Interspeech 2024
1-5 September 2024, Kos, Greece

Quantifying the Role of Textual Predictability in Automatic Speech Recognition

Sean Robertson, Gerald Penn, Ewan Dunbar

University of Toronto, Canada
sdrobert@cs.toronto.edu, gpenn@cs.toronto.edu, ewan.dunbar@utoronto.ca

Abstract

A long-standing question in automatic speech recognition research is how to attribute errors to the ability of a model to model the acoustics, versus its ability to leverage higher-order context (lexicon, morphology, syntax, semantics). We validate a novel approach which models error rates as a function of relative textual predictability, and yields a single number, k, which measures the effect of textual predictability on the recognizer. We use this method to demonstrate that a Wav2Vec 2.0-based model makes greater stronger use of textual context than a hybrid ASR model, in spite of not using an explicit language model, and also use it to shed light on recent results demonstrating poor performance of standard ASR systems on African-American English. We demonstrate that these mostly represent failures of acoustic–phonetic modelling. We show how this approach can be used straightforwardly in diagnosing and improving ASR.

Index Terms: speech recognition, perplexity, entropy, language model, acoustic model, accent-robust speech recognition, African American English

1. Introduction

Recent work has highlighted the difficulties automatic speech recognition (ASR) systems continue to have with minority and racialized language varieties. However, while all studies agree that the ultimate source of the problem is the change in domain—ASR training is generally on dominant language varieties—explanations for issues with African-American English in particular vary, with some arguing that many issues stem from morphological and vocabulary differences [1], while others that phonetic differences are the main source [2, 3]. These questions put into relief a long-standing question in ASR: how to assess how much a system relies on textual predictability (traditional "language modelling") versus modelling of the phonetic signal (traditional "acoustic modelling"). This has become difficult to resolve with the advent of powerful end-to-end models deploying context at long distances, reducing or eliminating the need for explicit language models.

We develop a new method for quantifying the role of textual predictability in ASR, starting from a psychoacoustic paradigm developed by Boothroyd and Nittrouer [4]. We validate the use of this framework for automatic (as opposed to human) speech recognition by demonstrating that utterances with different degrees of textual predictability yield increasing values of k. We also show that a more powerful explicit language model yields higher values of k, indicating stronger textual prediction.

We apply the method to comparing ASR models that we expect to have different intrinsic capacities for contextual predictability (GMM, TDNN, Wav2Vec 2.0-base, and Wav2Vec 2.0-large), demonstrating that k also increases with more powerful models. We also apply the method to an African-American English corpus [5], reaching a similar conclusion to previous works [2, 3]: the difficulties faced by ASR systems with these language varieties mainly reflect issues with acoustic modelling. We provide a recipe for using this method to diagnose issues and improve performance in ASR, and discuss its limitations. All of our code and results are open source and available at `https://github.com/sdrobert/kaldi-boothroyd`.

2. Background

2.1. ASR and textual predictability

Textual predictability in ASR is typically measured using perplexity as measured by some language model (LM). For the distribution Q induced by an LM, perplexity is the exponent of the negative log likelihood (NLL) H_y of a token sequence $y = y_1, y_2, \ldots, y_L$, formally:

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

Equation (1) is an estimate of the cross-entropy rate $\mathbb{E}_y[H_y]$ of Q relative to the population distribution P which generates y [6]. Since a lower H_y implies a higher $Q(y)$, the NLL measures how well Q predicts y, and, if we average H_y over a corpus drawn from P, how well Q predicts P.

We expect NLL calculated with respect to Q to be correlated with ASR accuracy: whether the ASR system uses an explicit language model following Q or not, assuming that the system is trained on data following P, the system has an implicit marginal textual distribution Q_{sys}: for a transcription y, and where $x \in \mathcal{X}$ is the set of all possible utterances:

$$Q_{sys}(y) = \sum_{x \in \mathcal{X}} Q_{sys}(y|x) P(x) \tag{2}$$

Because of the shared training data, we expect Q_{sys} to be fairly close to both P and to some LM distribution Q.

