Title
The Units of Gating and Access to Lexical Representations During Spoken Word Recognition

Authors
Antal, Caitlyn
de Almeida, Roberto G.

Publication Date
2023

Peer reviewed

The Units of Gating and Access to Lexical Representations During Spoken Word Recognition

Caitlyn Antal (caitlyn.antal@mail.mcgill.ca)
Department of Psychology, McGill University
Montreal, QC, H3A 1G1

Roberto G. de Almeida (roberto.dealmeida@concordia.ca)
Department of Psychology, Concordia University
Montreal, QC, H4B 1R6

Abstract

Word recognition models such as Cohort have long relied on the gating paradigm to investigate how acoustic-phonetic information maps onto lexical representations. We report on a methodological study investigating (a) whether the recognition point of a spoken word is affected by the speech variables employed in the gating paradigm, and (b) which distributional properties of a words' linguistic and social usage pattern affect its recognition point. We addressed the first question by contrasting the traditional "brute-force" gating paradigm (i.e., employing incremental segments of 50 ms) to "phonetically-driven" gating paradigms. Three methodologies were employed for determining phonemic segments: (1) articulatory measures, relying on the peak velocity of articulatory gestures, (2) acoustic measures, relying on the acoustic energy of consonants and vowels, and (3) brute-force measures, relying on 50 ms increments. We addressed the second question by relying on four social measures of lexical strength, which were attained from a corpus of 57 billion words from Reddit: word frequency (WF), contextual diversity (CD), discourse contextual diversity (DCD), and user contextual diversity (UCD). Results showed that the traditional brute-force gating method yielded significantly faster word recognition times, in comparison to articulatory and acoustically driven gating methods. Our results also showed that CD is a superior measure of lexical strength than WF, UCD, and DCD. Overall, our results suggest that the traditional gating paradigm is a reliable method for investigating spoken word recognition, given that spoken word recognition may rely on the gradual accumulation of phonetic information over time, rather than relying solely on the recovery of categorical phonetic features that are distributed non-linearly in time. We also suggest that the lexical system may be organized as a function of usage-based contextual measures of lexical items.

Keywords: spoken word recognition; gating paradigm; articulatory phonetics; acoustic segments; contextual diversity; lexical frequency.

Introduction

How do we understand spoken words as they unfold over time? Our phenomenological experience of speech processing is that of an unbroken, continuous sensory input. But how does the incoming speech signal ultimately get segmented and mapped onto different lexical items, triggering their own semantic representations, in such a rapid and effortless way? Is early word recognition achieved categorically, through the incremental processing of phonemes? Or is it achieved through the anticipatory co-articulation of phonetic segments, whereby co-articulatory information is reflected acoustically, and thus allowing hearers to predict upcoming speech sounds? These fundamental questions regarding the process by which the incoming acoustic-phonetic speech signals map onto the representations of word forms in the mental lexicon have been a matter of much dispute in the speech perception literature.

The process of spoken word recognition begins when the sensory input—or some abstract representation that is computed from this input speech signal—makes initial contact with the lexicon (Frauenfelder & Tyler, 1987; Zhang & Samuel, 2018). During this initial phase, the perceiver extracts acoustic-phonetic cues from the incoming speech signal and integrates these cues to generate a mental token that is associated with a higher-level semantic representation. It is thus not surprising that one of the long-standing concerns within the speech perception literature has pertained to the nature of the representations that make initial contact with the mental lexicon. This is so because the nature of their representations will have important consequences for how the phonetic-acoustic cues enter the lexicon, and thus, which lexical items will be initially retrieved. While several proposals have been put forward to account for early spoken word recognition (e.g., LAFS model [Klatt, 1980]; TRACE model [Elman & McClelland, 1984]; Search Model [Bradley & Forster, 1987]; Shortlist [Norris, 1994; Norris & McQueen, 2008]), one prominent theory remains the bottom-up, phoneme-based processing approach known as the Cohort model of word recognition.

According to the earliest iterations of the Cohort model (Marslen-Wilson, 1987; 1989; Marslen-Wilson & Tyler, 1980; Warren & Marslen-Wilson, 1987), the process of word recognition is a function of the accumulating speech input from an initial set of possible word candidates. When the first 100-200 milliseconds (ms) of acoustic-phonetic input associated with a word is heard, this activates a *word-initial cohort* of possible word candidates within the perceiver's mental lexicon. This initial cohort will include the target word (e.g., *candle*), but it will also include *competitors*—words that share the same initial acoustic-phonetic information as the target word (e.g., *cat, camel, cabbage*). As more of the speech signal is heard, an increasing number of competitors are no longer compatible with the signal and are subsequently removed from the cohort, until there is only one remaining candidate. The time point at which the target word

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

is distinguished from all other competitors is known as the *uniqueness point*.[1] It is at this stage that the semantic and syntactic properties of the target word are mapped onto the utterance representation. Consider the auditory recognition of the word *candle* in Table 1.

