Quantifying the Role of Textual Predictability in Automatic Speech Recognition

S. Robertson, G. Penn, E. Dunbar

University of Toronto

INTERSPEECH, Sept 2024

Table of Contents

UNIVERSITY OF TORONTO

What is textual predictability?

- Given *only* part of a transcription, how easy or hard it is to guess the rest.
- Such guesswork is the role of the language model (LM).
- The acoustic model (AM) models whatever is left.
- The textual predictability of transcript $y = y_1, y_2, \ldots, y_L$ may be estimated with the Negative Log Likelihood (NLL) an external LM Q assigns it:

$$H_y = -\frac{1}{L} \log Q(y).$$

Colourless green [redacted] sleep furiously

Why quantify it's role?

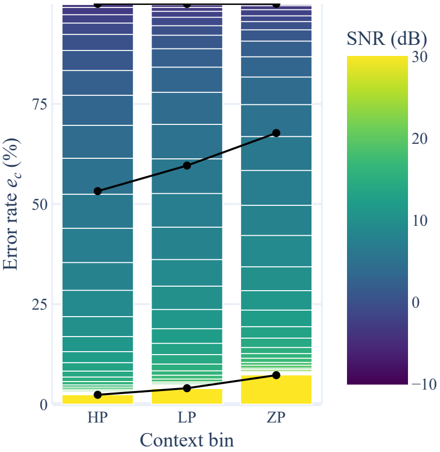
- Recent Automatic Speech Recognition (ASR) thinks it should have a very big role.
    - *E.g.* LLM-style ASR, contextual biasing, aggressive LM fusion...
- Is focusing on LMs enough?
    - The poor performance of ASR on African American English (AAE) was blamed on bad AMs, not LMs [1].
- An easy-to-compute number solves disputes and offers paths forward.

Error rates as a function of NLL

- Previous work focuses on an *absolute* relationship between H_y and e.
- Parametrized by $a, b \in \mathbb{R}^+$, they follow a power law [2]:

$$e_y = b \exp(aH_y).$$

- Fit depends on "acoustic conditions" [2].
  - a increases and b decreases with Signal-to-Noise Ratio (SNR).



UNIVERSITY OF TORONTO

Can we do better?

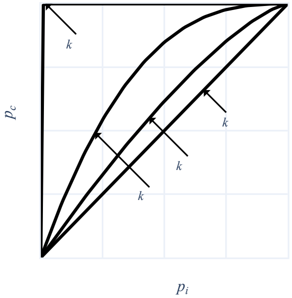
- ASR performance depends on *more* than textual predictability.
- Instead, we compare error rates e_i and e_c of less and more predictable utterances.
- Then the difference is due to predictability.

Relative improvements and k

- The relationship can be modelled with constant exponent $k \geq 1$ [3]:

$$e_c = e_i^k \text{ or } p_c = 1 - (1 - p_i)^k.$$

- k should increase with:
  - *greater predictability* of H_c, and/or
  - *greater reliance* on predictability.



Human subjects experiments

- The k model was verified for human listeners [3]:
  - Test utterances were partitioned into Zero, Low, and High Predictability (ZP, LP, and HP) bins.
  - Utterance i was drawn from the ZP bin; c from either LP or HP.
  - $k \approx 1.38$ between ZP-LP, but $k \approx 2.72$ between ZP-HP.
  - k was robust to adjustments of SNR (and thus error rates).
- k also increases with age [4].

The recipe

To do the same for ASR, we need to pick:

1. the corpora to draw utterances from,
2. an LM Q to bin utterances by predictability, and
3. our ASR systems to be subjects.