Indeed, NLL was proposed as a measure of the intrinsic difficulty of transcribing an utterance [7], with some attempts at modelling the relationship between ASR error rates e_y, $0 \leq e_y \leq 1$, and H_y [8, 9, 10]. Klakow and Peters [9] suggest the following power law relationship with fit coefficients $a, b \in \mathbb{R}$, that is, log error rates being proportional to H_y:

$$e_y = b \exp(a H_y). \tag{3}$$

Equation (3) would be a strong candidate for quantifying the role of textual predictability on ASR performance were it not sensitive to "acoustic conditions." As remarked by Klakow and

10.21437/Interspeech.2024-552

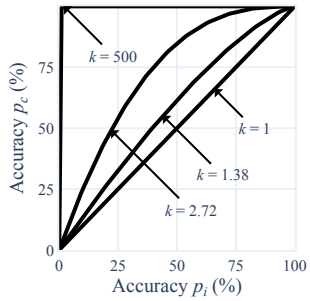


Figure 1: *Accuracy ratios across k from Equation* (4)*.*

Peters [9], the coefficient a decreases (while b grows) as acoustic conditions become more "challenging." Thus, Eq. (3) is unlikely to generalize across corpora. Rather than attempt to link NLL directly to performance, we propose to work using ratios, relating *relative* predictability to *relative* performance. Furthermore, we construct a measure which is aggregated over acoustic conditions of increasing difficulty, in an attempt to further factor out the role of acoustics.

2.2. Predictability and performance

Our method is based on the experimental paradigm of Boothroyd and Nittrouer [4], in which participants recognized sentences across three conditions: zero predictability (**ZP**)—words drawn randomly—low predictability (**LP**)—grammatical but semantically strange—and high predictability (**HP**). Error rates e and accuracies $p = 1 - e$ were computed per condition, inducing errors by masking the speech over a range of signal-to-noise ratios (SNRs). Treating **ZP** as the "isolated" condition i and either **LP** or **HP** as the "context" condition c, the authors found that error rates were related by a constant exponent k, regardless of the SNR range:

$$e_c = e_i^k, \text{ or } p_c = 1 - (1 - p_i)^k \tag{4}$$

Figure 1 illustrates the relation between p_c, p_i, and k, where variation in accuracy is induced by varying SNR. $k = 1$ means the listener is not using the additional predictability of condition c to compensate for acoustics, whereas for $k = 500$, the listener leverages so much context as to make acoustics irrelevant. In [4], a greater gap in predictability led to greater k: between **ZP**–**LP**, $k \approx 1.38$, and between **ZP**–**HP**, $k \approx 2.72$. The design is easily transposed to ASR. While the result that k is independent of SNR range has not always held up with human listeners [11, 12, 13, 14], we show it is a useful approximation for ASR.

Our method is as follows. Given an evaluation corpus, we split it into bins by textual predictability. We bin using the NLL of a LM trained on the same distribution as the target system. Call the training distribution P_{train}, the binning LM distribution Q_{bin}, and the marginal textual distribution of the target system Q_{sys}. We divide into three (or more) bins: one reference (i) bin (for which we continue the misnomer **ZP**), and two bins, **LP** (more predictable than **ZP**) and **HP** (much more predictable).

On in-domain data, we expect higher k when c is **HP** than when it is **LP**. Systems that, intuitively, "rely more on language modelling" are those for which Q_{sys} is closer to P_{train}. We expect a pronounced gap between **HP** and **LP** for these systems.

For evaluation data following an unknown distribution P_{test}, we keep the same NLL cuts, and continue to use Q_{bin} (trained on P_{train}). We consider two different cases. If the percentage of the new corpus in each bin is very different than an in-domain corpus, this suggests that Q_{bin} is severely mismatched to P_{test}. We can assume that Q_{sys} is also mismatched. The k calculated on in-domain data then tells us how sensitive the system should be to this textual domain shift: $k = 1$ should be insensitive, extremely high k should be catastrophic. On the other hand, it is possible that the bin frequencies reveal no major mismatch between Q_{bin} and P_{test}. Since we are extrapolating to P_{test}, this provides no guarantee that the ASR system is well-matched. However, if Q_{sys} is a poor match to P_{test}, we predict that k should be lower on out-of-domain than in-domain data, approaching 1.

