tilde{Z}_{\mathrm{high}} = P_{\mathrm{high}}\left(\left[Z_{\mathrm{low}},\, c_{\mathrm{high}},\, u_{\mathrm{global}},\, M_1, \ldots, M_{|\mathcal{R}|}\right]\right). \tag{50}$$

The final higher read-out adds a residual projection of Z_{low},

$$Z_{\mathrm{high}} = \mathrm{LN}\left(\tilde{Z}_{\mathrm{high}} + \alpha_{\mathrm{lh}}\, W_{\mathrm{lh}} Z_{\mathrm{low}}\right), \tag{51}$$

where $W_{\mathrm{lh}} \in \mathbb{R}^{D\times D}$ is learned and $\alpha_{\mathrm{lh}} = 0.1$ is the scale of the residual update.

All three states have width D and are defined at the frame level. Invalid frame positions are removed using the frame-validity mask m. After distillation, downstream evaluation uses the two read-outs Z_{low} and Z_{high}, while Z_{aux} remains a training-only state. A combination of read-outs [Z_{low};Z_{high}] (Fig. 1) is also used for evaluations, but not considered an exposed read-out as it is a combination of two optimized states and is not a separate entity on its own.

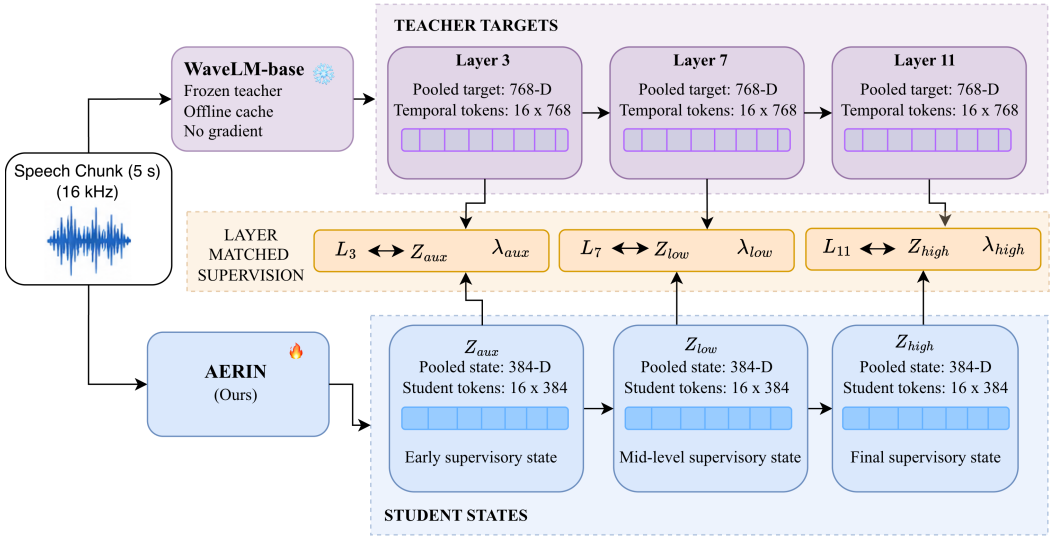


Figure 4: Knowledge-distillation setup used to train AERIN. Each chunk is processed in parallel by the frozen WavLM-base teacher and by AERIN. Teacher representations are cached offline and no gradient is propagated through WavLM. Three teacher depths are used as supervision targets: layer 3 for the auxiliary student state Z_{aux}, layer 7 for the intermediate state Z_{low}, and layer 11 for the final state Z_{high}. For each matched pair, supervision is applied to both the pooled representation and the corresponding sequence of 16 temporal tokens, after projection to the common distillation space. The three state pairs are combined with the state-specific weights λ_{aux}, λ_{low}, and λ_{high}. This layer-matched design provides progressively deeper teacher targets to the corresponding stages of the AERIN state ladder while keeping the teacher fixed throughout training.

D Training and Optimization Details

D.1 Full Distillation Objective

The final AERIN encoder is trained only by distillation from WavLM-base (Chen et al., 2022). The distillation mixture contains 539.1 hours of speech drawn from nine corpora: AMI, Switchboard, IEMOCAP, TED-LIUM, VoxPopuli, People's Speech, MELD, CREMA-D, and MSP-Podcast. Although VoxPopuli is a multilingual corpus, we use only a 40-hour English subset. Only the speech signal and cached WavLM targets are used by the final training objective. Task labels available in individual source corpora are not used. The DAIC-WOZ/E-DAIC evaluation corpus is excluded from the distillation mixture, and the in-domain distillation stage is disabled.

WavLM-base is kept fixed and its representations are cached before student training. We use teacher layers 3, 7, and 11. Layer 3 supervises the training-only auxiliary state Z_{aux}, layer 7 supervises Z_{low}, and layer 11 supervises Z_{high}. We denote the three training states by

$$\mathcal{S}_{\text{KD}} = \{\text{aux}, \text{low}, \text{high}\}. \tag{52}$$

Their state weights are

$$(\lambda_{\text{aux}}, \lambda_{\text{low}}, \lambda_{\text{high}}) = (0.10,\ 0.55,\ 0.35)\,. \tag{53}$$

Student and teacher representations are mapped through separate training-only projection heads before the distillation losses are evaluated. Each projection head applies feature normalization followed by a learned linear map into a shared distillation space. Let $s_i^{(\kappa)}$ and $t_i^{(\kappa)}$ denote a valid student–teacher pair for state $\kappa \in \mathcal{S}_{\text{KD}}$, after these projections. State-level alignment combines cosine distance, layer-normalized mean-squared error, and similarity-preserving relational alignment (Tung & Mori, 2019),

$$\mathcal{L}_{\text{align}} = \sum_{\kappa \in \mathcal{S}_{\text{KD}}} \lambda_\kappa \left(\mathcal{L}_{\text{cos}}^{(\kappa)} + \lambda_{mse}\, \mathcal{L}_{\text{mse}}^{(\kappa)} + \lambda_{sp}\, \mathcal{L}_{\text{sp}}^{(\kappa)} \right). \tag{54}$$

The cosine term is

$$\mathcal{L}_{\cos}^{(\kappa)} = 1 - \frac{1}{N}\sum_{i=1}^{N} \frac{s_i^{(\kappa)\top} t_i^{(\kappa)}}{\left\|s_i^{(\kappa)}\right\|_2 \left\|t_i^{(\kappa)}\right\|_2}, \tag{55}$$

and $\mathcal{L}_{\text{mse}}^{(\kappa)}$ is the mean-squared error between the layer-normalized projected student and teacher states. For the relational term, let $\bar{S}^{(\kappa)}$ and $\bar{T}^{(\kappa)}$ contain the ℓ_2-normalized projected states as rows. We use

$$\mathcal{L}_{\text{sp}}^{(\kappa)} = \frac{1}{N^2}\left\|\bar{S}^{(\kappa)}\bar{S}^{(\kappa)\top} - \bar{T}^{(\kappa)}\bar{T}^{(\kappa)\top}\right\|_F^2. \tag{56}$$

This term aligns pairwise similarities among valid examples rather than only matching each student state independently to its teacher target.

Temporal supervision is applied separately to ordered within-chunk tokens. For each of the three training states, valid frames are divided into K contiguous, nearly equal temporal segments and mean-pooled using the same segmentation rule described in Appendix C.3. The corresponding WavLM layer is pooled with the same rule when the teacher cache is constructed. Temporal-token alignment is

$$\mathcal{L}_{\text{temp}} = \sum_{\kappa\in\mathcal{S}_{\text{KD}}} \lambda_\kappa \left(\mathcal{L}_{\text{cos,tok}}^{(\kappa)} + \lambda_{mt}\,\mathcal{L}_{\text{mse,tok}}^{(\kappa)}\right). \tag{57}$$

The cosine and layer-normalized MSE terms are evaluated only at temporal-token positions that are valid for both the student and teacher.

