CrystalASR: Hierarchical Phoneme-Grounded Speech Decoding

Po-Ting Lin

***Abstract*—We investigate a design principle for automatic speech recognition where linguistic structure is explicitly enforced through intermediate representations. The resulting system, CrystalASR, decomposes decoding into three modular layers: a 3.3M parameter phoneme CTC head, a zero-parameter rule-based word decoder, and an optional language model (LM) for disambiguation. A defining constraint is strict upward information flow, ensuring higher layers modulate but never override lower-level acoustic evidence.**

Experiments on LibriSpeech dev-clean show that CrystalASR achieves 17.44% WER while requiring 21× fewer trainable parameters and 14× faster inference than a comparable end-to-end baseline. Error attribution reveals that word-level errors primarily originate from subtle phoneme inaccuracies amplified by downstream segmentation. Furthermore, a language model weight sweep reveals a sharp phase transition: beyond a narrow tiebreaker role ($w_{LM} > 0.03$), WER rises from 17% to 96% as the LM's score scale overwhelms acoustic evidence. These findings suggest that explicitly decoupling acoustic and lexical processing yields interpretable error diagnostics and substantial parameter savings, at a moderate accuracy cost relative to end-to-end models.

***Index Terms*—Speech recognition, phoneme decoding, hierarchical architecture, parameter efficiency, CTC**

I. Introduction

Modern ASR has converged on end-to-end (E2E) architectures that learn a direct mapping from acoustic features to characters or subwords. CTC-based models [1], [2] eliminate explicit alignment but assume conditional independence among outputs. Attention-based encoder–decoders [3] and Conformer-Transducers [4] relax this assumption through autoregressive decoding. Whisper [5] demonstrated that weak supervision at scale yields robust zero-shot performance. These systems encode all linguistic structure—phonotactic constraints, lexical identity, and contextual disambiguation—within a single learned function, requiring large decoder networks and extensive training data.

We investigate whether explicitly separating these levels of linguistic structure, rather than learning them jointly, offers practical advantages. Human speech perception proceeds through well-documented stages: acoustic–phonetic analysis identifies a small set of phonemic categories, lexical access maps phoneme sequences to words, and sentential integration resolves residual ambiguity [6]–[8]. A consistent finding across perceptual models is that higher-level expectations modulate but do not override lower-level acoustic evidence [9]. Classical HMM-based ASR pipelines once reflected this separation [10], but E2E systems largely abandoned it. Recent work has revisited phoneme representations via self-supervised learning [11] or phoneme CTC for low-resource settings [12], though typically as intermediate features within an E2E framework rather than as a strict interface between independently designed stages.

We present CrystalASR, a modular and interpretable alternative to E2E decoding based on a strict separation between acoustic, phonemic, and lexical processing. The system comprises three layers: (1) a **phoneme CTC head** (3.3M parameters) maps Whisper encoder features to 41 phonemes, (2) a **rule-based word decoder** (zero trainable parameters) segments phoneme sequences into words using the CMU Pronouncing Dictionary [13], a phoneme trie, and dynamic programming, and (3) an optional **sentence-level LM** [14] disambiguates homophones without overriding acoustic evidence. A defining constraint is that information flows strictly upward: each layer selects only among candidates compatible with the layer below.

Three experiments support this design. First, the zero-parameter word decoder achieves 7.14% WER when given ground-truth phonemes—with all errors falling into interpretable categories (OOV words, homophones, segmentation residuals)—confirming that the decoder is reliable and the primary bottleneck is phoneme accuracy. Second, layer-by-layer error attribution shows that the majority of word-level errors originate as subtle phoneme inaccuracies amplified by downstream segmentation, demonstrating that phoneme precision has outsized downstream impact. Third, a language model weight sweep reveals a sharp phase transition: at tiebreaker levels ($w \leq 0.03$) the LM is benign, but beyond this threshold WER rises from 17% to 96%, because the LM's score scale inherently dominates acoustic evidence. This is consistent with the psycholinguistic finding that lexical expectations cannot productively override acoustic–phonetic processing [8]. On LibriSpeech dev-clean, CrystalASR achieves 17.44% WER with $21\times$ fewer trainable parameters and $14\times$ faster inference than a comparable E2E CTC baseline trained on identical data.

