Anna Sophia Stein & Kevin Tang. Accepted. Modeling probabilistic reduction across domains with Naive Discriminative Learning. In Proc. Interspeech 2024. Kos, Greece. Forthcoming

Modeling probabilistic reduction across domains with Naive Discriminative Learning

Anna Stein[1], *Kevin Tang*[2,3]

[1]Institute of Linguistics / [2]Department of English Language and Linguistics, Institute of English and American studies, Faculty of Arts and Humanities, Heinrich Heine University Düsseldorf, Germany
[3]Department of Linguistics, University of Florida

anna.stein@hhu.de, kevin.tang@hhu.de

Abstract

The predictability of a word modulates its acoustic duration. Such probabilistic effects can compete across linguistic domains (segments, syllables and adjacent-word contexts e.g., frequent words with infrequent syllables) and across local and aggregate contexts (e.g., a generally unpredictable word in a predictable context). This study aims to tease apart competing effects using Naive Discriminative Learning, which incorporates cue competition. The model was trained on English conversational speech from the Buckeye Corpus, using words as outcomes and segments, syllables, and adjacent words as cues. The connections between cues and outcomes were used to predict acoustic word duration. Results show that a word's duration is more strongly predicted by its syllables than its segments, and a word's predictability aggregated over all contexts is a stronger predictor than its specific local contexts. Our study presents a unified approach to modeling competition in probabilistic reduction.

Index Terms: Discriminative Learning, Speech production, Acoustic Duration

1. Introduction

1.1. Probabilistic Reduction

Several probabilistic effects mediate acoustic duration in speech production. This phenomenon, called probabilistic reduction, has been observed at multiple levels of the speech signal. [1] describes probabilistic reduction as 'a general behavior over multiple linguistic levels', where the term 'level' or 'domain' can refer to words, but also other linguistic units like syllables and segments.

Generally, research has shown that unpredictable words are pronounced longer compared to predictable words, which are more likely to be reduced. This effect is traceable from various probabilistic measures: local ones, such as contextual predictability, as well as global ones, like average contextual predictability (informativity) and frequency. These probabilistic effects have been observed to influence duration at several levels of the speech signal: segments [2, 3], morpheme [4], syllable [5, 6], word [7, 8, 4].

Because probabilistic reduction manifests itself across multiple levels of speech elements, it is difficult to trace back to one clear source. For example, [3] found that in English, the probability of a segment affects its acoustic duration. However, due to the embedded structure of speech units - where words comprise syllables, and syllables consist of segments - any reduction that is visible at the word level could be accredited to reduction effects at the syllable or segment level.

These findings show that possibly conflicting probabilistic effects may manifest themselves across linguistic domains, such as a frequent word with infrequent syllables, and across local (e.g., contextual predictability) and global effects (e.g., informativity).

1.2. Models of probabilistic reduction effects

Count-based probabilistic measures, such as frequency, can account for probabilistic reduction effects; however, it is unlikely that speakers actively track word counts. This raises questions about how to integrate such measures into models of speech production. Traditional forward-sequential models, like the Levelt model [9], struggle to incorporate interactions between levels of encoding, limiting their ability to address probabilistic effects. Conversely, models such as the WEAVER model [10], which utilize 'activation' mechanisms for concepts or other representations, can more readily incorporate notions of frequency. The Directions Into Velocities of Articulators (DIVA) model [11], although focused on motor control, does not explicitly account for probabilistic effects in speech production.

Building on the challenges of traditional models, recent research has explored novel approaches grounded in discriminative learning [12, 13]. These approaches, exemplified by models such as those proposed by [14, 15, 16], have been successfully used in production studies [17, 13, 18] while being fully computationally implemented. Notably, these models do not inherently assume any feature representation, allowing the use of diverse feature sets, a crucial advantage in teasing apart combinatorial effects. This flexibility enables dynamic modeling of reduction effects based on learned cues rather than static, corpus-derived counts, presenting a psychologically and cognitively motivated approach for understanding on-line processes during speech production.