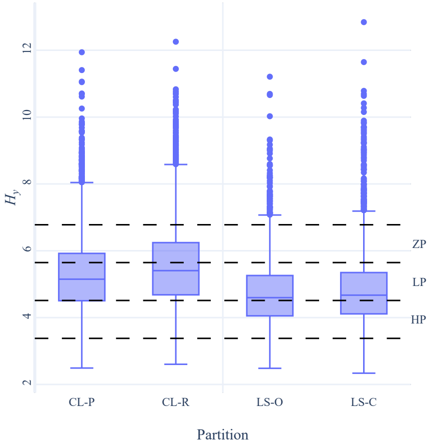
The corpora

- LibriSpeech [5] is our in-domain corpus.
  - Presumed standard/mainstream American English.
  - All ASR systems and LMs are trained on it.
  - Convenience and ecological validity.
  - Tested on dev-clean (LS-C) and dev-other (LS-O) partitions.
- CORAAL [6] is our out-of-domain corpus.
  - Regional AAE corpus from earlier studies [1], [7].
  - Tested on Rochester (CL-R) and Princeville (CL-P) partitions.
- **If ASR under-utilizes LMs on CORAAL, k should decrease.**

UNIVERSITY OF TORONTO

The language model

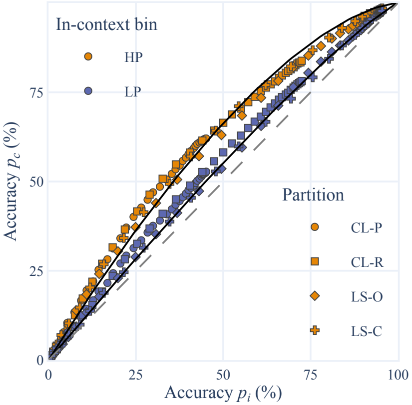
- Compared LMs from Kaldi *s5* recipe [8].
- Took RNN-LM as Q because it had the lowest average NLL.
- Used LS-C partition to make three equal intervals for ZP, LP, and HP bins.
- k **should be higher on ZP-HP pair than on ZP-LP pair.**

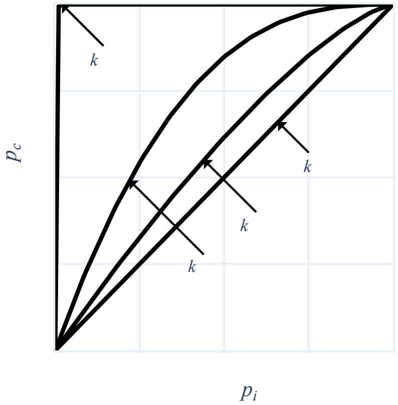


The ASR systems

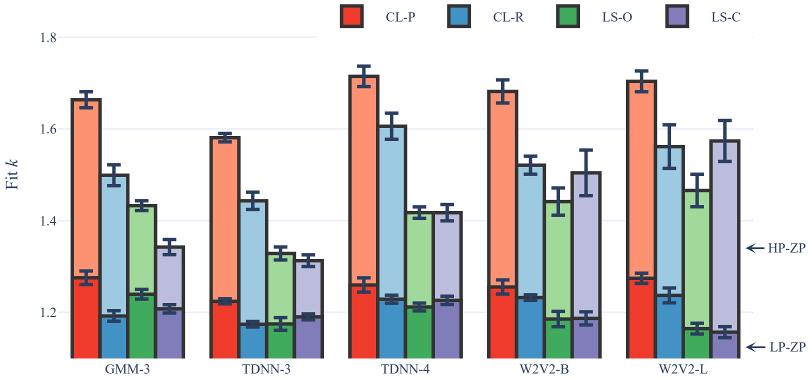
- Tested ASR systems from *s5* with explicit LMs:
  - GMM-3 = Gaussian Mixture Model + 3-gram, and
  - TDNN-{3,4} = Time-Delay Neural Network + {3,4}-gram.
- Tested E2E ASR systems with implicit LMs [9]:
  - W2V2-B = Wav2Vec 2-Base (smaller, LibriSpeech only), and
  - W2V2-L = Wav2Vec 2-Large (bigger, + LibriLight [10]).
- k **should increase with more sophisticated LMs.**

Plotting k


In-context bin
HP
LP
Partition
CL-P
CL-R
LS-O
LS-C
Accuracy p_c (%)
Accuracy p_i (%)
0
25
50
75
100


k
k
k
k
k
p_c
p_i

Fit k



- k reliably increases on ZP-HP comparisons.
- k reliably decreases on 3-gram-based models.
  - Less reliably: W2V2-L $>$ W2V2-B $>$ TDNN-4.
- k reliably increases on CORAAL data.

