ORIGINAL PAPER

Computational modelling of an auditory lexical decision experiment using jTRACE and TISK

Filip Nenadić[a] and Benjamin V. Tucker[a]

[a]Department of Linguistics, University of Alberta, Edmonton, Canada

ARTICLE HISTORY
Compiled April 29, 2020

ABSTRACT
We present a series of computational simulations of the auditory lexical decision task using the jTRACE and TISK models of spoken word recognition. Simulation 1 replicates high accuracy in word recognition and similar performance of these models using the small, default dictionary. Simulation 2 expands the set of words and phonemes, leading to issues in representing certain phonemes in jTRACE. Simulation 3 expands the lexicon of competitors and we find that TISK struggles to select the target word as the winner. Finally, Simulation 4 shows that the decision criteria employed leads to many false positives when pseudowords are presented to the model. None of the model estimates of the time cycle when the winner should be selected predicted participant response latency in the auditory lexical decision task. We discuss these findings and offer suggestions as to what a contemporary model of spoken word recognition should be able to do.

KEYWORDS
spoken word recognition; auditory lexical decision task; computational modelling; TRACE; TISK

1. Introduction

When someone calls your name or shouts a warning, you, as the listener, recognize the message in less than a second, duration of the acoustic signal included. This remarkable process of spoken word recognition has been an important topic of investigation within the field of psycholinguistics and numerous explanations of how it unfolds have been offered. Most current models of spoken word recognition adopt the metaphor of word activation — the notion that a signal stretch "activates" items in the lexicon based on their matching characteristics — from the so-called first-generation models, such as the logogen model (Morton, 1969) or the frequency ordered bin search model (Forster & Bednall, 1976; Taft & Forster, 1975). As the signal incrementally unfolds in time, the items compete in their activation, until finally a winner is selected.

In the past three decades, models of spoken word recognition have become increasingly detailed and complex. This increase in complexity has likely been enabled by the concurrent development of accessible computational power. In other words, models of spoken word recognition are now predominantly computational, rather than purely verbal models. However, computational models that allow simulation ordina-

CONTACT Filip Nenadić. Email: nenadic@ualberta.ca

rily received their most thorough testing in the very process of their creation, even though model testing is crucial to improve them and to generate hypotheses for behavioral experiments or corpus investigations. Furthermore, performing computational simulations may lead to simulation outcomes that were not intuitively expected based on the verbal theory and the computational setup (see Magnuson, Mirman, & Harris, 2012). Nonetheless, reports on large scale computational simulations are rare because (1) they were computationally demanding (as they still are today), (2) models usually lacked an approachable interface (many still do today), and (3) the data from behavioral experiments was limited in size and variety.

In this paper, we simulate human performance in the auditory lexical decision task using a computational model of spoken word recognition. We use the TRACE II model of spoken word recognition (in the remainder of the text referred to as TRACE; McClelland & Elman, 1986), or more precisely, its Java reimplementation called jTRACE (Strauss, Harris, & Magnuson, 2007) and the more recently developed TISK model (Hannagan, Magnuson, & Grainger, 2013; You & Magnuson, 2018) which is quite similar to the TRACE model. Both instantiations have a relatively accessible interface allowing for independent, third-party use. We compare model performance to the data collected in a large scale behavioral study called the Massive Auditory Lexical Decision (MALD) project (Tucker et al., 2019). To the best of our knowledge, these are the first simulations, and certainly of this scale, to test the performance of jTRACE and TISK in estimating how long the selection of the correct word should take depending on the activation-competition process. To that end, we link two hypotheses: (1) participant response latency in an auditory lexical decision task is taken as an indication of the time it takes for the process of selecting the winning candidate to completed, and (2) activation-competition models of SWR assume that a winning candidate should be selected from a group of competitors once its activation level is in some way significantly higher than the activation levels of other competitors. In other words, in the present paper we test whether jTRACE and TISK activation-competition patterns and isolation of a winning candidate are predictive of the assumed activation-competition process occurring in the listener when they perform an auditory lexical decision task.

