

How Much of Speech Recognition Must Be Learned?

A Parameter-Free Analysis of Lexical Decoding

Po-Ting Lin^[0009–0002–4825–6905]

Independent Researcher, Tainan, Taiwan
botimlin@gmail.com

Abstract. End-to-end recognisers learn one function from acoustics to text, which makes it impossible to ask where the difficulty sits. Separating the acoustic-phonetic stage from the lexical one, we find that given correct phonemes and no word boundaries a decoder with *zero trainable parameters* recovers words at 5.66% WER on LibriSpeech dev-clean – within 1.34 points of the floor that out-of-vocabulary words and homophones impose on any decoder over this lexicon, every residual error falling into one of three interpretable causes. Sweeping phoneme accuracy under controlled corruption yields a calibration curve, $\text{WER} = 9.3\% + 2.11 \times \text{PER}$ ($R^2 = 0.997$), validated to about a point by seven trained CTC heads: the relationship is linear, not amplifying. A frozen language model reordering phoneme-licensed candidates *improves* accuracy up to a broad optimum with no collapse at high weight, provided both scores are expressed on a common scale.

Keywords: Speech recognition · Error attribution · Lexical decoding.

1 Introduction

Modern speech recognition maps acoustics to text with a single learned function [2, 5]. The approach works, but it forecloses a question worth asking: of the difficulty in recognising speech, how much belongs to hearing sounds and how much to turning sounds into words? It is answerable only if the two stages are separated, and separation costs accuracy, which has discouraged it. We accept that cost: this is not a competitive recogniser but an attempt to measure one stage precisely enough to be useful to people building the other.

2 A Word Decoder with No Trainable Parameters

Let $p_{1..n}$ be a phoneme string with word boundaries removed. Decoding recovers the word sequence maximising $\sum_i [\log P(w_i) + \log(\ell_i + 1) - \lambda]$, where ℓ_i is the span of w_i in phonemes, λ a per-word penalty, and a phoneme absorbed into no word costs μ . This is MAP unigram segmentation [1] over pronunciations rather than characters; it needs no split penalty in principle, though we leave λ tunable

Table 1. Left: the ceiling on perfect phonemes and the irreducible part of it. Right: all 2,453 word errors, each assigned one mutually exclusive cause.

	WER (%)	Cause	Count	%
Zero-parameter decoder	5.66	Homophone ambiguity	942	38.4
irreducible: OOV	1.78	Out-of-vocabulary	774	31.6
irreducible: homophone	2.53	Segmentation	737	30.0
Floor for this lexicon	4.32	Phoneme corruption	0	0.0
Recoverable by a decoder	1.34	Unexplained	0	0.0

to test that. Lexical access walks a phoneme trie, optionally under a budget of k edits at cost β each; searching the trie under an explicit budget is exact and needs no $O(V^2)$ table of confusable words. Nothing is learned from speech: the five scalars come from a grid search on a disjoint tuning split, and every ablated configuration is re-tuned.

3 Setup

The lexicon is CMUdict [6]: 122,787 words, 132,690 pronunciations, 39 ARPAbet phonemes after stress removal. Unigram counts come from LibriSpeech train-clean-100 transcripts; that choice matters, since a general English frequency list prefers written abbreviations (MR over MISTER) that LibriSpeech never uses: in-domain counts alone move the result from 11.13% to 7.81%, and the further drop to 5.66% comes from the configuration selected in Sect. 4.3. We evaluate on dev-clean [3], holding out 20% of utterances for tuning and reporting on the remaining 2,163 (43,370 tokens); intervals are percentile bootstraps over utterances and ablation differences use a paired bootstrap. Only Sect. 4.2 needs audio: there we train a 3.17M-parameter BiLSTM CTC head over frozen Whisper-base [5] features, snapshotted seven times so one run supplies a range of phoneme qualities.

4 Results

4.1 The Ceiling, and What Lies Beneath It

Given correct phonemes and no boundaries, the parameter-free decoder reaches **5.66% WER**, CI [5.41, 5.92]. Words absent from the dictionary, and words sharing a pronunciation with a more frequent one, are irreducible for any decoder over this lexicon; together they account for 4.32 points, leaving **1.34 points** a better decoder could recover. Every error is assigned exactly one cause, and the categories sum to the errors the WER counts (Table 1).

