HuPER: A Human-Inspired Framework for Phonetic Perception

Chenxu Guo[1,2,*,†], **Jiachen Lian**[2,*], **Yisi Liu**[2], **Baihe Huang**[2], **Shriyaa Narayanan**[2], **Bixing Wu**[1], **Zoe Ezzes**[3], **Jet Vonk**[3], **Zachary Miller**[3], **Cheol Jun Cho**[2], **Maria Gorno-Tempini**[3], **Gopala Anumanchipalli**[2]

[1]Zhejiang University [2]University of California, Berkeley [3]University of California, San Francisco
{jiachenlian,gopala}@berkeley.edu

Abstract

We propose *HuPER*, a human-inspired framework that models phonetic perception as adaptive inference over acoustic–phonetic evidence and linguistic knowledge. HuPER first learns an acoustically grounded phone recognizer from limited human-annotated data and transcript-only speech: canonical G2P transcriptions are treated as auxiliary linguistic cues rather than ground truth, and are corrected into realized phone proxies for self-training. At inference time, HuPER supports multiple perceptual routes: it can rely on bottom-up phone evidence when the signal is clear, incorporate explicit expectations when a reference is available, or use lexical constraints when acoustic evidence is weak. With only 100 hours of training data, HuPER achieves state-of-the-art phonetic feature error rates on five English benchmarks and strong zero-shot transfer to 94 unseen languages. HuPER also improves robustness on weak-evidence and disordered speech, demonstrating the benefit of adaptive, multi-path phonetic perception under diverse acoustic and task conditions. All training data, models, and code are open-sourced. Code and demo are available at `https://github.com/Berkeley-Speech-Group/HuPER`.

1 Introduction

Phonetic modeling is fundamental to speech perception. Early speech perception systems were expert systems [1–5] designed to emulate human perception, but constrained in scalability. Recent advances have demonstrated that large-scale word-based ASR [6–9] can match or surpass human parity across domains. However, progress at the phonetic level remains limited, with little comparable gains despite similar scaling efforts [10]. These observations motivate a human-centered rethinking of phonetic foundation models.

The human phonetic processor extracts acoustic–phonetic cues [12, 13] and organizes them into hierarchical phonological representations for higher-level linguistic inference [14]. This process comprises a language-universal and an experience-dependent layer [15] encoding language-specific structure [16]. Both functions are dynamically shaped by bottom-up sensory evidence and top-down perceptual processes [13, 16].

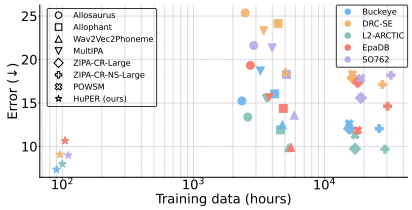


Figure 1: HuPER achieves highly data-efficient phonetic transcription on English variation benchmarks.

Early self-supervised speech learning (S3L) models [17] can be viewed as analogous to infants' early exposure to continuous speech [18], during which broad acoustic–phonetic representations [19] are initially formed. However, subsequent

*Equal contribution.
†Chenxu Guo was a Visiting Student Researcher at UC Berkeley during this work.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

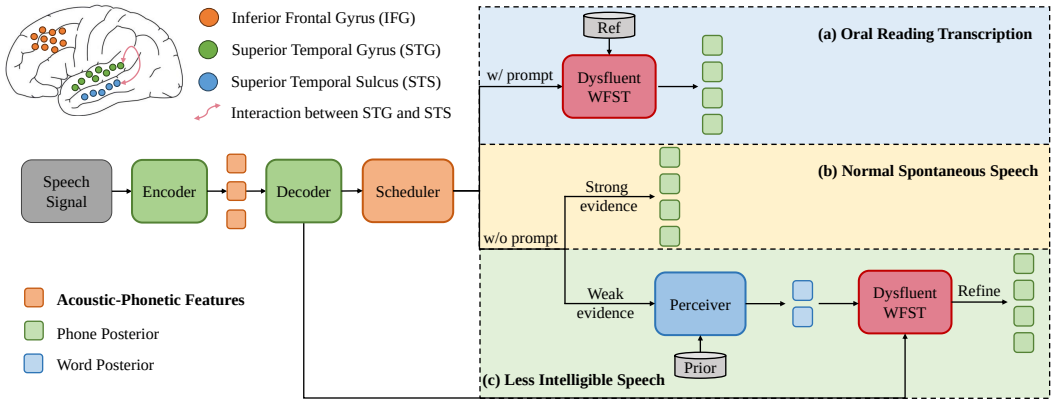


Figure 2: **HuPER overview: evidence-controlled multi-path speech perception.** Left: HuPER-Recognizer (Encoder/Decoder; mapped to STG) converts the speech signal into acoustic–phonetic features (orange) and phone posteriors (green). HuPER-Scheduler (mapped to IFG) monitors evidence strength and selects an inference route. Right: (a) *Oral reading transcription*: with an external prompt/reference, a Dysfluent WFST applies explicit top-down constraints to produce the transcript. (b) *Normal spontaneous speech*: when evidence is strong, the system trusts bottom-up phone evidence and outputs directly. (c) *Less intelligible speech*: when evidence is weak, HuPER-Perceiver combines phone evidence with a lexical prior to form word hypotheses (blue), which are then refined by the Dysfluent WFST [11].

fine-tuning [20, 21] typically introduces lexical supervision *prematurely*. For instance, in fast speech, "last Sunday" is often realized as "las[] Sunday." A developmentally plausible approach would supervise such surface forms first, allowing acoustic–phonetic representations to stabilize before higher-level abstraction. In contrast, most current phonetic models rely on G2P-derived canonical targets [22–24] (e.g., "last Sunday"), which encode phonological and grammatical regularities absent from the acoustic signal. This supervision mismatch prevents models from fully exploiting their acoustic–phonetic capacity and undermines subsequent phonological and lexical inference.

Beyond the premature introduction of lexical supervision, another open challenge is that human phonetic perception operates as a dynamic closed-loop system rather than a one-pass mapping from speech to phonemes [25–27]. Most phoneme recognition models implicitly assume a unidirectional, feedforward processing pipeline. However, under ambiguous or degraded conditions, human listeners often engage top-down inference, using lexical or contextual expectations to constrain phonetic interpretation [28–30], which in turn refines higher-level representations [25, 27]. Notably, even when top-down information is available, listeners can deliberately rely on bottom-up processing when instructed or required by task demands [31], indicating flexible cognitive control [32, 33]. Moreover, when reference text is available, as in performed or rehearsed speech, perception follows another pathway guided by explicit expectations [34]. In contrast, most current AI systems implement a single feedforward route and lack mechanisms for modeling such multi-path, closed-loop dynamics.

Given these limitations, we propose **HuPER (A Human-Inspired Framework for Phonetic Perception)**, a human-centered framework that models phonetic perception as adaptive inference integrating acoustic evidence and linguistic knowledge. HuPER departs from conventional end-to-end pipelines by explicitly coordinating bottom-up cues and top-down expectations for robust, context-aware perception. Our main contributions are summarized as follows:

(1) We present HuPER, the first unified and explicit computational framework for modeling human phonetic perception, which also provides a diagnostic perspective on existing phoneme recognition models.

(2) HuPER achieves state-of-the-art phonetic accuracy with only 100 hours of training data, reaching an average PFER of 8.82% on English benchmarks and demonstrating decent zero-shot multilingual transfer.

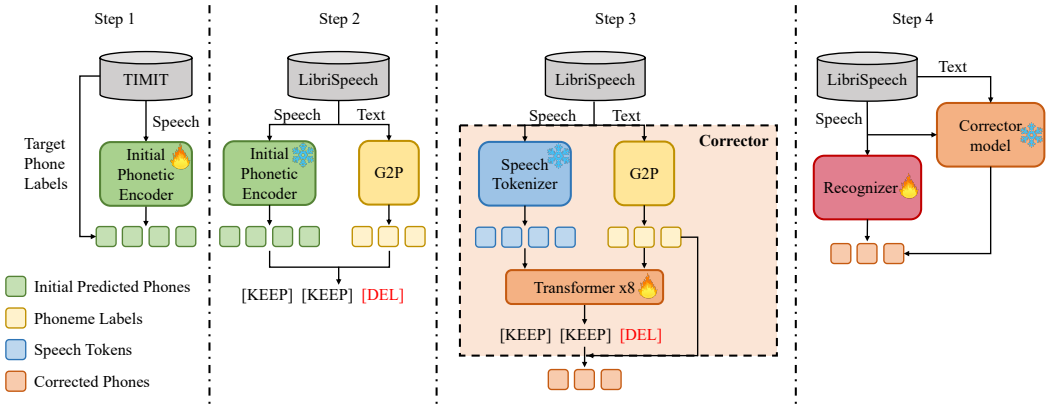


Figure 3: **HuPER-Recognizer self-learning pipeline**. HuPER uses an initial recognizer and G2P transcriptions to train a Corrector, which transforms canonical phoneme sequences into acoustically grounded phone proxies for retraining on LibriSpeech.

(3) We propose an adaptive multi-path inference mechanism for phonetic perception, enabling dynamic pathway selection and improved robustness under degraded and disordered speech conditions.

2 HuPER overview

We propose **HuPER (A Human-Inspired Framework for Phonetic Perception)**, a modular speech perception framework that supports adaptive multi-path inference by integrating bottom-up acoustic–phonetic modeling with explicit top-down constraints, serving as an approximation to human cognitive speech perception systems [25, 27]. As shown in Figure 2, HuPER consists of three functional modules and a central scheduler:

(1) HuPER-Recognizer (STG-like, Bottom-Up Acoustic–Phonetic Perception, Sec. 3). It extracts language-general acoustic–phonetic evidence and outputs spoken phones, providing the foundational representation for downstream modules and determining the upper bound of universal phonetic perception. We further develop a self-training procedure for HuPER-Recognizer, which is theoretically grounded in our DRRC framework. **(2) HuPER-Perceiver** (STS-like, Phonetic–Lexical Integration, Sec. 4). It combines acoustic–phonetic representations with explicit lexical and phonotactic priors to generate phonetic-enhanced word hypotheses. **(3) Dysfluent WFST** (explicit top-down constraints; [11]). It provides a human-inspired, top-down constraint mechanism by representing dysfluencies and pronunciation variants in a WFST constraint graph, which can be conditioned on external references or on hypotheses. **(4) HuPER-Scheduler** (IFG-like, Sec. 5). The scheduler selects inference pathways based on signal quality and task context. For clear speech, HuPER relies on bottom-up inference. Under degraded or ambiguous conditions, it integrates HuPER-Perceiver outputs with Dysfluent WFST constraints. In reference-guided scenarios [34], known intended text is incorporated through the constraint graph. We detail each part in the following sections.

3 HuPER-Recognizer

Task definition and outputs. The HuPER-Recognizer is a WavLM-Large model [21] fine-tuned for phone recognition. Given a speech signal X, it outputs phone posteriors and a decoded phone sequence $\hat{Y}_\theta$. We treat $\hat{Y}_\theta$ as the *spoken phone* (i.e., what a human would perceive from the acoustics), and we aim to learn phone evidence that generalizes across languages.

To scale beyond scarce human phone labels, a common approach is self-training: generate pseudo phone labels on a large transcript-only set and retrain the recognizer on them [35, 36]. However, naive pseudo-label training can amplify systematic errors (confirmation bias) and can be statistically biased when the availability of true phone labels is not random.

We address these issues by introducing **doubly robust risk correction (DRRC)**. Concretely, we (i) construct proxy phone supervision on $\mathcal{D}'$ via a phoneme→phone Corrector (Algorithm 1), and

(ii) view the resulting training problem as missing-label learning and analyze it through a DRRC objective (Sec. 3).

HuPER self-learning strategy. Figure 3 summarizes the recipe for constructing proxy phone supervision on a large transcript-only dataset. We use a small labeled set $\mathcal{D} = \{(X_i, Y_i)\}$ with human-verified phones and a transcript-only set $\mathcal{D}' = \{(X_j, T_j)\}$. For each $(X, T) \in \mathcal{D}'$, we compute a canonical phoneme sequence $Z = \mathrm{G2P}(T)$, obtain a teacher phone hypothesis $\bar{Y}$ from the current recognizer, and apply a Corrector to produce a corrected proxy phone sequence $\hat{Y}$. Algorithm 1 summarizes the iterative training procedure.

