Speakers Enhance Contextually Confusable Words

Eric Meinhardt
Department of Linguistics
UC San Diego
La Jolla, CA, USA 92093
emeinhardt@ucsd.edu

Eric Baković
Department of Linguistics
UC San Diego
La Jolla, CA, USA 92093
ebakovic@ucsd.edu

Leon Bergen
Department of Linguistics
UC San Diego
La Jolla, CA, USA 92093
lbergen@ucsd.edu

Abstract

Recent work has found evidence that natural languages are shaped by pressures for *efficient* communication — e.g. the more contextually predictable a word is, the fewer speech sounds or syllables it has (Piantadosi et al. 2011). Research on the degree to which speech and language are shaped by pressures for *effective* communication — robustness in the face of noise and uncertainty — has been more equivocal. We develop a measure of contextual confusability during word recognition based on psychoacoustic data. Applying this measure to naturalistic speech corpora, we find evidence suggesting that speakers alter their productions to make contextually more confusable words easier to understand.

1 Introduction

A major open question in the study of natural languages is the extent to which pressures for efficient communication shape the online production choices of speakers or the system of forms and form-meaning mappings. Zipf (1936, 1949) famously noted that highly frequent words tend to be shorter and hypothesized that this could be explained in terms of pressures for efficient communication: the average cost of producing a word is lower than it would be otherwise.

More recent work has formalized hypotheses about the effect of communicative pressures on language usage and design using tools from information theory (Shannon 1948, Cover and Thomas 2012) and rational analysis (Anderson 1990, 1991). This work has found evidence that meanings are allocated to word types in a way that minimizes speaker effort (Piantadosi et al. 2011, 2012), and that this appears to be at least partly explainable by online production choices (Mahowald et al. 2013).

While this research offers evidence that lexicons and the production choices of speakers are shaped by pressures for efficient communication, other work examining how much words and lexicons are shaped by pressures for ensuring effective communication in the face of noise and uncertainty has been more equivocal. This work has found evidence that words with greater *neighborhood size* or *density* — that is, words that have a greater number of similar-sounding neighbors — have faster onset of production, and have lower overall durations. Words with greater neighborhood density also take longer for listeners to recognize and comprehend, and have less acoustically distinctive vowels (Vitevitch 2002, Gahl et al. 2012; see Vitevitch and Luce 2016 for review).

This work provides a challenge for communicatively-oriented models of production: words with greater numbers of similar-sounding neighbors seem likely to be more confusable, and therefore speakers would be predicted to decrease the likelihood of noise by, e.g., increasing their duration. However, this work does not directly estimate word confusability, instead using neighborhood density or an acoustic similarity measure as a proxy. It remains possible that greater word confusability is associated with phonetic enhancement, and that a more direct measure of confusability would reveal this relationship.

In this paper, we present a measure of relative word confusability based on both a language model and psychoacoustic data, and we examine how well it predicts word durations in natural speech corpora. This measure differs from neighborhood density in three ways: 1) it is sensitive to edit type; 2) it considers words with edit distance greater than 1; and 3) it takes into account top-down expectations.

The structure of the paper is as follows. We first present a derivation of a Bayesian model of word recognition (broadly similar to Norris and McQueen 2008) that incorporates both linguistic context and a model of noise estimated from the

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1991–2002
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

gating data of Warner et al. (2014). We use this speech recognition model to define a measure of confusability, and apply this measure to content words in the NXT-annotated subset of the Switchboard corpus and in the Buckeye corpus (Calhoun et al. 2010, Pitt et al. 2005). We provide evidence that greater confusability is associated with longer duration.

1.1 Related work

A number of other studies have examined how language is shaped by pressures for communication in the presence of noise. Dautriche et al. (2017) examines whether the words of natural lexicons are dispersed, as would be predicted if these lexicons are optimized to prevent confusions between different words. This work finds that in fact lexicons exhibit clear tendencies towards being clumpier rather than dispersed.

The current study follows previous work in using the phenomena of *reduction* and *enhancement* to investigate whether communication is optimized for robustness to noise. Speech tokens that are produced with shorter than usual duration, or with parts omitted or made less distinctive, are said to be reduced, and those tokens produced with longer durations or produced more distinctively are enhanced.

Previous work has provided evidence that reduction and enhancement are influenced by contextual predictability. Words, syllables, and segments that are more contextually predictable tend to be reduced and those that are less contextually predictable tend to be enhanced (see e.g. Van Son et al. 1998, Van Son and Pols 2003, Jurafsky et al. 2001, Aylett and Turk 2004, 2006, Cohen Priva 2008, 2012, 2015, Seyfarth 2014, Demberg et al. 2012, Pate and Goldwater 2015, Buz et al. 2016, Turnbull et al. 2018; see Bell et al. 2009, Jaeger and Buz 2018 for reviews). According to a communicatively-oriented account, this is explainable as balancing efficiency against effectiveness: speakers economize on production cost the more that context facilitates accurate listener inference of the speaker's intent.

Other work has investigated the effects of environmental noise on speech production. This includes work investigating whether speakers modulate their productions in response to overt signals of communication difficulty, e.g. loud environments or talking to listeners who are children, elderly, or non-native speakers (Lombard 1911, Uther et al. 2007, Picheny et al. 1986).