4.2 What One Point of Phoneme Error Costs

We corrupt correct phoneme strings at a controlled rate under a confusion distribution driven by articulatory feature overlap, measuring the realised PER after

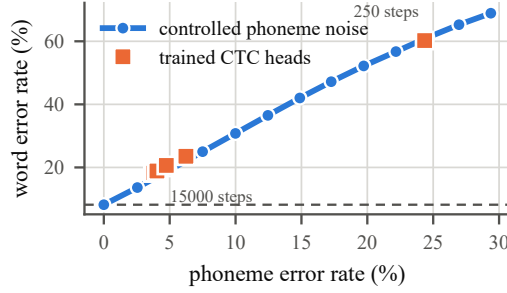


Fig. 1. Transfer from phoneme to word error. Squares are seven checkpoints of one CTC run, decoded through the identical path. The dashed line is this configuration on perfect phonemes, 8.18%; it is tuned under noise and so sits above the 5.66% of Table 1.

the fact since deletions and insertions shift the alignment. The relationship is linear throughout (Fig. 1): $WER = 9.3\% + 2.11 \times PER$, with $R^2 = 0.997$. The fitted intercept sits 1.1 points above the measured zero-noise value of 8.18%, the curve bending slightly near the origin.

One point of phoneme error costs slightly more than two of word error, at a rate constant from 0% to 29% PER. Since a single phoneme error can destroy a word and shift its neighbours’ boundaries, one might expect the cost to grow as errors interact; it does not, and that constancy is what makes the curve usable as a conversion. Seven trained CTC heads sit on it with residuals of -0.46 to $+1.35$ points.

4.3 Which Parameter-Free Components Earn Their Place

The unigram prior is the only component worth a large amount: without it the decoder loses 29 points (Table 2). The length term contributes exactly nothing once the objective is a log probability, since length is already implicit in a sum of per-word log probabilities; fragment merging is *harmful*, costing 2.16 points; and the heuristic objective costs 0.29. Most informative is bounded-edit access: worthless on correct phonemes (-0.01), worth 11 points under noise. It is a phoneme-error absorber rather than a lexical mechanism, and an ablation on clean input alone would have called it useless.

4.4 How Much Authority a Language Model May Hold

A frozen GPT-2 [4] rescores candidates in a beam search, restricted to words the phoneme evidence already licensed, over 400 utterances at $\epsilon = 0.05$. Sweeping its weight from 0 to 1 – adding its log probability directly, or standardising it within each candidate set first – word error falls monotonically from a 28.97% acoustic-only baseline to a broad optimum (25.28% at $w = 0.5$; 26.64% scale-matched) then degrades gently, to at worst 36.4%. The worry about granting a language

Table 2. Component ablation, each configuration separately re-tuned; Δ is against the reference, negative is better. The best configuration is the reference without fragment merging – the 5.66% of Table 1.

Configuration	perfect phonemes		noisy ($\epsilon = 0.08$)	
	WER (%)	Δ	WER (%)	Δ
Reference	7.81	—	36.26	—
without unigram prior	37.30	+29.49	62.15	+25.89
without length term	7.81	+0.00	36.51	+0.25
without fragment merging	5.66	−2.16	35.08	−1.19
with 1-edit lexical access	7.80	−0.01	25.23	−11.03

model authority is that lexical expectation overrides acoustic evidence and the system collapses into fluent nonsense. It does not here, for a reason separable from the architecture: such a collapse is what one should expect whenever the two scores live on incompatible scales, since a word score spanning $[0.1, 1.5]$ is swamped by a log probability spanning $[-15, 0]$ at any appreciable weight, leaving the weight to perform unit conversion rather than express authority. Under a log-probability objective both terms share a scale and no cliff can form.

5 Conclusion

The lexical stage needs no trainable parameters: given correct phonemes, a dictionary and a unigram prior come within 1.34 points of the floor vocabulary and homophony impose, and the residue is individually explainable. A phoneme error exchanges into word error at a stable two-to-one rate, making phoneme accuracy directly bankable. The limitation is scope: one read-speech corpus, one language, a lexicon whose 1.78% out-of-vocabulary rate is itself a third of the remaining error. Separating the stages once tells you where the work is.

Disclosure of Interests. The author has no competing interests.

References

1. Brent, M.R.: An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning* **34**(1–3), 71–105 (1999)
2. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: *ICML*, pp. 369–376 (2006)
3. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: LibriSpeech: an ASR corpus based on public domain audio books. In: *ICASSP*, pp. 5206–5210 (2015)
4. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI technical report (2019)
5. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: *ICML*, pp. 28492–28518 (2023)
6. Weide, R.: The CMU pronouncing dictionary, release 0.6 (1998)