DRRC perspective. To formalize our analysis, let $W = (X, Z, Y, \hat{Y}, A)$ denote one training example in the missing-label view: X is the speech input, $Z = \mathrm{G2P}(T)$ is the canonical phoneme sequence, Y is the latent human-verified phone label, $\hat{Y}$ is the always-observed proxy phone label, and $A \in \{0, 1\}$ indicates whether Y is observed. Let $e(k)$ be the one-hot vector for phone class k; thus $e(Y)$ and $e(\hat{Y})$ are the one-hot targets for the true and proxy labels. For a fixed HuPER-Recognizer parameter θ, we define the true population risk as

$$R(\theta) := \mathbb{E}\big[\ell_\theta(e(Y), X)\big], \tag{1}$$

Treating the canonical phoneme sequence $Z = \mathrm{G2P}(T)$ as an auxiliary covariate, we cast self-training as a *missing data problem* [37, 38] in which the observability of the true label Y is governed by Z via the true propensity function:

$$g^*(z, \hat{y}) := \mathbb{P}(A = 1 \mid Z = z, \hat{Y} = \hat{y}). \tag{2}$$

Practically, this formulation assumes that the transcript contains sufficient auxiliary information, distinct from the features captured by the teacher model f_T, to distinguish the pseudo-label $\hat{Y}$ from the true label Y (but not necessarily to fully recover the true label Y). We demonstrate that there exists a corrector such that the HuPER loss yields a *doubly robust* estimate of $R(\theta)$ on the self-training dataset. Specifically, the estimator is consistent if either (i) the propensity model for missing phone labels is correctly specified, or (ii) the proxy label $\hat{Y}$ is (asymptotically) accurate.

Theorem 3.1 (Informal version of Theorem B.5)**.** *For any measurable $g : \mathcal{Z} \times \{1, \ldots, K\} \to (0, 1]$, define*

$$C_g(W) := e(\hat{Y}) + \frac{A}{g(Z, \hat{Y})}\big(e(Y) - e(\hat{Y})\big). \tag{3}$$

Let $\hat{g}$ be a cross-fitted estimator clipped to $[\varepsilon, 1]$ and define

$$\hat{R}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \ell_\theta\big(C_{\hat{g}}(W_i), X_i\big). \tag{4}$$

*Then $\mathbb{E}[\ell_\theta(C_g(W), X)] = R(\theta)$ provided either **(Prop)** $g = g^*$ or **(Proxy)** $\hat{Y} = Y$. Furthermore, $\hat{R}_n(\theta) \to R(\theta)$ in probability provided either:*

$$\textbf{(Prop):}\ \mathbb{E}[\,|\hat{g}(Z, \hat{Y}) - g^*(Z, \hat{Y})|\,] \to 0, \text{ or } \textbf{(Proxy):}\ \mathbb{P}(\hat{Y} \neq Y) \to 0. \tag{5}$$

This theorem shows that the corrected target $C_g(W)$ recovers the true risk $R(\theta)$ in expectation, and the empirical estimator is consistent when either the proxy labels are accurate (**Proxy**) or the propensity model is correct (**Prop**). Thus, DRRC is robust to misspecification of either component alone.

In summary, we scale HuPER-Recognizer with a phoneme→phone Corrector (Algorithm 1) and analyze the resulting self-learning objective through DRRC, which is consistent under either accurate proxies or a correct propensity model. Next, HuPER-Perceiver converts the phone evidence into word transcripts using explicit lexical and LM constraints.

4 HuPER-Perceiver

HuPER-Perceiver converts HuPER-Recognizer's *phone evidence* into word transcripts by composing it with explicit, auditable language constraints, following the classic HMM–GMM decoding recipe (acoustic model + lexicon + language model) but with a phone recognizer as the evidence source [5, 39]. Given an utterance X and recognizer parameters θ, the output is a word sequence $\hat{T}$.

Acoustic evidence as a phone lattice. Rather than using only the 1-best phone sequence $\hat{Y}_\theta$, we represent the recognizer output as a weighted phone acceptor (phone lattice) $\Pi_\theta(X)$. Each path y in $\Pi_\theta(X)$ corresponds to a candidate phone sequence, and is assigned a cost $C_\theta(y \mid X)$ derived from the negative log evidence (e.g., frame-level phone posteriors or arc weights). The decoded phone sequence $\hat{Y}_\theta$ is simply the best path in this lattice, while $\Pi_\theta(X)$ retains uncertainty needed for downstream constrained search.

Imposing lexical and linguistic constraints. To map phone hypotheses to words, we introduce a phone-to-word transducer L (lexicon) and a word-level acceptor G_{LM} (language model). L restricts which phone sequences realize valid words, and G_{LM} provides sequence-level linguistic preferences. Both L and G_{LM} are modular constraints that can be swapped across domains without retraining the HuPER-Recognizer.

Search via WFST composition. Decoding is performed by composing the phone evidence with the constraints and extracting the shortest path in the unified search space:

$$\hat{T} \;=\; \mathrm{Output}(\mathrm{ShortestPath}(\Pi_\theta(X) \circ L \circ G_{\mathrm{LM}}))\,. \tag{6}$$

5 HuPER-Scheduler

The HuPER-Scheduler acts as the system's *planner*, orchestrating the flow between bottom-up phone evidence and top-down linguistic expectations. By evaluating an *evidence distortion score* $s(X)$ computed from HuPER-Recognizer emissions, it decides whether to decode phones directly, or to activate a reference-constrained refinement path. In the guided path, we compile a ***Dysfluent WFST*** constraint $\mathcal{H}(\cdot)$ from a reference word sequence (external R or Perceiver 1-best hypothesis $\hat{T}$), and refine phone decoding by composing $\mathcal{H}$ with the recognizer evidence [11, 39]. Algorithm 2 summarizes the routing logic.

Quantifying Evidence Distortion. To assess signal reliability, the Scheduler monitors the Recognizer's frame-level logits $\mathbf{z} \in \mathbb{R}^{N\times K}$ and posteriors $\mathbf{p}_t = \mathrm{softmax}(\mathbf{z}_t)$, where N is the number of acoustic frames and K is the phone-inventory size. We quantify uncertainty through the posterior margin m_t and normalized entropy h_t:

$$m_t := p_{t,(1)} - p_{t,(2)}, \quad h_t := \frac{-\sum_{k=1}^{K} p_{t,k} \log p_{t,k}}{\log K}. \tag{7}$$

Here $p_{t,(1)}$ and $p_{t,(2)}$ denote the largest and second-largest posterior probabilities at frame t. These are combined into a frame-level distortion proxy $d_t := \mathrm{clip}\big(\frac{1}{2}(1-m_t) + \frac{1}{2}h_t, 0, 1\big)$, and aggregated into an utterance-level score

$$s(X) := \frac{1}{N}\sum_{t=1}^{N} d_t, \tag{8}$$

which drives routing decisions.

Dysfluent WFST constraint $\mathcal{H}(\cdot)$. Given a reference word sequence U (either an external reference R or a hypothesis $\hat{T}$), the Dysfluent WFST compiles a phone-space constraint $\mathcal{H}(U)$, which encodes a bounded set of plausible *realized* phone sequences around the canonical pronunciation of U, while allowing dysfluent edits (insertions/deletions/substitutions) [11]. This constraint is then combined with the HuPER-Recognizer phone evidence for constrained shortest-path inference.

6 Experimental setup

As described in Sec. 5, the HuPER-Scheduler supports three tasks: (1) **Task-1:** Speech-only transcription with strong acoustic evidence, corresponding to standard phoneme recognition. (2) **Task-2:** Speech-only transcription under varying signal quality, where the scheduler adaptively selects between a purely bottom-up path (for strong evidence) and a combined bottom-up–top-down path (for weak evidence). (3) **Task-3:** Transcription with reference text provided, in which bottom-up and top-down perception are jointly performed. For **Task-1**, we conduct *phone recognition* experiments (Sec. 6.1). For **Task-2 and Task-3**, we conduct *multi-path speech perception* experiments (Sec. 6.2).

6.1 Phone recognition

Setup. We use WavLM-Large as the backbone and fine-tune it with a CTC [40] objective for phone recognition. Training follows our self-learning recipe. We first train an initial HuPER-Recognizer on TIMIT [41]. We then apply this initial model to LibriSpeech [42] to obtain teacher pseudo phone labels, and train a correction model (Sec. 3) that refines the canonical G2P [23, 24, 43] phone sequence using acoustic evidence. The Corrector is trained for 49 epochs (about 40 minutes) on $2\times$A6000 GPUs. Audio is tokenized with HuBERT [44] units. We use dropout 0.2, AdamW [45] with $\beta = (0.9, 0.999)$, and learning rate 2×10^{-4}. Finally, we train HuPER-Recognizer on the corrected pseudo labels for 100 epochs on $8\times$A6000 GPUs (batch size 12, learning rate 3×10^{-5}). We freeze the WavLM transformer for the first 24k updates and train only the linear CTC head, then unfreeze all WavLM layers for full fine-tuning. The final CTC training loss is 0.036.

Evaluation metric: PFER. We report **Phonetic Feature Error Rate (PFER)**, a PanPhon-based articulatory-feature edit distance widely used in multilingual phone recognition [46]. PFER computes a minimum-cost edit distance between hypothesis and reference phone sequences, where substitutions are weighted by distinctive-feature differences (thus giving partial credit to phonetically similar phones), and is normalized by the reference length. Unless otherwise stated, PFER is reported on a fractional scale (not multiplied by 100); Table 1 reports the same quantity as a percentage. Formal definition and costs are provided in Appx. E.

Baselines. We compare against widely used open-source universal phone recognizers (Table 9): **Allosaurus** [47], **Allophant** [48], **Wav2Vec2Phoneme** [20, 49, 50], **MultIPA** [51], **ZIPA** [52], and **POWSM** [10].

Evaluation datasets. We evaluate only on corpora with *human-annotated* or *human-verified* phone labels, so the error rates are grounded in human perception rather than automatic alignments. For analysis, we group the evaluation sets into *English variation* and *unseen-language transfer*. Dataset statistics and evaluation splits are summarized in Table 7.

Table 1: **PFER (%, ↓) on human-annotated / human-verified phone datasets.** All baseline numbers are reproduced using released checkpoints under a unified evaluation pipeline. Train (h) denotes the amount of training data reported/used by each model. DRC-SE = DoReCo South-England; SO762 = SpeechOcean762. *Allosaurus reports training scale in utterance counts; we convert to hours assuming an average utterance duration of 4 seconds.

Model	Train (h)	Buckeye	DRC-SE	L2-ARCTIC	EpaDB	SO762	Avg.
Allosaurus [47]	2,600*	44.03	25.36	13.03	12.82	16.73	22.39
Allophant [48]	4,628	35.04	24.13	11.91	14.03	18.11	20.64
Wav2Vec2Phoneme [20]	5,300	27.50	18.57	8.77	**9.59**	14.62	15.81
MultIPA [51]	3,600	18.69	23.31	15.52	15.64	21.34	18.90
ZIPA-CR-Large [52]	17,132	31.24	17.89	9.74	11.75	15.58	17.24
ZIPA-CR-NS-Large [52]	28,983	31.05	17.12	8.54	11.63	18.20	17.31
POWSM [10]	17,132	31.63	18.33	11.32	11.86	17.84	18.20
HuPER-Recognizer (ours)	**100**	**7.36**	**9.08**	**8.00**	10.66	**9.00**	**8.82**

English variation. We evaluate on Buckeye [53], which contains spontaneous conversational English. To probe dialectal variation, we additionally test on DRC-SE (DoReCo South-England), a dialect subset from DoReCo [54]. To measure robustness to non-native pronunciations, we evaluate on L2-ARCTIC [55], EpaDB [56], and SpeechOcean762 [57], which contain L2 English speech with verified phone-level annotations. For L2-ARCTIC, we use the manually annotated *perceived* transcriptions rather than dictionary/G2P pronunciations, so the reference reflects what speakers actually produced.

Unseen languages (zero-shot). We further test multilingual transfer on VoxAngeles [58], a post-processed version of the UCLA Phonetics Lab Archive with human-verified transcriptions spanning 94 languages. This setting evaluates zero-shot generalization to languages and phone inventories never seen during training.

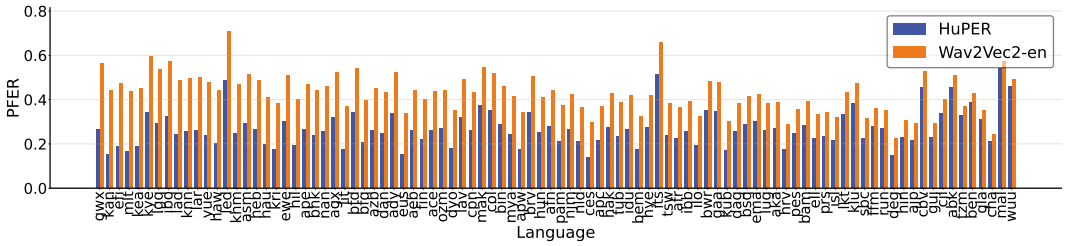


Figure 4: **Zero-shot multilingual phone recognition on VoxAngeles (94 languages).** Per-language phone error rate (PFER fraction, ↓) for HuPER-Recognizer (trained only on English) and an English-only public phoneme model `Wav2Vec2-en` (fine-tuned on G2P-generated phoneme labels from LJSpeech). HuPER improves on the majority of languages and reduces the macro-average PFER fraction from 0.35 to 0.19.