The state representations are also regularized to reduce dimensional collapse. Following the variance and covariance principles of VICReg (Bardes et al., 2022), each state receives a variance floor and an off-diagonal covariance penalty, together with an additional cosine-spread term,

$$\mathcal{L}_{\text{geo}} = \sum_{\kappa\in\mathcal{S}_{\text{KD}}} \lambda_\kappa \left(\mathcal{L}_{\text{var}}^{(\kappa)} + \lambda_{cov}\,\mathcal{L}_{\text{cov}}^{(\kappa)} + \lambda_{spread}\,\mathcal{L}_{\text{spread}}^{(\kappa)}\right). \tag{58}$$

For centered valid state vectors, the variance term is

$$\mathcal{L}_{\text{var}}^{(\kappa)} = \frac{1}{D}\sum_{d=1}^{D}\left[\gamma_\kappa - \sigma_d^{(\kappa)}\right]_+, \tag{59}$$

with state-specific variance floors

$$(\gamma_{\text{aux}}, \gamma_{\text{low}}, \gamma_{\text{high}}) = (0.20,\ 0.25,\ 0.20). \tag{60}$$

If $C^{(\kappa)}$ denotes the feature covariance matrix, then

$$\mathcal{L}_{\text{cov}}^{(\kappa)} = \frac{1}{D}\sum_{p\neq q}\left(C_{pq}^{(\kappa)}\right)^2. \tag{61}$$

The cosine-spread term is the mean squared cosine similarity between distinct valid centered state vectors,

$$\mathcal{L}_{\text{spread}}^{(\kappa)} = \frac{1}{N(N-1)}\sum_{i\neq j}\left(\bar{z}_i^{(\kappa)\top}\bar{z}_j^{(\kappa)}\right)^2, \tag{62}$$

where $\bar{z}_i^{(\kappa)}$ is the ℓ_2-normalized centered representation.

Because the higher read-out is constructed partly from the lower read-out, we add a between-state decorrelation term. Let $Z_{\text{low}}^{\text{std}}$ and $Z_{\text{high}}^{\text{std}}$ denote the valid lower and higher states after feature-wise centering and standardization. N_{valid} is the valid number of chunks. Their cross-correlation matrix is

$$C_{\text{lh}} = \frac{Z_{\text{low}}^{\text{std}\top} Z_{\text{high}}^{\text{std}}}{N_{\text{valid}}}. \tag{63}$$

We minimize

$$\mathcal{L}_{\text{sep}} = \frac{1}{D^2}\left\|C_{\text{lh}}\right\|_F^2. \tag{64}$$

This is related to cross-correlation regularization in Barlow Twins (Zbontar et al., 2021), but here the correlation between two different student states is driven toward zero rather than toward an identity matrix.

Combining the four components from Eq. 54, Eq. 57, Eq. 58 and Eq. 64 gives our final objective function described in Eq. 23.

Training is configured for 20 epochs with batch size 16. The encoder and training heads use a learning rate of 10^{-4} and weight decay 2×10^{-4}. The schedule includes five warm-up epochs, early-stopping patience of eight epochs, and gradient-norm clipping at 5.0. Mixed-precision training is enabled. The temporal distillation cache contains K ordered tokens per teacher layer. After distillation, the projection heads and all loss-specific components are discarded, and the frozen encoder alone is used for downstream evaluation.

D.2 Checkpointing and training diagnostics

Training uses a single optimization step per minibatch with training seed 42. The minimum valid speech duration is 0.25 s.

The following quantities are logged once per epoch from the valid student states and the corresponding cached teacher targets: teacher–student cosine alignment, temporal distillation loss, effective rank, top-eigenvalue ratio, feature standard deviation, random-pair cosine similarity, and low–high cross-correlation. These diagnostics are not additional terms in Eq. 23.

Teacher–student cosine alignment is recorded separately from the total distillation loss so that direct state-to-teacher agreement can be inspected without the contribution of the geometry and cross-state penalties. Temporal distillation loss is logged independently for the ordered K token sequences used by Eq. 57.

Effective rank, the top-eigenvalue ratio, feature standard deviation, and random-pair cosine similarity are recorded from the learned state representations to expose concentration or collapse during optimization. The effective-rank statistic measures how broadly variance is distributed across representation directions, while the top-eigenvalue ratio records the fraction concentrated in the dominant direction. Feature standard deviation is tracked together with these spectral statistics because the variance term in Eq. 59 acts directly on per-dimension dispersion. Random-pair cosine similarity provides a complementary sample-level view of whether valid representations become excessively aligned.

Low–high cross-correlation is logged from the standardized Z_{low} and Z_{high} states used in Eq. 63. It directly monitors the quantity penalized by $\mathcal{L}_{\mathrm{sep}}$ in Eq. 64. The diagnostic values are used only for training inspection; they do not introduce additional supervision or modify the loss weights in Eq. 23.

D.3 Model, training, and distillation configuration

Table 5 lists implementation-level model settings that are not all repeated in the main methodology. Table 6 summarizes the optimization and training configuration, while Table 7 records the correspondence between student states and the frozen WavLM teacher layers used by the distillation objective in Appendix D.1. These settings are fixed for the reported model unless stated otherwise.

The model-level settings in Table 5 specify the waveform framing, common representation width, periodicity analysis, shared-stem conditioning, contextual mixing, transformer configuration, and the final read-outs. Quantities already defined in Section 2 and Appendix C are listed here only to make the complete reported configuration available in one place.

Table 6 includes both optimizer settings and the scalar coefficients used by the final objective in Eq. 23. The encoder and training-only heads use the same base learning rate. The learning-rate schedule applies the five-epoch warmup before cosine decay, and gradient clipping is applied to the global gradient norm. Mixed-precision training is enabled throughout student optimization. Projection heads and other loss-specific modules are used only during distillation and are discarded after training.

Table 5: Model configuration used for the reported AERIN encoder. Architectural constants omitted from the main methodology are collected here.

Component / setting	**Value**
Sampling rate	16 kHz
Chunk duration	5 s
Frame window	256 samples
Frame hop	96 samples
Number of evidence streams	4
Common stream width D	384
Periodicity window Q	512 samples
Periodicity lag set	[40,53,67,80,100,133,160,200,267,320]
Stem width	320
Stem-conditioning scale α_s	0.25
Axial layers	3
Attention heads	4
Temporal window	64 frames
FFN multiplier	2
Attention dropout	0.10
Temporal segments K	16
token transformer internal width	576
token transformer FFN width	1152
token transformer layers	2
token transformer heads	4
Local prior window	2
Z_{low} to Z_{high} residual scale α_{lh}	0.10

Table 6: Optimization configuration for the reported encoder.

Setting	**Value**
Training epochs	20
Batch size	16
Gradient accumulation	1
Encoder learning rate	1×10^{-4}
Training-head learning rate	1×10^{-4}
Weight decay	2×10^{-4}
Warmup	5 epochs
Learning-rate schedule	cosine decay
Gradient clipping	5.0
Automatic mixed precision	enabled
Minimum valid speech	0.25 s
Training seed	42
KD weight	0.80
Temporal KD weight	0.45
Geometry / anti-collapse weight	0.20
Cross-state decorrelation weight	0.05
Temporal teacher tokens	16
Checkpoint frequency	every epoch

The state-to-teacher assignment in Table 7 is fixed throughout distillation. WavLM-base remains frozen, and layers 3, 7, and 11 supervise the auxiliary, lower, and higher student states, respectively. The listed weights are the state weights λ_κ used throughout the state-level, temporal, and geometry terms of Appendix D.1.