3. Experiments

3.1. Materials and systems

We experiment with LMs and ASR systems from Kaldi's *s5* recipe [15] and Wav2Vec 2.0 [16]. Kaldi LMs[1] include: a pruned, word-level, 3-gram LM with modified Kneser-Ney smoothing [10]; similarly, an un-pruned 4-gram and a word-level, recurrent neural network (RNN) [17]. Kaldi acoustic models include a speaker adaptive Gaussian mixture model (GMM) and a time-delay neural network (TDNN).[2] We denote GMM and TDNN ASR systems with lattices weighted by the 3-gram LM as **GMM-3** and **TDNN-3**, respectively. We denote the ASR system which re-scores **TDNN-3** lattices with the 4-gram LM as **TDNN-4**. We also apply two fine-tuned Wav2Vec 2.0 models available freely online. The "base" variant (**W2V2-B**), features 12, smaller Transformer layers and is trained on LibriSpeech [18].[3] The "large" variant, denoted **W2V2-L**, has 24, larger Transformer layers and has been additionally trained on LibriLight [19]. Both have been fine-tuned for ASR on LibriSpeech with a CTC objective [20]. We use greedy decoding without an external LM for simplicity. The hybrid models offer explicit control of the amount of language modelling the ASR system is doing, speaking directly to our hypotheses. Wav2Vec 2.0 allows us to explore the role of implicit textual prediction: since these networks use global self-attention [21], we expect them to use predictive context more aggressively.

Within-domain, we compute error rates and fit k values on LibriSpeech's *dev-clean* and *dev-other* partitions (**LS-C** and **LS-O** respectively). Following prior work [1, 2], we expect these systems to under-perform on utterances from the Corpus of Regional African American Language (CORAAL) [5, version 2023.06]. In particular, we focus on the utterances of speakers from Rochester, New York (**CL-R**) and Princeville, North Carolina (**CL-P**), on which [2] reported the lowest ($e_y = 0.20$) and highest ($e_y = 0.38$) error rates from the corpus, respectively.[4]

[1]Available at `https://kaldi-asr.org/models/m13` and `https://openslr.org/11/`, last accessed February 19, 2024.

[2]The TDNN is available at `https://kaldi-asr.org/models/m13`, last accessed February 19, 2024. We have uploaded our re-trained GMM, denoted *tri6b* in the Kaldi s5 recipe, to our repository.

[3]Available at `https://huggingface.co/facebook/wav2vec2-base-960h`, last accessed February 19, 2024.

[4]Compared to Koenecke et al. [2], we sanitize the partitions more aggressively to more closely resemble a standard ASR benchmark: any utterances containing restarts, fillers, unintelligible markers, non-speech noise, and so forth are excluded from consideration. After filtering, the CL-R and CL-P partitions contain roughly 4 and 3 hours of speech, respectively. Filtering is reproducible from our code base.

Table 1: *Word error rates e_y, reported as a percentage. Rows are grouped by partition and NLL bin; columns by model. The* all *rows contain the error rates over the entire partition, without binning.*

		GMM-3	TDNN-3	TDNN-4	W2V2-B	W2V2-L
LS-C	HP	8.4	3.7	2.4	2.2	**1.5**
	LP	11.1	4.9	3.5	3.3	**2.2**
	ZP	16.2	7.8	5.9	6.6	**4.4**
	all	10.5	4.7	3.3	3.3	**2.2**
LS-O	HP	22.1	10.0	6.5	6.3	**3.2**
	LP	28.4	13.1	9.7	10.0	**5.2**
	ZP	37.0	18.7	15.3	16.2	**8.5**
	all	26.1	12.2	8.7	8.8	**4.6**
CL-R	HP	45.8	31.2	26.4	25.2	**14.6**
	LP	54.0	38.5	35.2	31.5	**20.8**
	ZP	58.1	43.6	41.6	38.2	**26.0**
	all	53.9	37.3	33.9	32.8	**23.2**
CL-P	HP	73.5	56.2	52.4	50.6	**37.8**
	LP	79.2	64.1	62.1	59.2	**47.0**
	ZP	83.3	69.0	68.6	66.0	**54.3**
	all	78.7	61.9	59.4	58.2	**46.2**