6.2 Multi-path speech perception

Dataset. For both **Task-2** and **Task-3**, we evaluate on a primary progressive aphasia (PPA) reading dataset [59] (`nfvPPA`, 35 speakers). We manually annotated *spoken* phone sequences (1h1m12s total). Each utterance provides audio X, a passage text R (available but optionally hidden), and a human phone reference Y.

Experimental Configuration. For **Task-3**, we follow the Dysfluent-WFST pipeline [11], replacing the original phonetic encoder with our HuPER-Recognizer. For **Task-2** with dynamic planning, we first compute a distortion score $s(X)$ from HuPER-Recognizer posteriors (Sec. 5) to characterize recognition difficulty. We then analyze whether $s(X)$ correlates with recognition performance and whether hypothesis-guided constrained decoding provides greater benefits than bottom-up 1-best decoding under high distortion. For both tasks, we treat the switching threshold τ as an externally specified hyperparameter and report results across a sweep of τ. The distortion-based configuration for **Task-2** is presented as follows:

Distortion Computation (Task-2). For each utterance, we compute an utterance-level distortion score $s(X)$ by aggregating the frame-level proxy $\{d_t\}_{t=1}^{T}$ defined in Sec. 5. We compare two phonetic encoders (HuPER and Wav2Vec2Phoneme [20]) under three decoding modes: (i) **1-best**: CTC decoding from phone evidence $\Pi(X)$; (ii) **refine**: constrained decoding with a hypothesis-conditioned WFST $\mathcal{H}(\hat{T}(X))$ [11] built from Perceiver hypotheses. We report Spearman's rank correlation coefficient ρ between $s(X)$ and PFER across PPA utterances to validate $s(X)$ as an uncertainty proxy (Fig. 5a).

Dynamic Switching Analysis (Task-2). To quantify when constrained inference is beneficial under dynamic planning, we define a distortion-threshold switch between HuPER 1-best and HuPER refine decoding:

$$\hat{Y}_\tau(X) = \begin{cases} \hat{Y}_{\text{1best}}(X), & s(X) \le \tau, \\ \hat{Y}_{\text{refine}}(X), & s(X) > \tau, \end{cases} \tag{9}$$

We report the average PFER as a function of τ to identify distortion regimes in which constrained decoding yields the largest gains.

7 Results

7.1 English phone recognition (Task-1)

HuPER-Recognizer achieves the best average PFER with orders-of-magnitude less supervision and a more compact label space, demonstrating the effectiveness and transferability of our DRRC-based acoustic–phonetic learning. Table 1 reports PFER on five corpora with human-annotated (or human-verified) phone labels. HuPER-Recognizer achieves the **best average** PFER (8.82%) while using only 100 hours of English training data, whereas all compared baselines rely on orders-of-magnitude more supervision. This result highlights that our DRRC recipe (Sec. 3) can

Table 2: **Weak-evidence speech perception on `nfvPPA`.** Overall PFER for different decoding modes; HuPER-switched applies refinement only when $s(X) > \tau^\star$ (refine rate shown).

Method	Overall PFER ↓	Refine rate
Wav2Vec2Phoneme 1-best	0.46	–
HuPER-recognizer 1-best	0.44	–
HuPER switched ($\tau^\star$)	**0.38**	31.1%
Dysfluent WFST	0.35	100%
HuPER + reference	**0.32**	100%

turn a small amount of human phone annotation into strong, transferable acoustic–phonetic evidence. We note that HuPER uses a compact phone inventory with only 42 symbols, which is smaller than the vocabularies used by several baselines. A smaller inventory reduces representational resolution and can make certain fine-grained contrasts unexpressible, which may disadvantage HuPER under PFER (a feature-sensitive edit distance). Therefore, the average-best performance is achieved despite a conservative label space.

7.2 Zero-shot multilingual transfer (Task-1)

HuPER-Recognizer matches strong multilingual baselines in strict zero-shot cross-lingual transfer, despite being trained only on English. We evaluate cross-lingual generalization on VoxAngeles (94 languages) in a strict zero-shot setting. Trained only on English, HuPER-Recognizer achieves a macro-average PFER fraction of 0.19. In contrast, a public English-only phoneme model `Wav2Vec2-en`, fine-tuned on G2P-generated phoneme labels from LJSpeech, yields a much higher PFER fraction of 0.35. Figure 4 reports per-language results: HuPER improves over `Wav2Vec2-en` on the majority of languages, indicating robust transfer to unseen phone inventories. Moreover, HuPER's macro-average PFER fraction of 0.19 is strongly competitive with the best multilingual baselines, despite HuPER being trained only on English. Full per-language results are provided in Table 8.

7.3 Flexible speech perception (Task-2, Task-3)

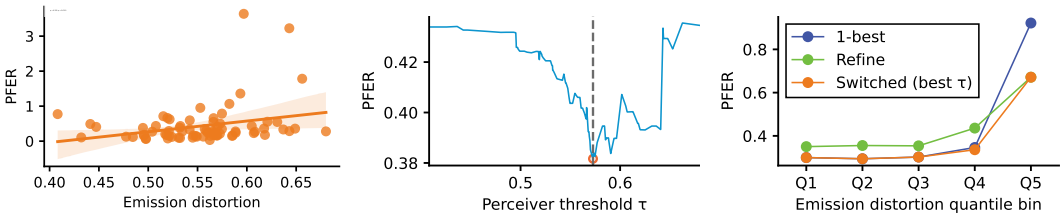
Task-2: HuPER achieves substantial gains under weak acoustic evidence through distortion-guided activation of top-down constraints. On `nfvPPA`, the distortion score correlates positively with HuPER 1-best PFER (Figure 5a), validating it as an uncertainty proxy. While HuPER 1-best already improves over Wav2Vec2Phoneme, distortion-controlled switching yields the main gains, substantially reducing overall PFER (Table 2). Threshold analysis and distortion-bin breakdown (Figures 5b, 5c) show that improvements concentrate in high-distortion segments, where constrained decoding is most effective. Failure cases are discussed in Appendix F.

Task-3: HuPER-Recognizer consistently outperforms existing phonetic encoders on disordered speech, indicating more robust acoustic–phonetic representations. With reference text, constrained decoding further improves robustness, and HuPER + reference achieves the best overall PFER on `nfvPPA` (Table 2), outperforming the Dysfluent-WFST baseline.

8 Understanding HuPER-Recognizer Gains

8.1 Embedding analysis: acoustic–phonetic geometry

HuPER maintains stronger alignment to distinctive-feature geometry in mid/late layers than an English G2P-supervised baseline. We measure layer-wise centroid RSA between phone representations and PanPhon distinctive-feature distances [46] on TIMIT. Figure 6 shows that all models peak in early layers (around layer 2–4), but **HuPER remains more aligned than *WavLM-Libri* deeper in the network**, suggesting DRRC-style acoustic correction helps preserve realized-phone structure beyond shallow acoustic features. Full RSA setup is in Appendix G.1, and additional acoustic/articulatory reference analyses are in Appendix D.



(a) 1-best PFER vs. distortion. (b) Threshold sweep for switching. (c) Refine helps in high-distortion bins.

Figure 5: **Distortion as a control signal for refinement on PPA. (a)** Using HuPER 1-best throughout, emission distortion is positively correlated with PFER, indicating it tracks weak-evidence difficulty. **(b)** Sweeping the switching threshold τ reveals an optimal region for triggering refinement: we output 1-best when $s(X) \leq \tau$, and otherwise invoke the perceiver to decode words from phone evidence and refine the phone sequence. **(c)** Stratifying utterances by distortion explains why switching helps: refinement yields the largest gains in high-distortion regimes, while low-distortion cases are best handled by direct decoding.

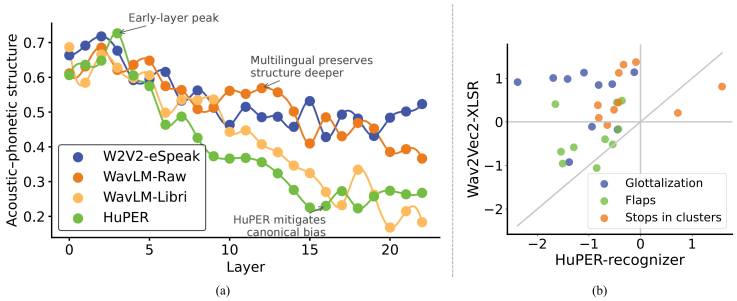


Figure 6: (a) **Centroid RSA to acoustic–phonetic geometry.** Spearman correlation between pairwise phone-centroid cosine distances and PanPhon distinctive-feature distances across layers. Higher values indicate stronger acoustic–phonetic organization. (b) **Emission-level canonical-restoration diagnostic.** Scatter of canonical-vs-realized preference scores on identical utterances: HuPER (x-axis) vs. XLSR (y-axis), colored by category. Points above the diagonal ($y = x$) indicate stronger canonical restoration by the baseline. In panel (a), the label *W2V2-eSpeak* denotes the *Wav2Vec2Phoneme* checkpoint listed in Table 9.

8.2 Emission analysis: canonical restoration

HuPER's emissions consistently favor the realized sequence on the same acoustics, indicating less emission-level canonical "auto-correction" than an XLSR baseline. We evaluate controlled contrasts in three settings where canonical restoration is tempting: **glottalization**, **flaps**, and **stop reductions in clusters**. For each waveform, we compare CTC evidence for a canonical G2P sequence versus a manually verified realized sequence using a length-normalized preference score. Figure 6 shows that the XLSR baseline is more often canonical-restoring (especially for glottalization and cluster reductions), while **HuPER remains more acoustic-faithful**. This matches our design goal: bottom-up emissions should track available cues, and any canonical restoration should be applied explicitly by higher-level constraints when needed. Diagnostic-set construction, the exact score definition, and the full text/label pairs are provided in Appendix G.2.

9 Conclusion and Limitations

In this work, we view phonetic perception as an adaptive and controllable inference process and introduce HuPER as an initial step toward explicit, human-centered modeling. Our results highlight the value of combining phonetic representations, structured constraints, and dynamic control for robust and interpretable perception. However, the current system is still limited by the small amount of human-verified spoken-phone labels, relies on heuristic and rule-based routing, and models only a small subset of the inference pathways involved in real human perception. Addressing these

limitations with larger annotated datasets, learned control policies, and more expressive perceptual models is an important direction for future work.

Acknowledgments and Disclosure of Funding

We thank the UC Noyce Initiative, the Society of Hellman Fellows, NIH/NIDCD, and the Schwab Innovation Fund for their support. We thank the UCSF team for providing access to part of the clinical dataset used in this study.

References

[1] Davis, K. H., R. Biddulph, S. Balashek. Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, 24(6):637–642, 1952.

[2] Erman, L. D., F. Hayes-Roth, V. R. Lesser, et al. The hearsay-ii speech-understanding system: Integrating knowledge to resolve uncertainty. *ACM Computing Surveys (CSUR)*, 12(2):213–253, 1980.

[3] Lowerre, B. T. *The harpy speech recognition system.* Carnegie Mellon University, 1976.

[4] Baker, J. The dragon system–an overview. *IEEE Transactions on Acoustics, speech, and signal Processing*, 23(1):24–29, 1975.

[5] Rabiner, L. R. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 2002.

[6] Radford, A., J. W. Kim, T. Xu, et al. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[7] Zhang, Y., W. Han, J. Qin, et al. Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*, 2023.

[8] Pratap, V., A. Tjandra, B. Shi, et al. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52, 2024.

[9] Omnilingual, A., G. Keren, A. Kozhevnikov, et al. Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages. *arXiv preprint arXiv:2511.09690*, 2025.

[10] Li, C.-J., K. Chang, S. Bharadwaj, et al. Powsm: A phonetic open whisper-style speech foundation model. *arXiv preprint arXiv:2510.24992*, 2025.

[11] Guo, C., J. Lian, X. Zhou, et al. Dysfluent wfst: A framework for zero-shot speech dysfluency transcription and detection. In O. Scharenborg, C. Oertel, K. Truong, eds., *INTERSPEECH*. ISCA, 2025.

[12] Stevens, K. N. *Acoustic phonetics*, vol. 30. MIT press, 2000.

[13] Mesgarani, N., C. Cheung, K. Johnson, et al. Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174):1006–1010, 2014.

[14] Johnson, K. *Acoustic and auditory phonetics*. John Wiley & Sons, 2011.

[15] Werker, J. F., C. E. Lalonde. Cross-language speech perception: Initial capabilities and developmental change. *Developmental psychology*, 24(5):672, 1988.

[16] Bhaya-Grossman, I., M. K. Leonard, Y. Zhang, et al. Shared and language-specific phonological processing in the human temporal lobe. *Nature*, 649(8095):140–151, 2026.

[17] Mohamed, A., H.-y. Lee, L. Borgholt, et al. Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1179–1210, 2022.