1.1. The TRACE model

The TRACE model of spoken word recognition was developed by McClelland and Elman (1986). TRACE accepts mock-speech input as a string of phonemes. Each phoneme in the language is described in terms of its values on seven acoustic pseudofeatures (such as voiced, vocalic, or burst), forming the *feature* level of the model. As the signal unfolds in discrete time slices, pseudofeature values are registered at each time slice, forming a spatial trace of activation. Based on the pseudofeature values registered at the feature level, phoneme units at the *phoneme* level are activated and compete, forming a trace of their own. By default, every phoneme takes up 12 time slices. At the same time (or more precisely, space), activation at the *word* level is contingent on the activation of phoneme units. Finally, traces of word activations are formed across the time slices. During the activation-competition process, even competitors that did not match the beginning of the target word are considered (e.g., both *cabin* and *handle* are competitors to *candle*), which is in contrast with another notable model, COHORT (Marslen-Wilson & Tyler, 1980; Marslen-Wilson & Welsh, 1978). Every unit on the *phoneme* and the *word* level is duplicated many times in order to account for the incremental characteristic of the mock-speech input. Besides excitatory connections between the lower and the upper levels (and similar top-down connections which are by default excluded), TRACE also includes lateral inhibition on all levels.

The TRACE model has been used to simulate a variety of experimental findings since it was first introduced, including the original publication (McClelland & Elman, 1986). Notable independent simulations include, for example, reports on lexical segmentation simulations (Frauenfelder & Peeters, 1990) and the impact competitors have on the recognition point, i.e., the time slice in which the word is recognized (Frauenfelder & Peeters, 1998). However, these initial simulations were performed on a small number of example items as proofs of concept. Since then, the model was used to simulate other language phenomena, and is probably best known for successfully simulating eye-movement data from experiments utilizing the visual world paradigm task (e.g., Allopenna, Magnuson, & Tanenhaus, 1998; Dahan, Magnuson, & Tanenhaus, 2001; Dahan, Magnuson, Tanenhaus, & Hogan, 2001).

The model was not without criticism. For example, certain authors argued against conceptual solutions used by TRACE, such as the existence of feedback, i.e., top-down effects between the word and the phoneme level (Marslen-Wilson & Warren, 1994; Norris, McQueen, & Cutler, 2000), or at least reported findings that the model does not fully account for (see, e.g., Chan & Vitevitch, 2009; Frauenfelder & Content, 2000; Gaskell, Quinlan, Tamminen, & Cleland, 2008; McMurray, Tanenhaus, & Aslin, 2009; Smith, Monaghan, & Huettig, 2017). The biggest issue with TRACE, however, is simply how computationally unfeasible it is due to its complex architecture, an issue stressed by the creators of Shortlist (Norris, 1994). Duplicating units to capture their order in "time" creates a very complex network which has difficulties supporting more than a highly limited set of phonemes and words. Even so, "the original TRACE model, with 14 phonemes and 212 words would require 15,000 units and 45 million connections" (Hannagan et al., 2013, pp. 4), and the model is unable to successfully handle lexicons containing more than 1,000 words.

Regardless of its limitations, TRACE is a powerful tool, and it is still one of the most developed models of spoken word recognition. In the past three decades, the model has remained influential. It is without exception described in overviews of models of spoken word recognition (see, e.g., Jusczyk & Luce, 2002; Magnuson et al., 2012; McQueen, 2007; Protopapas, 1999; Scharenborg & Boves, 2010; Vitevitch, Siew, & Castro, 2018; Weber & Scharenborg, 2012) and is widely used to contextualize or explain experimental findings. Still, the vast majority of hundreds of publications referencing TRACE only briefly mention the model: as of 2011 less than 40 papers report an actual simulation (Chawla & Chillcock, 2019). Most simulations in fact appeared once the model became more accessible as it received its computational implementation in Java (Strauss et al., 2007). This instantiation is named jTRACE and it maintained near-identical performance to the original. Easier use also allowed researchers to even expand some of its options, such as by including a larger set of phonemes (Mayor & Plunkett, 2014) or Mandarin tone (Shuai & Malins, 2017).