II. Proposed Method

CrystalASR decomposes speech recognition into three layers, each targeting one level of linguistic abstraction. Fig. 1 illustrates the full pipeline.

A. Layer 0: Acoustic Encoder

We use the Whisper base encoder [5] (74M parameters) as a frozen acoustic front-end. The encoder was previously fine-tuned on LibriSpeech train-clean-100 for character-level CTC. It produces 512-dimensional feature vectors at 50 fps.

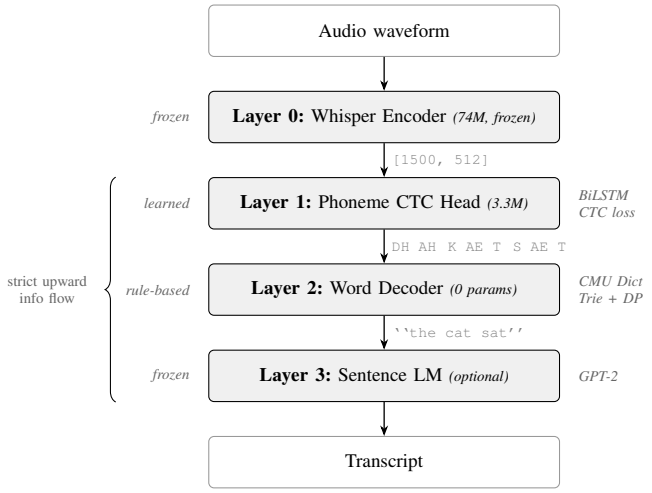

Fig. 1. CrystalASR architecture. A frozen Whisper encoder extracts acoustic features. The phoneme CTC head (3.3M trainable params) decodes a phoneme sequence via CTC. The word decoder (zero trainable params) segments phonemes into words using the CMU Pronouncing Dictionary with trie lookup and dynamic programming. An optional sentence LM disambiguates homophones without overriding acoustic evidence.

During CrystalASR training, only the phoneme CTC head is optimized; encoder weights are frozen.

B. Layer 1: Phoneme CTC Head

A two-layer bidirectional LSTM (hidden size 256) followed by a linear projection maps each encoder frame to a distribution over 42 classes (41 ARPAbet phonemes plus a CTC blank token). The model is trained with CTC loss [1], requiring only the phoneme sequence order—not frame-level alignment. At inference, greedy CTC decoding produces a phoneme sequence of approximately 50–80 symbols per utterance. Phoneme targets are generated by converting transcript words to phoneme sequences via the CMU Pronouncing Dictionary [13], with stress markers removed. Total: 3.3M trainable parameters.

C. Layer 2: Word Decoder

The word decoder converts a phoneme sequence into a word sequence with zero trainable parameters, using four components.

Phoneme trie. All 123,455 entries from the CMU Pronouncing Dictionary are stored in a trie indexed by phoneme sequence, enabling $O(L)$ lookup.

Confusable expansion. For each trie match, words with similar pronunciations (phoneme edit distance ≤ 2) are added from a precomputed pool, handling common CTC errors.

DP segmentation. The highest-scoring segmentation of the phoneme sequence into dictionary words is found via dynamic programming. Each candidate word w spanning phonemes $p_i \dots p_j$ is scored as:

$$s(w) = \big(0.1 + f(w)\big) \cdot \sqrt{j - i + 1} \tag{1}$$

where $f(w) \in [0, 1]$ is the normalized log frequency of w. The DP objective is:

$$\mathrm{dp}[t] = \max_{(w,\, i,\, t) \in \mathcal{C}} \big(\mathrm{dp}[i] + s(w) - \lambda\big) \tag{2}$$

where $\lambda = 1.5$ is a split penalty discouraging over-segmentation, and a skip penalty $\mu = 0.5$ per phoneme absorbs isolated CTC errors.

Fragment merging. Consecutive words whose combined pronunciation exactly matches a single dictionary entry are merged (e.g., “manage” + “aerial” $\rightarrow$ “managerial”).

D. Layer 3: Sentence LM (Optional)

A frozen GPT-2 model [14] provides disambiguation for homophones only (e.g., “there”/“their”/“they’re”), selecting among acoustically identical candidates using sentence context. It cannot substitute, insert, or delete words, and cannot override any non-homophone decision from Layer 2. We evaluate the consequences of relaxing this constraint in Section III-E.