Building upon this framework, previous studies [18, 19] have explored reduction effects through computational implementations of discriminative learning, such as investigating the impact of morphology and context on word-final <s> in English. They found that the duration of <s> is linked to its morphological function. Their model incorporates cues of mixed domains, e.g., diphones and word lemmas [18]. However, to the authors' best knowledge, there is no research on the explicit comparison of several domains predicting word duration.

1.3. The current study

In this study, we use Naive Discriminative Learning (NDL) to model probabilistic reduction effects dynamically, reflecting the competitive nature of cue interactions. Specifically, we train an NDL model using transcribed interviews from the Buckeye

corpus of conversational speech[1]. During training, the model learns associations between a word and its segments, syllables, and local context. Throughout learning, the associations are iteratively updated by the learning algorithm of NDL, which includes a cue competition mechanism. To discern the relative contribution of the domains (segments, syllables, context), we derive local and global predictors from the association weights for a regression analysis to predict the word duration from the Buckeye corpus. Our findings reveal that local predictors exhibit a larger effect size compared to global predictors. Furthermore, regarding the domains, we show that the segment predictors have the smallest effect size, while the syllable and context domains perform similarly. Our results contribute to clarifying the importance of certain elements of speech in production, in particular, the significance of the segment, syllable, and context domains, therefore further characterising on-line probabilistic reduction effects.

2. Materials and Method

All scripts for the data processing and statistical analyses are publicly available on GitHub[2].

2.1. Corpus data

Word tokens, along with their respective duration and context, were extracted from the Buckeye corpus of conversational speech. The corpus consists of conversational speech from interviews with 40 speakers from Columbus, Ohio. Each speaker has up to four interview tracks with different conversations, all containing time-aligned phonetic labels and part-of-speech tags. All word tokens for all 40 speakers were extracted using the *Buckeye* package (v1.3)[3], which provides word and pause tags for the corresponding tags in the Buckeye corpus. Pause entries that received the word class were omitted.

2.2. Naive Discriminative Learning

Naive Discriminative Learning (NDL) is a type of discriminative learning that aims to model the effects of implicit or low-level learning, as opposed to logical reasoning processes used by adults [16]. NDL's learning algorithm is based on Pavlovian conditioning, which is rooted in human and animal learning. The model is trained by learning events, which consist of cues connected to an outcome. Cues are indicators that the outcome is going to follow. In each learning event, cues can either be present or absent. This binary representation determines whether the connection strength of the cues to the outcome in the learning event will be lowered or increased (see [20, 16] for an in-depth explanation of the learning algorithm).

Additionally, cue competition influences the connection strength of a cue and an outcome. Cues compete against each other to be the best (most informative) predictor of an outcome. When one cue appears with an outcome in a given learning event, but the other cues that have previously appeared with the outcome do not, their connection strength decreases.

[1] https://buckeyecorpus.osu.edu/BuckeyeCorpusmanual.pdf

[2] https://github.com/ansost/ModelingProbabilisticReduction

[3] https://github.com/scjs/buckeye

2.3. Model training and feature extraction

We use *pyndl* [15], the Python implementation of NDL. This implementation considers words as outcomes, with their respective syllables, segments, and adjacent context words serving as cues. Each learning event within our model comprises the word as an outcome, along with its associated segments, syllables, and context (including one word before and after). Since the objective of our study is not to enhance the accuracy of an NDL model in predicting words but rather to explore the model's learning end state, we trained the model on the entire dataset without any splits for testing or development. The default parameters of NDL were used as it is the standard practice for NDL modeling [18]: learning rate ($\alpha = 0.001$), maximum connection strength ($\lambda = 1.0$), and the amount of increase and decrease ($\beta_i = 0.1$ (reward), $\beta_j = 0.1$ (punishment)).

A finished NDL model is represented by a weight matrix, wherein each cell represents the connection strength (weight) from a cue to an outcome. This weight matrix yields various measures, which have previously been used to predict diverse phenomena, ranging from decision time latencies to word duration and other linguistic behaviors (see [16] for an overview).