According to the Cohort model, the recognition of the lexical item *candle* is said to obey the following steps. When a comprehender hears the first two phonemes of the lexical item (i.e., /k/ and /æ/), the hearer will activate the target word *candle*, along with a multitude of competitors that share the same initial phonetic information, such as *candy*, *cattle*, *can*, *camel*, to name a few. However, as more phonetic information gets introduced within the speech signal, this then restricts the domain of possible word candidates. For instance, as the phonemes /n/ and /d/ are subsequently introduced within the speech signal, the competitors *cattle* and *can*, respectively, become incompatible with the input and are excluded from the cohort. Finally, as the last phoneme /l/ is introduced, the uniqueness point arises as there remains only one possible candidate within the cohort, the target word *candle*, and the word can thus be recognized.

Table 1: Schematic example of the spoken word recognition process for the lexical item *candle*, according to the Cohort model.

Phonetic Unit	Competitors
/k/	candy, cattle, can, key, candle, king
/kæ/	candy, cattle, can, candle, camel
/kæn/	candy, can, candle
/kænd/	candy, candle
/kændl/	candle

The Gating Paradigm

The gating paradigm has been the most prominent method for investigating the Cohort model, and spoken word recognition more broadly (Grosjean, 1980; 1996; Tyler, 1984; Tyler & Marslen-Wilson, 1986; Warren & Marslen-Wilson, 1987; Van Petten et al., 1999). This method has been used to investigate topics ranging from the neuronal correlates underlying spoken word recognition (Kocagoncu et al., 2018), the effects of sentential-semantic context on spoken-word processing with cross-modal priming (Zwitserlood, 1989; Dossey et al., 2022), to the effect of stress (McAllister, 1991), word length (Grosjean, 1980), and frequency (Tyler, 1984; Marslen-Wilson, 1990) on the recognition of spoken words in isolation.

This paradigm is used to address the amount of acoustic-phonetic information that is needed for individuals to correctly identify the lexical item associated with a given acoustic input. In a gating task, participants are presented with a spoken language stimulus in segments of increasing duration. After each increment, participants indicate what word they think corresponds to the acoustic segment. The first segment is usually very short (e.g., 20-30 ms), while the last segment corresponds to the entire stimulus (Grosjean, 1996). Variants of the task differ on the increment size of the gates—usually ranging between 20-50 ms, with 50 ms being the norm (e.g., Zwitserlood, 1989; Van Petten et al., 1999; Warren & Marslen-Wilson, 1987). In addition to guessing the word, participants are also asked to give a confidence rating following each segment. The confidence ratings of interest are those at the moment of the isolation point (i.e., time point at which the word is correctly identified).

While the gating paradigm seems to be a well-established method for determining the time point of spoken word recognition, its underlying rationale for slicing the acoustic-phonetic information into bins of arbitrary length is questionable, given that this slicing does not obey any natural boundary or phonetic cue. It is also possible that these arbitrary slices artificially create phonetic material by segmenting the acoustic signal midway through the articulatory process. And if spoken word recognition is dependent on the acoustic-phonetic information available at a given bin segment, the arbitrary slicing may compromise the retrieval of semantic representation more broadly. It is thus under scrutiny whether the gating paradigm can be taken to accurately reflect the process of spoken word recognition, raising concerns about the validity of experimental results obtained with the use of this paradigm.

It has been shown, for instance, that adverse conditions, such as degradations in the speech signal due to single phonetic unit deletion or reduction, negatively affect word recognition (e.g., Ernestus & Warner, 2011; Bürki et al., 2011; Mattys et al., 2012; Ernestus, 2014; van de Ven & Ernestus, 2018). In particular, studies have shown that highly reduced pronunciations of word forms take significantly longer to be recognized in isolation, in comparison to highly reduced words embedded within context (Ernestus et al., 2002; Brouwer et al., 2012, 2013). Similar results have also been obtained in cases of syllable reductions, whereby words take significantly longer to identify when the initial syllable is reduced (Racine & Grosjean, 1997; van de Ven & Ernestus, 2018). Together, these studies suggest that spoken word recognition might rely on phoneme-based cues, given that interfering with particular phonemes seems to decrease recognition accuracy and yield longer recognition times. Using monotonic increases, then, may mask some essential acoustic-phonetic details that are necessary for word recognition. Crucially, these results suggest that the activation of words in the lexicon may depend on the non-linear distribution of acoustic-phonetic information over time, rather than a linear increase over time (i.e., increasing gates of 50 ms).