[18] Lavechin, M., Y. Sy, H. Titeux, et al. Babyslm: language-acquisition-friendly benchmark of self-supervised spoken language models. In *Interspeech 2023*, pages 4588–4592. 2023.

[19] Choi, K., A. Pasad, T. Nakamura, et al. Self-Supervised Speech Representations are More Phonetic than Semantic. In *Interspeech 2024*, pages 4578–4582. 2024.

[20] Xu, Q., A. Baevski, M. Auli. Simple and effective zero-shot cross-lingual phoneme recognition. In H. Ko, J. H. L. Hansen, eds., *INTERSPEECH*, pages 2113–2117. ISCA, 2022.

[21] Chen, S., C. Wang, Z. Chen, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. 2021.

[22] Rudnicky, A. The cmu pronouncing dictionary, 1993. Accessed October 2, 2025.

[23] Mortensen, D. R., S. Dalmia, P. Littell. Epitran: Precision g2p for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.

[24] Black, A. W., K. A. Lenzo. Flite: a small fast run-time synthesis engine. In *SSW*, page 204. 2001.

[25] McClelland, J. L., J. L. Elman. The trace model of speech perception. *Cognitive psychology*, 18(1):1–86, 1986.

[26] Hickok, G., D. Poeppel. The cortical organization of speech processing. *Nature reviews neuroscience*, 8(5):393–402, 2007.

[27] Norris, D., J. M. McQueen, A. Cutler. Perceptual learning in speech. *Cognitive Psychology*, 47(2):204–238, 2003.

[28] Warren, R. M. Perceptual restoration of missing speech sounds. *Science*, 167(3917):392–393, 1970.

[29] Ganong, W. F. Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, 6(1):110–125, 1980.

[30] Samuel, A. G. Knowing a word affects the fundamental perception of the sounds within it. *Psychological Science*, 12(4):348–351, 2001.

[31] Cutler, A. *Native listening: Language experience and the recognition of spoken words*. Mit Press, 2012.

[32] Posner, M. I., S. E. Petersen. The attention system of the human brain. 1989.

[33] Norman, D. A., T. Shallice. Attention to action: Willed and automatic control of behavior. In *Consciousness and self-regulation: Advances in research and theory volume 4*, pages 1–18. Springer, 1986.

[34] Lian, J., X. Zhou, Z. Ezzes, et al. Ssdm: Scalable speech dysfluency modeling. *Advances in neural information processing systems*, 37:101818–101855, 2024.

[35] Lee, D.-H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop on Challenges in Representation Learning*. 2013.

[36] Xie, Q., M.-T. Luong, E. Hovy, et al. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

[37] Robins, J. M., A. Rotnitzky, L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

[38] Bang, H., J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[39] Mohri, M., F. Pereira, M. Riley. Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, 16(1):69–88, 2002.

[40] Graves, A., S. Fernández, F. Gomez, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376. 2006.

[41] Garofolo, J. S., N. I. of Standards, T. U.S., et al. TIMIT : acoustic-phonetic continuous speech corpus., 1993.

[42] Panayotov, V., G. Chen, D. Povey, et al. Librispeech: an asr corpus based on public domain audio books. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5206–5210. IEEE, 2015.

[43] Zhu, J., C. Zhang, D. Jurgens. Byt5 model for massively multilingual grapheme-to-phoneme conversion. In H. Ko, J. H. L. Hansen, eds., *INTERSPEECH*, pages 446–450. ISCA, 2022.

[44] Hsu, W.-N., B. Bolte, Y.-H. H. Tsai, et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

[45] Loshchilov, I., F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[46] Mortensen, D. R., P. Littell, A. Bharadwaj, et al. Panphon: A resource for mapping ipa segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484. 2016.

[47] Li, X., S. Dalmia, J. Li, et al. Universal phone recognition with a multilingual allophone system. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8249–8253. IEEE, 2020.

[48] Glocker, K., A. Herygers, M. Georges. Allophant: Cross-lingual phoneme recognition with articulatory attributes. In *Interspeech 2023*, pages 2258–2262. 2023.

[49] Baevski, A., Y. Zhou, A. Mohamed, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[50] Conneau, A., A. Baevski, R. Collobert, et al. Unsupervised cross-lingual representation learning for speech recognition. In H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, P. Motlícek, eds., *Interspeech*, pages 2426–2430. ISCA, 2021.

[51] Chen, Y.-W., Z. Yu, J. Hirschberg. MultiPA: A Multi-task Speech Pronunciation Assessment Model for Open Response Scenarios. In *Interspeech 2024*, pages 297–301. 2024.

[52] Zhu, J., F. Samir, E. Chodroff, et al. Zipa: A family of efficient models for multilingual phone recognition. In W. Che, J. Nabende, E. Shutova, M. T. Pilehvar, eds., *ACL (1)*, pages 19568–19585. Association for Computational Linguistics, 2025.

[53] Pitt, M. A., K. Johnson, E. Hume, et al. The buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95, 2005.

[54] Paschen, L., et al. Building a time-aligned cross-linguistic reference corpus from language documentation data (doreco). In *Proc. LREC*. 2020.

[55] Zhao, G., S. Sonsaat, A. Silpachai, et al. L2-arctic: A non-native english speech corpus. In *Proc. Interspeech*. 2018.

[56] Vidal, J., L. Ferrer, L. Brambilla. A database for development of pronunciation assessment systems for english learners. In *Proc. Interspeech*. 2019.

[57] Zhang, J., Z. Zhang, Y. Wang, et al. Speechocean762: An open-source non-native english speech corpus for pronunciation assessment. In *Proc. Interspeech*. 2021.

[58] Chodroff, E., B. Pazon, A. Baker, et al. Phonetic segmentation of the ucla phonetics lab archive. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue, eds., *LREC/COLING*, pages 12724–12733. ELRA and ICCL, 2024.

[59] Gorno-Tempini, M. L., A. E. Hillis, S. Weintraub, et al. Classification of primary progressive aphasia and its variants. *Neurology*, 76(11):1006–1014, 2011.

[60] Mark, G., Y. Steve. The application of hidden markov models in speech recognition. *Foundations and Trends® in Signal Processing*, 1(3):195–304, 2024.

[61] Hinton, G., L. Deng, D. Yu, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

[62] Liao, S.-H. Expert system methodologies and applications—a decade review from 1995 to 2004. *Expert systems with applications*, 28(1):93–103, 2005.

[63] Angelaki, D., B. Benson, J. Benson, et al. A brain-wide map of neural activity during complex behaviour. *Nature*, 645(8079):177–191, 2025.

[64] AlKhamissi, B., C. N. De Sabbata, G. Tuckute, et al. Mixture of cognitive reasoners: Modular reasoning with brain-like specialization. *arXiv preprint arXiv:2506.13331*, 2025.

[65] Huang, R., M. Li, D. Yang, et al. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pages 23802–23804. 2024.

[66] Liu, X., Z. Zhu, H. Liu, et al. Wavjourney: Compositional audio creation with large language models. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[67] Zhou, X., J. Lian, H. Hong, et al. Speech world model: Causal state-action planning with explicit reasoning for speech. *International Conference on Learning Representations*, 2026.

[68] Kang, J. D., J. L. Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. 2007.

[69] Wager, S., S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[70] Angelopoulos, A. N., S. Bates, C. Fannjiang, et al. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.

A Related Work

Phonetic recognition models Traditional phonetic models have been closely co-developed with ASR paradigms and are primarily data-driven, relying on lexicon-derived targets and large-scale supervision, evolving from HMM–GMM [5, 60] systems to hybrid HMM–DNN [61] and end-to-end architectures [40]. Such systems [47, 48] assign canonical phonemes rather than realized phones, and progress in phone recognition has therefore depended heavily on costly manual annotations. In parallel, linguistically driven approaches incorporate phonological and articulatory priors [46], such as distinctive features and allophonic modeling, to improve generalization and cross-lingual transfer, but rely on curated linguistic resources and handcrafted representations, limiting their scalability. From a cognitive perspective, phoneme perception itself is a multi-path, closed-loop process that integrates bottom-up acoustic evidence with top-down linguistic and contextual expectations [25, 27]. To date, these three paradigms remain largely disconnected.

Explicit modeling of human-inspired speech perception Early expert systems [62] attempted to emulate human cognition through rigid pipelines, which were later shown not to be faithful proxies [63], as cognition emerges from dynamic interactions among functional modules. While cognitive language models, such as [64], introduce brain-inspired modular specialization, these designs remain domain-specific and are not tailored to speech perception. In the context of speech modeling, audio agents [65, 66] focus on high-level task orchestration and act mainly as wrappers around existing models, without explicitly modeling perceptual mechanisms. Speech world models [67] represent early attempts to modularize speech perception, but operate primarily at the prompt level rather than directly decomposing perceptual representations.

B Doubly Robust Consistency of HuPER Corrector

B.1 Setup and notation

Fix $K \geq 2$. Let $W = (X, Z, Y, \hat{Y}, A)$ be a generic draw, where: (i) X are features used by a predictive model $p_\theta(\cdot \mid X)$, (ii) Z are covariates governing label missingness, (iii) $Y \in \{1, \ldots, K\}$ is the true class label, (iv) $\hat{Y} \in \{1, \ldots, K\}$ is an always-observed proxy label, (v) $A \in \{0, 1\}$ indicates whether Y is observed (so AY is observed, but Y may be missing when $A = 0$). Let $e(k) \in \mathbb{R}^K$ denote the one-hot vector with a 1 in coordinate k. Write $e(Y)$ and $e(\hat{Y})$ for the corresponding (random) one-hot vectors of Y and $\hat{Y}$. Define the multiclass log-loss for any label vector $q \in \mathbb{R}^K$ by

$$\ell_\theta(q, x) := -\sum_{k=1}^{K} q_k \log p_\theta(k \mid x).$$

In particular, $\ell_\theta(e(Y), X) = -\log p_\theta(Y \mid X)$. The target population risk is

$$R(\theta) := \mathbb{E}\big[\ell_\theta(e(Y), X)\big].$$

Define the score vector $s_\theta(X) \in \mathbb{R}^K$ by $s_{\theta,k}(X) := -\log p_\theta(k \mid X)$, so that

$$\ell_\theta(q, X) = q^\top s_\theta(X).$$

Missingness model and nuisance functions. Define the true propensity (missingness) function

$$g^*(z, \hat{y}) := \mathbb{P}(A = 1 \mid Z = z, \hat{Y} = \hat{y}).$$

Define also the conditional class distribution given the always-observed variables $V := (X, Z, \hat{Y})$:

$$f^*(x, z, \hat{y}) := \mathbb{E}\big[e(Y) \mid X = x, Z = z, \hat{Y} = \hat{y}\big] \in \Delta_K,$$

where $\Delta_K := \big\{q \in \mathbb{R}^K : q_k \geq 0,\ \sum_{k=1}^{K} q_k = 1\big\}$.

Proxy-baseline AIPW corrector. For any measurable $g : \mathcal{Z} \times \{1, \ldots, K\} \to (0, 1]$, define

$$C_g(W) := e(\hat{Y}) + \frac{A}{g(Z, \hat{Y})}\big(e(Y) - e(\hat{Y})\big).$$

Note that $C_g(W)$ need not lie in Δ_K (it can have negative entries when $A = 1$ and $g(Z, \hat{Y}) < 1$), but $\ell_\theta(\cdot, X)$ is linear in its first argument, so $\ell_\theta(C_g(W), X)$ is well-defined.

Cross-fitted risk estimator. Let $W_1, \dots, W_n$ be i.i.d. copies of W. Split $\{1, \dots, n\}$ into $J \geq 2$ folds $I_1, \dots, I_J$ with $|I_j| \to \infty$ and J fixed. For each fold j, fit an estimator $\hat{g}^{(-j)}$ using only data in $\{W_i : i \notin I_j\}$, and enforce $\hat{g}^{(-j)}(z, \hat{y}) \in [\varepsilon, 1]$ for some fixed $\varepsilon \in (0, 1)$ (e.g. by clipping). For each $i \in I_j$, set $\hat{g}_i := \hat{g}^{(-j)}$ and $C_i := C_{\hat{g}_i}(W_i)$. Define the empirical risk

$$\hat{R}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell_\theta(C_i, X_i).$$

B.2 Main result

We will work on the following assumptions standard in semi-parametric statistics [68–70].

Assumption B.1 (MAR depending only on $(Z, \hat{Y})$)**.** $\mathbb{P}(A = 1 \mid X, Z, Y, \hat{Y}) = g^*(Z, \hat{Y})$ almost surely.

Assumption B.2 (Positivity)**.** There exists $\varepsilon > 0$ such that $g^*(Z, \hat{Y}) \geq \varepsilon$ almost surely.

Assumption B.3 (Bounded log score at fixed θ)**.** There exists $M_\theta < \infty$ such that $\max_{1 \leq k \leq K} |s_{\theta,k}(X)| \leq M_\theta$ almost surely. Equivalently, $\max_{1 \leq k \leq K} |\log p_\theta(k \mid X)| \leq M_\theta$ almost surely.