2 A model of word confusability

We propose a simplified model of word confusability, in which there are two factors that will make word w in context c more vs. less confusable. On the one hand, a listener who has observed context c has some 'top-down' beliefs and expectations about what w will be before the speaker produces any acoustics for w. On the other hand, once the speaker has produced acoustics for w, there will be (in general ambiguous) 'bottom-up' acoustic cues that will usually underdetermine what the speaker's choice of w actually was. The goal of the listener is then to combine their top-down expectations with their bottom-up observations to reason about which words are more vs. less likely to have been what the speaker intended.[1]

We operationalize the perceptibility of word w as the probability that the listener accurately recovers this word in situations where the speaker uses it; the confusability of a word is inversely related to its perceptibility. If a speaker has a model of the expected confusability of a given word, they can then decide to lengthen or shorten their particular production of the word token, balancing listener comprehension and their own effort.

2.1 Model definition

To model the in-context confusability of word tokens, we model the task of word recognition as one of Bayesian inference, with the following underlying generative process for the speaker:

1. At some point in time, the speaker has already produced some existing sentential context c, consisting of a sequence of orthographic words. We assume for simplicity and tractability that the listener knows exactly what this context is at each timestep.
2. The speaker produces the current word w — e.g. *cigarette*. We model this as sampling according to a language model p_L: $w \sim p_L(\cdot|c)$.
3. The speaker determines the segment sequence $x_{1:f} = (x_1, ..., x_f)$ corresponding to their word choice. For example, the speaker will determine that the segments [sɪgəɹɛt] correspond to the word *cigarette*.

[1] Note that of the two basic factors integrated here, previous probabilistic work on reduction has been limited to using only 'top-down' expectations.

In our corpora, there is a unique correct segment sequence for a given orthographic word. For ease of exposition, we therefore identify $x_{1:f}$ with its corresponding orthographic form v. Abusing notation, we will write $p_L(x_{1:f}|c)$ for the distribution over segmental forms induced by the language model.[2]

4. The listener receives a segment sequence $y_{1:f} = (y_1, ..., y_f)$ — e.g. [ʃɪgəɹɛt] ('*shigarette*') — drawn from a channel distribution p_N conditioned on the speaker's intended segment sequence: $y_{1:f} \sim p_N(\cdot|x_{1:f})$. This represents the effects of noise on the signal received by the listener.

The task of the listener is to then combine their observation (represented here by $y_{1:f}$) with their prior expectations about which words are likely given the context. The listener tries to determine how likely each wordform in the lexicon is to have been the one intended by the speaker. Their posterior belief p_{LISTENER} about which segmental wordform $x_{1:f}$ was intended is described by Bayes' rule:

$$p_{\text{LISTENER}}(x_{1:f}|y_{1:f}, c) \tag{1}$$

$$= \frac{p_N(y_{1:f}|x_{1:f})p_L(x_{1:f}|c)}{p(y_{1:f}|c)} \tag{2}$$

$$= \frac{p_N(y_{1:f}|x_{1:f})p_L(x_{1:f}|c)}{\sum_{x'_{1:f}} p_N(y_{1:f}|x'_{1:f})p_L(x'_{1:f}|c)} \tag{3}$$

Suppose for example that the listener perceives $y_{1:f}$ =[ʃɪgəɹɛt]. Their beliefs about the lexicon $p_L(X_{1:f}|C)$ will tell them that this is not a valid segmental wordform, but that [sɪgəɹɛt] *is* a valid wordform. Their beliefs about the noise distribution for the language $p_N(Y_{1:f}|X_{1:f})$ tell them that x_j =[s] is a plausible segment to be misperceived as y_j =[ʃ]; together this suggests that a good explanation of their percept is the intended wordform $x_{1:f}$ =[sɪgəɹɛt].

Equation 1 allows us to measure how accurately the listener will be able to reconstruct the speaker's intended message, given a perceived segmental wordform $y_{1:f}$. However, this is not sufficient to determine the confusability of an intended wordform. In general, an intended wordform $x_{1:f}$ may give rise to many different perceived wordforms $y_{1:f}$ as a result of noise. In order to measure its confusability, we therefore need to marginalize over the possible perceived segment sequences.

We define the contextual perceptibility of a segmental wordform $x_{1:f}$ in context c to be the expected probability that the listener accurately recovers it:

$$\mathop{\mathbb{E}}_{y_{1:f}\sim p_N(\cdot|x_{1:f})} p_{\text{LISTENER}}(x_{1:f}|y_{1:f}, c) \tag{4}$$

$$= \sum_{y_{1:f}} p_{\text{LISTENER}}(x_{1:f}|y_{1:f}, c)p_N(y_{1:f}|x_{1:f}) \tag{5}$$

The space of all possible channel strings $y_{1:f}$ grows exponentially in sequence length f. However, each segment is only substantially confusable with a small number of other segments and the probability of more than a small number of channel errors is small. We therefore approximated Eq. 4 with a Monte Carlo estimator:

$$\mathop{\mathbb{E}}_{y_{1:f}\sim p_N(\cdot|x_{1:f})} p_{\text{LISTENER}}(x_{1:f}|y_{1:f}, c) \tag{6}$$

$$\approx \frac{1}{n}\sum_{i=1}^{n} p_{\text{LISTENER}}(x_{1:f}|y^i_{1:f}, c) \tag{7}$$

$$y^i_{1:f} \sim p_N(\cdot|x_{1:f}) \tag{8}$$

We choose n =1000 to balance the variance and computational feasibility of the estimator.

Finally, following the reasoning given in Levy (2005, 2008b), we take the negative logarithm of this quantity and arrive at a surprisal, which represents the contextual confusability of segment sequence $x_{1:f}$ in context c:[3]

$$h(x_{1:f}|x_{1:f}, c) \tag{9}$$

$$= -\log \mathop{\mathbb{E}}_{y_{1:f}\sim p_N(\cdot|x_{1:f})} p_{\text{LISTENER}}(x_{1:f}|y_{1:f}, c) \tag{10}$$

3 Materials and methods

We make use of two types of data: psychoacoustic gating data for estimating a noise model, and several corpora of natural speech for evaluating whether individuals increase the duration of more confusable words.