1.2. The TISK model

The Time-Invariant String Kernel (TISK) model was introduced by Hannagan et al. (2013). The model was designed to correspond to TRACE and be able to match its performance, but with one important change — whereas TRACE solves the issue of the signal being incremental in time by creating time-specific duplicates of phoneme and word nodes (effectively translating time into space), TISK uses time-invariant nodes which are essentially combinations of two phones (diphones). This change allows TISK to sidestep the already noted inefficiency of TRACE caused by a huge number of connections needed for realistic phoneme inventories and lexicon sizes (see McClelland & Elman, 1986; Norris, 1994; Strauss et al., 2007).

With TISK, input units are directly translated into temporally-ordered phonemes which are then mapped to atemporal single phones and all possible diphone combinations given the input string. For example, the word "bit" creates the phoneme level b - i - t which activates atemporal phones /b/, /i/, and /t/, but also diphone combinations /bi/, /bt/, /it/, /ti/, /tb/, /ib/. This means that certain words, for example "dog" and "god", activate exactly the same diphones. In order to avoid such overlap, the model gives higher weights to diphone combinations that match input order, so diphone /do/ would receive higher activation in the word "dog" than in "god", and /sn/ would receive higher activation in the word "snap" versus the word "naps" (for more detail see Hannagan et al., 2013). Phones and diphones then activate atemporal unique lexical units (words). Lateral inhibition is present at the phone/diphone and at the word level.

Initial testing of TISK was performed using the same 14 phonemes and the 212-word lexicon (called *slex*) from TRACE and jTRACE. Besides successfully simulating visual world paradigm data, the authors also simulated and compared free single word recognition in the two models. Three criteria for winner selection were used: (1) absolute activation threshold, where the winner is the first word to reach certain activation level (You & Magnuson, 2018, report that the value used in the simulation was .75), (2) relative activation threshold, where the winner is the first word to have an activation higher by .05 than the runner-up, and (3) a time-dependent criterion, in which the winner is the first word that had the highest activation for 10 consecutive cycles. The authors found that both jTRACE and TISK had accuracy rates higher than 95% in free word recognition, except for TISK with the absolute activation threshold criterion, which was accurate in 88% of words. Additionally, the correlations between the time cycle in which the winner was selected were moderate to high for the two models, being .68, .83, and .88 for each criterion respectively. In short, TISK performs quite similarly to jTRACE in some key simulations. Unfortunately, simulation estimates were not compared to actual participant responses.

You and Magnuson (2018) implemented TISK in Python 3, offering detailed guidelines to its use. To the best of our knowledge and up until the time this paper has been finalized, TISK has only been implemented once, even if many more mentions of the new model have been made. Magnuson and You (2018) showed that top-down effects can also be implemented in TISK and expanded the parameter set to include word-to-phoneme weights. The simulations were performed using the same lexicons adopted from TRACE and jTRACE, and the authors found patterns for which they claimed match the findings of previous empirical studies. Furthermore, the authors introduced changes to the parameter set values which did not significantly affect the relationship between jTRACE and TISK simulations.

1.3. The present study

One of the staple experimental tasks used to investigate spoken word recognition is the auditory lexical decision task. This task is a straightforward way to assess whether a certain stimulus or participant characteristic plays a role in the process of spoken word recognition by observing whether it is predictive of response accuracy and latency. Findings from experiments using the auditory lexical decision task have been used for decades to drive the discussion about the spoken word recognition process (for an overview including earlier studies see Goldinger, 1996), and the task, although sometimes augmented by including, e.g., noise, context, or additional online measures, continues to be used (e.g., Balling & Baayen, 2008; Goldstein & Vitevitch, 2017; Sauval, Perre, & Casalis, 2018; Ventura, Morais, Pattamadilok, & Kolinsky, 2004).

Recently, researchers started to more directly address an issue present in the process of item selection in psycholinguistic studies. Since stimuli for the lexical decision (and many other) tasks are selected from the population of words (or other items) in a language, no control over their characteristics can be imposed — effectively making many psycholinguistic studies quasi-experiments. Ordinarily, this forced researchers to carefully select items so that they are equal in a large number of relevant characteristics and different only in the characteristic under investigation. This procedure made the item sets small and potentially special in comparison to the breadth and variability found in the language from which these items were hand-picked. Further limitations were created by the attention span of an average participant (limited session time) and the sheer number of available participants. Auditory lexical decision studies were not exempt.