Table 7: Student states and their WavLM teacher targets.

Paper state	Teacher layer	Loss weight
Z_{aux}	3	0.10
Z_{low}	7	0.55
Z_{high}	11	0.35

D.4 Knowledge-Distillation and Routing Diagnostics

We analyze the evolution of the student representations during distillation using diagnostics computed once per epoch on held-out distillation chunks. Epoch 1 denotes the model after its first training epoch, rather than at initialization. These quantities are used only for analysis and are not additional terms in the training objective.

The optimization curves converge smoothly and are effectively flat by the end of training (Fig. 5a). Both pooled and temporal teacher-student alignment improve consistently and then saturate (Fig. 5b, c). The checkpoint used for downstream evaluation lies on this late-training plateau, indicating that its selection does not correspond to a transient stage of the distillation process.

These cosine similarities are measured in the shared projected distillation space. They therefore show that the student states preserve information needed to predict the corresponding teacher targets, but do not imply that the raw AERIN representations reproduce WavLM features. The stream interventions in Section 3.2 and the routing intervention in Section 3.3 instead evaluate the internal organization of the student directly.

The representation geometry also changes substantially during training. Effective rank increases while both leading-eigenvalue concentration and random-pair similarity decrease (Fig. 5d, f). Thus, stronger teacher alignment is accompanied by a broader use of the representation space rather than concentration into a small number of dominant directions.

The low and high-level states remain related but become less redundant during training (Fig. 5e). This residual dependence is expected because Z_{high} is constructed partly from Z_{low} (Eq. 51). The separation objective therefore reduces, rather than eliminates, shared structure between the two read-outs.

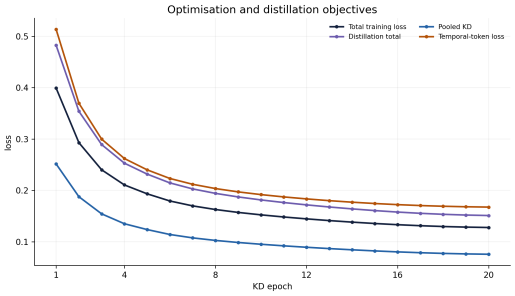
Figure 6 shows how the competitive router evolves during the same training process. The average stream allocation changes early in training and subsequently stabilizes (Fig. 6a). Phonation and texture receive somewhat larger average shares, while all four evidence streams remain active. At the same time, the entropy of the mean allocation approaches its maximum (Fig. 6b), showing that the router does not collapse globally onto a single evidence stream.

Importantly, high entropy of the *mean* routing distribution does not imply diffuse routing at every frame and feature coordinate. The gate can make locally concentrated assignments while different streams win at different positions. The corpus-level routing analysis in Table 11 captures this complementary view.

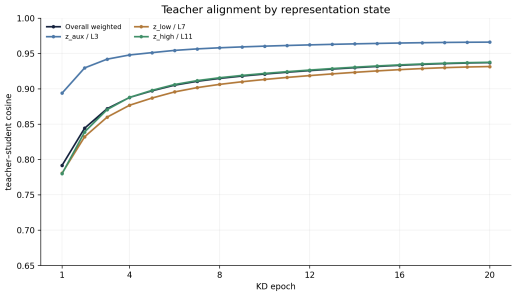
E Panel evaluation protocol

This appendix specifies the complete downstream evaluation protocol used for AERIN and for all thirteen baseline systems. The same probe family, optimization settings, split construction, seed set, selection metric, and uncertainty procedure are used across systems unless a corpus-specific constraint is stated explicitly below.

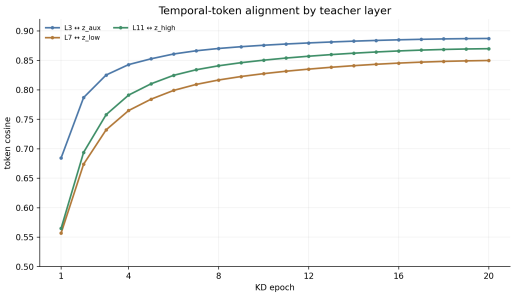
The encoder is frozen before downstream evaluation. No encoder parameter is updated on any evaluation corpus. Each recording is first passed through the frozen encoder, the required read-out states are cached, and all subsequent training is restricted to the probe. This separates representation evaluation from encoder fine-tuning.



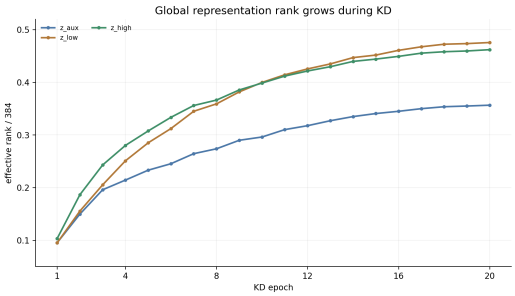
(a) Training objective and distillation components.



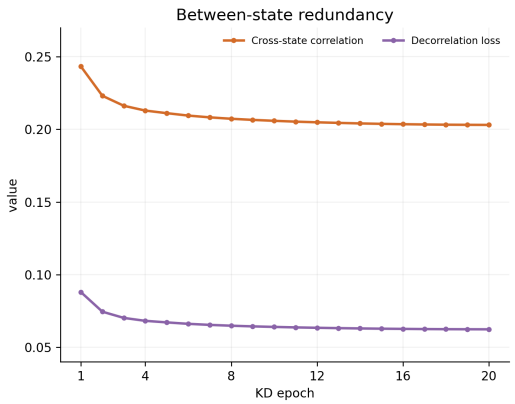
(b) Pooled teacher–student alignment.



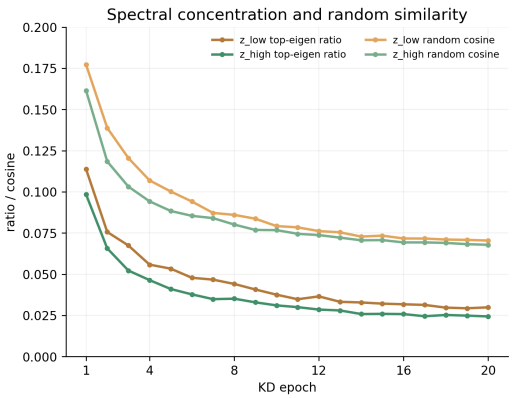
(c) Temporal-token alignment.



(d) Effective rank of each student state.



(e) Low–high state redundancy.



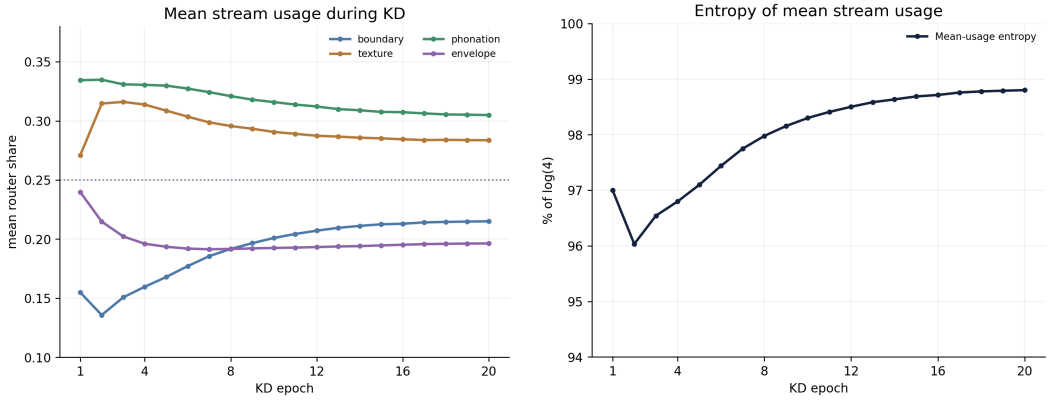
(f) Spectral concentration and random similarity.