Assumption B.4 (Cross-fitted range constraint)**.** For each fold j, the fitted $\hat{g}^{(-j)}$ satisfies $\hat{g}^{(-j)}(z, \hat{y}) \in [\varepsilon, 1]$ for all $(z, \hat{y})$.

We will show consistency under either of the following alternative conditions:

$$\text{(Prop) Propensity consistency:} \quad \mathbb{E}\big[|\hat{g}^{(-j)}(Z, \hat{Y}) - g^*(Z, \hat{Y})|\big] \to 0 \text{ for each } j,$$
$$\text{(Proxy) Proxy label consistency:} \quad \mathbb{P}(\hat{Y} \neq Y) \to 0.$$

Theorem B.5 (Proxy-baseline AIPW corrector is doubly robust)**.** *Assume B.1–B.4 and fix θ.*

1. Exact population bias identity (double robustness). *For any measurable $g : \mathcal{Z} \times \{1, \dots, K\} \to [\varepsilon, 1]$,*

$$\mathbb{E}\big[\ell_\theta(C_g(W), X)\big] - R(\theta) = \mathbb{E}\left[\frac{g(Z, \hat{Y}) - g^*(Z, \hat{Y})}{g(Z, \hat{Y})} \big(e(\hat{Y}) - f^*(X, Z, \hat{Y})\big)^\top s_\theta(X)\right].$$

In particular:

- *If $g = g^*$ almost surely, then $\mathbb{E}[\ell_\theta(C_g(W), X)] = R(\theta)$ exactly.*
- *If $\hat{Y} = Y$ almost surely, then $f^*(X, Z, \hat{Y}) = e(\hat{Y})$ almost surely and hence $\mathbb{E}[\ell_\theta(C_g(W), X)] = R(\theta)$ for any g bounded away from 0.*

2. Doubly robust consistency of the cross-fitted risk estimator. *The cross-fitted estimator $\hat{R}_n(\theta)$ satisfies*

$$\hat{R}_n(\theta) - R(\theta) \to 0 \quad \textit{in probability as } n \to \infty,$$

provided either (Prop) holds or (Proxy) holds.

Proof. By definition of $s_\theta(X)$, we write

$$\ell_\theta(q, X) = q^\top s_\theta(X) \quad \text{for all } q \in \mathbb{R}^K.$$

Using the definition of $C_g(W)$ and the above, we have

$$\ell_\theta(C_g(W), X) = \ell_\theta(e(\hat{Y}), X) + \frac{A}{g(Z, \hat{Y})} \Big(\ell_\theta(e(Y), X) - \ell_\theta(e(\hat{Y}), X)\Big). \tag{10}$$

Let $V := (X, Z, \hat{Y})$ and define

$$\Delta := \ell_\theta(e(Y), X) - \ell_\theta(e(\hat{Y}), X).$$

Under Assumption B.1,

$$\mathbb{E}[A \mid V, Y] = \mathbb{E}[A \mid X, Z, Y, \hat{Y}] = g^*(Z, \hat{Y}),$$

which implies A is conditionally independent of Y given V. Therefore, by iterated expectations,

$$\mathbb{E}[A\Delta \mid V] = \mathbb{E}\big[\mathbb{E}[A\Delta \mid V, Y] \mid V\big] = \mathbb{E}\big[\mathbb{E}[A \mid V, Y]\Delta \mid V\big] = g^*(Z, \hat{Y})\mathbb{E}[\Delta \mid V].$$

Hence,

$$\mathbb{E}\left[\frac{A}{g(Z, \hat{Y})}\Delta \,\middle|\, V\right] = \frac{g^*(Z, \hat{Y})}{g(Z, \hat{Y})}\mathbb{E}[\Delta \mid V]. \tag{11}$$

Since $s_\theta(X)$ is V-measurable (it depends only on X),

$$\mathbb{E}\big[\ell_\theta(e(Y), X) \mid V\big] = \mathbb{E}\big[e(Y)^\top s_\theta(X) \mid V\big] = \mathbb{E}[e(Y) \mid V]^\top s_\theta(X) = f^*(V)^\top s_\theta(X) = \ell_\theta(f^*(V), X).$$

Consequently,

$$R(\theta) = \mathbb{E}\big[\ell_\theta(e(Y), X)\big] = \mathbb{E}\big[\mathbb{E}[\ell_\theta(e(Y), X) \mid V]\big] = \mathbb{E}\big[\ell_\theta(f^*(V), X)\big]. \tag{12}$$

Take conditional expectations of the expansion in Eq. (10) given V and use Eq. (11):

$$\mathbb{E}[\ell_\theta(C_g(W), X) \mid V] = \ell_\theta(e(\hat{Y}), X) + \frac{g^*(Z, \hat{Y})}{g(Z, \hat{Y})}\Big(\mathbb{E}[\ell_\theta(e(Y), X) \mid V] - \ell_\theta(e(\hat{Y}), X)\Big).$$

Using Eq. (12) to substitute $\mathbb{E}[\ell_\theta(e(Y), X) \mid V] = \ell_\theta(f^*(V), X)$ gives

$$\begin{aligned}\mathbb{E}[\ell_\theta(C_g(W), X) \mid V] &= \ell_\theta(e(\hat{Y}), X) + \frac{g^*(Z, \hat{Y})}{g(Z, \hat{Y})}\Big(\ell_\theta(f^*(V), X) - \ell_\theta(e(\hat{Y}), X)\Big) \\ &= \ell_\theta(f^*(V), X) + \left(1 - \frac{g^*(Z, \hat{Y})}{g(Z, \hat{Y})}\right)\Big(\ell_\theta(e(\hat{Y}), X) - \ell_\theta(f^*(V), X)\Big).\end{aligned}$$

By linearity,

$$\ell_\theta(e(\hat{Y}), X) - \ell_\theta(f^*(V), X) = \big(e(\hat{Y}) - f^*(V)\big)^\top s_\theta(X).$$

Therefore,

$$\mathbb{E}[\ell_\theta(C_g(W), X) \mid V] - \ell_\theta(f^*(V), X) = \frac{g(Z, \hat{Y}) - g^*(Z, \hat{Y})}{g(Z, \hat{Y})}\big(e(\hat{Y}) - f^*(V)\big)^\top s_\theta(X). \tag{13}$$

Taking unconditional expectations and using $R(\theta) = \mathbb{E}[\ell_\theta(f^*(V), X)]$ yields the stated bias identity.

The first claim follows immediately: if $g = g^*$ then the multiplicative factor is 0 almost surely; if $\hat{Y} = Y$ almost surely then Y is V-measurable and hence $f^*(V) = \mathbb{E}[e(Y) \mid V] = e(\hat{Y})$ almost surely.

Under Assumption B.4, $g(Z, \hat{Y}) \geq \varepsilon$. Since $\|e(Y)\|_1 = 1$ and $\|e(Y) - e(\hat{Y})\|_1 \leq 2$,

$$\|C_g(W)\|_1 \leq \|e(\hat{Y})\|_1 + \frac{A}{g(Z, \hat{Y})}\|e(Y) - e(\hat{Y})\|_1 \leq 1 + \frac{2}{\varepsilon}.$$

By Assumption B.3, $\|s_\theta(X)\|_\infty \leq M_\theta$ almost surely, so

$$|\ell_\theta(C_g(W), X)| = |C_g(W)^\top s_\theta(X)| \leq \|C_g(W)\|_1 \|s_\theta(X)\|_\infty \leq \left(1 + \frac{2}{\varepsilon}\right) M_\theta. \tag{14}$$

Fix a fold j and write the fold average

$$\hat{R}_{n,j}(\theta) := \frac{1}{|I_j|}\sum_{i \in I_j} \ell_\theta(C_{\hat{g}^{(-j)}}(W_i), X_i).$$

Condition on the training data used to fit $\hat{g}^{(-j)}$ (i.e. on $\{W_i : i \notin I_j\}$). By cross-fitting, the summands for $i \in I_j$ are i.i.d. given the training data, and bounded by Eq. (14). Therefore, Chebyshev's inequality (conditional on the training data) yields

$$\hat{R}_{n,j}(\theta) - \mathbb{E}\big[\ell_\theta(C_{\hat{g}^{(-j)}}(W), X) \mid \{W_i : i \notin I_j\}\big] \to 0 \quad \text{in probability.}$$

Averaging over $j = 1, \ldots, J$ and using that J is fixed gives

$$\hat{R}_n(\theta) - \bar{\mu}_n(\theta) \to 0 \quad \text{in probability,} \tag{15}$$

where

$$\bar{\mu}_n(\theta) := \sum_{j=1}^{J} \frac{|I_j|}{n} \mathbb{E}\big[\ell_\theta(C_{\hat{g}^{(-j)}}(W), X) \mid \{W_i : i \notin I_j\}\big].$$

Fix j. Conditional on the training data, $\hat{g}^{(-j)}$ is deterministic and belongs to $[\varepsilon, 1]$. Apply the population identity from Eq. (13) with $g = \hat{g}^{(-j)}$ and then take absolute values:

$$\begin{aligned}
&\Big|\mathbb{E}\big[\ell_\theta(C_{\hat{g}^{(-j)}}(W), X) \mid \{W_i : i \notin I_j\}\big] - R(\theta)\Big| \\
&\le \mathbb{E}\left[\left|\frac{\hat{g}^{(-j)}(Z, \hat{Y}) - g^*(Z, \hat{Y})}{\hat{g}^{(-j)}(Z, \hat{Y})}\right| \Big|\big(e(\hat{Y}) - f^*(V)\big)^\top s_\theta(X)\Big| \;\Big|\; \{W_i : i \notin I_j\}\right] \\
&\le \frac{M_\theta}{\varepsilon} \mathbb{E}\left[|\hat{g}^{(-j)}(Z, \hat{Y}) - g^*(Z, \hat{Y})| \|e(\hat{Y}) - f^*(V)\|_1 \;\Big|\; \{W_i : i \notin I_j\}\right],
\end{aligned}$$

using $\|s_\theta(X)\|_\infty \le M_\theta$ and $\hat{g}^{(-j)} \ge \varepsilon$. Since $\|e(\hat{Y}) - f^*(V)\|_1 \le 2$, we further have

$$\Big|\mathbb{E}\big[\ell_\theta(C_{\hat{g}^{(-j)}}(W), X) \mid \{W_i : i \notin I_j\}\big] - R(\theta)\Big| \le \frac{2M_\theta}{\varepsilon} \mathbb{E}\left[|\hat{g}^{(-j)}(Z, \hat{Y}) - g^*(Z, \hat{Y})| \;\Big|\; \{W_i : i \notin I_j\}\right].$$

Thus, under condition (Prop) we get the conditional mean converges to $R(\theta)$ in probability.

Under condition (Proxy), we instead use the *exact* identity

$$\mathbb{E}\big[\|e(\hat{Y}) - f^*(V)\|_1\big] = 2\mathbb{P}(\hat{Y} \neq Y),$$

due to the following: letting $j := \hat{Y}$ and writing $f_k^*(V) = \mathbb{P}(Y = k \mid V)$,

$$\begin{aligned}
\|e(\hat{Y}) - f^*(V)\|_1 &= |1 - f_j^*(V)| + \sum_{k \neq j} |0 - f_k^*(V)| = (1 - f_j^*(V)) + \sum_{k \neq j} f_k^*(V) = 2(1 - f_j^*(V)) \\
&= 2\big(1 - \mathbb{P}(Y = \hat{Y} \mid V)\big).
\end{aligned}$$

Taking expectations yields $2(1 - \mathbb{E}[\mathbb{P}(Y = \hat{Y} \mid V)]) = 2(1 - \mathbb{P}(Y = \hat{Y})) = 2\mathbb{P}(\hat{Y} \neq Y)$. Hence (Proxy) implies $\mathbb{E}[\|e(\hat{Y}) - f^*(V)\|_1] \to 0$. Returning to the bound, the same consistency statement follows due to $|\hat{g}^{(-j)} - g^*| \le 1$ and $\hat{g}^{(-j)} \ge \varepsilon$. Combining across folds leads to $\bar{\mu}_n(\theta) - R(\theta) \to 0$ in probability under either (Prop) or (Proxy).