According to Pisoni & Luce (1987), "phonemes are rarely realized in the speech waveform as a linearly ordered sequence of discrete acoustic-events" in time (p. 23). They

[1] Marslen-Wilson (1989) takes the uniqueness point as referring to the *word recognition point*; however, we will refrain from using the latter term, given that the psychological validity on the concept of *recognition point* is not well established and that there is no general consensus that a lexical items' uniqueness point actually reflects its recognition point. It is not implausible to suggest that word recognition for a given lexical item might occur before its uniqueness point (see Grosjean, 1996, for discussion).

suggest that this is due to the coarticulation of neighbouring phonemes, whereby the acoustic signal of one phoneme is impacted by the articulation of its adjacent phoneme. Although one can use strict acoustic criteria for segmenting phonemes (see, e.g., Ladefoged & Johnson, 2015), these criteria usually lead to a greater number of acoustic segments than there are phonemes in the speech utterance. Thus, the continuity of the articulatory gestures, as well as the information from anticipatory coarticulations, may better identify potential word candidates—and upcoming phonetic segments, more broadly (Warren & Marlsen-Wilson, 1987)—than acoustic signals, suggesting that spoken word recognition may rely on the representation of individual articulatory gestures accumulated over time.

In the present study, we investigated whether the speed of spoken word recognition is affected by the factors underlying the gating paradigm by comparing the traditional gating method (i.e., employing incremental segments of 50 ms), to "phonetically-driven" gating methods. We thus employed three methods for determining phonetic segments: (1) articulatory measures, relying on the peak velocity of articulatory gestures, (2) acoustic measures, relying on the acoustic energy of consonants and vowels, and (3) 'brute-force' measures, relying on 50 ms increments. If spoken word recognition relies on the retrieval of acoustic-phonetic information distributed in time non-linearly, we predicted faster recognition times for either the acoustic or articulatory gating paradigms. However, if spoken word recognition relies on the lexical activation of words as a function of a linear increase over time, we predicted faster recognition times for the traditional brute-force gating paradigm.

We also conducted hierarchical linear regressions to investigate whether spoken word recognition may be influenced by the social usage of the lexical item. It has been a tradition in gating studies, as well as in studies involving some versions of the visual-world paradigm (e.g., Dahan, Magnuson, & Tanenhaus, 2001), to rely on frequency as a modulating factor in the speed at which a word is recognized (e.g., Grosjean, 1980; Tyler, 1984; Marslen-Wilson, 1990; Zhuang et al., 2014). In the present study, we used four social usage measures of lexical strength obtained from a corpus of 57 billion words from Reddit (Johns, 2021): word frequency (WF), contextual diversity (CD), user contextual diversity (UCD), and discourse contextual diversity (DCD). We reasoned that the social usage frequency of a word across contexts, relying on large corpora, may constitute a better modulating factor in its recognition speed and its consequent mapping to semantic representation (see, e.g., Johns, 2021; Johns et al., 2012).

Method

Participants

A total of 43 participants (31 females), between the ages of 19 and 56 (M = 26, SD = 7) were presented with the gating task. They were all native speakers of English (i.e., learned English before the age of 3) and used it as a dominant language.

Materials

We used the Wisconsin X-ray microbeam database (XRMB) as the basis for the stimuli (Westbury et al., 1994). The XRMB database includes naturally spoken utterances from 57 different speakers, with most of them speaking an Upper Midwest dialect of American English. The utterances were gathered from a set of 56 different tasks, ranging from eliciting vowels, to producing words in citation form, and reading large paragraphs of multiple sentences from a book. For each task, the database includes the recording of participants' acoustic speech wave, simultaneously with the motion of eight articulatory pellets, at a sampling rate ranging from 40 to 160 Hz (average of 74 Hz). The articulatory pellets were recorded in the midsagittal plane of the vocal tract and include the *upper lip* (UL), *lower lip* (LL), *tongue tip* (T1), *tongue front* (T2), *tongue back* (T3), *tongue root* (T4), *lower jaw* (MNI), *and upper jaw* (MNM). Given that the XRMB database includes both articulatory and acoustic data measurements, all materials for the current experiment were computed from this database.

Stimuli

Stimuli for this experiment consisted of 396 token items, which were derived from 12 target words produced in citation form. The target words were divided into segments using three different gating methods, resulting in 6 to 20 segments for each word. Token items were generated from the recordings of speaker "JW25" in the XRMB database. The speaker was a 24-year-old female from Madison, Wisconsin, who was a native speaker of English with no formal knowledge of any language other than English. The primary criterion when choosing target words was that each phonetic unit be produced by independent articulators as much as possible, in order to ensure that all phonetic units within a given word could be teased apart. Given that words differed in number of syllables, syntactic category, word length, and frequency, these lexical properties were entered as covariates during statistical analyses.