Figure 5: **Representation diagnostics during distillation.** Teacher-student cosine similarities are measured after the training-only projection heads described in Appendix D.1. Z_{aux}, Z_{low}, and Z_{high} are supervised from WavLM-base layers 3, 7, and 11, respectively. Effective rank, spectral concentration, random-pair similarity, and between-state correlation track the geometry of the learned student representations throughout training.

E.1 Evaluation order

For each encoder-corpus pair, evaluation follows the same sequence:

1. Every chunk from every recording is encoded once using the frozen encoder. The selected read-out states are cached for all chunks, with no chunk subsampling. All chunks belonging to the same recording are then aggregated into a single recording-level representation before probe fitting.
2. Dataset partitions are constructed at the speaker level wherever speaker identity is available. The resulting speaker assignment is propagated to every recording from that speaker, so that recordings from the same speaker cannot appear in both training and evaluation partitions.



(a) Mean routing share of each evidence stream.

(b) Entropy of the mean routing allocation.

Figure 6: **Evolution of evidence routing during distillation.** (a) Mean routing share, averaged over valid frames and feature dimensions; the dotted line denotes the uniform allocation of $1/4$. (b) Entropy of the mean allocation, normalized by $\log 4$. These plots describe average routing behavior; local gate concentration is reported separately in Table 11.

Table 8: **Distillation diagnostics at the beginning and end of training.** Cosines are measured in the shared projected distillation space. Effective rank is normalized by the representation dimension. Low–high correlation summarizes the cross-correlation between standardized Z_{low} and Z_{high} states. Stream shares are averaged over valid frames and feature dimensions.

Quantity	**State**	**E1**	**E20**
Total loss	—	0.399	0.128
Pooled cosine	Z_{aux}	0.894	0.966
	Z_{low}	0.781	0.931
	Z_{high}	0.780	0.937
Temporal-token cosine	Z_{aux}	0.684	0.887
	Z_{low}	0.557	0.850
	Z_{high}	0.565	0.870
Effective rank	Z_{aux}	0.095	0.356
	Z_{low}	0.095	0.476
	Z_{high}	0.103	0.462
Leading-eigen fraction	Z_{aux}	—	0.038
	Z_{low}	0.114	0.030
	Z_{high}	0.098	0.024
Random-pair cosine	Z_{aux}	—	0.081
	Z_{low}	0.177	0.070
	Z_{high}	0.162	0.068
Low–high correlation	—	0.243	0.203
$\mathcal{L}_{\text{sep}}$	—	0.088	0.062
Stream share	Phonation	0.334	0.305
	Texture	0.271	0.284
	Boundary	0.155	0.215
	Envelope	0.240	0.196
Entropy of mean share (% of $\log 4$)	—	97.0	98.8

3. For corpora evaluated by cross-validation, the reported performance is obtained with speaker-disjoint outer folds. Model-selection decisions, including the read-out and training duration, are made only within the corresponding outer-training partition using an inner speaker-disjoint validation split. For corpora with an official held-out evaluation split, the same selection procedure is performed only within the official training partition.

4. The probe is first trained on the inner-training data and the stopping epoch is selected from the inner validation performance. Once the training duration and read-out have been fixed, the probe is refitted using the complete outer-training partition for the selected number of epochs. The corresponding outer test fold or official held-out split is evaluated only after this refit.
5. For binary tasks, the operating threshold is estimated exclusively from training-partition predictions. Where possible, we use speaker-disjoint out-of-fold predictions from the outer-training partition to estimate this threshold. The resulting threshold is fixed before the outer test predictions are scored and is never optimized on the reported evaluation split.
6. The procedure is repeated for each probe seed. Predictions are first combined across seeds at the scored-unit level. Confidence intervals are then estimated using clustered bootstrap resampling: speakers are resampled for speaker-based corpora, with all recordings from a sampled speaker kept together. Recording-level resampling is used only when recordings are the independent evaluation units.

E.2 Probe architecture

The probe operates at the recording level. Let $\{u_j\}_{j=1}^{M}$ denote the valid chunk representations of one recording, with $u_j \in \mathbb{R}^D$. The chunk sequence is reduced using three statistics computed independently for every feature dimension:

$$p = [\mathrm{mean}_j(u_j),\ \mathrm{std}_j(u_j),\ \max_j(u_j)] \in \mathbb{R}^{3D}. \tag{65}$$

Only valid chunks contribute to these statistics. The pooled vector is then processed by

$$\mathrm{LN}(3D) \to \mathrm{Linear}(3D, 128) \to \mathrm{GELU} \to \mathrm{Dropout}(0.20) \to \mathrm{Linear}(128, C), \tag{66}$$

where C is the number of target classes. At $D = 384$, the probe contains approximately 1.5×10^5 trainable parameters.

The three-statistic pooling is retained because recordings can differ not only in their average chunk representation but also in how that representation varies over time. Mean pooling alone removes this variation. The standard deviation term preserves across-chunk dispersion, while the maximum term retains strongly expressed local evidence that may occur in only part of a recording.

E.3 Probe optimization and model selection

All probe configurations use the same training setup for AERIN and every baseline encoder. Encoder parameters remain frozen, and optimization is restricted to the recording-level probe defined in Appendix E.2.

For a recording i, let $g_i \in \mathbb{R}^C$ denote the logits produced by the probe and let $y_i \in \{1, \ldots, C\}$ denote its target class. Probe parameters are optimized with categorical cross-entropy,

$$\mathcal{L}_{\mathrm{probe}} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(g_{i,y_i})}{\sum_{c=1}^{C} \exp(g_{i,c})}, \tag{67}$$

where B is the number of recordings in the minibatch. The loss is evaluated at the recording level after the chunk representations have been aggregated by Eq. 65; individual chunks do not receive separate supervision.

The same objective is used for binary, categorical, and ordinal corpora. For a binary task, $C = 2$. Ordinal targets are likewise optimized as a C-class classification problem; their ordering is used only when evaluating quadratic weighted kappa. AUC, macro-F_1, and quadratic weighted kappa are therefore *selection metrics*, not differentiable training objectives. No auxiliary loss, representation regularizer, or encoder-level loss is active during downstream probe fitting.

The shared optimization settings are listed in Table 9. Only the maximum epoch budget and, for the two small corpora described in Appendix E.4, the source of the stopping signal vary across datasets.

The task-specific validation metric is used both for early stopping and for selecting among candidate read-outs within a corpus. Binary tasks use area under the ROC curve (AUC), categorical tasks use

Table 9: Training configuration used for all downstream probes. The same probe objective and optimization settings are used for AERIN and all baseline encoders.