Although there is no way to exert strict control over natural language, another option is to collect data from a large number of participants responding to a large number of stimuli, with few restrictions in participant and stimulus sampling. This so-called megastudy approach allows for more comfort when generalizing the findings, statistical control of relevant variables, and impartial testing of findings obtained through targeted experiments (Balota, Yap, Hutchison, & Cortese, 2012; Keuleers & Balota, 2015; Kuperman, 2015). Megastudies collecting data from lexical decision tasks now exist for both visual (e.g., Balota et al., 2007; Ferrand et al., 2010; Keuleers, Lacey, Rastle, & Brysbaert, 2012) and auditory modalities (e.g., Ernestus & Cutler, 2015; Ferrand et al., 2018; Tucker et al., 2019). Megastudies have another useful purpose: they are well-suited to be used as benchmarks for computational models, since they represent an impartial dataset of participant behavior that is also large enough to include much more variety than a targeted experiment.

We have seen that TRACE has extensively been used to simulate certain findings from psycholinguistic experiments, such as the time-course of word activation in the visual world paradigm experiments (e.g., Allopenna et al., 1998; Dahan, Magnuson, & Tanenhaus, 2001; Dahan, Magnuson, Tanenhaus, & Hogan, 2001). However, TRACE simulations that directly compare model estimates to participant response accuracy and response latency to particular words in the auditory lexical decision task are rare. McClelland and Elman (1986) show the time course of word recognition on the example of a single word "product", with a small, unrealistic number of competitors ("produce" being the closest competitor and "products" not being included). Other than this, we found only two targeted simulations in which TRACE output was compared to actual behavioral data from a lexical decision task. Chan and Vitevitch (2009) only mention jTRACE simulations in the discussion section to convey that using the model on a small number of items does not distinguish between two particular groups of word stimuli while participants in a behavioral experiment do. Marslen-Wilson and Warren (1994) used a lexical decision task alongside two other tasks to investigate whether subcategorical mismatches affect spoken word recognition in spliced stimuli. The authors also presented TRACE simulations complementing their behavioral experiments. Due to limitations imposed by the phonemes described in TRACE, the lexical decision simulations were performed on 5 sets of words and 5 sets of pseudowords only (a total of 30 different items). Their results showed that the unfolding of TRACE activations (i.e., response probabilities) did not match the patterns in responses to three different types of spliced words used in the lexical decision study. The purpose of the simulation was to investigate patterns of activations of specific kinds of word/pseudoword splices, averaging across conditions. A decision criterion for the word/pseudoword decision was never defined. Another interesting finding was that in the case of spliced pseudowords where the first part of the pseudoword was taken from an actual word, TRACE continued to highly activate that word, which would potentially lead to a high number of false positive responses.

The literature seems to favor the visual-world paradigm over other experimental paradigms, such as the auditory lexical decision task. One reason for favoring the visual world paradigm over the auditory lexical decision task might be that in the visual world paradigm the participants choose their response from one of a few options, allowing the simulation lexicon to be limited only to the presented options. Similarly, in the simulation reported by Marslen-Wilson and Warren (1994), the focus was on observing differences between three stimulus (splice) types of the same word; by design, only the activation of a single word candidate was considered for each item. This in turn does not require large lexicons or complete phoneme sets in computational simulations, neither of which could be supported by (j)TRACE. Contrary to that, a stimulus presented in an auditory lexical decision experiment can be any word (or even a pseudoword!) of the language, and the competition processes includes all plausible candidates at any given point in time as the acoustic signal unfolds. Ultimately, the lack of simulations that allow for realistic unrestricted competition, and furthermore the lack of direct comparison with actual participant data, means that it has not been reported in the literature how well TRACE and its instantiations can match the competition process occurring when actual human listeners perform the task. For comparison, simulations using lexicons of more than 20,000 words have been reported for other notable models of SWR, such as Shortlist A and B (Norris & McQueen, 2008; Norris, McQueen, & Cutler, 1995) or DIANA (ten Bosch, Boves, & Ernestus, 2015). We believe that simulations using large (realistic) lexicons are extremely important in the investigation of spoken word recognition.