Articulatory Segments. The articulatory segments for each of the target words were parsed using MVIEW (Tiede, 2005). MVIEW is a MATLAB-based program, which displays the positional signal of the eight articulatory pellets, together with the acoustic signal.

Target words were segmented into their individual phonetic units by following these steps: (a) we first determined the primary articulator for each consonant, within each target, using the *findgest* function in MVIEW; (b) this primary articulatory was then used as the landmark for that given consonant; and finally (c) we then used the peak velocity (i.e., the timepoint at which the movement of the articulatory landmark reaches its highest speed during the production of a sound) as the segmenting criterion to identify the boundaries between phonetic units.

We sliced each consonant into two bins. We used the timepoint of *gestural onset* (GONS) of the primary articulator as the bin onset time value, and we used the timepoint of *peak velocity towards* (PVEL-to) the point of maximum constriction as the bin offset time value. Determining the

onset and offset timepoints of the second bins was more complex, given that the primary articulators for different phonetic units within a target word often overlap. While the onset of the second bin always corresponded to the offset timepoint of the first bin (i.e., the timepoint of PVEL-to), the offset of the second bin either corresponded to (a) the timepoint of *peak velocity from* (PVEL-fr) the release point of maximum constriction, or (b) the gestural onset time of the following phonetic unit. In cases where the GONS of the following phonetic unit preceded the PVEL-fr of the current consonant (see Figure 1), the GONS was selected as the offset timepoint of the second bin. The offset of the second bin for the last phonetic unit of a target word always corresponded to the gesture offset time (GOFFS) of that unit.

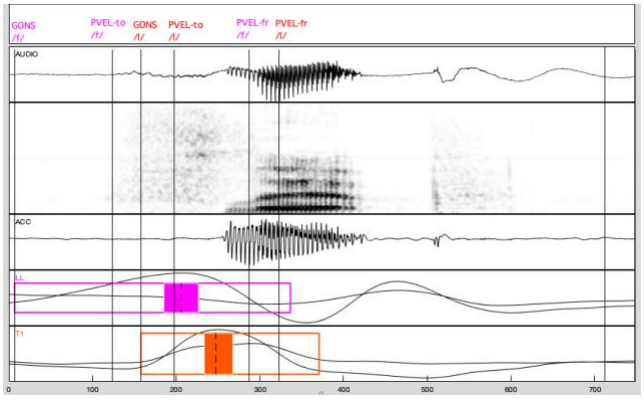

Figure 1: Articulatory segmentation of the first two consonants in the target word /flɪp/. The GONS value of the tongue tip for the /l/ consonant (orange) precedes the PVEL-fr of the lower lip for the /f/ consonant (pink).

We relied on peak velocity as the primary segment measure, rather than the point of maximum constriction, because "timing relationships expressed in terms of time-points of peak velocities are more stable than position-based ones", such as maximum constriction (Hoole & Pouplier, 2015, p. 136; see also Kollia et al, 1995). The timepoint of peak velocity is the point at which much of the acoustic-phonetic information related to the change of phonetic units is available, and thus, may be the time-point that individuals associate as being the uniqueness point for a given consonant during spoken word recognition.

Given that vowels are produced with minimal constriction within the oral cavity—and given that the sagittal midline remains unobstructed during production—they are difficult to segment in accordance to primary articulators (Beauman-Waengler, 2016). Thus, we naturally segmented vowels by using the PVEL-fr from the preceding consonant and the GONS of the following consonant as boundaries. We then divided this segment in half to create two bins.

Acoustic Segments. The acoustic data was segmented using PRAAT (2001). The audio files were extracted from MVIEW using the MATLAB-based program *mat2wav*. We segmented consonants and vowels based on the acoustic criteria described by Ladefoged and Johnson (2015). Vowels were primarily distinguished from consonants by analyzing the patterns of acoustic energy—the onset and offset of voicing, in particular—displayed in the spectrogram. Given that vowels are associated with less obstruction in the oral cavity, they display higher amplitude and greater periodic energy waves than consonants (Beauman-Waengler, 2015). Consonants were segmented by using several forms of acoustic evidence for consonant identity, such as voice bar, aspiration, formant values, anti-formants, formant transitions, noise frequency, periodicity, and aperiodicity of wave energy (Ladefoged & Johnson, 2015).

Target words were segmented into their phonetic units using PRAAT (2001) and were subsequently labelled with text grids. The experimenter recorded the limits of each phonetic unit and then measured the midpoint for each of these units by computing the average between the two boundary measures. Each phonetic unit was then segmented into two bins: (a) the first bin used the onset of a phonetic unit as the bin onset and the midpoint of a phonetic unit as the bin offset; while (b) the second bin used the midpoint value as the bin onset and the onset of the following phonetic unit as the bin offset (see Figure 2).