E. Design Principle: Strict Upward Information Flow

A defining constraint is that no layer may override a lower layer’s output. The word decoder selects only among words whose pronunciations appear in the phoneme CTC output. The sentence LM selects only among words proposed by the word decoder. This preserves acoustic grounding throughout the pipeline and ensures that errors are attributable to specific layers—a property we exploit in Sections III-D–III-E.

III. Experiments

A. Setup

Data. All models are trained on LibriSpeech train-clean-100 (28,539 utterances, 100 hours) and evaluated on dev-clean (2,703 utterances). Phoneme targets are derived from transcripts via the CMU Pronouncing Dictionary [13] (41 phonemes, stress markers removed).

Acoustic encoder. Whisper base encoder [5] (74M parameters), top four transformer blocks fine-tuned for character-level CTC (30k steps), then frozen for all subsequent experiments.

Phoneme CTC head. Two-layer BiLSTM (hidden 256, bidirectional), linear projection to 42 classes, trained for 50k steps (batch 8, LR 3×10^{-4}, cosine annealing). Total: 3.3M parameters.

Word decoder. CMU Dict (123,455 entries), phoneme trie, DP segmentation ($\lambda = 1.5$, $\mu = 0.5$), confusable expansion, fragment merging. Zero trainable parameters.

Baseline. E2E character-level CTC using the same Whisper encoder with a BiLSTM CTC head (71M trainable parameters), trained on identical data with regularization (SpecAugment, feature dropout, weight decay).

B. Main Results

Table I compares CrystalASR against the E2E CTC baseline. Under identical training data and encoder, CrystalASR achieves 17.44% WER with $21\times$ fewer trainable parameters and $14\times$ faster inference.

Table II contextualizes this result against systems trained on substantially more data. The WER gap is primarily attributable to training data scale (6,800$\times$): scaling the phoneme CTC head to larger corpora requires no architectural change, as the word decoder and its 123k-entry dictionary are data-independent.

TABLE I
MAIN RESULTS ON LIBRISPEECH DEV-CLEAN (200 SAMPLES). ALL SYSTEMS USE THE SAME WHISPER BASE ENCODER AND TRAIN-CLEAN-100 DATA.

System	WER	Params	Speed
Whisper base (decode)	0.19%	74M	∼1 s
E2E CTC (char)	14.79%	71M	∼1 s
CrystalASR	**17.44%**	**3.3M**	**0.07 s**

TABLE II
INDIRECT COMPARISON (DIFFERENT TRAINING DATA, NOT DIRECTLY COMPARABLE). *TRAINABLE DECODER PARAMS ONLY; ENCODER IS FROZEN.

System	WER	Data (hrs)	Params
Whisper large-v3	<2%	680,000	1.5B
wav2vec 2.0 (large)	∼3%	60,000	317M
Conformer-T	∼2%	60,000	120M
CrystalASR (ours)	**17.44%**	**100**	**3.3M***

C. Layer 2 Reliability: Ground-Truth Phoneme Experiment

To isolate the word decoder's intrinsic error rate, we bypass Layer 1 entirely: ground-truth transcripts are converted to phoneme sequences via the CMU Dictionary and fed directly into the word decoder.

TABLE III
WORD DECODER PERFORMANCE WITH GROUND-TRUTH VS. CTC PHONEMES. THE 10.3 PP GAP IS ENTIRELY ATTRIBUTABLE TO LAYER 1 PHONEME NOISE.

Input	WER	S	D	I
CTC phonemes (noisy)	17.44%	445	125	109
GT phonemes (perfect)	**7.14%**	154	110	14

With perfect phonemes, the zero-parameter word decoder achieves 7.14% WER (Table III). Crucially, all residual errors fall into interpretable categories: out-of-vocabulary words absent from the CMU Dictionary (1.8% of reference words, e.g., proper nouns *Hurstwood*, *Drouet*), homophone confusion (e.g., "knew"→"new", "know"→"no"), and DP segmentation residuals. No error is unexplained. The 10.3 percentage point gap between 17.44% and 7.14% is entirely attributable to Layer 1 phoneme inaccuracies, confirming that phoneme precision is the system's primary bottleneck.