3.2. Procedure

Utterances are first corrupted by noise over a range of SNRs, and decoded by each ASR system, on each corpus partition. We follow the procedure of Zhang et al. [22] for introducing noise to utterances. Each recording is first normalized to a fixed reference power and 0 DC. Then, white noise of SNRs between -10 and 30 dB is added to each recording. As Zhang et al. found that different types of generated noise lead to similar accuracies at similar SNRs, we did not experiment with different types of noise. For the sake of our analysis, it is sufficient that noise degrades acoustic conditions consistently across NLL bins.

For binning, following Section 2.1, an LM with a low NLL is considered close to the training distribution P_{train}. We used the RNN LM to generate the bins as it produced the lowest NLL on **LS-C** and **LS-O**. The cutpoints (same across corpora) were obtained by evenly splitting per-utterance NLL from **LS-C** into three intervals. Because the tails of the distribution were long, we dropped the top and bottom 5% of NLLs before constructing the bins. The **HP** bin covers $H_y \in (3.4, 4.5]$, the **LP** bin $H_y \in (4.5, 5.6]$, and the **ZP** bin (i condition) $H_y \in (5.6, 6.8]$.

To estimate k, we calculate a single error rate e per system/corpus/bin triplet. We take the **ZP** rate e_i as our "isolated" or "no-context" condition and either the **LP** or **HP** rate e_c as our "context" condition, fitting k to Eq. (4). We perform non-linear least-squares regression to fit $e_c = e_i^k$. We could perform ordinary least squares on $\ln e_c = k \ln e_i$ instead, but because the residuals are smaller when $e_c \approx 1$, we found this biased the fit to the lowest values of e_c. To compute 95% confidence intervals for each fit of k, we rely on the Wild bootstrap [23]: for each of $B = 9999$ iterations, we resample the log-space residuals $\widehat{\epsilon} = \epsilon V$, where $V \sim \mathcal{N}(0, 1)$, re-compute $\ln \widehat{e}_c = k \ln e_i + \widehat{\epsilon}$, and re-fit $\widehat{k}$. The log-space ensures $\widehat{e}_c > 0$; multiplication ϵV maintains heteroskedasticity of the residuals.

4. Results

Table 1 lists average error rates, without noise. In general, **GMM-3** has the most errors, then **TDNN-3**, **TDNN-4**, **W2V2-B**, and **W2V2-L** the fewest. Further, as expected [7, 9], error rates increase as a function of NLL. Finally, we note wide disparity between LibriSpeech and CORAAL.

Table 2: *Estimated k and bootstrapped 95% confidence intervals. The first block lists fit k values per partition, averaged over models. The second block is per model, averaged over partitions. The* all *row aggregates all models per partition.*

		HP		LP	
		k	CI	k	CI
LS-C	GMM-3	1.34	[1.33, 1.36]	1.21	[1.20, 1.22]
	TDNN-3	1.31	[1.30, 1.33]	1.19	[1.18, 1.20]
	TDNN-4	1.42	[1.40, 1.43]	1.23	[1.22, 1.24]
	W2V2-B	1.50	[1.45, 1.56]	1.19	[1.17, 1.20]
	W2V2-L	1.57	[1.53, 1.62]	1.16	[1.14, 1.17]
	all	1.40	[1.38, 1.42]	1.20	[1.19, 1.20]
LS-O	GMM-3	1.43	[1.42, 1.44]	1.24	[1.23, 1.25]
	TDNN-3	1.33	[1.31, 1.34]	1.17	[1.16, 1.19]
	TDNN-4	1.42	[1.41, 1.43]	1.21	[1.20, 1.22]
	W2V2-B	1.44	[1.41, 1.47]	1.19	[1.17, 1.20]
	W2V2-L	1.47	[1.43, 1.50]	1.16	[1.15, 1.18]
	all	1.41	[1.40, 1.42]	1.20	[1.19, 1.20]
CL-R	GMM-3	1.50	[1.48, 1.52]	1.19	[1.18, 1.20]
	TDNN-3	1.44	[1.42, 1.46]	1.17	[1.17, 1.18]
	TDNN-4	1.61	[1.58, 1.63]	1.23	[1.22, 1.24]
	W2V2-B	1.52	[1.50, 1.54]	1.23	[1.23, 1.24]
	W2V2-L	1.56	[1.51, 1.61]	1.24	[1.22, 1.25]
	all	1.53	[1.51, 1.54]	1.22	[1.21, 1.22]
CL-P	GMM-3	1.66	[1.65, 1.68]	1.28	[1.26, 1.29]
	TDNN-3	1.58	[1.57, 1.59]	1.22	[1.22, 1.23]
	TDNN-4	1.71	[1.69, 1.74]	1.26	[1.24, 1.28]
	W2V2-B	1.68	[1.66, 1.71]	1.26	[1.24, 1.27]
	W2V2-L	1.70	[1.68, 1.73]	1.27	[1.26, 1.29]
	all	1.67	[1.66, 1.69]	1.26	[1.25, 1.26]