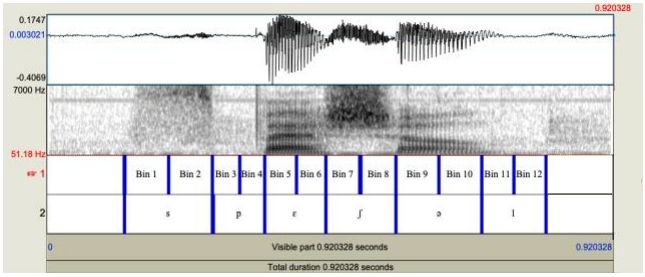

Figure 2: Acoustic segmentation of the target word /spɛʃəl/ and the corresponding bins.

Brute-Force Segments. The length of the brute-force bins was determined by computing the average duration of all articulatory and acoustic bin segments. Thus, target words were segmented into incremental bins of 51.06 ms (SD = 20.11) until the whole word was presented.

Lexical Properties of Corpora Measures. We relied on four social measures of lexical strength, which were computed from a corpus of 57 billion words from Reddit (Johns, 2021): word frequency (WF), contextual diversity (CD), discourse contextual diversity (DCD), and user contextual diversity (UCD). Word frequency is the number of occurrences of a word across all comments in the Reddit corpus. CD is the number of comments a word occurred in (roughly analogous to a context the size of a paragraph). The DCD count is the number of discourses (operationally defined here as a subReddit) that a word was used in—with a maximum value of 30,327, which is the total number of subReddits contained in the corpus. Finally, UCD is the total number of users who used a word in their comments, with a maximum value of 334,345, which is the total number of users in the corpus. Each variable used in the analysis was reduced with a natural logarithm, consistent with past research (Adelman & Brown, 2006; Jones et al., 2012).

Procedure

Participants were seated in a dimly lit room, with noise-cancelling headphones. They were instructed that they would be presented aurally with a word of English, in segments of incremental duration. Their task was to write down what word they thought was being presented and to indicate how confident they were about each guess, on a scale ranging from 0% (completely unsure) to 100% (completely sure), following each segment. The token items (i.e., words segmented into the three gating methods) were counterbalanced among three lists. All lists began with two practice items. Lists were administered online, through Qualtrics (2015), and each list took approximately 20 minutes to complete.

Data Analyses

We first explored how spoken word recognition may be influenced by the social-usage properties of lexical items by conducting hierarchical linear regressions to determine the unique contribution of each of the lexical strength measures. We measured the percentage of change (ΔR^2) engendered by each of the diversity models (UCD, DCD), while controlling for the effect of WF and CD. The percent of ΔR^2 reflects the proportion of variance explained in word recognition times in the gating task that a given diversity predictor engenders over that of WF and CD. We included the following covariates as they significantly improved model fits: phonological neighborhood density, orthographic neighborhood density, word category, word length, and number of syllables.

We then investigated whether different gating paradigms affect (a) the speed at which spoken words are recognized, and (b) participants' confidence ratings at the point of word recognition. To address these questions, we conducted linear mixed effects models (Baayen et al., 2008) using the *lme4* package (Bates et al., 2013) for the R statistical programming environment (R Dev. Core Team, 2014). Our fully fitted models included random intercepts for participants and items as random factors, and the interaction between gating paradigm and CD as fixed factors.[2] We also included two sets of covariates—one for each LME model—as they significantly improved model fits. The word recognition model included the following covariates: phonological neighborhood density, orthographic neighborhood density, word category, and word length. And the confidence ratings model included the following covariates: phonological neighborhood density, orthographic neighborhood density, and word length. We derived *p*-values using the Likelihood Ratio Test by comparing the full model to a nested model excluding the relevant term (Winter, 2013, 2019). Planned comparisons were conducted using *emmeans* with Tukey's correction (Lenth et al., 2018). Effect size measures were derived using the pooled SD between two groups as the standardizer.

[2] We included CD in our LME models given that this variable was shown to explain a greater proportion of variance in recognition times in our dataset than WF, UCD, and DCD.

Results and Discussion

Social Usage Measures on Word Recognition

Before evaluating the individual contribution of each lexical strength measure, we first conducted linear regressions to examine the fit of the individual social-usage measures through their overall R^2 measures. As shown in Figure 3(A), while results revealed similar model fits across the four social-usage measures, diversity measures yielded numerically greater fits to the data than WF.