Setting	Value
Training objective	categorical cross-entropy
Supervision level	recording
Learning rate	1×10^{-3}
Weight decay	1×10^{-3}
Batch size	8 recordings
Batch shuffling	enabled
Probe dropout	0.20
Gradient clipping	1.0 (global norm)
Early-stopping patience	5 epochs
Probe seeds	42, 43, 44
Inner-validation fraction	20% of training recordings
Binary selection metric	AUC
Categorical selection metric	macro-F_1
Ordinal selection metric	quadratic weighted kappa
Binary decision threshold	fitted on inner validation split

macro-F_1, and ordinal tasks use quadratic weighted kappa. These metrics do not enter Eq. 67. For binary tasks, the decision threshold is estimated from the corresponding inner-validation predictions after probe fitting and is then applied unchanged to the reported split, as detailed in Appendix E.1.

The corpora differ substantially in recording count, so the same maximum number of training epochs is not imposed on every dataset. The epoch ceiling limits the optimization budget; it is therefore not the selection epoch. Whenever an independent inner-validation split is available, early stopping determines the selected epoch within that ceiling. The corpus-specific epoch budgets and stopping rules are given in Table 10.

Table 10: Per-corpus probe budgets and stopping protocol. Recording counts refer to recording-level probe examples rather than waveform chunks.

Corpus	Family	Recordings	Epoch ceiling	Stopping rule
DAIC-WOZ	clinical	106 / 35	20	inner val
MODMA	clinical	52	10	reported split
IPVS	clinical	65	10	reported split
NeuroVoz (HC/PD)	clinical	108	20	inner val
SAND	clinical	273 / 67	20	inner val
NeuroVoz-MinPair	phonetic	108	25	inner val
NeuroVoz-Prompt	phonetic	108	25	inner val
NeuroVoz-Vowel	phonetic	1035	30	inner val
TESS	paralinguistic	1223	30	inner val, word-disjoint
RAVDESS-Speech	paralinguistic	1440	30	inner val
ShEMO	paralinguistic	2999	30	inner val

E.4 Split construction and early stopping

Where the corpus size permits it, early stopping uses an inner validation set containing 20% of the training recordings. The inner split is speaker disjoint from the remaining training subset. The probe is first fitted on the reduced training subset and monitored on this inner validation set. Once the stopping epoch is selected, a new probe is fitted from initialization on the complete training partition for that number of epochs, after which the reported split is scored once.

Two of the clinical corpora are too small for this procedure to provide a stable inner validation estimate. MODMA contains 52 recordings and IPVS contains 65. A 20% validation fraction would leave only about 10 and 13 recordings, respectively, in the validation subset. For these two corpora, stopping is therefore performed directly on the split that is later reported.

This exception introduces selection on the reported split. If the score at each evaluated epoch has standard deviation σ, selecting the maximum over N evaluations produces the usual winner's-curse inflation. Using the approximation

$$\Delta_{\mathrm{sel}} \approx 0.43\,\sigma\sqrt{2\ln N}, \tag{68}$$

the approximate inflation is

N	Approximate inflation
10	$0.92\,\sigma$
20	$1.05\,\sigma$
30	$1.12\,\sigma$

For this reason, the two affected corpora use an epoch ceiling of 10 rather than 30. The shorter ceiling reduces the number of opportunities for selection on the reported split, but does not remove the bias. Their grid-evaluation scores should therefore be interpreted as carrying an additional optimistic component of approximately $0.9\,\sigma$ relative to corpora for which stopping uses an independent inner validation subset.

This exception concerns the fixed-split grid only. When cross-validation is reported, every fold uses its own inner validation subset and therefore does not use the held-out fold for early stopping.

E.5 Cross-validation protocol

Cross-validation results use five folds. Fold assignment is performed at the speaker level. Each speaker is assigned the majority label among that speaker's recordings, and folds are stratified using these speaker-level labels. The resulting speaker assignment is then propagated to all recordings from that speaker. No speaker therefore appears in both the training and held-out portions of a fold.

Within each fold, 20% of the fold-specific training recordings are reserved as an inner validation subset using the same speaker-disjoint construction. Early stopping is subject to a minimum training floor of five epochs, and the validation trajectory is smoothed over three epochs before the stopping decision is made. The fold assignment seed is held fixed across all encoders so that every system is evaluated on the same train/validation/test partition at each fold.

For binary corpora, threshold fitting is also performed within the fold-specific inner validation subset. The resulting threshold is applied unchanged to the held-out fold.

E.6 Read-out search budget

The number of candidate read-outs evaluated for the baseline encoders is not constant across all corpora. The large baseline models sweep across all of their layers.

The AERIN encoder exposes two predefined read-outs. Its search budget is therefore smaller than the baseline sweep. No numerical correction is applied for these different search-space sizes. Reported results use the best configuration within the search budget assigned to each encoder by the evaluation panel.

This distinction matters when interpreting the read-out ladder. The ladder is part of the representation design, but its evaluation also exposes two candidate states to downstream selection. Conversely, a baseline allowed to search many internal layers receives a larger downstream selection surface. The panel reports the resulting scores directly rather than treating the number of candidate read-outs as equal across systems. The model from the final distillation epoch is used for evaluations.

E.7 Protocol limitations

The evaluation protocol has the following known limitations that affect interpretation of the reported numbers.

1. **One corpus cannot be split by speaker.** The corpus contains only two actors, so a speaker-disjoint partition is not possible. It is instead partitioned by carrier word. This prevents the

same lexical item from occurring on both sides of the split, but does not test generalization to unseen speakers. Results on this corpus are therefore within-speaker results.

2. **Two corpora use the reported split for early stopping.** MODMA and IPVS do not provide enough recordings for a stable 20% inner validation subset under the protocol used for the larger corpora. Their fixed-split grid scores consequently contain the selection effect described in Appendix E.4.

These limitations do not change the mechanics of the reported evaluation, but they determine what each number supports. In particular, the affected fixed-split results should not be interpreted as estimates obtained under a fully nested selection protocol, and the two-actor corpus should not be interpreted as unseen-speaker evaluation.

F Extended results and analysis

This section extends the results in Section 3 with analyses that are useful for interpreting AERIN but are too detailed for the main paper. The tables below use the same frozen-encoder evaluation protocol as Table 1 unless stated otherwise. The appendix focuses on per-corpus routing behavior, stream ablations, read-out selection, metric sensitivity, and results on an additional corpus.

F.1 Routing analysis

The routing statistics are reported separately for each corpus in Table 11. The table summarizes the same learned gate at two levels. The *per-element* statistics are computed at the stream, frame, and feature positions where routing is actually applied. The final four columns first average the gate over valid frames and feature dimensions, giving one corpus-level mean weight for each evidence stream. All statistics exclude padded frames through the waveform-validity mask.

Table 11: **Per-corpus routing behavior of AERIN.** The *per-element* block measures the gate at the granularity at which routing is applied. Entropy is normalized by $\log R$, $\max g$ is the mean largest stream weight at each valid frame–feature position, and the next two columns report the fraction of positions for which that maximum exceeds 0.5 or 0.8. Lower entropy and larger values in the other three columns indicate stronger local routing concentration. The final block reports the same gate after averaging over valid frames and feature dimensions, yielding one mean weight for Boundary (bnd), Envelope (env), Phonation (pho), and Texture (tex). *Valid* is the fraction of frames retained by the validity mask, and C is the number of classes in the corresponding routing condition. The NeuroVoz rows are different task or elicitation conditions drawn from the same underlying corpus rather than separate datasets. *Vowel ID* uses sustained-vowel recordings, *spontaneous* denotes spontaneous speech, *HC/PD* is the healthy-control versus Parkinson's disease task, *word* denotes the GANGA/MANGA condition, and *DDK* denotes the three-class DDK condition.