Prior The Prior of an outcome serves as a metric of its prior availability or entrenchment within the model [20]. Previous studies have identified a correlation between Prior and frequency [20] and have effectively used Prior to predict duration reduction [19, 18]. For our regression analysis, we compute four Prior predictors: *Prior Syllable*, *Prior Segment*, and *Prior Context* are derived by summing the absolute connection strengths of all cues from the respective domains to the given outcome, similar to a column 1-norm [20]. Additionally, *Prior All* is computed by summing all cues, regardless of domain, for a particular outcome.

Activation The Activation of an outcome shows how strongly a given set of cues supports it. In contrast to the Prior measures, Activation only considers local support. This predictor has also successfully been used to predict duration reduction [19, 18]. Activation is calculated by summing the connection strengths of cues to an outcome in a given learning event, encompassing the connection strengths of cues from each domain. The four resulting predictors are *Activation All*, *Activation Segment*, *Activation Syllable*, and *Activation Context*.

3. Regression analysis

3.1. Dependent variable: Duration

The dependent variable is the word duration generated by the timestamps in the Buckeye corpus. Impossible duration values were removed ($< 0\,\text{s}$ or $> 10\,\text{s}$). The variable in milliseconds was log-transformed to the base of 10.

3.2. Fixed effect variables: Control variables

A number of control variables were included as they have been found to influence word duration independently (e.g., [18]).

Word length Word length, as measured by the number of segments and syllables, was included to serve as the baseline duration. The segment transcriptions were generated using *DeepPhonemizer* (v0.0.17)[4] and then syllabified using a modified version of the syllabifier from the P2K toolkit[5].

Speaker and interviewer data Interviewer gender, speaker gender, and speaker age ('young': < 40 and 'old': > 40) were

[4] https://github.com/as-ideas/DeepPhonemizer

[5] https://sourceforge.net/projects/p2tk/

included in the Buckeye documentation [21]. They were coded as binary variables – interview gender (reference level: female), speaker gender (reference level: female), and speaker age (reference level: young), with the contrast coding (-0.5, 0.5).

Part of speech The syntactic category for each token was coded using the provided part-of-speech tags from [21]. After excluding grammatical categories, four lexical categories (adjectives, verbs, nouns, and adverbs) remained. This variable was coded using the target encoding scheme [22], which takes the mean of the dependent variable (duration) for each category to yield a single continuous variable.

Speech rate Speech rate was calculated as the number of syllables per utterance divided by the total duration of the utterance. An utterance is defined as one conversational turn marked by pauses or other interruptions. Pauses tagged in the corpus and the pauses omitted from the word token were used as utterance boundaries.

3.3. Fixed effect variables: NDL variables

The variables outlined in Section 2.3 were used as fixed effects.

3.4. Statistical procedures

A linear mixed-effects model was used to examine how well the NDL variables predict word duration using the *lme4* package (v1.1.31) [23] in *R* (Version 4.2.2) [24]. Speaker and Word were included as random effects. All continuous variables were *z*-transformed to enable us to compare the relative effect size β of the predictors. First, two models were fit to test whether there is an *a-priori* difference between Prior and *Activation* split by domain and Prior and *Activation* over all domains. The by-domain model, which is more fine-grained, has a better fit in AIC model selection (ΔAIC=1032.4). We, therefore, focused on examining the by-domain model.

Starting with the most complex model, a series of nested model comparisons was performed to determine the best model structure. *Prior Segments* was the only NDL variable that did not significantly improve model fit (ΔAIC=2)[6] and was excluded. The regression structure of the best model is given below in *lmer* syntax: *Word duration* $\sim$ *(1 | Speaker) + (1 | Word) + Segment count + Syllable count + Speaker gender + Interviewer gender + Speaker age + Part-of-Speech + Speech rate + Activation Context + Activation Syllables + Activation Segment + Prior Context + Prior Syllables*

The final model underwent model criticism as follows. 2.1% of the data points were excluded as their residuals were 2.5 standard deviations above and below the mean residual value. The fixed effects of the best model are reported in Table 2. In order to evaluate the collinearity of our predictors, we computed the Variance Inflation Factor (VIF). Our VIF is < 10, which indicates no serious issues of collinearity [26]. Table 1 summarizes the pairwise correlations between all NDL variables and word duration. They revealed that all NDL variables, as expected, negatively correlate with word duration, ranging from -0.10 to -0.51, suggesting a reduction effect.