We then investigated the proportion of unique variance that the UCD and DCD variables explain over that of WF and CD. Given that UCD and DCD are competing models, their contributions were measured through separate regression analyses. We also computed the proportion of unique variance that the WF and CD variables explain over each other, while controlling for UCD and DCD. Results showed that the UCD and DCD variables did not provide significant improvements in performance over that of WF and CD (UCD: $F(1, 482) = 0.57$, $p = .45$, $\Delta R^2 = 0.04$; DCD: $F(1, 482) = 0.57$, $p = .45$, $\Delta R^2 = 0.04$—see Figure 3B and 3C, respectively). There was, however, a noticeable advantage for the WF and CD variables, over those of UCD (WF: $F(1, 482) = 85.3$, $p < .001$, $\Delta R^2 = 6.64$; CD: $F(1, 482) = 186.00$, $p < .001$, $\Delta R^2 = 14.47$—see Figure 3B) and DCD (WF: $F(1, 482) = 87.6$, $p < .001$, $\Delta R^2 = 6.82$; CD: $F(1, 482) = 184.00$, $p < .001$, $\Delta R^2 = 14.29$—see Figure 3C). The CD advantage was of appreciable magnitude considering that CD typically engenders a 6% increase in variance over that of WF (e.g., Adelman et al., 2006). The lack of advantage for the UCD and DCD variables is surprising, given that they have been shown to provide better explanatory power than WF and CD in lexical decision (Johns, 2021) and spoken word recognition tasks (Johns et al., 2012).

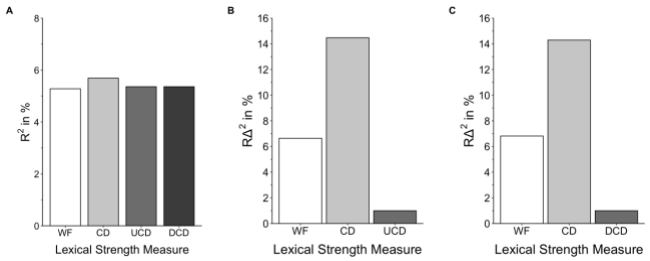

Figure 3. Results from the hierarchical regression analyses for (A) the four lexical strength measures, (B) the UCD variable, and (C) the DCD variable. Panel (A) represents the proportion of variance explained by each of the lexical measures (R^2). For panels (B) and (C), each bar represents the unique variance explained by a given variable, while controlling for the other two variables (ΔR^2).

Together, our results suggest that spoken word recognition may be affected by properties from the social environment. Particularly, words with greater contextual

diversity (i.e., words which occur more frequently, across many contexts) may lead to a facilitation in recognition due to their increased likelihood of being used in future contexts.

Word Recognition Times

The full model was compared to a null model consisting of only random predictors and was found to provide a statistically significant better fit to the data, $\chi^2(14) = 48.30$, $p < .01$, $R^2 = 0.87$, 95% CI [0.71, 0.91]. Results also showed main effects of gating method ($\chi^2(2) = 21.6$, $p < .001$) and contextual diversity frequency ($\chi^2(1) = 5.48$, $p = .02$), but no significant interaction between the two factors. As can be seen in Figure 4, results from planned comparisons revealed that the brute-force gating method yielded significantly faster word recognition times than the articulatory (M_{diff} = -37.81, 95% CI[-57.60, -18.02], $p < .001$, d = -0.19) and acoustic (M_{diff} = -29.72, 95% CI[-49.50, -9.92], $p = .001$, d = -0.14) gating methods. There was no difference in word recognition time between acoustic and articulatory data types (M_{diff} = -8.09, 95% CI[-27.90, 11.73], $p = .60$, d = -0.04).

One potential explanation for the brute-force advantage is that this classic version of the gating paradigm generates bin segments, which contain a myriad of phonetic-acoustic information, rather than isolated, categorical phonetic units. Thus, the traditional brute-force paradigm may have facilitated the selection of the appropriate lexical entry, in comparison to the purely acoustic-phonetically based forms of gating paradigms, due to the gradual accumulation of phonetic information over time. This would suggest that the spoken word recognition system considers the categorical representation of phonemes and the incremental *change of phonemes* as its perceptual natural kinds.

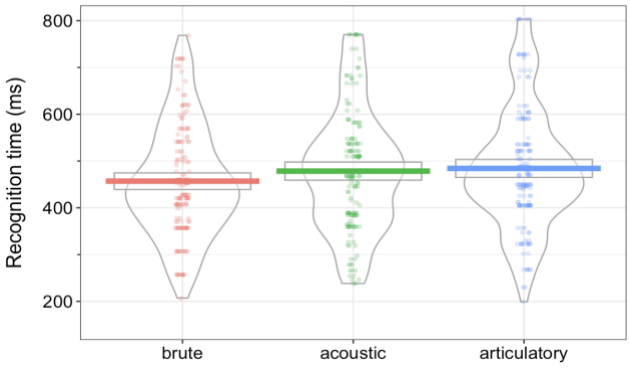

Figure 4. Mean word recognition times by gating type. Individual points represent raw data; central bar represents central tendencies; rectangular band represents the 95% confidence intervals around central tendencies.