In this report, we present a series of simulations of participant performance in an auditory lexical decision task using jTRACE and TISK. To the best of our knowledge, these are the first such simulations using large lexicons in these two prominent models. Estimates generated by jTRACE and TISK should simulate the activation-competition process and therefore be predictive of participant response latency. The main goal of the study is to learn about the process of spoken word recognition and to inform TRACE/TISK and other models of spoken word recognition by observing how these models perform when used to simulate a large scale auditory lexical decision study.

In the first simulation, we attempt to replicate the basic finding from Hannagan et al. (2013) that jTRACE and TISK are successful and provide similar estimates in free word recognition when the default dictionary of 212 words (*slex*) is used. We augment this replication by comparing model estimates to actual behavioral data. In the second simulation, we use a different set of 442 words for which we have a larger number of participant responses, making the central tendency estimates for human responses more reliable. An increase in the number of words and their variety also expanded the phoneme set beyond the 14 default phonemes described in TRACE's *slex*. We investigate how jTRACE and TISK perform with a larger phoneme inventory, while still being confined to a relatively small word set. In the third simulation, we put word competition under stricter scrutiny. The default dictionaries do not include a large number of close competitors for every target. Therefore, we preselect close competitors and create separate lexicons for every target to observe close competition. Finally, in the fourth simulation we test model performance when an input string (a pseudoword) is not present in its lexicon. A general discussion brings together the findings from these simulations and offers suggestions as to what a contemporary model of spoken word recognition should be able to do.

All of the data from behavioral experiments, materials (lexicons) used for simulations, simulation scripts for jTRACE and TISK, and R scripts used for data preparation and analysis are available as supplementary material at https://doi.org/10.7939/r3-52m3-a502.

2. Behavioral experiments

The data used in our simulations comes from the Massive Auditory Lexical Decision (MALD) project. MALD is described in Tucker et al. (2019), including detailed information about the participants, stimuli and their recording procedure, and the experimental procedure. Here, we only provide the most important information. Besides the main dataset described below, we also use the data from a branch of the project which was created to replicate and extend the findings from the Goh, Yap, Lau, Ng, and Tan (2016) study. The full datasets are also available at this link: mald.artsrn.ualberta.ca.

2.1. MALD1 experiment

The MALD project includes responses by many participants to many auditory recordings of actual English words and phonotactically licit pseudowords. We use data from the *MALD1* database, which includes responses from native monolingual English listeners only.

2.1.1. Sample

The MALD1 participants were 231 monolingual English listeners recruited from the University of Alberta (180 females, 51 males; age M = 20.11, SD = 2.39). The participants received partial course credit for participation in the experiment.

2.1.2. Stimuli

Stimuli were recordings made by one 28-year-old male speaker of Western Canadian English. A total 26,800 words and 9,600 pseudowords were split into 67 word and 24 pseudoword sets each containing 400 unique items. Each word set was then paired with two different pseudoword sets to create a total of 134 experimental lists containing 800 items (400 words + 400 pseudowords each).

2.1.3. Procedure

The experiment was conducted in sound-attenuated booths equipped with a computer monitor, headphones, and a button box. The participants were presented with stimuli using the E-Prime experimental software (Schneider, Eschman, & Zuccolotto, 2012). Each stimulus was preceded by a 500 ms fixation cross. The task for the participants was to decide whether the stimulus they heard was a word of English or not by pressing one of two designated buttons on the button box. The participants made the "word" response with their dominant hand and the "non-word" response with their non-dominant hand. Responses could be made during stimulus presentation, which would interrupt it and the experiment would proceed to the next fixation cross and stimulus. The participants had three seconds to respond and if no response was registered in this time the experiment would proceed to the next fixation cross and stimulus. Stimulus order was randomized per participant.

Each participant completed a single experimental list during the session. However, the participants could return for up to three sessions, each time responding to a new experimental list which did not contain word and pseudoword sets they have already encountered. A total of 284 sessions (experimental lists) were completed.