3.1 Words duration data

Word durations were analyzed separately in two spoken corpora of American English: the Buckeye Corpus of Conversational Speech (Pitt et al.

[2] This notation ignores homophony, though the model is in fact sensitive to this.

[3] Compare Equations 4–9 with Eq. VII of Levy (2008a), a study of sentence-level confusability.

2005) and the NXT Switchboard Annotations (Calhoun et al. 2010), a richly annotated subset of Switchboard-1 Release 2 (Godfrey and Holliman 1997).

The Buckeye Corpus contains about 300,000 word tokens, taken from interviews with 40 speakers from central Ohio. Word durations for the present study were taken from the timestamps provided for word-level annotations. Each word token had a broad transcription uniform across all instances of the word type and a second, token-specific close transcription created by a human annotator.

The Switchboard Corpus contains transcripts of telephone conversations between strangers. The NXT annotated subset includes about 830,000 word tokens from 642 conversations between 358 speakers recruited from all areas of the United States. Word durations for the present study were taken from the 'phonological word'-level timestamps; these were the result of annotator-checked and -corrected timestamps initially made by alignment software. Each phonological word was also associated with a segmental transcription that was uniform across all instances of the word type.

Exclusion criteria almost exactly follow Seyfarth (2014) for the reasons cited there. These criteria are mainly designed to exclude non-content words and words whose pronunciation is likely affected by disfluencies or prosodic structure. Our criteria only diverge in the following manner: Word tokens were excluded if the utterance speech rate (total number of syllables / length of the utterance in seconds) was more than 3 standard deviations from the speaker mean (vs. 2.5 in Seyfarth 2014). After exclusion criteria were applied, about 44,000 (4,900) and 113,000 (8,900) word tokens (word types) remained in the Buckeye and NXT Switchboard corpora, respectively.

3.2 Diphone gating data

The model of word confusability was based on the diphone gating experiment data of Warner et al. (2014). Participants listened to gated intervals of every phonotactically licit diphone of (western) American English and attempted to identify the full diphone they thought was being produced during the interval. Along with earlier work by some of the same researchers on Dutch (Smits et al. 2003, Warner et al. 2005), this represents by far the richest and most comprehensive acoustic confusion matrix data of its kind.

Warner et al. (2014) identified all adjacent pairs of segments within and between words based on an electronic pronouncing dictionary of about 20,000 American English wordforms. A set of approximately 2,000 phonotactically licit diphones were extracted from this transcribed lexicon. At least one stimulus nonsense word was created per diphone by inserting the diphone into an environment consisting of at most one syllable on the left and at most one syllable on the right.

A recording of each stimulus wordform was then marked up with (generally) six temporal gates. For each stimulus wordform, one recording was created for each gate, starting at the beginning of the original recording and going all the way up to a gate location, followed by a ramping procedure (rather than truncation or white noise) to avoid systematically biasing confusion data.

In each trial, participants heard a gated stimulus recording.[4] If the recording included a preceding context, this context was displayed on the screen. The participant then selected the stimulus diphone they thought was in the recording (i.e. not including context).

From this response data, each gate of each stimulus diphone can be associated with a frequency distribution over response diphones. Only the response data for gates corresponding to the end of each segment of the diphone were used in the current study. For each of Buckeye and NXT Switchboard, the segment inventories of the gating data and of each speech corpus had to be projected down to a common set of segments. In each case, this involved collapsing the distinction in the corpora between syllabic and non-syllabic nasal stops. For reasons of data sparsity, the distinction between stressed and unstressed versions of any given vowel was also collapsed.

3.3 Language model

Our measure of contextual confusability uses a language model to compute the prior probability of a word in context. We estimate a language model from the Fisher corpus (Cieri et al. 2004), a speech corpus matched for genre and register to Buckeye and Switchboard. This corpus contains about 12 million (orthographic) word tokens taken from nearly 6000 short conversations, each on one of

[4] See Grosjean (1980) for reference on the gating paradigm.

about 100 topics.

We estimated n-gram models of several orders from the Fisher corpus using KenLM (Heafield 2011).[5] The n-gram order was treated as a hyperparameter, and selected on the Training Set, as described below. An add-1 smoothed unigram model was also created from word frequencies in the Fisher corpus using SRILM (Stolcke 2002, Stolcke et al. 2011).

3.4 Channel model

The *channel model* describes the conditional distribution $p_N(Y_{1:f}|X_{1:f})$ over what sequence of segments $y_{1:f}$ a listener will perceive (e.g. [ʃɪgəɹɛt], *shigarette*) given the full intended sequence $x_{1:f}$ (e.g. [sɪgəɹɛt], *cigarette*). We estimate this distribution using the diphone gating data in Section 3.2. We make the simplifying assumption that the channel distribution for segment y_i is conditionally independent of all other y_j ($j \neq i$) given intended segments x_{i-1}, x_i, x_{i+1}.

By conditioning on adjacent segments, we can capture some effects of coarticulation on confusability. For example, nasals before oral stops are systematically likely to be misheard as having the same place of articulation as the stop: $x_{1:f}$ =[ɑnpɑ] (alveolar nasal before labial stop) is more likely to be misperceived as $y_{1:f}$ =[ɑmpɑ] (a labial nasal) than the reverse, and a confusion of [n] for [m] is comparatively less likely when [n] is between vowels as in [ɑnɑ] (Ohala 1990).