Table 3: *Proportion of partition captured by each NLL bin (%). The* total *column sums each row.*

	HP	LP	ZP	total
LS-C	37.2	40.1	12.6	89.9
LS-O	39.9	39.5	11.2	90.6
CL-R	18.3	38.7	26.9	84.0
CL-P	23.8	41.9	23.4	89.1

Table 2 lists k by partition, model, and in aggregate. Columns represent the choice of "context" bin e_c, which is either **LP** or **HP**; the "isolated" bin is always **ZP**. We concentrate on in-domain (**LS**) first. In all cells, $k > 1$, and confidence intervals do not include $k = 1$: textual predictability plays a role in error rates. Furthermore, k is higher when fit to the **HP** error rates than the **LP** rates—k increases as a function of predictability. Finally, on **HP**, there is a divide between models using a 3-gram LM versus the more sophisticated models, particularly **W2V2-L**, with the latter yielding higher k values.

Next we consider the out-of-domain (**CL**) data. First, Table 3 tabulates the proportion of utterances per partition captured in each bin. The two **LS** partitions have similar proportions in each bin. On CORAAL, the vast majority of the data remain in the **HP** and **LP** bins, in line with the observation of Koenecke et al. [2] that CORAAL and LibriSpeech transcriptions are more similar than different. Nevertheless, the mass shifts toward the **ZP** bin, raising the possibility that CORAAL error rates could be affected by textual predictions, as per Martin and Tang [1]. In the absence of previous studies, it is difficult to say how big such an effect could be: while in-domain k values are greater than 1, they are far from the catastrophic $k = 500$ case. Returning, then, to Table 2, we recall that, in

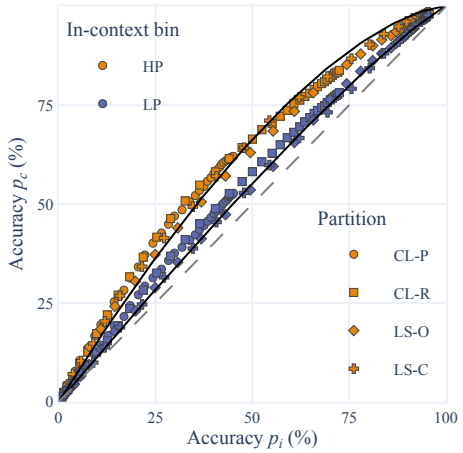


Figure 2: *In-context* vs. *isolated accuracies of* ***W2V2-L****. The grey, dashed line is $y = x$. Black lines mark the interpolated fits over* ***LS-C*** *from Table 2: the shallow curve is* ***LP****; the steep curve is* ***HP****.*