D. Error Attribution Across Layers

To quantify how errors propagate through the hierarchy, we classify each of the 679 word-level errors (200 samples, 3,894 reference words) by origin layer. An error is attributed to Layer 1 if the reference word's phoneme sequence is absent from the CTC output (edit distance > 1); to Layer 2 if the phonemes are present but the DP selected an incorrect segmentation; to Layer 3 if the predicted word is a homophone of the reference; and to OOV if the reference word is absent from the dictionary.

The dominant category is Layer 2 segmentation errors (69.2%, Table IV). However, this does not indicate a defect in the word decoder—which achieves 7.14% WER in isolation (Section III-C). Rather, these are *cascade errors*: subtle phoneme inaccuracies that pass the fuzzy-match threshold (edit distance ≤ 1) but alter the DP's scoring landscape enough to select an incorrect word. For instance, CTC drops the initial /B/ from "blight" (/**B** L AY T/ → /L AY T/); the remaining phonemes exactly match "light," a higher-frequency word that the DP prefers. The phoneme error is small (one deletion), but the word-level consequence is total (wrong word).

TABLE IV
ERROR ATTRIBUTION BY ORIGIN LAYER (679 TOTAL ERRORS).

Layer	Count	%	Example
Layer 1 (phoneme)	75	11.0	"managerial"→"aerial"
Layer 2 (segmentation)	470	69.2	"blight"→"light"
Layer 3 (homophone)	65	9.6	"there"→"their"
OOV	69	10.2	"Hurstwood"→"would"

This cascade effect explains why phoneme precision has outsized downstream impact: even marginal improvements in Layer 1 accuracy eliminate disproportionately many Layer 2 errors.

E. The Cost of Overriding Acoustic Evidence

A central claim of our design is that higher layers must not override lower-layer decisions. We test this by progressively increasing the language model's authority in the DP scoring:

$$s'(w) = s(w) + w_{\text{LM}} \cdot \log P_{\text{LM}}(w \mid \text{context}) \qquad (3)$$

where $s(w)$ is the acoustic score (Eq. 1) and $\log P_{\text{LM}}$ is the conditional log probability from GPT-2 [14].

TABLE V
WER AS A FUNCTION OF LM WEIGHT w_{LM}. THE LM'S CONDITIONAL LOG PROBABILITY SCALE (∼0 TO −15) INHERENTLY DOMINATES THE ACOUSTIC SCORE RANGE (∼0.1 TO 1.5). BEYOND THE TIEBREAKER REGIME ($w_{\text{LM}} \leq 0.03$), THE LM EFFECTIVELY OVERRIDES ACOUSTIC EVIDENCE.

w_{LM}	WER	S	D	Regime
0.00	**17.44%**	445	125	Acoustic only
0.01	17.67%	441	132	Tiebreaker
0.03	17.69%	440	130	Tiebreaker
0.05	18.21%	453	133	Transition
0.10	22.39%	506	180	Override onset
0.20	59.53%	1,218	721	Full override
0.50	95.25%	1,282	1,440	LM dominant
1.00	95.99%	1,277	1,461	LM dominant

Table V reveals a sharp phase transition. At $w_{\text{LM}} \leq 0.03$, the LM acts as a tiebreaker and WER is essentially unchanged. At $w_{\text{LM}} = 0.10$, WER begins to rise; by $w_{\text{LM}} = 0.20$ it has tripled; and at $w_{\text{LM}} \geq 0.50$ the system degenerates to 96% WER—worse than random. The transition is explained by a scale mismatch: the LM's conditional log probabilities range from ∼0 to −15 (span ≈ 14), while the acoustic score ranges from 0.1 to 1.5 (span ≈ 1.4). At $w_{\text{LM}} = 0.10$, the LM's effective swing equals the acoustic range, and the LM begins to dominate.

This result is consistent with the psycholinguistic finding that lexical expectations modulate but cannot override

prelexical acoustic processing [8], [9]. In both the human perceptual system and our engineered system, granting higher-level processes the authority to override lower-level acoustic evidence degrades performance. The practical implication is that language model integration in hierarchical ASR must be carefully constrained to a tiebreaker role—any greater authority risks the LM's score scale overwhelming acoustic grounding.

F. Ablation: Phoneme Recognition Method

Table VI shows the impact of the phoneme recognition method, with all other components held constant. Switching from per-frame classification to CTC was the single largest improvement (-40.8 pp), as CTC eliminates the need for precise frame-level alignment labels.