For each gate $g \in \{3, 6\}$ and for each diphone x_1x_2, the response data from Section 3.2 induce a conditional frequency distribution over channel diphones $f_g(y_1, y_2|x_1, x_2)$. These frequency distributions were smoothed by adding a pseudocount to every channel diphone in every distribution; the distributions were then normalized to define a smoothed pair of diphone-to-diphone channel distributions $p_g(y_1, y_2|x_1, x_2)$. From the marginals of these distributions we constructed an approximation (Eq. 11) of the triphone-to-uniphone channel distribution via their geometric mean:[6]

$$\widetilde{p}_t(y_i|x_{i-1}, x_i, x_{i+1}) \propto \sqrt{p_3(y_i|x_{i-1}, x_i) \cdot p_6(y_i|x_i, x_{i+1})} \quad (11)$$

[5] We do not use lower-perplexity neural language models due to intractability resulting from the normalizing constant in Equations 3 and 4.

[6] We stop short of utilizing a full triphone-to-triphone channel distribution for tractability.

With the simplifying assumption that only substitution errors are possible,[7] we obtain a preliminary string-to-string channel model:

$$\widetilde{p}_N(y_{1:f}|x_{1:f}) = \prod_{j=1}^{j=f} \widetilde{p}_t(y_j|x_{j-1}, x_j, x_{j+1}) \quad (12)$$

We are primarily interested in using the channel model to define a ranking on the confusability of words, i.e. to determine which words are more or less confusable than others. This makes the channel model defined by Equations 11 and 12 not fully adequate.

The diphone gating data were collected in a laboratory setting with rates of noise lower than for naturalistic speech. As a result, when the noise model is estimated from this data, it implies the absolute rate of accurate perception (as defined by Equation 3) is close to 1 for most words. This makes it hard for the Monte Carlo estimator defined in Equation 7 to determine stable rankings of confusability. In order to estimate rankings in a more stable manner, we introduce a model hyperparameter $0 < \lambda \leq 1$, and define a new triphone-to-uniphone channel distribution by:

$$p_t(y_i|x_{i-1}, x_i, x_{i+1}) \quad (13)$$

$$= \begin{Bmatrix} \lambda \cdot \widetilde{p}_t(y_i|x_{i-1}, x_i, x_{i+1}), & y_i = x_i \\ \beta \cdot \widetilde{p}_t(y_i|x_{i-1}, x_i, x_{i+1}), & y_i \neq x_i \end{Bmatrix} \quad (14)$$

Here $\beta \geq 1$ is used to normalize the distributions; it is fully determined by λ for a particular distribution $p_t(\cdot|x_{i-1}, x_i, x_{i+1})$. The term λ is used to increase the noise rate in the channel distributions. Note that two important features of the original triphone-to-uniphone distributions $\widetilde{p}_t$ are maintained in the new model. First, the ratios of outcome probabilities within a single triphone distribution remain the same:

$$\frac{p_t(y_i|x_{i-1}, x_i, x_{i+1})}{p_t(y_i'|x_{i-1}, x_i, x_{i+1})} = \frac{\widetilde{p}_t(y_i|x_{i-1}, x_i, x_{i+1})}{\widetilde{p}_t(y_i'|x_{i-1}, x_i, x_{i+1})} \quad (15)$$

for segments $y_i, y_i' \neq x_i$. Second, the relative probability of accurate perception is preserved across triphone distributions:

$$\frac{p_t(x_i|x_{i-1}, x_i, x_{i+1})}{p_t(x_i'|x_{i-1}', x_i', x_{i+1}')} = \frac{\widetilde{p}_t(x_i|x_{i-1}, x_i, x_{i+1})}{\widetilde{p}_t(x_i'|x_{i-1}', x_i', x_{i+1}')} \quad (16)$$

The new model maximally agrees with the experimentally estimated distribution, differing only in the absolute amount of noise implied.

[7] The gating data does not provide information for estimating the probability of deletion or insertion errors.

The final string-to-string channel model is defined by:

$$p_N(y_{1:f}|x_{1:f}) = \prod_{j=1}^{j=f} p_t(y_j|x_{j-1}, x_j, x_{j+1}) \quad (17)$$

This new channel model has an increased noise rate, making it easier to estimate stable rankings of confusability across words.

The most similar previous channel model (Norris and McQueen 2008) was based on Dutch gating data (Smits et al. 2003) comparable to that used here. Norris and McQueen (2008) did not construct a triphone-to-uniphone channel model, but made use of all gates and also allowed investigation of word boundary identification.

3.5 Statistical methods

Prior to any analyses, the Switchboard and Buckeye corpora were each randomly divided into evenly-sized Training and Test sets. The Training sets were used for exploratory statistical analyses, and for determining the values of several model hyperparameters. Following this, all parameters and statistical analyses were frozen, and preregistered with the Open Science Foundation.[8]

We perform several linear regressions in order to determine the effect of confusability on word duration. Contextual confusability is defined throughout using Equation 9. Word durations are log-transformed. The following covariates are standard in the literature, and are included in our analyses: speaker identity; part of speech; unigram prior surprisal; speech rate (the average rate of speech, in syllables per second, of the utterance containing the target word); word length (measured by number of segments and syllables). Several covariates that are included are more non-trivial, and are discussed in more detail below: segmental inventory factors; forward and backward surprisal; neighborhood size and log weighted neighborhood density; and unigram confusability.

The segmental inventory variables code each word as a 'bag-of-segments.' A separate variable is defined for each phoneme in the segmental lexicon of the corpus. Each variable counts the number of times the corresponding phoneme occurs in the word. This is a variant of the baseline model used in previous work (Bell et al. 2009, Gahl et al. 2012).