4. Summary of the Results

The effect sizes (β) of all predictors can be seen in Table 2. The predictor *Activation Syllable* has the largest effect size (β = -3.115 $\times 10^{-2}$, $p < 0.001$), followed by *Prior Context* (β = -2.880$\times 10^{-2}$, $p < 0.001$) and *Activation Context* (β = -1.980$\times 10^{-2}$, $p < 0.001$). *Activation Segment* (β = 7.930$\times 10^{-3}$, $p < 0.001$) and *Prior Syllable* (β = -1.810$\times 10^{-2}$, $p = 0.0633$) have the smallest effect size. The positive effect on duration by *Activation Segment* is likely caused by a suppressor effect[7] since the correlation analysis shows that it is negatively correlated with word duration (-0.22, see Table 1).

5. Discussion

5.1. The role of distinctiveness

Both the segments and syllables are phonological properties of the word that they predict, and therefore, one could expect them to have similar predictive power. However, syllable predictors have a stronger effect on duration than segment predictors for both Activation and Prior measures. This may be due to the fact that there is a smaller segment inventory than syllable inventory. As a result, segments are less discriminatory when predicting words since they have many connections, but few of them are particularly distinctive. Conversely, because syllables are more discriminatory than segments, they are less affected by cue competition and more strongly predict a given word. For languages with a different syllable and segment inventory size, we expect this effect to vary accordingly.

Between the syllable and context predictors, it is difficult to discern which domain is a 'stronger predictor' of word duration. Overall, there are far more context cues than there are syllable cues. Following the reasoning of the previous paragraph, this would mean that the context would be more distinctive than the syllables. However, if cues are too distinctive, it equally presents a problem as when they are not distinctive enough. In this case, there may not be enough meaningful/strong connections to the word.

5.2. Local and global contexts

As outlined in the introduction, previous research suggests a difference between predictability measures from local and global contexts. This is mirrored in the NDL predictors Prior and Activation. Prior takes into account all cues the model has seen and is therefore very broad, whereas Activation considers the specific context and the phonetic makeup of a word. Since Activation is more specific than Prior, one could expect it to yield better predictions than Prior. It is unsurprising that using more specific cues leads to more accurate predictions, making an Activation measure the strongest predictor. This is shown in the performance of the segment and Activation predictors. Despite the poor predictive power of the segment level in general, *Activation Segment* is stronger than *Prior Segment*. Similarly, this may explain why, even though both *Prior Context* and *Activation Context* have stronger predictive power than *Prior Syllable*, *Activation Syllable* emerges as the strongest predictor.

Taken together, these observations suggest that the syllable domain is the strongest predictor of word duration, followed by the context domain and then the segment domain.

5.3. Limitations and future directions

While our study has uncovered a stronger impact of syllable predictors compared to segment predictors, it is important to acknowledge the potential influence of the design constraints

[6] see for example [25]

[7] One diagnostic of a suppressor effect is whether the model estimate is in the same or opposite direction as the correlation between the dependent and independent variable.

Table 1: *Correlation values all NDL measures and duration.*

	Activation Segment	**Activation Syllable**	**Activation Context**	**Prior Segment**	**Prior Syllable**	**Prior Context**	**Duration**
Activation Segment	1	0.56	-0.14	0.41	0.43	-0.02	-0.22
Activation Syllable	0.56	1	-0.13	0.55	0.84	0.34	-0.51
Activation Context	-0.14	-0.13	1	0.16	0.02	0.21	-0.10
Prior Segment	0.41	0.55	0.16	1	0.82	0.76	-0.35
Prior Syllable	0.43	0.84	0.02	0.82	1	0.68	-0.50
Prior Context	-0.02	0.34	0.21	0.76	0.68	1	-0.35
Duration	-0.22	-0.51	-0.10	-0.35	-0.50	-0.35	1