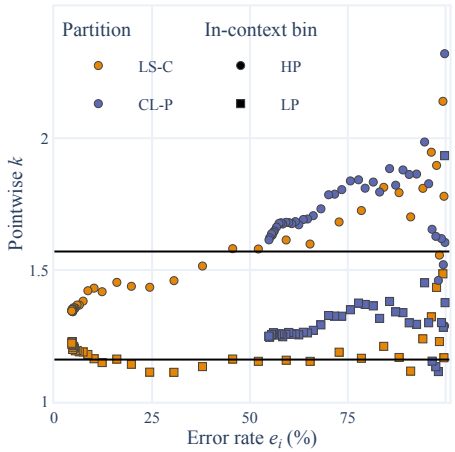


Figure 3: *Point-wise estimates of* $k = \ln e_c / \ln e_i$ vs. *error rates* e_i *of* ***W2V2-L****. Each point is paired by SNR and partition. Black lines mark the interpolated fits from Table 2.*

cases of mismatch between Q_{sys} and the new domain P_{test}, we expect k to go down, approaching 1. In fact, in general, k values are *higher* on CORAAL. Thus, we reach a similar conclusion to Koenecke et al. [2]: when applying (these) ASR systems to African-American English, the effect of textual domain shift is limited—at least as measured by our approach. Further research should be done to explore the relation between in-domain k and out-of-domain performance.

As discussed above, we know that, for human listeners, k as calculated using Equation (4), does not give a perfect fit to the data. Figure 2 plots accuracy in the isolated versus the context condition at a fixed SNR on **W2V2-L**.[5] Colour and shape distinguish context bin and data partition, respectively. Black lines mark the fit k to Eq. (4) on **LS-C**. Data from all partitions follow a similar curve. The fitted k is in broad agreement with this curve, though it overestimates at low isolated accuracies and underestimates at high isolated accuracies.

To better illustrate this, in Figure 3 we compare the point-wise estimates of k versus isolated error rates e_i on **W2V2-L**. The two fitted values of k to **LS-C** are shown as lines. Were Eq. (4) a perfect fit, the point-wise estimates would follow horizontal lines: rather, point-wise k changes as a function of e_i (SNR), peaking at around 80-90% errors. Generally, the fitted k tends to match the point-wise estimates for $e_i \approx 0.5$. We reason that point-wise k in this region are the best approximation for the global fit. Since error rates rarely fall below 50% on CORAAL, its k values in Table 2 are likely inflated. Nonetheless, as Fig. 3 illustrates, point-wise k on **CL** routinely match or exceed **LS**, indicating no less an impact of textual predictability.

5. Limitations

Our profile of a small subset of ASR systems and types of noise may lead to an incomplete picture of the behaviour and utility of k. The choice of the LM measuring predictability may not be anodyne: mismatch between LM and ASR induced by different vocabularies (words, sub-words, characters) is not uncommon, and certainly worth exploring. Though k fits the data well as a single-parameter, interpretable estimate of the effects of predictability on ASR performance, as mentioned in Sections 2 and 4, the model itself is simplistic. k fails to capture the fact that increased SNRs lead to higher k, in both humans and ASR systems. More complicated models of predictability involving combinatorics of errors could be fit to account for the failures of k [see 13]. Indeed, it could be the case that, as e_i approaches 0, so, too, do differences between ASR systems.

[5] Analogous plots to Figs. 2 and 3 for other models are included in the supplementary material.

6. Summary and Discussion

It has been long understood that textual predictability plays an important role in ASR performance [7], but little has been done to quantify this. We have shown that the impact of textual predictability on ASR performance can be quantified by estimating a global ratio of log errors, k, across a range of acoustic conditions, showing a reliable increase in k as the gap in predictability rises. We also see k increase as a function of the ASR systems' (implicit or explicit) language modelling capacity. For example, as Wav2vec 2.0-Large (**W2V2-L**) [16] has $k \approx 1.6$, greater than the other systems tested, we conclude that it depends strongly on textual predictability for its performance.

When applied to the Corpus of Regional African-American Language [5], all systems' k values increased. Though these out-of-domain data were less predictable to LMs trained on in-domain data (Librispeech, [18]), *pace* Martin and Tang [1], higher k indicate that this disagreement is slight. We interpret this as supporting the notion that, for this case, improvements to ASR should focus on acoustic modelling [2, 3].

We propose k as a crucial complement to error rates in ASR research. We recognize that, as a general-purpose tool, the calculation of k by decoding on a wide range of SNRs can be cumbersome. The following simplified recipe may be employed: **(1)** split a corpus into high and low NLL based on an LM trained on textually similar data; **(2)** add noise until the high-NLL condition yields an error rate of around 50%; **(3)** estimate k as the ratio of point-wise log error rates. The 50% point follows from Fig. 3, but any reference rate may be used that is large enough to permit improvement in error rates. We hope this straightforward recipe will push researchers to carefully weigh their options when choosing what aspects of ASR models are most worth improving, and which are already close to being optimal.