Currently, the MALD1 dataset includes responses from well over 200 participants. However, since each participant responds to a smaller subset of a large number of words, the number of responses to a particular word rarely exceeds five. When only correct responses are taken into consideration, estimates of a general tendency (mean) of participant response latencies become less reliable.

2.2. MALD_semrich experiment

In contrast to the MALD1 dataset, MALD_semrich dataset, collected to replicate the Goh et al. (2016) study, offers responses from 27 participants to all the stimuli in the experiment. This allows for greater reliability of mean response latency estimation, but still uses a large-enough set of 442 English nouns and 442 MALD pseudowords, enabling calculations of correlation between behavioral tendencies in responses and model estimates. Logged frequency distribution from the Corpus of Contemporary American English (COCA; Davies, 2009) in the two word sets (*slex* and MALD_semrich) had a similar, near-normal distribution, although the mean logged frequency in the MALD_semrich set was slightly lower than in *slex* words.

2.2.1. Sample

Twenty-seven monolingual native speakers of Canadian English (15 females, 12 males; age M = 20.67, SD = 2.79) participated in the experiment. The participants were students at the University of Alberta and received partial course credit for participating in the experiment.

2.2.2. Stimuli

Stimuli were word and pseudoword recordings created as part of the MALD project (Tucker et al., 2019) described above. Out of 468 nouns used in Goh et al. (2016) study, 442 were available within MALD stimuli. We randomly selected 442 MALD pseudoword recordings to complement the word stimuli.

2.2.3. Procedure

The same procedure was followed as for the MALD1 experiment. The only differences were that the list included 884 items in total, instead of 800, and that the participants completed only this list in a single session.

3. Central tendencies in participant response latencies

A computational model of spoken word recognition simulating an auditory lexical decision experiment is attempting to predict per-item general tendencies in participant responses, i.e., resemble an average performance on a certain item. There are many ways in which an "average performance" could be calculated, but also a number of factors that affect participant responses which are not necessarily considered in the computational model. We decided to represent general tendencies in behavioral data in three ways, each of which takes into account an additional source of variation in participant response latency — potentially assisting the model in making better predictions.

First, we use the most simple measure of mean logged response latency per item. Only correct responses are included in the calculation and the response latencies are logged to approximate a normal distribution. This measure removes some of the individual variation between participants and also some random variation between particular responses, giving a more general estimate of how much time it takes to recognize a certain item. In the remainder of the text, we will refer to this measure as mRT.

Second, we take into account the so-called "local effects" by de-trending participant responses (ten Bosch, Ernestus, & Boves, 2018). Local effects encompass variation that happens due to the participant's state, rather than their longer-lasting characteristics. Some of these effects include fatigue, attention fluctuation, but also the aftereffects of being exposed to the previous experimental stimuli. These effects have traditionally been taken into account by including the response latency to the previous stimulus as a predictor of the current response latency. More recently, researchers rely on novel statistical techniques, such as calculating and accounting for autocorrelation when using generalized additive mixed modelling (see, e.g., Baayen, Vasishth, Kliegl, & Bates, 2017).

A model of spoken word recognition is not susceptible to local effects in the manner a participant would be, as a model does not get tired, learn, strategize, or have its mind wander. For example, Mirman, McClelland, Holt, and Magnuson (2008) had to specifically label a two-level attention manipulation in order to simulate an ambiguous phoneme identification experiment that was investigating attention effects. When there is no clear manipulation of attention, TRACE and other computational models of spoken word recognition are unable to account for it, and that variation becomes strictly noise.