			Per-element routing				**Mean gate over valid** T, D			
Corpus	C	**Valid**	$H(g)/\log R$	$\max g$	>0.5	>0.8	**bnd**	**env**	**pho**	**tex**
RAVDESS-Speech	8	0.75	41.8%	0.770	88.0%	51.6%	0.191	0.238	0.287	0.284
NeuroVoz vowel ID	5	0.84	45.5%	0.752	87.1%	47.4%	0.214	0.176	0.333	0.277
TESS	7	0.41	46.9%	0.743	85.8%	45.7%	0.213	0.191	0.304	0.292
IPVS	2	1.00	47.7%	0.737	85.2%	44.2%	0.204	0.198	0.317	0.281
MODMA	2	0.99	49.4%	0.726	84.1%	41.6%	0.202	0.218	0.298	0.283
NeuroVoz spontaneous	2	1.00	49.5%	0.726	83.9%	41.8%	0.203	0.200	0.307	0.290
NeuroVoz (HC/PD)	2	0.78	50.0%	0.724	83.7%	41.3%	0.208	0.193	0.316	0.284
DAIC-WOZ	2	0.78	50.9%	0.718	83.2%	39.9%	0.210	0.198	0.308	0.283
NeuroVoz word	2	0.72	51.1%	0.717	82.8%	39.7%	0.206	0.200	0.311	0.283
NeuroVoz DDK	3	1.00	52.0%	0.711	82.1%	38.5%	0.211	0.199	0.314	0.275
ShEMO	6	0.74	53.0%	0.706	81.4%	37.3%	0.212	0.197	0.298	0.293

AERIN exhibits concentrated routing on every corpus in Table 11. The mean per-element maximum gate remains between 0.706 and 0.770, and a single stream receives more than half of the routing mass at 81.4%–88.0% of valid frame–feature positions. The stronger threshold gives the same

result: 37.3%–51.6% of valid positions assign more than 0.8 of the mass to one stream. Normalized entropy, $H(g)/logR$, varies inversely, from 41.8% to 53.0% of $\log R$. Thus, the router is not operating near an element-wise uniform mixture on any of the measured corpora.

The corpus-level mean weights give a different view of the same gate. Boundary remains between 0.191 and 0.214, Envelope between 0.176 and 0.238, Phonation between 0.287 and 0.333, and Texture between 0.275 and 0.293. These averages are substantially less variable than the per-element statistics. In particular, a corpus can have mean stream weights that appear relatively balanced while still containing strongly selective routing at individual frame–feature positions. For AERIN, routing is therefore better characterized by where the gate concentrates locally than by a single global stream weight averaged over an utterance or corpus.

The variation in concentration is not explained by the number of classes. RAVDESS-Speech is the most concentrated row, whereas ShEMO is the least concentrated. TESS lies between them, while several binary corpora occupy both the middle and upper end of the entropy range. The gate statistics therefore do not follow a monotonic relation with downstream class count.

The NeuroVoz conditions provide a more controlled view because several elicitation settings are measured within the same underlying cohort. Normalized entropy changes from 45.5% for vowel identification to 49.5% for spontaneous speech, 51.1% for word speech, and 52.0% for DDK. The high-confidence fraction $\Pr(\max_r g_r > 0.8)$ correspondingly falls from 47.4% to 41.8%, 39.7%, and 38.5%. At the same time, the corpus-averaged stream weights move only modestly. This separation is consistent with the intended behavior of AERIN; the routing decision changes at the local evidence level rather than being reducible to a fixed corpus-specific mixture of the four streams.

The mean stream weights also show that Phonation and Texture receive somewhat larger aggregate mass than Boundary and Envelope on most corpora, but neither stream globally dominates the router. The largest Phonation mean is 0.333 on NeuroVoz vowel identification, while the largest Texture mean is 0.293 on ShEMO. These values remain far below the local maxima observed at individual positions, again showing that the identity of the dominant stream changes across the representation rather than remaining fixed.

F.2 Stream ablation and sufficiency analysis

Table 1 evaluates the intact encoder. Table 12 instead perturbs the evidence interface at inference time. A *w/o* condition zeros one private evidence stream while leaving the shared waveform stem and the remaining streams unchanged. An *only* condition keeps one private stream and zeros the other three. These are standard component-ablation and single-component-sufficiency tests; they are not retrained models.

Table 12: Stream ablation and single-stream sufficiency under the reported AERIN evaluation protocol. Scores are F_1 for binary corpora and macro-F_1 for categorical corpora. The intact row reproduces the reported AERIN result in Table 1. The $1/n$ row gives the score resolution associated with one scored unit and is included because the smallest corpora have coarse evaluation grids. No ablation value is highlighted as "best": an ablated condition exceeding the intact score is not evidence of a better model and must be interpreted relative to the evaluation resolution and probe variability.

Condition	DAIC	MODMA	IPVS	NVoz	NV-Vw	NV-Mp	NV-Pr	TESS	RAV	ShEMO
Scored units n	35	52	65	108	1 035	214	1 690	1 223	1 440	2 999
Score unit $1/n$	0.0286	0.0192	0.0154	0.0093	0.0010	0.0047	0.0006	0.0008	0.0007	0.0003
Intact AERIN	0.361	0.705	0.982	0.892	0.942	0.980	0.972	0.893	0.685	0.661
w/o Boundary	0.417	0.728	0.965	0.905	0.928	0.980	0.968	0.890	0.692	0.663
w/o Envelope	0.444	0.753	0.951	0.890	0.929	0.983	0.968	0.893	0.684	0.645
w/o Phonation	0.417	0.722	0.969	0.860	0.895	0.977	0.957	0.871	0.574	0.575
w/o Texture	0.361	0.714	0.982	0.801	0.882	0.973	0.913	0.886	0.600	0.591
Boundary only	0.361	0.662	0.945	0.827	0.894	0.922	0.812	0.857	0.426	0.495
Envelope only	0.500	0.724	0.922	0.770	0.855	0.900	0.801	0.878	0.449	0.518
Phonation only	0.500	0.632	0.917	0.788	0.850	0.898	0.853	0.888	0.566	0.543
Texture only	0.417	0.722	0.945	0.826	0.899	0.972	0.946	0.874	0.565	0.547

The most consistent ablation effect appears on the larger multi-class corpora. On RAVDESS-Speech, removing Phonation reduces macro-F_1 from 0.685 to 0.574 (-0.111), while removing Texture reduces it to 0.600 (-0.085). On ShEMO, the corresponding reductions are 0.086 for Phonation and 0.070 for Texture. Texture is also strongly used on NeuroVoz HC/PD ($0.892 \rightarrow 0.801$), vowel identification ($0.942 \rightarrow 0.882$), and prompt identification ($0.972 \rightarrow 0.913$). These effects are substantially larger than the $1/n$ resolution of the respective evaluations.

Boundary and Envelope do not show the same uniform direction. Their removal is close to neutral on several larger corpora and can slightly increase the reported score. This is not inconsistent with the router result: AERIN can use different evidence at different frame–feature positions, and a stream that is useful locally need not produce a positive marginal effect after the complete recording-level probe and thresholding procedure. The ablation table therefore measures downstream sensitivity to each input stream; it is not an additive decomposition of the representation.

The single-stream rows show the complementary side of the same result. Texture alone retains $0.946/0.972 = 97.3\%$ of the intact score on NeuroVoz prompt ID, whereas on RAVDESS-Speech the best single stream retains only $0.566/0.685 = 82.6\%$. TESS is different again: Phonation alone reaches 0.888 against an intact score of 0.893. AERIN therefore does not exhibit a single fixed "essential" stream across tasks; the amount of evidence that can be discarded depends on the downstream discrimination problem.