TABLE VI
PHONEME RECOGNITION METHOD VS. END-TO-END WER.

Method	Params	Phoneme metric	WER
Template (per-frame)	93K	85.5% frame acc	58.22%
Template + CNN	1.2M	88.1% frame acc	43.27%
CTC (BiLSTM)	**3.3M**	**0.17% PER**	**17.44%**

G. Interpretable Error Trace

Fig. 2 illustrates layer-by-layer error diagnosis on a representative utterance, demonstrating the interpretability enabled by the hierarchical architecture.

Reference:	*"…owing to the blight his wife's action threatened to cast upon …"*
Layer 1	`...OW IH NG T UW DH AH L AY T HH IH Z ...`
(Phoneme CTC)	← Missing /B/: "blight" **/B** L AY T/ → /L AY T/
Layer 2	"…owing to the light his wife's action that and to cast upon …"
(Word Decoder)	← /L AY T/ → "light" (valid, higher-frequency match) ← /TH R EH T AH N D/ over-segmented to "that" + "and"
Diagnosis:	Both errors originate in Layer 1 phoneme output. Layer 2 correctly maps the (incorrect) phonemes to valid dictionary words. Fixing phoneme recall for initial consonant clusters would resolve both errors.

Fig. 2. Layer-by-layer error trace. Errors are localizable to specific phoneme failures in Layer 1; Layer 2 faithfully propagates them. This diagnosis is unavailable in E2E systems, where the same output error could arise from any component.

IV. CONCLUSION

We presented CrystalASR, a hierarchical phoneme-grounded decoder that enforces linguistic structure through explicit intermediate representations. Three experimental findings support the constrained decoding principle. The zero-parameter word decoder achieves 7.14% WER from ground-truth phonemes, with all residual errors in interpretable categories, confirming that phoneme accuracy is the dominant bottleneck. Layer-by-layer error attribution reveals that subtle phoneme inaccuracies produce cascade effects through downstream segmentation, explaining why marginal improvements in phoneme precision yield outsized WER reductions. A language model weight sweep demonstrates a sharp phase transition: beyond a narrow tiebreaker regime, the LM's inherently larger score scale overwhelms acoustic evidence, degrading WER from 17% to 96%—a computational parallel to the psycholinguistic observation that lexical expectations cannot productively override acoustic–phonetic processing.

These findings have a practical consequence for language model integration in ASR: the LM's contribution must be architecturally constrained to a tiebreaker role, not merely tuned to a small weight. As a secondary outcome, the hierarchical separation yields $21\times$ parameter reduction and $14\times$ inference speedup relative to an E2E CTC baseline trained on identical data, because linguistic knowledge encoded in the pronunciation dictionary substitutes for learned decoder capacity.

The primary limitation is phoneme CTC accuracy under limited training data (100 hours). Since the word decoder and dictionary are data-independent, scaling the phoneme head to larger corpora requires no architectural change. Other directions include syllable-level matching for improved long-word handling and character-level fallback for out-of-vocabulary words.

REFERENCES

[1] A. Graves, "Connectionist temporal classification," in *Supervised Sequence Labelling with Recurrent Neural Networks*. Berlin, Heidelberg: Springer, 2012, pp. 61–93.

[2] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[3] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960–4964.

[4] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[5] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[6] W. D. Marslen-Wilson, "Functional parallelism in spoken word-recognition," *Cognition*, vol. 25, no. 1, pp. 71–102, 1987.

[7] J. L. McClelland and J. L. Elman, "The TRACE model of speech perception," *Cognitive Psychology*, vol. 18, no. 1, pp. 1–86, 1986.

[8] D. Norris, J. M. McQueen, and A. Cutler, "Merging information in speech recognition: Feedback is never necessary," *Behavioral and Brain Sciences*, vol. 23, no. 3, pp. 299–325, 2000.

[9] W. F. Ganong, "Phonetic categorization in auditory word perception," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 6, no. 1, pp. 110–125, 1980.

[10] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[11] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460.

[12] X. Li, S. Dalmia, J. Li, M. Lee, P. Littell, J. Yao, A. Anastasopoulos, D. R. Mortensen, G. Neubig, A. W. Black, and F. Metze, "Universal phone recognition with a multilingual allophone system," in *ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8249–8253.
[13] R. Weide, "The CMU pronunciation dictionary, release 0.6," 1998.
[14] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.