7. Acknowledgements

This research is funded by the Data Sciences Institute at the University of Toronto and by the Natural Sciences and Engineering Research Council of Canada (NSERC) RGPIN-2022-04431. It was also enabled in part by support provided by Compute Ontario (`https://www.computeontario.ca/`) and the Digital Research Alliance of Canada (`https://alliancecan.ca/`).

8. References

[1] J. L. Martin and K. Tang, "Understanding racial disparities in automatic speech recognition: The case of habitual "be"," in *Proc. Interspeech 2020*, 2020, pp. 626–630.

[2] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, Apr. 2020.

[3] A. B. Wassink, C. Gansen, and I. Bartholomew, "Uneven success: automatic speech recognition and ethnicity-related dialects," *Speech Communication*, vol. 140, pp. 50–70, 2022.

[4] A. Boothroyd and S. Nittrouer, "Mathematical treatment of context effects in phoneme and word recognition," *JASA*, vol. 84, no. 1, pp. 101–114, Jul. 1988.

[5] T. Kendall and C. Farrington, "The corpus of regional African American language," 2023.

[6] C. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[7] L. R. Bahl, F. Jelinek, and R. L. Mercer, "A maximum likelihood approach to continuous speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-5, no. 2, pp. 179–190, 1983.

[8] H. Printz and P. A. Olsen, "Theory and practice of acoustic confusability," *CSL*, vol. 16, no. 1, pp. 131–164, Jan. 2002.

[9] D. Klakow and J. Peters, "Testing the correlation of word error rate and perplexity," *Speech Communication*, vol. 38, no. 1, pp. 19–28, 2002.

[10] S. F. Chen, D. Beeferman, and R. Rosenfeld, "Evaluation metrics for language models," Jan. 2008.

[11] S. Nittrouer and A. Boothroyd, "Context effects in phoneme and word recognition by young children and older adults," *JASA*, vol. 87, no. 6, pp. 2705–2715, Jun. 1990.

[12] K. W. Grant and P. F. Seitz, "The recognition of isolated words and words in sentences: Individual variability in the use of sentence context," *JASA*, vol. 107, no. 2, pp. 1000–1011, Feb. 2000.

[13] A. W. Bronkhorst, A. J. Bosman, and G. F. Smoorenburg, "A model for context effects in speech recognition," *JASA*, vol. 93, no. 1, pp. 499–509, Jan. 1993.

[14] A. W. Bronkhorst, T. Brand, and K. Wagener, "Evaluation of context effects in sentence recognition," *JASA*, vol. 111, no. 6, pp. 2874–2886, Jun. 2002.

[15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *ASRU*. Hilton Waikoloa Village, Big Island, Hawaii, US: IEEE Signal Processing Society, Dec. 2011.

[16] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460.

[17] H. Xu, K. Li, Y. Wang, J. Wang, S. Kang, X. Chen, D. Povey, and S. Khudanpur, "Neural network language modeling with letter-based features and importance sampling," in *ICASSP*, 2018, pp. 6109–6113.

[18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP*, Apr. 2015, pp. 5206–5210.

[19] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, "Libri-light: A benchmark for ASR with limited or no supervision," in *ICASSP*, 2020, pp. 7669–7673.

[20] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling unsegmented sequence data with recurrent neural networks," in *ICML*. New York, NY, USA: ACM, 2006, pp. 369–376.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 5998–6008.

[22] P. Zhang, Y. Huang, C. Yang, and W. Jiang, "Estimate the noise effect on automatic speech recognition accuracy for mandarin by an approach associating articulation index," *Applied Acoustics*, vol. 203, p. 109217, 2023.

[23] C.-F. J. Wu, "Jackknife, bootstrap and other resampling methods in regression analysis," *The Annals of Statistics*, vol. 14, no. 4, pp. 1261–1295, Dec. 1986.