Certain segments take longer to pronounce than others, and the baseline model is used in case the confusability scores contain information about segment identities within a word. Note, however, that this is a conservative baseline, as segment identity has an effect on confusability; certain segments are, individually, harder to perceive than others. The model will be used to predict word durations after these segmental effects have been factored out.

The forward language-model surprisal of a word is the surprisal of the word given preceding words in the context, and its backward surprisal is the surprisal given the following words in the context. Previous work in English has found backward surprisal to be a stronger predictor of spoken word duration than forward surprisal (Bell et al. 2009, Seyfarth 2014). Word confusability is expected to be correlated with surprisal, as more surprising words will be more difficult for the listener to recover in the presence of noise.

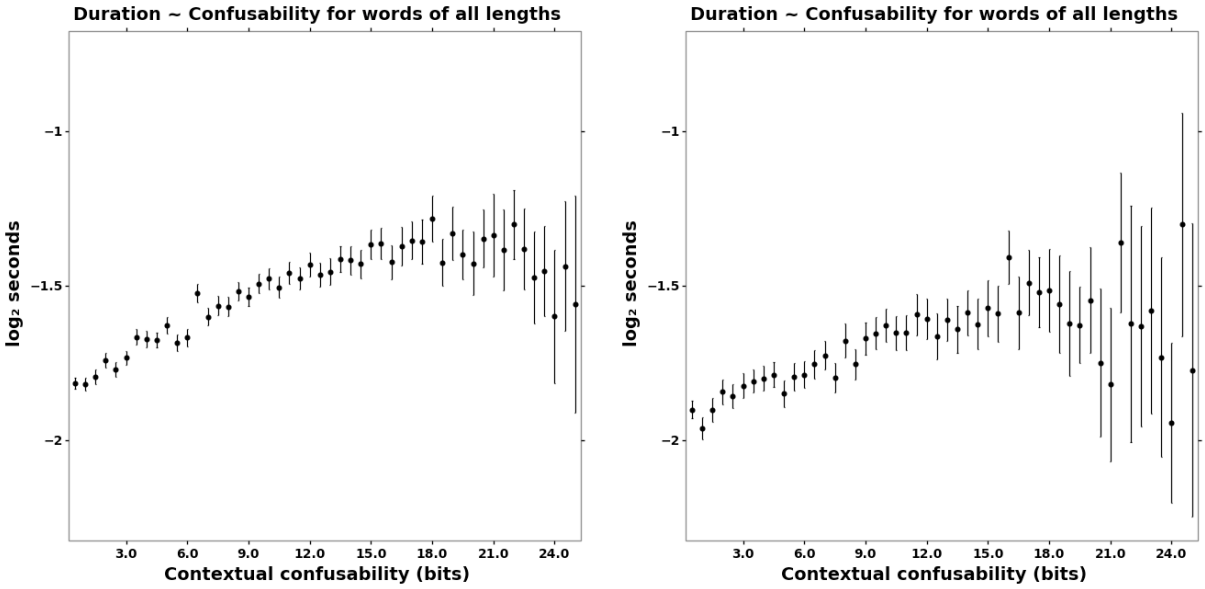
Neighborhood size and log weighted neighborhood density are measures of the number of words adjacent (within Levenshtein distance 1) to a target word. These measures have been extensively studied as explanatory variables for word duration (see Gahl et al. 2012, Vitevitch and Luce 2016 for review), and are expected to correlate with word confusability: words with more neighbors are expected to be more confusable. We evaluate whether there is any residual effect of confusability beyond its impact on these variables.

Unigram confusability measures the confusability of a word (Equation 9) given a unigram (word frequency) language model. This is a measure of the out-of-context confusability of a word, as discussed below.

All variables are treated as fixed effects, and OLS is used for regressions. Confidence intervals and p-values are calculated using the bias-corrected bootstrap. Bootstrapping is used to address possible heteroskedasticity in the data. Random effects are not used due to potential issues arising in observational studies like the current one. In particular, random effects may correlate with predictors in an observational study, leading to incorrect estimates of uncertainty and the potential for bias (Bafumi and Gelman 2006, Wooldridge 2010).[9]

[8] The preregistered analyses are available at the following link: https://osf.io/gj3ph/?view_only=6c5bd9b1211e4b798d2268fb8a8f5842

[9] While Bafumi and Gelman (2006) propose a solution to



(a) Switchboard

(b) Buckeye

Figure 1: Confusability vs. log duration on the Test sets of the Switchboard and Buckeye corpora. Error bars are 95% confidence intervals (non-bootstrapped). As illustrated in Figure 2, data are sparse beyond 18 bits, resulting in large confidence intervals in this range.

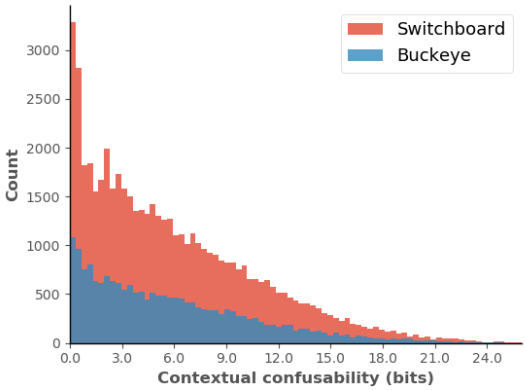


Figure 2: Histogram of contextual confusability scores on the Test sets.