F.2.1 Small-cohort ablation diagnostics

DAIC-WOZ and MODMA require separate inspection because several ablated conditions score above the intact configuration in Table 12. Table 13 reports the underlying binary metrics for the intact model and the four single-stream ablations. Weighted F_1 is omitted here because its refreshed panel value is reported consistently in Table 17 to diagnose the operating point of the binary probe.

Table 13: Binary-metric diagnostics for the two smallest clinical corpora. Each row uses the same predictions as the corresponding F_1 entry in Table 12. P, R, Acc, AUC, and MCC denote precision, recall, accuracy, ROC-AUC, and Matthews correlation coefficient. The table is included to explain why some stream ablations increase F_1 on these small evaluation sets; it is not used to rank evidence streams.

Corpus	Condition	F_1	P	R	Acc	AUC	MCC
DAIC-WOZ	Intact	0.361	0.133	0.056	0.629	0.606	-0.066
	w/o Boundary	0.417	0.411	0.389	0.600	0.551	0.100
	w/o Envelope	0.444	0.439	0.611	0.600	0.671	0.197
	w/o Phonation	0.417	0.167	0.083	0.657	0.679	0.151
	w/o Texture	0.361	0.295	0.306	0.600	0.576	0.116
MODMA	Intact	0.705	0.629	0.826	0.699	0.811	0.444
	w/o Boundary	0.728	0.790	0.696	0.776	0.861	0.555
	w/o Envelope	0.753	0.651	0.899	0.737	0.840	0.521
	w/o Phonation	0.722	0.611	0.899	0.686	0.816	0.442
	w/o Texture	0.714	0.747	0.696	0.756	0.849	0.510

On DAIC-WOZ, the intact probe is operating at very low recall (0.056). Removing Envelope, for example, raises recall to 0.611 while accuracy falls from 0.629 to 0.600. The increase from F_1 0.361 to 0.444 is therefore driven by a large movement of the decision operating point rather than by a uniform improvement across metrics. MODMA is less extreme, but all four single-stream ablations still exceed the intact F_1 on only 52 scored recordings. These two corpora are consequently not used to infer a negative contribution from Boundary, Envelope, Phonation, or Texture.

F.2.2 Controlled class-count analysis

The cross-corpus ablations mix class count, corpus size, language, and task. Table 14 isolates one of these factors by staying entirely within RAVDESS-Speech, keeping all 24 speakers, and merging the label set to construct $C = 8$, 4, and 2 classification problems.

The fraction of the intact score recoverable from one stream increases from 82.5% at eight classes to 90.7% at two classes. The result is consistent with the ablation pattern in Table 12: as the label space becomes simpler, one evidence family can account for a larger fraction of the discriminative signal.

Table 14: Single-stream sufficiency as the class count is reduced within RAVDESS-Speech. All 24 speakers are retained at every setting. *Best single stream* is the highest macro-F_1 obtained when only one private evidence stream is kept, and *retention* is that score divided by the intact AERIN score. Texture is the best single stream at all three class counts. Because the macro-F_1 chance floor changes with C, the table supports the monotonic within-corpus trend but not an effect-size claim attributable solely to class count.

C	Intact	Best single stream	Retention	Stream
8	0.703	0.580	0.825	Texture
4	0.751	0.668	0.890	Texture
2	0.830	0.753	0.907	Texture

We do not attach a correlation statistic to the three settings; the controlled result is the observed monotonic trend.

F.3 Read-out ladder analysis

AERIN offers three candidate read-outs to a downstream probe: Z_{low}, Z_{high}, and their concatenation $[Z_{\text{low}}; Z_{\text{high}}]$. For each evaluation, the read-out is chosen with the selection metric of Appendix E.3 (AUC for binary tasks, macro-F_1 for categorical tasks), computed only on the speaker-disjoint inner-validation data and averaged over probe seeds. Ties are broken in the order Z_{high}, Z_{low}, $[Z_{\text{low}}; Z_{\text{high}}]$, which was fixed before any held-out score was computed, Z_{high} is the final distilled state, and the concatenation is not an exposed read-out (Appendix C.4). Held-out scores play no part in the choice.

Table 15 reports both sides of this decision. The left block is the information the selection rule saw. The right block reports, for diagnosis only, the held-out score each read-out would have obtained under the protocol of Table 1. The final column gives the *regret* of the rule: the best held-out score in the row minus the held-out score of the selected read-out.

Table 15: Read-out selection and its consequence. C is the number of target classes; $C = 2$ rows are binary evaluations. *Left:* selection scores on the inner-validation data (AUC for binary evaluations, macro-F_1 for categorical ones), averaged over probe seeds 42–44. *Right:* held-out scores in the metric of Table 1 (F_1 or macro-F_1) for every read-out, obtained with the same protocol; these values are never used for selection. The two blocks therefore use different metrics on binary rows and should not be compared numerically. **Bold** marks the maximum within each block (all tied maxima are marked); underline marks the selected read-out, whose held-out value is the one reported in Table 1. *Sel.* gives the selected read-out (L $= Z_{\text{low}}$, H $= Z_{\text{high}}$, LH $= [Z_{\text{low}}; Z_{\text{high}}]$), with * where a tie was broken by the pre-declared order. *Regret* is the best held-out score in the row minus the selected one. SAND is omitted because it is scored only by the challenge server.

		Selection (inner val.)			Held-out (diagnostic)				
Evaluation	C	Z_{low}	$[Z_{\text{low}}; Z_{\text{high}}]$	Z_{high}	Z_{low}	$[Z_{\text{low}}; Z_{\text{high}}]$	Z_{high}	**Sel.**	**Regret**
DAIC-WOZ	2	0.583	0.597	**0.646**	**0.361**	0.333	0.333	H	0.028
MODMA	2	**0.924**	0.896	**0.924**	**0.715**	0.662	0.705	H*	0.010
IPVS	2	**1.000**	**1.000**	**1.000**	0.970	0.970	**0.982**	H*	0.000
NV-H/P	2	**0.968**	**0.968**	0.963	0.887	0.867	**0.892**	L*	0.005
NV-SV	5	0.970	**0.974**	0.965	0.937	**0.945**	0.942	LH	0.000
NV-GM	2	**1.000**	**1.000**	**1.000**	0.981	**0.983**	0.980	H*	0.003
NV-LR	16	0.970	**0.976**	0.974	0.971	0.969	**0.972**	LH	0.003
TESS	7	**0.908**	0.907	0.899	0.893	0.904	**0.906**	L	0.013
RAVDESS-Speech	8	**0.786**	0.783	0.721	**0.685**	**0.685**	0.645	L	0.000
ShEMO	6	0.662	**0.678**	0.649	**0.665**	0.661	0.630	LH	0.004

The selection scores saturate at 1.000 for every read-out on IPVS and NV-GM, and are tied between two read-outs on MODMA and NV-H/P, so the pre-declared order decides four of the ten choices. On NV-SV, NV-LR, and TESS the full selection spread is at most 0.009. Only three evaluations separate the read-outs materially on the selection data: RAVDESS-Speech (0.065), DAIC-WOZ

(0.063), and ShEMO (0.029). No read-out is preferred across the panel: Z_{high} is selected four times, and Z_{low} and the concatenation three times each.

Where the selection data separate the read-outs, the selected one is also the best or near-best on held-out data: RAVDESS-Speech selects Z_{low}, which attains the held-out maximum, and ShEMO selects the concatenation, 0.004 below it. The largest regret occurs on DAIC-WOZ (0.028), where the held-out split has 35 participants and one scored unit corresponds to $1/35 \approx 0.029$ (Table 12). Across the ten evaluations the regret is at most 0.028 and averages 0.007.