In this study, we follow the procedure from ten Bosch et al. (2018), who proposed a method of accounting for local effects by de-trending the data ordered by trial. Taking the logged response latencies, the calculation estimates the optimal number of previous responses (trials) that should be considered when estimating the "true" latency of the current trial response (Equation 1). The "predicted" reaction time ($predRT$) represents a weighted average of a number of previous stimuli. A parameter α determines the number of previous stimuli that have an impact on the predicted reaction time. If $\alpha = 1$ then only the first preceding response latency is used, and smaller fractions of 1 indicate a larger number of previous stimuli being taken into account. Finally, the de-trended response latency (dRT) for a particular response r is calculated as the difference between the predicted ($predRT_r$) and the recorded (RT_r) response latency.

$$\begin{aligned} predRT_1 &= RT_1 \\ \forall r > 1 : predRT_r &= \alpha \cdot RT_{r-1} + (1 - \alpha) \cdot predRT_{r-1} \\ dRT_r &= RT_r - predRT_r \end{aligned} \tag{1}$$

The optimal value of parameter α is selected by estimating average pairwise correlations of participant response latencies to the same stimuli. Since de-trending removes some of the variation due to, for example, attention loss or fatigue, correlations between participant responses should increase after the procedure has been applied. In other words, the de-trending procedure eliminates some of the variation stemming from the fact that participants tend to respond with similar speed to consecutive trials. The highest average correlations between participant response latencies were $r = .19$ in MALD1 and $r = .23$ in MALD_semrich for $\alpha = .1$, indicating that responses to ten previous stimuli should be taken into account. We used this value to calculate mean de-trended response latencies to particular stimuli, and we refer to this measure as dRT.

Third, a number of item characteristics have been shown to predict participant response latencies in auditory lexical decision tasks. Effects of some of those predictors can be expected to emerge independently in an incremental activation-competition model given the lexicon of competitors. Such predictors are, for example, phonological neighborhood density, uniqueness point, or the number of phonemes or syllables (word length/duration). Others, however, probably would not — the number of morphemes a word has, its frequency (if not included in the model), and a host of other semantic variables are not included in the simulation, but shape participant responses. Not considering their values makes it more difficult for the computational model of spoken word recognition to match participant performance.

Therefore, we also created statistical linear models to predict dRT. We include jTRACE/TISK estimates as predictors and observe whether their addition increases the linear model fit. In the case of MALD1, the only variable that was considered alongside jTRACE/TISK estimates was logged frequency from COCA (Davies, 2009). The number of morphemes was not included as nearly all *slex* words are monomorphemic. The effects of phonological neighborhood density, phonological uniqueness point, and word "length" variables (number of syllables, number of phonemes, and the duration of the stimulus in milliseconds) are expected to emerge from the competition process. However, since jTRACE and TISK are supposed to simulate the activation-competition process, not just word length, we also tested whether their estimates contribute more to predicting dRT than a simple length variable does. We chose the variable number of phonemes for this purpose, as all phonemes in jTRACE are of equal "duration" in terms of time-slices, and since the phoneme is the basic unit used in TISK.

In the case of MALD_semrich, the model also included the number of morphemes and three semantic richness variables that are significant predictors of response latency to these items (see Goh et al., 2016): concreteness (Brysbaert, Warriner, & Kuperman, 2014), valence (Warriner, Kuperman, & Brysbaert, 2013), and the number of semantic features (McRae, Cree, Seidenberg, & McNorgan, 2005). These variables were not considered in MALD1 as they are only available for a limited number of MALD words, but for all MALD_semrich words.

To summarize, we estimated how well jTRACE/TISK estimates match participant responses in three ways: (1) by comparing them to mRT, which is the mean logged response latency for each item, (2) by comparing them to dRT, which is the mean de-trended logged response latency for each item, and (3) by observing whether a jTRACE/TISK estimate is predictive of dRT alongside other important predictors in a statistical linear model. To check whether using data from both MALD1 and MALD_semrich is warranted, we correlated mRT and dRT estimates for the 442 words appearing in both sets. In the case of mRT, the correlation was $r = .47$, while in the case of dRT it was expectedly higher and equaled $r = .55$. In both cases, the correlation was only moderate, meaning that the central tendency estimates were somewhat different in the two data sets.

4. Simulation 1

In the first simulation we wanted to replicate the findings from Hannagan et al. (2013) regarding the successfulness and performance similarity of jTRACE and TISK in spoken word recognition. Crucially, we expand the simulation by also comparing estimates obtained from the two models to participant response latencies from the MALD1 dataset.