With the above step and Eq. (15), we have

$$\hat{R}_n(\theta) - R(\theta) = \big(\hat{R}_n(\theta) - \bar{\mu}_n(\theta)\big) + \big(\bar{\mu}_n(\theta) - R(\theta)\big) \to 0 \quad \text{in probability,}$$

thus establishing the consistency claim. □

C Algorithm

Algorithm 1 HuPER self-learning strategy

Input: human-labeled phone dataset $\mathcal{D} = \{(X_i, Y_i)\}$; transcript-only dataset $\mathcal{D}' = \{(X_j, T_j)\}$; G2P function $\mathrm{G2P}(\cdot)$; Corrector $C(\cdot)$; number of rounds R
Output: HuPER-Recognizer $f^{(R)}$
Train initial recognizer $f^{(0)}$ on $\mathcal{D}$ with CTC loss.
for $r = 0$ **to** $R - 1$ **do**
  Initialize pseudo-labeled set $\tilde{\mathcal{D}}'^{(r)} = \emptyset$.
  for each (X, T) **in** $\mathcal{D}'$ **do**
    $Z \leftarrow \mathrm{G2P}(T)$ (canonical phoneme sequence)
    $\bar{Y}^{(r)} \leftarrow f^{(r)}(X)$ (teacher phone prediction)
    $\tilde{Y}^{(r)} \leftarrow C(X, Z; \bar{Y}^{(r)})$ (corrected pseudo phones; $\bar{Y}^{(r)}$ optional)
    $\tilde{\mathcal{D}}'^{(r)} \leftarrow \tilde{\mathcal{D}}'^{(r)} \cup \{(X, \tilde{Y}^{(r)})\}$
  end for
  Train $f^{(r+1)}$ on $\tilde{\mathcal{D}}'^{(r)}$ (optionally together with $\mathcal{D}$).
end for
return $f^{(R)}$

Algorithm 2 HuPER-Scheduler: distortion-controlled reference-constrained phone refinement

Input: utterance X; phone evidence graph $\Pi_\theta(X)$; lexicon L; LM G_{LM}; threshold τ; optional external reference R
Output: final phone sequence $\hat{Y}$
Compute $s(X)$ using Equation 8.
if $s(X) \leq \tau$ **then**
  $\hat{Y} \leftarrow \mathrm{Output}(\mathrm{ShortestPath}(\Pi_\theta(X)))$.
else
  if R is available **then**
    $U \leftarrow R$.
  else
    $\hat{T} \leftarrow \mathrm{Output}(\mathrm{ShortestPath}(\Pi_\theta(X) \circ L \circ G_{\mathrm{LM}}))$.
    $U \leftarrow \hat{T}$.
  end if
  Compile Dysfluent WFST constraint $\mathcal{H} \leftarrow \mathcal{H}(U)$.
  $\hat{Y} \leftarrow \mathrm{Output}(\mathrm{ShortestPath}(\Pi_\theta(X) \circ \mathcal{H}))$.
end if
return $\hat{Y}$

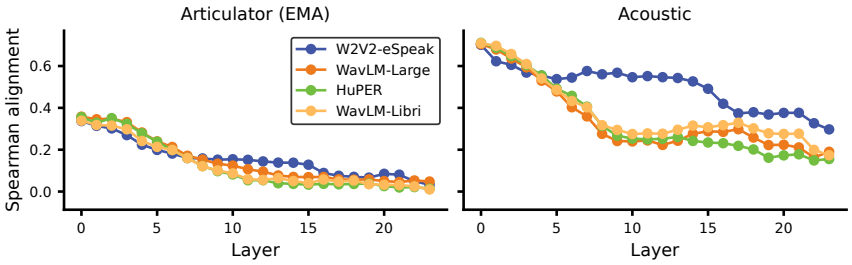
D Zero-shot alignment to acoustic and articulatory references

To complement Sec. 8.1, we run a segment-level, zero-shot reference-alignment analysis on MOCHA-TIMIT[1], which provides paired audio and EMA measurements. This analysis asks whether model embeddings preserve similarity structure induced by (i) raw acoustics and (ii) articulatory trajectories.

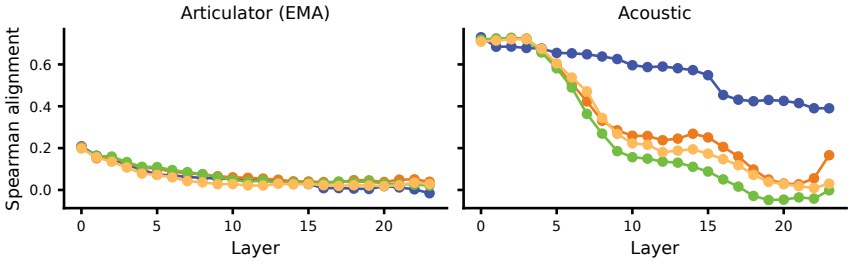
Reference spaces. For each phone segment, we compute an **acoustic** reference vector by averaging an 80-dim log-mel spectrogram over frames. We compute an **articulatory (EMA)** reference vector by averaging the 20-dim EMA coordinates within the segment (after utterance-level de-meaning and keeping frames marked as present). We ℓ_2-normalize reference vectors and use cosine distance.

Embedding distances and RSA. For each model and transformer layer, we mean-pool hidden states over frames inside each phone segment to obtain a segment embedding, then ℓ_2-normalize it

[1] `https://www.cstr.ed.ac.uk/research/projects/artic/mocha.html`



(a) **Within-phone (same phone).** Segment pairs are sampled from the same phone label. We report layer-wise Spearman correlation between embedding cosine distances and reference cosine distances (EMA vs. acoustic), after subtracting a shuffled-reference baseline.



(b) **Overall (mostly different phones).** Segment pairs are sampled uniformly from all segments (thus mainly different phone labels). Same metric as above.

Figure 7: **Zero-shot alignment to acoustic and articulatory references on MOCHA-TIMIT.** Articulatory (EMA) alignment is weak and not diagnostic, and raw acoustic alignment provides limited separation among models.

and compute cosine distances on sampled segment pairs. We report Spearman correlation between embedding distances and reference distances, and subtract a shuffled-reference correlation as a small random baseline (so values near zero indicate no reliable alignment).

Within-phone vs. overall pairs. We evaluate two pairing regimes: **within-phone** samples pairs of segments that share the *same* phone label (probing variability across different realizations of the same phone), while **overall** samples pairs uniformly from all segments (mostly *different* phones, probing global geometry).

Key observations. Figure 7a–7b supports two takeaways.

First, **articulatory (EMA) alignment is weak**. Across both within-phone and overall regimes, EMA alignment is modest in the earliest layers and quickly collapses toward zero in mid/late layers, with little separation between models. This indicates that, under a simple segment-mean EMA summary, embeddings do not preserve a stable articulatory similarity geometry across depth.

Second, **raw acoustic alignment is also not diagnostic for comparing models**. All models show similarly high early-layer alignment to log-mel similarity, and the curves provide limited separation among HuPER, WavLM-Raw, and WavLM-Libri (even when trends diverge later). As a result, a pure acoustic reference does not clearly reveal where HuPER's improvements come from. This motivates our main-paper focus on a more controlled *acoustic–phonetic* reference (PanPhon distinctive-feature geometry), which is less entangled with nuisance factors and yields clearer, more interpretable cross-model differences.

E Phonetic Feature Error Rate (PFER)

Following prior work, we use PFER [46], an articulatory-feature edit distance based on PanPhon distinctive features. Let $\mathrm{feat}(p) \in \{0,1\}^{24}$ denote the 24-dimensional binary distinctive-feature vector for phone p. For a substitution $p \to q$, the cost is the normalized Hamming distance in feature

space:

$$d_{\mathrm{sub}}(p, q) = \frac{1}{24} \left\| \mathrm{feat}(p) - \mathrm{feat}(q) \right\|_1 . \tag{16}$$

Insertions and deletions each have unit cost. Let $D(\hat{Y}, Y)$ be the minimum-cost Levenshtein distance between the predicted phone sequence $\hat{Y}$ and the reference sequence Y under these costs. We normalize by the reference length:

$$\mathrm{PFER}(\hat{Y}, Y) = \frac{1}{|Y|} D(\hat{Y}, Y). \tag{17}$$

This definition yields a PFER fraction. We multiply it by 100 when reporting percentages in Table 1; the remaining PFER tables and figures report the fraction directly. Compared to exact-match PER, PFER assigns partial credit to substitutions that differ in only a small number of distinctive features (e.g., voicing).

F Failure cases study

We analyze failure modes on PPA to understand when (i) bottom-up phone evidence collapses, (ii) WFST-based top-down refinement hurts, and (iii) distortion-based switching makes suboptimal routing decisions. For each utterance we consider three routes available at test time: *1-best* (bottom-up), *refine* (WFST refinement conditioned on the predicted transcript), and *switch* (distortion-controlled selection between 1-best and refine). For diagnosis only, we additionally report *refine*$_{given}$, i.e., refinement conditioned on an external reference transcript (not available at test time).

F.1 Case selection protocol

We select representative cases using three reproducible criteria: (1) highest PFER under 1-best, (2) largest degradation from refinement ($\Delta = \mathrm{PFER}_{ref} - \mathrm{PFER}_{1b}$), and (3) largest *switching regret* defined as $\mathrm{PFER}_{sw} - \min(\mathrm{PFER}_{1b}, \mathrm{PFER}_{ref})$. To compute the switching route in this appendix, we use a fixed threshold $\tau = 0.573$ on the distortion score (chosen to minimize mean PFER on this evaluation set). We omit utterance identifiers and audio references for privacy.

F.2 Failure taxonomy

We categorize failures into four practical types: **(A) Extreme evidence failures**, where 1-best exhibits large insertion bursts or very short phone references make the metric unstable; **(B) Wrong hypothesis conditioning**, where refine hurts but refine$_{given}$ would help, indicating that the guiding hypothesis (predicted transcript) is incorrect; **(C) Over-constrained top-down**, where both refine and refine$_{given}$ hurt, suggesting pronunciation coverage or LM bias issues; and **(D) Distortion/scheduler outliers**, where distortion-based routing makes noticeable mistakes (false positives/negatives) or distortion is weakly aligned with PFER for that sample.

F.3 Representative examples and aggregate summary

Table 4 lists representative examples for each failure type, and Table 3 summarizes aggregate statistics by category.

F.4 Implications

Across (B) and (C), a recurring pattern is that hard constraints can hurt unless the guiding hypothesis is reliable and the pronunciation model has sufficient coverage for weak/atypical realizations. In practice, we found the following mitigations useful: (i) hypothesis ensembling (e.g., ASR N-best) when conditioning refinement; (ii) softening constraints by tuning LM weight/word insertion penalty and expanding pronunciation variants; and (iii) sanity checks to flag extremely short references or insertion bursts before interpreting PFER.

Category	n	Dist.	PFER_{1b}	PFER_{ref}	Δ	Regret
Extreme evidence failures	5	0.614	2.214	1.023	-1.191	0.000
Wrong hypothesis conditioning	11	0.570	0.277	0.551	0.274	0.093
Over-constrained top-down	16	0.563	0.223	0.422	0.199	0.089
Distortion/scheduler outliers	5	0.512	0.751	0.585	-0.167	0.174
Other	37	0.543	0.290	0.305	0.016	0.018

Table 3: Aggregate statistics by failure category. $\Delta = \text{PFER}_{ref} - \text{PFER}_{1b}$. Regret is computed for distortion switching with $\tau = 0.573$.

Case	Ref text (abbr.)	Dist.	PFER_{1b}	PFER_{ref}	PFER_{given}	PFER_{sw}	Diagnosis
A. Extreme evidence failures							
A1	Twice each day he plays skillfully and with zest …	0.597	3.638	2.350	2.258	2.350	1-best insertion burst; refine stabilizes.
A2	yet he still thinks as swiftly as ever	0.643	3.225	0.250	1.800	0.250	Short phone ref; insertion dominates.
A3	Grandfather likes to be modern in his language	0.656	1.783	1.022	0.395	1.022	1-best insertion burst; refine stabilizes.
A4	giving those who observe him a pronounced feeling of …	0.593	1.359	0.915	0.859	0.915	Both routes fail; weak evidence.
B. Wrong hypothesis conditioning							
B1	Well he is nearly ninety three years old	0.649	0.357	0.663	0.312	0.663	Hypothesis mismatch: ref-conditioned refine hurts; Given ref helps.
B2	When he speaks	0.496	0.163	0.483	0.208	0.163	Hypothesis mismatch: ref-conditioned refine hurts; Given ref helps.
B3	He dresses himself in an old black frock coat	0.568	0.243	0.635	0.364	0.243	Hypothesis mismatch: ref-conditioned refine hurts; Given ref helps.
B4	He dresses himself in an old black frock coat	0.605	0.430	0.765	0.500	0.765	Hypothesis mismatch: ref-conditioned refine hurts; Given ref helps.
C. Over-constrained top-down							
C1	but he always answers Banana oil	0.585	0.150	0.556	0.679	0.556	Over-constraint: refine hurts even with Given ref.
C2	but he always answers Banana oil	0.617	0.361	0.703	0.696	0.703	Over-constraint: refine hurts even with Given ref.
C3	A long beard clings to his chin	0.560	0.242	0.554	0.554	0.242	Over-constraint: refine hurts even with Given ref.
C4	Yet he still thinks as swiftly as ever	0.643	0.286	0.591	0.535	0.591	Over-constraint: refine hurts even with Given ref.
D. Distortion/scheduler outliers							
D1	usually several buttons are missing	0.589	0.240	0.619	0.619	0.619	Scheduler FP: high distortion but refine hurts.
D2	but he always answers Banana oil	0.566	0.657	0.361	0.404	0.657	Scheduler FN: low distortion but refine would help.
D3	giving those who observe him a pronounced feeling of …	0.553	0.951	0.656	0.719	0.951	Scheduler FN: low distortion but refine would help.
D4	he slowly takes a short walk in the open …	0.519	0.663	0.422	0.422	0.663	Scheduler FN: low distortion but refine would help.