UNIVERSITY OF TORONTO

Discussion

- k captures *main* effects of textual predictability on error rates.
- Models with fancy LMs rely more on textual predictability.
- Evidence that poor AAE performance is not due to LM under-utilization.
  - k *may* increase, but certainly not *decrease*.
- $k \approx \frac{\log e_c}{\log e_i}$ can substitute for fit.
  - k stabilizes as $e \to 0$.

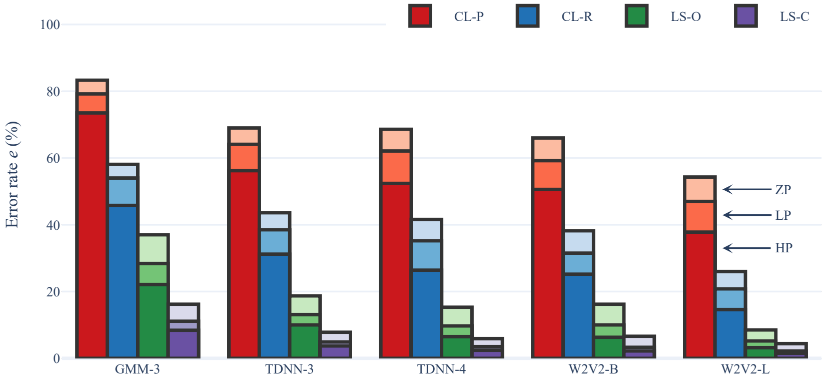
Thank you!

sdrobert@cs.toronto.edu gpenn@cs.toronto.edu ewan.dunbar@utoronto.ca

Repo + slides:

UNIVERSITY OF TORONTO

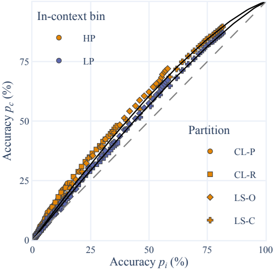
Error rates



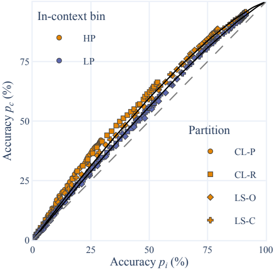
- Error rates decrease with more sophisticated LMs.
- Error rates decrease with increasing predictability.
- Error rates increase on CORAAL.