Confidence Ratings

Lastly, we also investigated whether participants' confidence ratings at the point of recognition significantly differed as a function of gating paradigm. Our fully fitted model provided a statistically significant better fit to the data than a model consisting of only random predictors, $\chi^2(11) = 22.50$, $p = .02$, $R^2 = 0.31$, 95% CI [0.00, 0.46]. While there was no significant interaction effect, there were significant main effects of gating method ($\chi^2(2) = 6.24$, $p = .04$) and contextual diversity frequency ($\chi^2(1) = 5.02$, $p = .03$). Planned comparisons revealed that confidence ratings at the point of word recognition were significantly lower for the brute-force gating method than the articulatory gating method (M_{diff} = -0.33, 95% CI[-0.56, 0.00], $p = .05$, d = -0.31). There was no difference in confidence ratings between the other gating methods. Thus, although participants recognize spoken words faster when these words are presented through the brute-force gating method, participants are less confident that they have tokened the correct lexical item (Figure 5).

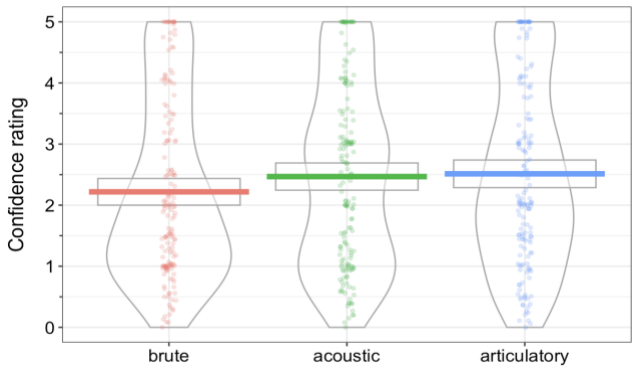

Figure 5. Mean word confidence ratings by gating type.

Conclusion

The goal of the present study was to investigate the units of spoken word recognition relying on the gating paradigm, and the distributional properties of a words' linguistic and social usage pattern. Our results showed that the traditional brute-force gating method yielded significantly earlier recognition points than the articulatory and acoustically driven gating methods. We also found that contextual diversity measures of language use—namely, words that occur frequently, and across many context—seem to be stronger predictors than the traditional notion of word frequency. The results from the present study have important implications for psycholinguistic studies. For one, our results suggest that the brute-force gating paradigm is an efficient way of probing spoken word recognition, thus supporting studies that have employed this paradigm in language perception. Our results also suggest that the usage pattern of words may be abstracted from the social environment and affect word recognition. Particularly, words with greater contextual diversity may be more available within the lexicon and, consequently, may readily interface with the input analysis, thus yielding faster spoken recognition. Taken together, our analyses suggest that the word recognition system operates on multiple constraints, with (a) bottom-up activation of several candidates—akin to what is proposed by the Cohort model—based on the accumulation of a myriad of phonetic-acoustic types of information—and (b) top-down, with usage information making likely candidates more readily available to match the incoming input.

Acknowledgments

We would like to thank Eleonora Albano for raising important questions—long time ago!—that motivated the present experiment. This paper is dedicated to her. Thanks are also due to two anonymous reviewers for their helpful comments on the first draft of this paper, and to Jason Shaw for his comments during the early stages of this project. This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) and from the Social Sciences and Humanities Research Council of Canada (SSHRC).

References

Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, *17*(9), 814-823.

Bates, D., Maechler, M., & Bolker, B. (2011). lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-39

Bauman-Waengler, J. A. (2016). *Articulation and phonology in speech sound disorders: A clinical focus*. London, UK: Pearson.

Bayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390–412.

Bradley, D. C., & Forster, K. I. (1987). A reader's view of listening. *Cognition*, *25*(1-2), 103-134.

Brouwer, S., Mitterer, H., & Huettig, F. (2012). Speech reductions change the dynamics of competition during spoken word recognition. *Language and Cognitive Processes*, *27*(4), 539-571.

Brouwer, S., Mitterer, H., & Huettig, F. (2013). Discourse context and the recognition of reduced and canonical spoken words. *Applied Psycholinguistics*, *34*(3), 519-539.

Bürki, A., Fougeron, C., Gendrot, C., & Frauenfelder, U. H. (2011). Phonetic reduction versus phonological deletion of French schwa: Some methodological issues. *Journal of Phonetics*, *39*(3), 279-288.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology*, *42*(4), 317–367.

Dossey, E., Jones, Z., & Clopper, C. G. (2022). Relative contributions of social, contextual, and lexical factors in speech processing. *Language and Speech*, 1-32.

Elman, J.L. & McClelland, J.L. (1984). Speech perception as a cognitive process: The interactive activation model. In NJ. Lass (Ed.), *Speech and language: Advances in basic research and practice* (pp. 337-374) New York, NY: Academic Press.