Table 4: Representative failure cases on PPA. Dist. is the distortion score. PFER_{1b} is HuPER 1-best, PFER_{ref} is refinement conditioned on the predicted transcript, PFER_{given} is refinement conditioned on an external reference transcript, and PFER_{sw} is distortion-controlled switching with $\tau = 0.573$.

F.5 Concrete error snippets

Red marks phones/words that are not aligned to the reference (substitutions/insertions). Deleted reference phones are shown with on the GT line. A red underscore indicates an empty counterpart in the alignment.

Case A1. **Reference text:** Twice each day he plays skillfully and with zest upon a small organ
Predicted hypothesis (for refinement): twice a day he place skilly and with zeppa ponce mogin
Distortion: 0.597. **PFER_{1b}**=3.638, **PFER_{ref}**=2.350, **PFER_{given}**=2.258.
Extreme evidence failure with a large insertion burst; both routes struggle.

```
GT:       _ _ EH _ _ _ _ N _ _ _ _ _ _ HH IY S _ P
1-best:   +K +IH EH +L +L +IY +IH N +D +W +IH +TH +Z EH P S +AH P
refine:   _ _ _ _ _ _ _ _ _ _ _ _ _ N D _ _ _ _
```

Case A2. **Reference text:** yet he still thinks as swiftly as ever
Predicted hypothesis (for refinement): citizen
Distortion: 0.643. **PFER**$_{1b}$=3.225, **PFER**$_{ref}$=0.250, **PFER**$_{given}$=1.800.
Very short phone reference; refinement can sharply reduce insertion-driven PFER but may be unstable.

GT:	Y EH T HH IY S T IH L TH IH NG K S AE Z S W IH F T L IY AE Z EH V ER
1-best:	+S +IH +T +IH +Z +AH +N +CH +AH +N +T +S +AH +N +CH +AH +N +T
refine:	Y EH T HH IY S T IH L TH IH NG K S AE Z S W IH F T L IY AE Z EH V ER

Case B2. **Reference text:** When he speaks
Predicted hypothesis (for refinement): he spake
Distortion: 0.496. **PFER**$_{1b}$=0.163, **PFER**$_{ref}$=0.483, **PFER**$_{given}$=0.208.
Wrong-hypothesis conditioning: refinement is guided by an incorrect word form ("spake"), distorting constraints.

GT:	HH EH N HH IY S P IY K S
1-best:	HH IY Z S P IY K S
refine:	HH IY S P EY K

Case B1. **Reference text:** Well he is nearly ninety three years old
Predicted hypothesis (for refinement): well nigh thistle
Distortion: 0.649. **PFER**$_{1b}$=0.357, **PFER**$_{ref}$=0.663, **PFER**$_{given}$=0.312.
Wrong-hypothesis conditioning: hypothesis is far from the reference, so refinement degrades phone accuracy.

GT:	W EH L HH IY IH Z N IH R L IY N AY N T IY TH R IY Y IH R Z OW L D
1-best:	W EH L HH IY IH Z N IH R L IY N AY N S IY TH R IY S AH L D
refine:	W EH L N AY TH AH S AH L

Case C1. **Reference text:** but he always answers Banana oil
Predicted hypothesis (for refinement): business boil
Distortion: 0.585. **PFER**$_{1b}$=0.150, **PFER**$_{ref}$=0.556, **PFER**$_{given}$=0.679.
Over-constrained top-down: even with the correct transcript, refinement hurts (pronunciation/LM bias).

GT:	B AH T HH IY AO L W EY Z AE N S ER Z B AH N AE N AH OY L
1-best:	B AH T HH IY AO L W EY Z AE N S ER Z B AH N AE N AH OY L
refine:	B IH Z N AH S B OY L

Case C3. **Reference text:** A long beard clings to his chin
Predicted hypothesis (for refinement): henri kisses chin
Distortion: 0.560. **PFER**$_{1b}$=0.242, **PFER**$_{ref}$=0.554, **PFER**$_{given}$=0.554.
Over-constrained top-down on a short phrase; refinement introduces consistent substitutions.

GT:	AH L AO NG B IH R D K L IH NG Z T UW HH IH Z CH IH N
1-best:	AH L AO NG B IH R D K L IH NG Z T UW HH IH Z CH IH N
refine:	HH EH N R IY K IH S AH Z CH IH N

Case D2. **Reference text:** but he always answers Banana oil
Predicted hypothesis (for refinement): but he has he was answers banana
Distortion: 0.566. **PFER**$_{1b}$=0.657, **PFER**$_{ref}$=0.361.
Scheduler false negative at $\tau = 0.573$: distortion is below threshold so switching keeps 1-best, but refinement would help.

GT:	B AH T HH IY AO L W EY Z AE N S ER Z B AH N AE N AH OY L
1-best:	S P B AH DX HH IY HH AE Z HH IY W AA Z AE N S ER Z B AH N AE N AH
refine:	B AH T HH IY HH AE Z HH IY W AA Z AE N S ER Z B AH N AE N AH OY L

Case D3. **Reference text:** giving those who observe him a pronounced feeling of the utmost respect
Predicted hypothesis (for refinement): giving those who observe him announced filling of the

upmost respect
Distortion: 0.553. **PFER**$_{1b}$=0.951, **PFER**$_{ref}$=0.656.
Scheduler false negative at τ = 0.573: switching keeps 1-best but refinement reduces PFER under weak evidence.
GT: G IH V IH NG DH OW Z HH UW AH B Z ER V HH IH M EY P R AH N AW N S T F IY L IH NG AH V DH AH AH T ER M OW S T R IH S P EH K T
1-best: G IH V IH NG N DH OW Z HH UW AH B Z ER V HH IH M AH N UH N S T F IY L IH NG AH V DH AH
refine: G IH V IH NG DH OW Z HH UW AH B Z ER V HH IH M EY P R AH N AW N S T F IY L IH NG AH V DH AH

G Analysis details for understanding HuPER-Recognizer gains

G.1 Centroid RSA: setup and implementation

Goal. We quantify whether an encoder's phone representations are organized by broad acoustic–phonetic similarity, using PanPhon distinctive-feature distances as a controlled proxy [46].

Data. We run the analysis on TIMIT, which provides time-stamped phone segments. Each segment is associated with a phone label.

Phone representations. At each transformer layer ℓ, we extract hidden states and mean-pool within each phone segment to obtain a segment embedding. We then average segment embeddings of the same phone to form a phone centroid $\mathbf{c}^{(\ell)}(p)$.

Distances and RSA score. We compute pairwise cosine distances between phone centroids to form an embedding-distance matrix $D^{(\ell)}_{\text{emb}}$. We also compute a PanPhon feature-distance matrix D_{pan} over the same phone set. The layer-wise RSA score is the Spearman correlation between vectorized upper triangles of $D^{(\ell)}_{\text{emb}}$ and D_{pan}.

Models compared. We include (i) *Wav2Vec2Phoneme* (multilingual phone recognition baseline), (ii) *WavLM-Raw* (pretrained WavLM-Large), (iii) *WavLM-Libri* (English-only fine-tuning on LibriSpeech with G2P labels), and (iv) HuPER-Recognizer.

G.2 Emission diagnostic: construction, scoring, and text inputs

We use a controlled diagnostic to test whether CTC emissions prefer the *canonical* (G2P) phone sequence or the *realized* phone sequence supported by the acoustics.

Diagnostic set. We focus on three cases where canonical restoration is tempting in casual English speech: **glottalization**, **flaps**, and **stops in consonant clusters**. For each synthesized utterance waveform x, we create a matched contrast $(x, y_{\text{can}}, y_{\text{real}})$: y_{can} is the canonical G2P phone sequence, and y_{real} is a realized phone sequence verified by listening.

CTC evidence. Given a CTC recognizer, we compute the marginal log-likelihood $\log P(y \mid x)$ via the forward algorithm in log-space. To compare sequences of different lengths, we use a per-phone normalized preference score:

$$\Delta_{\text{norm}}(x) = \frac{\log P(y_{\text{can}} \mid x)}{|y_{\text{can}}|} - \frac{\log P(y_{\text{real}} \mid x)}{|y_{\text{real}}|}. \tag{18}$$

$\Delta_{\text{norm}} > 0$ indicates canonical-restoring emissions, while $\Delta_{\text{norm}} < 0$ indicates acoustic-faithful emissions favoring the realized sequence.

Reporting. In the main paper, we plot per-utterance scores for HuPER vs. the XLSR baseline (Fig. 6). We also report a compact category-wise summary:

Text inputs (plain text only). This appendix also records the exact text inputs used to synthesize the diagnostic waveforms referenced in Sec. 8.2. All inputs are plain text (no style tags or special TTS

Category	HuPER-Recognizer (ours) median Δ_{norm}	HuPER-Recognizer (ours) $\Pr(\Delta_{\text{norm}} > 0)$	XLSR baseline median Δ_{norm}	XLSR baseline $\Pr(\Delta_{\text{norm}} > 0)$	Paired diff (XLSR – Ours) median Δ_{norm}	Paired diff (XLSR – Ours) $\Pr(> 0)$
Glottalization	-1.018	0.00	$+0.889$	0.70	$+1.535$	1.00
Flaps	-0.768	0.00	-0.459	0.30	$+0.630$	0.90
Stops in clusters	-0.424	0.20	$+0.413$	0.90	$+0.886$	0.80

Table 5: **Emission-level diagnostic using Δ_{norm}.** $\Delta_{\text{norm}} > 0$ indicates that emissions prefer the canonical phone sequence over the realized sequence. The paired-difference columns compare the two models on the same utterances; positive values mean the XLSR baseline is more canonical-restoring than HuPER-Recognizer.

instructions). We generate short, TTS-friendly phrases using a single prompt template (run once per category):

```
Generate 10 short English phrases (2-4 words) that are likely
to be pronounced in casual speech with the following phenomenon:
{PHENOMENON}.
Constraints:  common words, no proper nouns, keep it short and
natural for TTS.
Return only the 10 phrases, one per line.
```

The resulting phrases (10 per category) are listed in Table 6. After audio generation, each item is manually checked and paired with $(y_{\text{can}}, y_{\text{real}})$ for the emission test.

Glottalization	Flaps	Stops in consonant clusters
a button	a better idea	last Sunday
my kitten	a little later	next day
that mountain	water bottle	just say
in Britain	get it	best friend
a little bit	put it away	first time
not now	what a day	most people
can't go	write it down	west side
get back	I need it	old man
sit down	go to bed	hand bag
at night	it is ready	asked to

Table 6: Text inputs used to synthesize the emission diagnostic set in Sec. 8.2 (plain text only; 10 per category).

Table 7: Evaluation datasets for phone recognition.

Dataset	Description
Buckeye [53]	Natural English conversational speech with human phonetic annotation.
DRC-SE (DoReCo South-England) [54]	English dialectal-variation subset from DoReCo (South England).
L2-ARCTIC-Perceived [55]	L2 English speech corpus with human-verified / annotated pronunciations.
EpaDB [56]	L2 English (Spanish-accented) speech with detailed phonetic annotations (e.g., mispronunciations).
SpeechOcean762 [57]	Large-scale L2 English corpus with human annotations / verification.
VoxAngeles [58]	Multilingual word recordings; evaluate on languages such as Chamorro, Degema, Lakota, Pampanga, Iloko, etc.