All analyses were performed in two ways: using the raw values for each variable, and with rank-transformed values for the continuous variables. The rank-transformed analyses provide a test of the papers hypothesis that greater (i.e. higher-rank) confusability is associated with longer (higher-rank) duration. The analyses eliminate the potentially questionable parametric assumption of a linear relationship between confusability (in bits) and duration (in log seconds). The rank-transformed analyses are intended as sensitivity analyses for the non-transformed analyses; if the two analyses provide different results, this provides evidence of a problem with the statistical methods.[10]

4 Results

Four model hyperparameters were selected using the Switchboard and Buckeye Training sets: the order and direction of the n-gram model, the diphone-to-diphone channel pseudocounts, and the noise factor λ.[11] Backward bigram language models were found to perform best on the Training sets, possibly due to distributional differences between these corpora and the Fisher corpus, which was used for language model estimation. This is consistent with prior work in the area (e.g. Bell et al. 2009, Seyfarth 2014). Pseudocounts were set to 0.01, and the term λ was set to 2^{-6}.

Figure 2 shows the frequency of model-computed confusability scores on the Switchboard and Buckeye Test sets. Figure 1 shows the relationship between confusability and word duration on the Test sets.

The first set of analyses include all of the co-

this problem by decorrelating the fixed effect from random effects, the method produces identical estimates for the fixed effect, and is primarily useful when the random effect estimates themselves are of interest.

[10] Model and analysis code is available at: https://github.com/emeinhardt/wr

[11] The language model order was the same across all covariates where it was used.

Dataset	Rank	β	95% CI	p-value
SWBD	No	0.006	(0.004, 0.008)	0.001
SWBD	Yes	0.086	(0.067, 0.109)	0.001
Buckeye	No	0.005	(0.001, 0.008)	0.01
Buckeye	Yes	0.123	(0.080, 0.130)	0.001

Table 1: Effect of contextual confusability on log word duration, not controlling for unigram confusability. Estimates from the Test sets. Rank indicates whether continuous variables were rank-transformed. p-values are upper-bounds.

Dataset	Rank	β	95% CI	p-value
SWBD	No	0.009	(0.006, 0.011)	0.001
SWBD	Yes	0.132	(0.095, 0.130)	0.001
Buckeye	No	0.007	(0.003, 0.011)	0.001
Buckeye	Yes	0.148	(0.106, 0.164)	0.001

Table 2: Effect of contextual confusability on log word duration, controlling for unigram confusability. Estimates from the Test sets.

variates from Section 3.5, except for unigram confusability. This allows us to determine whether there is an effect of word confusability on duration, independent of whether this effect is sensitive to context. Greater confusability is associated with longer word durations on both the Switchboard and Buckeye Training sets ($p<0.001$ for all analyses). Table 1 shows results of the same analyses performed on the Test sets. The effects replicate on the Test sets, and are qualitatively similar when continuous variables are rank-transformed.

These analyses provide evidence that higher confusability is associated with longer word duration. In the second set of analyses, we investigate whether a context-sensitive measure of confusability is necessary for explaining this effect, or whether an out-of-context measure suffices. In order to do this, we include unigram confusability as a covariate in the analyses, in addition to the previous covariates. Unigram confusability is identical to our target measure of word confusability, except that the language model is replaced with a unigram model. The measure calculates a word's confusability based on its acoustic properties and its phonological similarity to other words. It therefore does not take into account top-down expectations based on a word's context.

After controlling for unigram confusability, contextual confusability remains associated with longer word durations on both the Switchboard and Buckeye Training sets ($p<0.001$ for all analyses). Table 2 shows the same analyses on the Test sets. The effects replicate on both Test sets, and similarly for the rank-transformed analyses.

4.1 Neighborhood density

We report the results of several unplanned analyses. Confidence intervals and p-values reported in this section are non-bootstrapped.

We evaluate the effect of neighborhood density on word duration in the Test sets. Weighted neighborhood density is associated with lower word duration in all analyses. (See Appendix B.) The results provide evidence that the neighborhood density effects identified in previous work remain qualitatively similar, after adjusting for contextual confusability.

5 Discussion

We draw two main conclusions from our results. First, we provide evidence that speakers lengthen words that are more confusable. This supports the hypothesis that variation and structure in natural languages are shaped not only by pressures for efficient signals, but also pressures for effective communication of the speaker's intended message in the face of noise and uncertainty (Lindblom 1990, Lindblom et al. 1995, Hall et al. 2018).

Second, we provide large scale, naturalistic evidence for reduction and enhancement driven by contextual confusability. Conversational context may make a speaker's intended message easier or harder to recover from ambiguous acoustics. The results suggest that speakers modulate their utterances in a manner that is sensitive to this effect of context, increasing duration when context makes the intended utterance harder to recover.

The results complement previous work which demonstrates reduction and enhancement driven by contextual predictability (see e.g. Seyfarth 2014). They also complement work which shows confusability-driven reduction and enhancement in targeted experimental manipulations (see e.g. Kirov and Wilson 2012, Schertz 2013, Seyfarth et al. 2016, Buz et al. 2016).

The study may help to resolve questions raised by previous work examining the effects of neighborhood density. That work found negative or null

associations between word duration and neighborhood density and related measures (e.g. Gahl et al. 2012, Gahl and Strand 2016). The proposed confusability measure differs from neighborhood density in three ways: it is sensitive to edit type, words greater than two edits away, and top-down effects.

These differences may account for the discrepancy in the effects of neighborhood density and confusability. Under one hypothesis, neighborhood density effects reflect spillover of activation between words with overlapping subsequences of speech sounds (e.g. Gahl and Strand (2016), Chen and Mirman (2012), Dell (1986), Vitevitch and Luce (2016)). This spillover is potentially sensitive only to Levenshtein distance. In contrast, confusability is sensitive to fine-grained perceptual structure. When lexical neighbors differ in perceptually distinct segments, they will typically be non-confusable.