Table 2: *Fixed effect summary for the best model.*

Variable	β (10^{-2})	**SE** (10^{-3})	***p*-value**
Intercept	244.8	3.502	< 0.001
Segment count	6.120	1.494	< 0.001
Syllable count	3.462	1.409	< 0.001
Speaker gender	-0.745	5.287	0.1672
Interviewer Gender	-0.315	5.286	0.5552
Speaker Age	0.751	5.286	0.1641
POS	1.445	0.668	< 0.001
Speech Rate	-7.139	0.359	< 0.001
Activation Context	-1.980	0.483	< 0.001
Activation Syllables	-3.115	6.824	< 0.001
Activation Segments	0.793	1.854	< 0.001
Prior Context	-2.880	5.894	< 0.001
Prior Syllables	-1.810	9.746	0.0633

inherent in Naive Discriminative Learning. NDL's binary representation of cues does not account for multiple occurrences of the same cue within a learning event, which may adversely affect the predictive power of the segment cues. Future research could address this limitation by encoding the position of each ngram (similar to 'positional segment frequency' by [27]).

To further examine cues from other domains, Linear Discriminative Learning [16] can be used to examine the semantic domain by using semantic vectors as cues. Furthermore, lower-level linguistic units can be included, such as using distinctive features [28].

Additionally, expanding the scope of investigation by using multiple and larger corpora could enhance the replicability and generalizability of the findings presented in this paper. This includes an extension of this analysis to other languages where probabilistic reduction effects have been observed but are morphologically more complex, such as Japanese [1] and Kaqchikel [29], which have different syllable-to-segment inventory ratios compared to English and a different definition of wordhood. Specifically, they could offer valuable insights to disentangle effects stemming from these linguistic domains.

Finally, Prior and Activation were used as local and global measures in this study. It is to be examined whether our findings can be replicated using count-based probabilistic measures.

6. Conclusion

This study presents a cognitively motivated approach to modeling probabilistic reduction effects dynamically, as result of learning. Leveraging predictors derived from a Naive Discriminative Learning model, we successfully predicted word duration in the Buckeye corpus.

Our analysis revealed that local predictors exhibited a stronger influence on word duration prediction compared to global predictors. Furthermore, when examining the segment, syllable, and context domains, predictors derived from segment cues displayed the least explanatory power for predicting word duration compared to other cues.

Our results contribute to clarifying the importance of certain elements of speech in probabilistic reduction in production, in particular, the significance of the segment, syllable, and context domains, therefore further characterising on-line probabilistic reduction effects. Our results could be incorporated in models of speech production by, for example, weighting the contribution of specific cues for shortening duration.

7. References

[1] D. Hashimoto, "Probabilistic reduction and mental accumulation in Japanese: Frequency, contextual predictability, and average predictability," *Journal of Phonetics*, vol. 87, p. 101061, 2021.

[2] R. Turnbull, "Patterns of probabilistic segment deletion/reduction in english and japanese," *Linguistics Vanguard*, vol. 4, no. s2, p. 20170033, 2018.

[3] U. Cohen Priva, "Informativity affects consonant duration and deletion rates," *Laboratory phonology*, vol. 6, no. 2, pp. 243–278, 2015.

[4] K. Tang and R. Bennett, "Contextual predictability influences word and morpheme duration in a morphologically complex language (Kaqchikel Mayan)," *The Journal of the Acoustical Society of America*, vol. 144, no. 2, pp. 997–1017, 2018.

[5] M. Aylett and A. Turk, "The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech," *Language and speech*, vol. 47, no. 1, pp. 31–56, 2004.

[6] ——, "Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei," *The Journal of the Acoustical Society of America*, vol. 119, no. 5, pp. 3048–3058, 2006.