Table 8: **Zero-shot multilingual phone recognition on VoxAngeles (94 languages).** Per-language PFER is reported as a fraction (lower is better). Entries of 1.00 mean the model lacks support for the corresponding language/inventory. Red highlights languages where HuPER attains the best (lowest) PFER across all baselines.

language	Allosaurus	Wav2Vec2Phoneme	MultIPA	ZIPA	Allophant	HuPER
abk	0.61	0.40	0.32	0.44	1.00	0.35
ace	0.24	0.18	0.15	0.15	1.00	0.29
ady	0.36	0.32	0.35	0.32	1.00	**0.30**
aeb	0.32	0.17	0.17	0.25	1.00	0.17
afn	0.20	0.10	0.11	0.21	1.00	0.14
afr	0.23	0.11	0.14	0.16	1.00	**0.10**
agx	0.25	0.20	0.17	0.16	1.00	0.30
ajp	0.38	0.12	0.13	0.21	1.00	**0.12**
aka	0.20	0.10	0.13	0.18	1.00	0.14
apc	0.25	0.14	0.13	0.14	1.00	0.18
ape	0.34	0.17	0.14	0.15	1.00	0.19
apw	0.30	0.19	0.18	0.21	1.00	**0.10**
asm	0.21	0.10	0.09	0.18	1.00	0.20
azb	0.25	0.15	0.12	0.18	1.00	0.20
bam	0.27	0.22	0.23	0.32	1.00	0.28
bem	0.14	0.08	0.05	0.07	1.00	0.08
ben	0.32	0.27	0.24	0.28	0.18	0.40
bfd	0.35	0.17	0.17	0.21	1.00	0.27
bfq	0.21	0.20	0.15	0.31	1.00	0.20
bhk	0.28	0.14	0.12	0.13	1.00	0.14
bin	0.19	0.08	0.10	0.17	1.00	0.12
brv	0.36	0.28	0.29	0.26	1.00	0.27
bsq	0.33	0.20	0.22	0.30	1.00	**0.05**
bwr	0.24	0.10	0.15	0.14	1.00	0.10
cbv	0.30	0.25	0.25	0.25	1.00	0.47

Continued on next page

Table 8: **Zero-shot multilingual phone recognition on VoxAngeles (94 languages).** Per-language PFER is reported as a fraction (lower is better). Entries of 1.00 mean the model lacks support for the corresponding language/inventory. Red highlights languages where HuPER attains the best (lowest) PFER across all baselines. (Continued)

ces	0.21	0.10	0.08	0.13	0.09	0.08
cha	0.16	0.05	0.06	0.15	1.00	0.06
cji	0.40	0.30	0.21	0.47	1.00	0.32
col	0.38	0.27	0.30	0.31	1.00	**0.25**
cpn	0.26	0.17	0.19	0.35	1.00	0.17
dag	0.30	0.17	0.14	0.34	1.00	0.16
dan	0.24	0.16	0.19	0.23	0.24	0.18
deg	0.13	0.07	0.07	0.06	1.00	0.07
dyo	0.17	0.07	0.12	0.16	1.00	0.13
efi	0.22	0.09	0.10	0.20	1.00	0.20
ell	0.18	0.07	0.05	0.08	0.13	0.13
ema	0.26	0.10	0.21	0.17	1.00	0.24
eus	0.40	0.07	0.08	0.09	0.05	0.08
ewe	0.36	0.14	0.18	0.33	1.00	0.22
ffm	0.22	0.12	0.12	0.16	1.00	0.17
fin	0.21	0.13	0.11	0.15	0.14	0.12
fub	0.19	0.06	0.09	0.16	1.00	0.16
gaa	0.25	0.17	0.20	0.23	1.00	0.26
gla	0.32	0.14	0.16	0.30	1.00	0.18
guj	0.18	0.13	0.14	0.16	1.00	0.18
gwx	0.54	0.21	0.24	0.24	1.00	**0.16**
hak	0.30	0.17	0.17	0.23	1.00	0.18
hau	0.12	0.11	0.07	0.15	1.00	0.16
haw	0.21	0.12	0.14	0.14	1.00	0.15
heb	0.26	0.11	0.16	0.15	1.00	0.21
hil	0.21	0.11	0.12	0.15	1.00	0.12
hin	0.27	0.12	0.07	0.17	0.09	0.15
hni	0.37	0.19	0.22	0.27	1.00	0.57
hrv	0.21	0.09	0.11	0.13	1.00	0.16
hun	0.30	0.15	0.15	0.26	0.16	**0.10**
hye	0.26	0.10	0.11	0.13	1.00	0.17
ibb	0.20	0.12	0.15	0.22	1.00	0.23
ibo	0.30	0.15	0.14	0.22	1.00	0.30
idu	0.23	0.13	0.12	0.16	1.00	0.19
ilo	0.17	0.10	0.09	0.13	1.00	0.12
isl	0.25	0.15	0.12	0.16	1.00	0.15
its	0.64	0.11	0.18	0.29	1.00	0.20
kan	0.18	0.10	0.07	0.12	1.00	**0.07**
kea	0.31	0.24	0.20	0.25	1.00	**0.13**

Continued on next page

Table 8: **Zero-shot multilingual phone recognition on VoxAngeles (94 languages).** Per-language PFER is reported as a fraction (lower is better). Entries of 1.00 mean the model lacks support for the corresponding language/inventory. Red highlights languages where HuPER attains the best (lowest) PFER across all baselines. (Continued)

khm	0.27	0.22	0.21	0.24	1.00	**0.15**
klu	0.37	0.17	0.21	0.28	1.00	0.22
knn	0.20	0.13	0.11	0.25	1.00	0.20
kri	0.15	0.08	0.09	0.10	1.00	0.38
kub	0.17	0.09	0.10	0.16	1.00	0.15
kye	0.25	0.20	0.18	0.28	1.00	0.29
lad	0.25	0.15	0.13	0.12	1.00	0.16
lar	0.48	0.14	0.18	0.30	1.00	**0.11**
lav	0.35	0.20	0.19	0.22	1.00	0.22
led	0.42	0.23	0.15	0.21	1.00	0.25
lgq	0.25	0.11	0.16	0.41	1.00	0.23
lit	0.26	0.14	0.15	0.18	0.14	**0.12**
lkt	0.26	0.18	0.17	0.17	1.00	0.21
lug	0.16	0.10	0.15	0.17	1.00	0.16
mak	0.29	0.17	0.19	0.18	1.00	0.25
mal	0.37	0.19	0.18	0.22	1.00	0.46
mlt	0.27	0.13	0.09	0.17	0.19	0.16
mya	0.28	0.24	0.25	0.25	1.00	**0.16**
njm	0.26	0.17	0.09	0.38	1.00	0.21
nld	0.20	0.13	0.13	0.14	0.20	0.13
ozm	0.22	0.14	0.16	0.18	1.00	0.19
pam	0.24	0.16	0.19	0.18	1.00	**0.08**
pes	0.23	0.14	0.15	0.17	1.00	0.19
prs	0.27	0.16	0.17	0.20	0.99	0.17
run	0.27	0.09	0.13	0.11	1.00	0.21
sbc	0.32	0.22	0.12	0.18	1.00	0.18
tsw	0.30	0.10	0.11	0.21	1.00	0.14
tzm	0.33	0.16	0.15	0.21	1.00	**0.14**
wuu	0.45	0.36	0.33	0.31	1.00	**0.22**
yue	0.29	0.14	0.15	0.21	1.00	0.16
Avg.	0.28	0.15	0.15	0.21	0.90	0.19

Table 9: Baseline model checkpoints

Model	Checkpoint
Allosaurus	https://github.com/xinjli/allosaurus
Allophant	https://github.com/kgnlp/allophant
Wav2Vec2Phoneme	https://huggingface.co/facebook/wav2vec2-xlsr-53-espeak-cv-ft
MultIPA	https://huggingface.co/ctaguchi/wav2vec2-large-xlsr-japlmthufielta-ipa1000-ns
ZIPA	https://github.com/lingjzhu/zipa
POWSM	https://huggingface.co/espnet/powsm
Wav2Vec2-en	https://huggingface.co/Bluecast/wav2vec2-Phoneme

H Broader Impacts

This work aims to advance speech and language technologies by improving the reliability and interpretability of phonetic representations. More accurate and acoustically grounded phonetic modeling can substantially benefit assistive applications in education, healthcare, and accessibility, where transcription errors and hallucinated outputs often undermine user trust and efficiency. By reducing such errors, HuPER has the potential to support more reliable screening, assessment, and communication tools for diverse populations. Beyond immediate applications, HuPER provides a scalable phonetic foundation that may facilitate the development of more expressive and flexible speech generation systems, including text-to-speech models based on acoustically meaningful tokens. More broadly, by framing speech perception as an adaptive and modular inference process, this work points toward future speech foundation systems that integrate perception, reasoning, and control, rather than isolated task-specific models.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the scope of HuPER as a human-inspired framework for adaptive phonetic perception over acoustic–phonetic evidence and linguistic knowledge. The stated contributions correspond to the proposed HuPER-Recognizer, HuPER-Perceiver, HuPER-Scheduler, and Dysfluent WFST based inference routes. The empirical claims are restricted to the evaluated English variation benchmarks, zero-shot transfer on VoxAngeles, and nfvPPA experiments under weak acoustic evidence or with available reference text.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section Conclusion and Limitations discusses the main limitations of the current system, including the limited amount of human-verified spoken-phone supervision, the heuristic nature of the current Scheduler, and the fact that HuPER models only part of the inference pathways involved in human phonetic perception. The paper presents HuPER as a human-inspired computational framework rather than a complete cognitive or neural simulation.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: HuPER-Recognizer introduces the DRRC formulation, the missing-label view of self-training, and the consistency conditions used by the main theorem. Appendix B provides the formal setup, complete theorem statements, assumptions, and proofs. The theorem in the main text is presented as an informal summary of the full appendix result.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the model backbone, training pipeline, proxy-label construction, phoneme-to-phone correction procedure, inference routes, evaluation metrics, datasets, baselines, and switching analysis. Algorithms 1–2, Appendix E, and the released code, labels, and model artifacts provide the information needed to reproduce or verify the main experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The HuPER repository is publicly available and provides the core code for the phone recognizer and phoneme-to-phone Corrector, together with runnable notebooks and a minimal inference example for the Corrector. The repository documents the main code layout and usage entry points. Third-party raw datasets are referenced through their original sources and are not redistributed unless permitted by their access conditions. We will continue to maintain the released repository with the scripts, checkpoints, and instructions needed to reproduce the main evaluations.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Experimental Setup specifies the WavLM-Large backbone, CTC objective, TIMIT-to-LibriSpeech self-learning recipe, Corrector training configuration, optimizer, learning rates, batch size, freezing and unfreezing schedule, evaluation datasets, and PFER metric. Appendix E and the dataset and baseline appendices further document the evaluation protocol, label normalization, and benchmark construction details.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper reports a statistical test for the distortion-based scheduling analysis, including the Spearman correlation between emission distortion and PFER and the corresponding p-value. It also reports per-dataset and per-language results under fixed evaluation splits and a unified evaluation pipeline. However, the main phone-recognition benchmark tables do not yet include run-level error bars, confidence intervals, or significance tests, because the principal comparisons are reported as deterministic evaluations of released checkpoints on fixed test sets.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Experimental Setup reports the main compute used for the HuPER training pipeline, including 2×A6000 GPUs for Corrector training and 8×A6000 GPUs for HuPER-Recognizer training, together with the number of epochs, batch size, learning rates, and freezing schedule. The released instructions provide the remaining environment and execution details needed to reproduce the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and believe the work conforms to it. The study uses existing speech datasets, released model assets, and de-identified or properly accessed evaluation resources under their intended research use. The paper reports limitations and potential risks, and does not involve deceptive data collection, unsafe deployment, or undisclosed intervention with human participants.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The Appendix Broader Impacts discusses both the potential benefits and risks of HuPER. HuPER may support robust speech processing, pronunciation variation analysis, low-resource phonetic modeling, and speech accessibility, but phonetic transcription systems could also be misused for unwanted speech analysis or for drawing unreliable conclusions about individual speakers or speaker groups. The paper therefore presents HuPER as a research system evaluated on controlled benchmarks, with explicit limitations rather than deployment-ready claims.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The released HuPER artifacts consist of research code, phone-recognition components, Corrector code, and usage examples rather than generative models, large language models, surveillance systems, or scraped unsafe datasets. The release does not redistribute third-party raw speech datasets unless permitted by their original access conditions. The paper presents HuPER as a research system and discusses limitations and potential misuse in the Impact Statement.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the original sources for the datasets, pretrained models, G2P tools, baseline recognizers, and WFST-related components used in the experiments. Existing assets are used through their original access channels and under their intended research-use conditions. The release focuses on HuPER code, trained models, processed labels, and evaluation scripts, and does not redistribute third-party raw datasets unless redistribution is permitted by the corresponding license or data-use agreement.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

Justification: The paper introduces HuPER-related assets, including the released codebase, phoneme-to-phone Corrector, phone-recognition training code, and runnable examples. The repository documents the basic usage, minimal Corrector inference procedure, and code organization for the released HuPER components. The release distinguishes HuPER implementation artifacts from third-party datasets, pretrained models, and external tools, which are referenced through their original sources.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The work does not conduct crowdsourcing experiments or recruit new human participants. The phonetic annotation and evaluation procedures are performed on existing speech data, and no participant-facing instructions, screenshots, or compensation procedures are involved in this study.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: The nfvPPA experiments use a private clinical speech dataset accessed under IRB-approved and data-use restricted conditions. The data are de-identified before analysis, used only for controlled research evaluation, and are not redistributed. We report only aggregate experimental results and omit utterance identifiers or audio references in failure-case analysis to avoid disclosure risk. No new participants were recruited and no new participant-facing intervention was conducted for this study.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as an important, original, or non-standard component of HuPER's core methodology. The proposed framework relies on acoustic–phonetic recognition, corrected proxy-phone supervision, explicit constraint graphs, and distortion-controlled routing rather than LLM-based inference.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.