A second hypothesis is that the discrepancy arises from the role of top-down expectations in confusability. Neighborhood effects are type-level phenomena: a word has the same neighbors no matter what context it appears in. Confusability, on the other hand, is a token-level phenomenon: contextual expectations will change the confusability of a word. Stable properties of the lexicon may determine which segment sequences undergo frequent articulatory rehearsal, and are reduced as a consequence. The confusability measure picks up on context-dependent variation, which rehearsal processes in the articulatory system may not be sensitive to.

The study suggests several directions for future work. First, while there are advantages of using naturalistic speech data (Gahl et al. 2012), it would be desirable to have experimental validation of the confusability measure and its relationship to speaker reduction. Second, a lower-perplexity neural language model would provide better estimates of a word's confusability, but would first need to be validated on speech data. Third, a more sophisticated channel model would allow for insertions and deletions, and better capture transitional coarticulatory cues (Wright 2004). Because speakers enhance or reduce their speech in ways other than changing duration (see e.g. Kirov and Wilson 2012, Schertz 2013, Seyfarth et al. 2016, Buz et al. 2016), such a model would permit investigation of targeted enhancement and reduction in naturalistic data.

Acknowledgements

We thank Uriel Cohen Priva and Scott Seyfarth for help reproducing their analyses. We also thank Silas Horton, Todd Williams, and Thanh Nguyen for computing support. The Titan V used for this research was donated by the NVIDIA Corporation.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Erlbaum, Hillsdale, NJ.

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, 14:471–517.

Aylett, M. and Turk, A. (2004). The smooth signal redundancy hypothesis: a functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and speech*, 47(Pt 1):31–56.

Aylett, M. and Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *The Journal of the Acoustical Society of America*, 119(5 Pt 1):3048–3058.

Bafumi, J. and Gelman, A. (2006). Fitting multilevel models when predictors and group effects correlate. *Available at SSRN 1010095*.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., and Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60(1):92–111.

Buz, E., Tanenhaus, M. K., and Jaeger, T. F. (2016). Dynamically adapted context-specific hyperarticulation: Feedback from interlocutors affects speakers ' subsequent pronunciations. *Journal of Memory and Language*, 89:68–86.

Calhoun, S., Carletta, J., Brenier, J. M., Mayo, N., Jurafsky, D., Steedman, M., and Beaver, D. (2010). The NXT-format Switchboard Corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. In *Language Resources and Evaluation*, volume 44, pages 387–419.

Chen, Q. and Mirman, D. (2012). Competition and cooperation among similar representations: toward a unified account of facilitative and inhibitory effects of lexical neighbors. *Psychological review*, 119(2):417.

Cieri, C., Miller, D., and Walker, K. (2004). The Fisher corpus: a Resource for the Next Generations of Speech-to-Text. *Language Resources and Evaluation*, 4:69–71.

Cohen Priva, U. (2008). Using information content to predict phone deletion. *Proceedings of the 27th West Coast Conference on Formal Linguistics*, pages 90–98.

Cohen Priva, U. (2012). *Sign and Signal Deriving Linguistic Generalizations From Information Utility*. Doctoral dissertation, Stanford University.

Cohen Priva, U. (2015). Informativity affects consonant duration and deletion rates. *Laboratory Phonology*, 6(2):243–278.

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Dautriche, I., Mahowald, K., Gibson, E., Christophe, A., and Piantadosi, S. T. (2017). Words cluster phonetically beyond phonotactic regularities. *Cognition*, 163:128–145.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological review*, 93(3):283.

Demberg, V., Sayeed, A. B., Gorinski, P. J., and Engonopoulos, N. (2012). Syntactic surprisal affects spoken word duration in conversational contexts. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, (July):356–367.

Gahl, S. and Strand, J. F. (2016). Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception. *Journal of Memory and Language*, 89:162–178.

Gahl, S., Yao, Y., and Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, 66(4):789–806.

Godfrey, J. J. and Holliman, E. (1997). Switchboard-1 Release 2. Technical report, Linguistic Data Consortium.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & psychophysics*, 28(4):267–283.

Hall, K. C., Hume, E., Jaeger, T. F., and Wedel, A. (2018). The Role of Predictability in Shaping Phonological Patterns. *Linguistic Vanguard*, 4.

Heafield, K. (2011). KenLM: Faster and Smaller Language Model Queries. *Proceedings of the Sixth Workshop on Statistical Machine Translation*, (2009):187–197.

Jaeger, T. F. and Buz, E. (2018). Signal Reduction and Linguistic Encoding. In *The Handbook of Psycholinguistics*, pages 38–81. Wiley-Blackwell.

Jurafsky, D., Bell, A., Gregory, M., and Raymond, W. D. (2001). Probabilistic Relations between Words: Evidence from Reduction in Lexical Production. *Frequency and the emergence of linguistic structure*, pages 229–254.

Kirov, C. and Wilson, C. (2012). The Specificity of Online Variation in Speech Production. *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, pages 587–592.

Levy, R. (2005). *Probabilistic Models of Word Order and Syntactic Discontinuity*. Doctoral dissertation, Stanford University.

Levy, R. (2008a). A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 234–243.

Levy, R. (2008b). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In *Speech Production and Speech Modelling*, pages 403–439.

Lindblom, B., Guion, S., Hura, S., Moon, S.-J., and Willerman, R. (1995). Is sound change adaptive? *Rivista di Linguistica*, 7:5–37.

Lombard, E. (1911). Le signe de l'elevation de la voix. *Ann. Mal. de L'Oreille et du Larynx*, pages 101–119.