Ernestus, M. (2014). Acoustic reduction and the roles of abstractions and exemplars in speech processing. *Lingua*, *142*, 27-41.

Ernestus, M., & Warner, N. (2011). An introduction to reduced pronunciation variants. *Journal of Phonetics*, *39*(SI), 253-260.

Ernestus, M., Baayen, H., & Schreuder, R. (2002). The recognition of reduced word forms. *Brain and language*, *81*(1-3), 162-173.

Frauenfelder, U. H., & Tyler, L. K. (1987). The process of spoken word recognition: An introduction. *Cognition*, *25*(1-2), 1-20.

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

Grosjean, F. (1996). Gating. *Language and cognitive processes*, *11*(6), 597-604.

Hoole, P., & Pouplier, M. (2015). Interarticulatory coordination: Speech sounds. *The handbook of speech production*, 131-157.

Johns, B. T. (2021). Disentangling contextual diversity: Communicative need as a lexical organizer. *Psychological Review*, *128*(3), 525.

Johns, B. T., Gruenenfelder, T. M., Pisoni, D. B., & Jones, M. N. (2012). Effects of word frequency, contextual diversity, and semantic distinctiveness on spoken word recognition. *The Journal of the Acoustical Society of America*, *132*(2), EL74-EL80.

Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology*, *66*(2), 115-124.

Klatt, D.H. (1980). Speech perception: A model of acoustic-phonetic analysis and lexical access. In R.A. Cole (Ed.), *Perception and production of fluent speech*. Hillsdale, N.J: Erlbaum.

Kollia, H. B., Gracco, V. L., & Harris, K. S. (1995). Articulatory organization of mandibular, labial, and velar movements during speech. *The Journal of the Acoustical Society of America*, *98*(3), 1313-1324.

Kocagoncu, E., Clarke, A., Devereux, B. J., & Tyler, L. K. (2017). Decoding the cortical dynamics of sound-meaning mapping. *Journal of Neuroscience*, *37*(5), 1312-1319.

Ladefoged, P., & Johnson, K. (2015). *A course in phonetics* (7th Ed.). Toronto, ON: Nelson Education.

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2018). Package "Emmeans". R Package Version 4.0-3.

Marslen-Wilson, W. (1989). Access and integration: Projecting sound onto meaning. In W. Marslen-Wilson (Ed.), *Lexical representation and process* (pp. 3–24). Cambridge, MA: MIT Press.

Marslen-Wilson, W. (1990). Activation, competition, and frequency in lexical access. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 148–172). Cambridge, MA: MIT Press.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word recognition. *Cognition*, *25*(1-2), 71-102.

Marslen-Wilson, W., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, *8*(1), 1-71.

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*(7-8), 953-978.

McAllister, J. (1991). The processing of lexically stressed syllables in read and spontaneous speech. *Language and Speech*, 34, 1–26.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*(3), 189-234.

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

Pisoni, D. B., & Luce, P. A. (1987). Acoustic phonetic representations in word recognition. *Cognition*, *25*(1-2), 21-52

Qualtrics Platform (2015). Qualtrics, Provo, Utah, USA.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for statistical computing, Vienna, Austria. Available at: http:// www.R-project.org/

Racine, I. & Grosjean, F. La reconnaissance des mots en parole continue: Effacement du schwa et frontier lexicale. *Actes des Journees d'Etudes Linguistiques*, Nantes, 1997

Tiede, M. (2005). MVIEW: Software for visualization and analysis of concurrently recorded movement data. New Haven, CT: Haskins Laboratories.

Tyler, L. K. (1984). The structure of the initial cohort: Evidence from gating. *Perception & Psychophysics*, *36*(5), 417-427.

Van de Ven, M., & Ernestus, M. (2018). The role of segmental and durational cues in the processing of reduced words. *Language and Speech*, *61*(3), 358-383.

Van Petten, C., Coulson, S., Rubin, S., Plante, E., & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 394.

Warren, P., & Marslen-Wilson, W. (1987). Continuous uptake of acoustic cues in spoken word recognition. *Perception & Psychophysics*, *41*(3), 262-275.

Westbury, J. R., Turner, G., & Dembowski, J. (1994). X-ray microbeam speech production database user's handbook. *University of Wisconsin*.

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. *arXiv preprint arXiv:1308.5499*.

Winter, B. (2019). *Statistics for linguists: An introduction using R.* Abington, UK: Routledge.

Zhang, X., & Samuel, A. G. (2018). Is speech recognition automatic? Lexical competition, but not initial lexical access, requires cognitive resources. *Journal of Memory and Language*, *100*, 32-50.

Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition*, *32*(1), 25-64.