[7] G. K. Zipf, "Relative frequency as a determinant of phonetic change," *Harvard Studies in Classical Philology*, vol. 40, pp. 1–95, 1929.

[8] C. E. Wright, "Duration differences between rare and common words and their implications for the interpretation of word frequency effects," *Memory & Cognition*, vol. 7, no. 6, pp. 411–419, 1979.

[9] W. J. Levelt, *Producing spoken language: a blueprint of the speaker*. Oxford University Press, 1999, p. 83–122.

[10] A. Roelofs, "The weaver model of word-form encoding in speech production," *Cognition*, vol. 64, no. 3, pp. 249–284, 1997.

[11] F. H. Guenther and T. Vladusich, "A neural theory of speech acquisition and production," *Journal of neurolinguistics*, vol. 25, no. 5, pp. 408–422, 2012.

[12] R. A. Rescorla, "Pavlovian conditioning: It's not what you think it is." *American psychologist*, vol. 43, no. 3, p. 151, 1988.

[13] M. Ramscar and D. Yarlett, "Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition," *Cognitive science*, vol. 31, no. 6, pp. 927–960, 2007.

[14] R. H. Baayen, Y.-Y. Chuang, E. Shafaei-Bajestan, and B. J. P., "The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning," *Complexity*, vol. 2019, no. 4895891, p. 39, 2019. [Online]. Available: https://doi.org/10.1155/2019/4895891

[15] K. Sering, M. Weitz, E. Shafaei-Bajestan, and D.-E. Künstle, "Pyndl: Naive discriminative learning in python," *Journal of Open Source Software*, vol. 7, no. 80, p. 4515, 2022.

[16] Y.-Y. Chuang and R. H. Baayen, "Discriminative learning and the lexicon: Ndl and ldl," in *Oxford Research Encyclopedia of Linguistics*. Oxford University Press Oxford, 2021.

[17] R. H. Baayen and E. Smolka, "Modeling morphological priming in german with naive discriminative learning," *Frontiers in Communication*, vol. 5, p. 17, 2020.

[18] F. Tomaschek, I. Plag, M. Ernestus, and R. H. Baayen, "Phonetic effects of morphology and context: Modeling the duration of word-final s in english with naïve discriminative learning," *Journal of Linguistics*, vol. 57, no. 1, pp. 123–161, 2021.

[19] B. V. Tucker, M. Sims, and R. H. Baayen, "Opposing forces on acoustic duration," 2019, https://doi.org/10.31234/osf.io/jc97w.

[20] P. Milin, L. B. Feldman, M. Ramscar, P. Hendrix, and R. H. Baayen, "Discrimination in lexical decision," *PloS one*, vol. 12, no. 2, p. e0171935, 2017.

[21] S. Kiesling, L. Dilley, and W. D. Raymond, "The variation in conversation (vic) project: Creation of the buckeye corpus of conversational speech," *Language Variation and Change*, pp. 55–97, 2006.

[22] D. Micci-Barreca, "A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems," *SIGKDD Explorations Newsletter*, vol. 3, no. 1, p. 27–32, jul 2001. [Online]. Available: https://doi.org/10.1145/507533.507538

[23] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[24] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[25] M. Wieling, "Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between l1 and l2 speakers of english," *Journal of Phonetics*, vol. 70, pp. 86–116, 2018.

[26] S. Chatterjee and A. S. Hadi, *Regression analysis by example*. John Wiley & Sons, 2015.

[27] M. S. Vitevitch and P. A. Luce, "A web-based interface to calculate phonotactic probability for words and nonwords in english," *Behavior Research Methods, Instruments, & Computers*, vol. 36, no. 3, pp. 481–487, 2004.

[28] K. Tang and D. Baer-Henney, "Modelling l1 and the artificial language during artificial language learning," *Laboratory Phonology*, vol. 14, no. 1, 2023.

[29] K. Tang and R. Bennett, "Contextual predictability influences word and morpheme duration in a morphologically complex language (kaqchikel mayan)," *The Journal of the Acoustical Society of America*, vol. 144, no. 2, pp. 997–1017, 2018.