Mahowald, K., Fedorenko, E., Piantadosi, S. T., and Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126(2):313–318.

Norris, D. and McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological Review*, 115(2):357–395.

Ohala, J. J. (1990). The phonetics and phonology of aspects of assimilation. In Kingston, J. and Beckman, M. E., editors, *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, chapter 14, pages 258–282.

Pate, J. K. and Goldwater, S. (2015). Talkers account for listener and channel characteristics to communicate efficiently. *Journal of Memory and Language*, 78.

Piantadosi, S. T., Tily, H., and Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Piantadosi, S. T., Tily, H., and Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, 122(3):280–291.

Picheny, M. A., Durlach, N. I., and Braida, L. D. (1986). Speaking clearly for the hard of hearing ii: Acoustic characteristics of clear and conversational speech. *Journal of Speech, Language, and Hearing Research*, 29(4):434–446.

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., and Raymond, W. (2005). The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95.

Schertz, J. (2013). Exaggeration of featural contrasts in clarifications of misheard speech in English. *Journal of Phonetics*, 41(3-4):249–263.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, 133(1):140–155.

Seyfarth, S., Buz, E., and Jaeger, T. F. (2016). Dynamic hyperarticulation of coda voicing contrasts. *The Journal of the Acoustical Society of America*, 139(2).

Shannon, C. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423.

Smits, R., Warner, N., McQueen, J. M., and Cutler, A. (2003). Unfolding of phonetic information over time: a database of Dutch diphone perception. *The Journal of the Acoustical Society of America*, 113(January):563–574.

Stolcke, A. (2002). SRILM-An Extensible Language Modeling Toolkit. In *8th International Conference on Spoken Language Processing (INTERSPEECH 2002)*, volume 2, pages 901–904.

Stolcke, A., Zheng, J., Wang, W., and Abrash, V. (2011). SRILM at Sixteen: Update and Outlook. In *Proceedings - IEEE Automatic Speech Recognition and Understanding Workshop*.

Turnbull, R., Seyfarth, S., Hume, E., and Jaeger, T. F. (2018). Nasal place assimilation trades off inferrability of both target and trigger words. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 9(1).

Uther, M., Knoll, M. A., and Burnham, D. (2007). Do you speak e-ng-li-sh? a comparison of foreigner- and infant-directed speech. *Speech communication*, 49(1):2–7.

Van Son, R. J. J. and Pols, L. C. W. (2003). How efficient is speech? In *Proceedings of the Institute of Phonetic Sciences*, pages 171–184.

Van Son, R. J. J. H., Koopmans-van Beinum, F. J., and Pols, L. C. W. (1998). Efficiency As An Organizing Principle Of Natural Speech. In *Fifth International Conference on Spoken Language Processing*.

Vitevitch, M. S. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology, Learning, Memory, and Cognition*, 28(4):735–747.

Vitevitch, M. S. and Luce, P. A. (2016). Phonological Neighborhood Effects in Spoken Word Perception and Production. *Annual Review of Linguistics*, 2:75–94.

Warner, N., McQueen, J. M., and Cutler, A. (2014). Tracking perception of the sounds of English. *The Journal of the Acoustical Society of America*, 135(5):2995–3006.

Warner, N., Smits, R., McQueen, J. M., and Cutler, A. (2005). Phonological and statistical effects on timing of speech perception: Insights from a database of Dutch diphone perception. *Speech Communication*, 46(1):53–72.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Wright, R. (2004). A review of perceptual cues and cue robustness. In Hayes, B., Kirchner, R., and Steriade, D., editors, *Phonetically based phonology*, chapter 2. Cambridge University Press.

Zipf, G. K. (1936). *The Psychobiology of Language*. Routledge, London.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press.

A Sensitivity analyses

In this section we present the results of several sensitivity analyses. These analyses are post-hoc, and were not pre-registered with OSF. They are performed in order to assess the sensitivity of the findings to the bootstrapping method that was used for calculating p-values.

The analyses are intended to evaluate the effect of contextual confusability on word duration, and are identical to the analyses in Section 4, except that p-values are calculated using a likelihood ratio test. Each likelihood ratio test compares a pair of OLS models: one model containing contextual confusability as a covariate, and an ablated model which does not use this covariate, but is otherwise identical. The tests evaluate whether the inclusion of contextual confusability improves the prediction of word duration, beyond the contributions of other covariates.

Table 3 and Table 4 show results without and with unigram confusability included as a covariate. All comparisons performed in Section 4 remain significant with the likelihood ratio test.

Dataset	Rank	Likelihood ratio	p-value
SWBD	No	35.4	$3x10^{-9}$
SWBD	Yes	91.8	$3x10^{-22}$
Buckeye	No	7.23	0.007
Buckeye	Yes	64.9	$8x10^{-16}$

Table 3: Likelihood ratio tests, evaluating whether contextual confusability improves OLS model fit on the test set. No control for unigram confusability included.

Dataset	Rank	Likelihood ratio	p-value
SWBD	No	51.3	$8x10^{-13}$
SWBD	Yes	160.6	$8x10^{-37}$
Buckeye	No	12.0	0.0005
Buckeye	Yes	70.1	$6x10^{-17}$

Table 4: Likelihood ratio evaluation of contextual confusability, controlling for unigram confusability.

B Neighborhood density analyses

Table 5 shows the effect of log weighted neighborhood density on log word duration. Confidence intervals and p-values are non-bootstrapped.

Dataset	β	95% CI	p-value
SWBD	-4.27	(-4.96, -3.58)	0.001
Buckeye	-1.91	(-2.88, -0.94)	0.001

Table 5: Effect of log weighted neighborhood density on log word duration.