Table 16 asks whether the read-out choice drives the panel conclusion. It recomputes the mean rank of Table 1 with AERIN restricted to a single read-out on every evaluation, alongside the selection rule and an oracle that chooses the best held-out read-out. The oracle is shown only as an upper bound and is not a valid estimate. AERIN has the lowest mean rank under every rule, including each fixed read-out chosen without any selection. Note also that the baselines retain their full layer sweeps in every row (Appendix E.6), so fixing AERIN to one read-out makes the comparison stricter for AERIN, not looser.

Table 16: Sensitivity of the panel mean rank (Table 1, eleven evaluations) to the AERIN read-out rule. Baseline scores are unchanged in every row. The SAND score is held fixed because its read-out cannot be re-scored locally. The *oracle* uses held-out information and is reported only as an upper bound.

AERIN read-out rule	**AERIN mean rank**	**Best baseline (WavLM-large)**
Selection rule (reported)	4.09	4.59
Fixed Z_{low}	4.00	4.68
Fixed $[Z_{\text{low}}; Z_{\text{high}}]$	4.18	4.59
Fixed Z_{high}	3.64	4.68
Oracle (upper bound)	*3.23*	*4.77*

The ladder should therefore be read as a small, predefined set of candidate representations and not as a taxonomy that assigns one level of abstraction to each task family. Most selection margins are too small to support such a reading, and the one evaluation family that shows a consistent pattern (the affective corpora, where Z_{high} is worst on both selection and held-out data for RAVDESS-Speech and ShEMO) is not matched by TESS, where Z_{high} is best on held-out data. Because the ladder adds three candidates to the downstream search while each large baseline is swept over all of its layers, the read-out choice gives AERIN a smaller selection surface than the baselines, not a larger one.

F.4 Metric and coverage sensitivity

F.4.1 Weighted-F_1 sensitivity

Table 17 recomputes the ten locally scored benchmark corpora using weighted F_1. This is a metric-sensitivity check on Table 1, not a second benchmark.

For AERIN, changing from the primary class-balanced metric to weighted F_1 moves the mean rank only from 3.90 to 4.00 within this fixed comparison subset. The benchmark conclusion on this subset is therefore not determined by whether class support is used in the F_1 aggregation.

The corpus-level differences are more informative than the rank change. DAIC-WOZ moves from primary F_1 0.361 to weighted F_1 0.562, and ShEMO moves from macro-F_1 0.661 to weighted F_1 0.790. These large increases are consistent with class imbalance: support weighting gives the majority class more influence than the primary metric used in Table 1. MODMA, by contrast, changes only from 0.705 to 0.692, and several near-ceiling phonetic tasks change by only a few thousandths. This is why the primary metric remains the paper's main comparison and weighted F_1 is kept as a sensitivity analysis.

F.4.2 Speech Commands coverage

Speech Commands v2 (Warden, 2018) is reported as an additional evaluation and is not included in the mean-rank comparison of Table 1. The main reason is that complete baseline coverage is not

Table 17: Weighted-F_1 scores on the ten locally scored corpora. Binary predictions use each seed's out-of-sample decision threshold; categorical predictions use the class argmax. The two rank columns are computed over the *same 12 systems shown in this table* and the same ten local corpora: *wF1 rank* uses weighted F_1, whereas *primary rank* uses the reported F_1/macro-F_1 metric. SAND is absent because it has no local prediction files, and the two openSMILE systems are absent because complete predictions are not available on all ten corpora. Consequently, *primary rank* is not the headline rank in Table 1. In particular, both ranks shown here are calculated only from the participants present. **Bold** marks the best weighted-F_1 value in each corpus.

	Weighted F_1										Mean rank	
System	**DAIC**	**MODMA**	**IPVS**	**NVoz**	**NV-Vw**	**NV-Mp**	**NV-Pr**	**TESS**	**RAV**	**ShEMO**	**wF1**	**primary**
HuBERT-base	**0.643**	0.649	0.979	0.880	0.906	0.984	0.972	0.911	0.696	0.753	4.40	4.30
HuBERT-large	0.361	0.696	0.903	0.889	0.861	0.961	0.943	0.895	0.733	0.755	6.40	6.40
wav2vec2-large-robust	0.406	0.643	0.887	0.499	0.660	0.373	0.087	0.307	0.629	0.394	10.80	10.80
WavLM-base	0.318	0.696	**0.990**	0.861	0.866	0.980	0.974	**0.918**	0.700	0.757	4.80	4.40
WavLM-base+	0.492	**0.711**	0.979	0.812	0.901	0.978	0.973	0.903	0.727	0.768	4.90	4.80
data2vec-audio-large	0.521	0.691	0.944	0.659	0.462	0.895	0.952	0.828	0.648	0.604	8.80	9.00
XLS-R-300m	0.538	0.707	0.865	0.836	0.896	0.845	0.713	0.875	0.681	0.714	7.40	8.30
WavLM-large	0.345	0.669	0.867	0.829	0.893	**0.986**	0.976	0.905	0.755	0.808	5.00	4.10
w2v-BERT 2.0	0.519	0.608	0.698	0.805	0.611	0.941	0.802	0.881	0.734	0.677	9.00	9.40
Whisper-large-v3 enc.	0.530	0.666	0.856	0.839	0.933	0.981	**0.981**	0.897	0.752	0.840	4.40	5.00
emotion2vec+ large	0.508	0.620	0.850	0.709	0.731	0.643	0.223	0.892	**0.870**	**0.884**	8.00	7.50
AERIN (ours)	0.562	0.692	0.985	**0.892**	**0.947**	0.980	0.972	0.902	0.692	0.790	**4.00**	**3.90**

available for this corpus. Evaluating every intermediate layer of the large speech encoders is also considerably more expensive on Speech Commands v2 than on the other datasets in the evaluation panel because of its much larger size. We therefore report results only on a small subset of the baselines.

For the available baselines, the representation is taken from the final encoder layer. AERIN is evaluated using its predefined Z_{high} read-out. These results follow the same evaluation procedure described in Appendix E. The openSMILE systems are included as non-neural reference representations using their corresponding descriptor sets. Speech Commands v2 differs from the clinical, phonetic, and affective tasks in the main evaluation panel because it directly evaluates spoken-word recognition. It therefore provides an additional test of whether the frozen AERIN representation preserves lexical information from the teacher model even though the encoder is not trained specifically for this downstream task.

Table 18: Speech Commands v2 results under partial baseline coverage. Scores are macro-F_1 on the complete official test partition and are averaged over the reported probe seeds.

System	**Params (M)**	**Macro-F_1**	**Representation**
HuBERT-large	315.4	0.643	final layer
WavLM-base+	94.4	0.747	final layer
WavLM-large	315.5	**0.821**	final layer
openSMILE eGeMAPS	—	0.338	descriptor set
openSMILE ComParE-16	—	0.671	descriptor set
AERIN (ours)	25.66	0.768	Z_{high}

AERIN obtains a macro-F_1 of 0.768. Among the baselines, it outperforms WavLM-base+ and HuBERT-large, while WavLM-large achieves the highest score at 0.821. AERIN reaches this result with a substantially smaller encoder than the large HuBERT and WavLM models. The result therefore shows that the frozen AERIN representation retains useful information for lexical recognition in addition to the clinical, phonetic, and affective information evaluated elsewhere in the paper.

This experiment should not, however, be interpreted as a complete benchmark comparison on Speech Commands v2. Results are available for only part of the baseline set, so the corpus is kept outside the cross-dataset ranking and is used only as an additional evaluation of the learned representation.