Regression plots

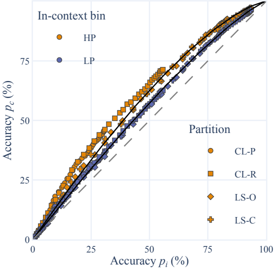
GMM-3



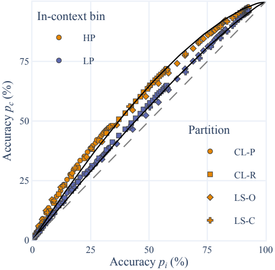
TDNN-3



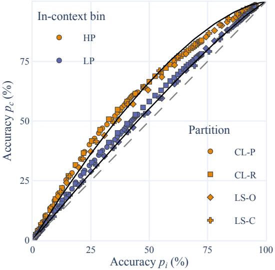
TDNN-4



W2V2-B



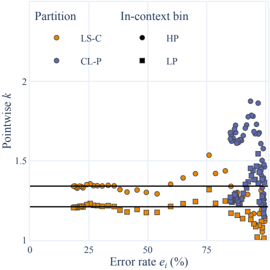
W2V2-L



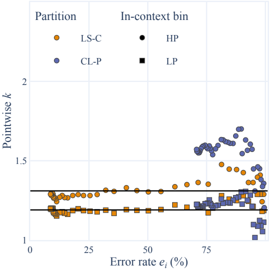
UNIVERSITY OF TORONTO

Point-wise estimates

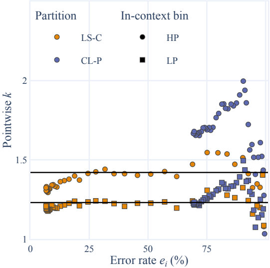
GMM-3



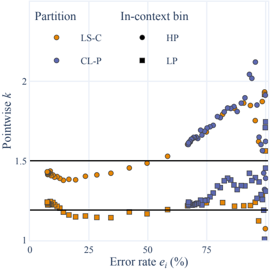
TDNN-3



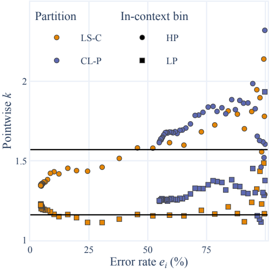
TDNN-4



W2V2-B



W2V2-L



UNIVERSITY OF TORONTO

Looking forward

- Like with humans, k increases as a function of e [4], [11].
- Plenty of other models, settings to explore.
- Other forms of "predictability."

UNIVERSITY OF TORONTO

References I

[1] Allison Koenecke, Andrew Nam, Emily Lake, et al. "Racial disparities in automated speech recognition". In: *Proceedings of the National Academy of Sciences* 117.14 (Apr. 2020), pp. 7684–7689. DOI: 10.1073/pnas.1915768117. (Visited on 12/20/2022).

[2] Dietrich Klakow and Jochen Peters. "Testing the correlation of word error rate and perplexity". In: *Speech Communication* 38.1 (2002), pp. 19–28. ISSN: 0167-6393. DOI: 10.1016/S0167-6393(01)00041-3.

[3] Arthur Boothroyd and Susan Nittrouer. "Mathematical treatment of context effects in phoneme and word recognition". In: *JASA* 84.1 (July 1988), pp. 101–114. ISSN: 0001-4966. DOI: 10.1121/1.396976.

[4] Susan Nittrouer and Arthur Boothroyd. "Context effects in phoneme and word recognition by young children and older adults". In: *JASA* 87.6 (June 1990), pp. 2705–2715. ISSN: 0001-4966. DOI: 10.1121/1.399061. (Visited on 02/13/2024).

UNIVERSITY OF TORONTO

References II

[5] Vassil Panayotov, Guoguo Chen, Daniel Povey, et al. "Librispeech: An ASR corpus based on public domain audio books". In: *ICASSP*. 2015, pp. 5206–5210. ISBN: 2379-190X. DOI: 10.1109/ICASSP.2015.7178964.

[6] Tyler Kendall and Charlie Farrington. *The corpus of regional African American language*. 2023. DOI: 10.7264/1ad5-6t35.

[7] Joshua L. Martin and Kevin Tang. "Understanding racial disparities in automatic speech recognition: The case of habitual "be"". In: *Proc. Interspeech 2020*. 2020, pp. 626–630. DOI: 10.21437/Interspeech.2020-2893.

[8] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, et al. "The Kaldi speech recognition toolkit". In: *ASRU*. Hilton Waikoloa Village, Big Island, Hawaii, US: IEEE Signal Processing Society, 2011.

UNIVERSITY OF TORONTO

References III

[9] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations". In: *NeurIPS*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, et al. Vol. 33. Curran Associates, Inc., 2020, pp. 12449–12460.

[10] J. Kahn, M. Rivière, W. Zheng, et al. "Libri-light: A benchmark for ASR with limited or no supervision". In: *ICASSP*. 2020, pp. 7669–7673. DOI: 10.1109/ICASSP40776.2020.9052942.

[11] Adelbert W. Bronkhorst, Thomas Brand, and Kirsten Wagener. "Evaluation of context effects in sentence recognition". In: *JASA* 111.6 (June 2002), pp. 2874–2886. ISSN: 0001-4966. DOI: 10.1121/1.1458025. (Visited on 02/02/2024).

UNIVERSITY OF TORONTO