4.1. Simulation setup

4.1.1. jTRACE setup

Hannagan et al. (2013) and You and Magnuson (2018) did not report the parameter values used in their simulations comparing jTRACE and TISK. In Simulation 1, we used four different sets of parameters for jTRACE. These four sets of parameters were selected by observing the default values of jTRACE parameters, the values reported in the original TRACE paper (McClelland & Elman, 1986), and also the parameter values from a simulation provided in the jTRACE gallery called "word recognition". The parameters recorded in these sources varied in two regards. First, the *alignment* was set to either "specified" with the time slice equal to 4 or to "MAX-ADHOC". Details about the two alignments can be found in an appendix to the jTRACE user manual. Second, the value of the resting word activation (*rest.w*) was set to either -.01 or -.1. Table 1 shows the values of these two parameters in the four jTRACE parameter sets we created. All other parameters were set to their default jTRACE values and are available in our supplementary material. Our decision was further supported by the simulations conducted by Magnuson, Mirman, Luthra, Strauss, and Harris (2018), where the authors claim that the parameters used are robust, and also by a comment made by Strauss et al. (2007, pp. 4) stating that "in most simulations, most or all parameters are left at their default values".

Table 1. The variation in the four jTRACE parameter sets used. All other parameter values were set to jTRACE default values.

Parameter set	alignment	rest.w
jTRACE-A	specified	-.01
jTRACE-B	specified	-.1
jTRACE-C	MAX-ADHOC	-.01
jTRACE-D	MAX-ADHOC	-.1

The default phoneme set of 14 phonemes and the default lexicon of 212 words (*slex*) were used in the current simulation. The 212 words were both the target words and the lexicon of competitors considered for each word. After consulting the results figures from the original simulations, the number of cycles for simulating each word was set to 100. We extracted activation values for the top 20 competitors at every time cycle, and then calculated what the winning word should be. Since TRACE has no built-in moment of recognition (Strauss et al., 2007), we used the same criteria as Hannagan et al. (2013): (1) absolute criterion that selects the first word to reach activation level .75, (2) relative criterion that selects the first leading candidate to have an activation level higher than the runner-up by .05, and (3) time-dependent criterion that selects the first word to have the highest activation for 10 consecutive cycles as the winner. For all three criteria we noted the time slice in which the winning candidate was selected.

4.1.2. TISK setup

The TISK simulation also used the default dictionary of 212 words called *slex* and the 14 phonemes that occur in it. The same criteria for selecting the winner as in the jTRACE simulation were used. The simulation parameters were taken from three sources. The first set of parameter values (TISK-A) came from the example code provided by You and Magnuson (2018). The second and third sets were retrieved from Magnuson and You (2018), and we used both the set without feedback (TISK-B) and with feedback included (TISK-C). The exact values of these parameters are given in a table in the supplementary material.

4.1.3. RT comparison

Model estimates of the time cycle when the winner should be selected were compared to mRT, dRT, and used as predictors in the previously described statistical linear models. Out of 212 *slex* pronunciations, 189 were recorded in MALD so although the simulations included the entirety of *slex*, our response latency comparison was restricted to these 189 words. We also noted that there is a number of phonemic homophones in the MALD stimuli that are present in *slex*. Words such as “ark” and “arc” or “troop” and “troupe” have the same pronunciation as recorded in the CMU dictionary which we used for pronunciation referencing (Weide, 2005). Since we cannot know which homophone was intended to be a part of *slex*, and since we do not want to assume that the MALD audio recordings for these homophones are identical, we simply picked the word with higher frequency in COCA (Davies, 2009) when comparing MALD data to simulation estimates.

To reiterate, we used jTRACE with four parameter sets (A, B, C, and D) and TISK with three parameter sets (A, B, and C). In all cases, we used three decision criteria (the absolute, relative, and time-dependent criterion). The estimates generated in these simulations were compared to general tendencies in MALD1 data represented by mRT, dRT, and also in a statistical linear model.

4.2. Results

We first performed a visual inspection of model performance by creating activation-competition plots for both jTRACE and TISK simulations. Figure 1 shows example plots generated based on jTRACE and TISK activations in time. As can be seen, the simulations adequately matched the predictions shared by TRACE and most contemporary models of spoken word recognition — a number of competitors rise in activation

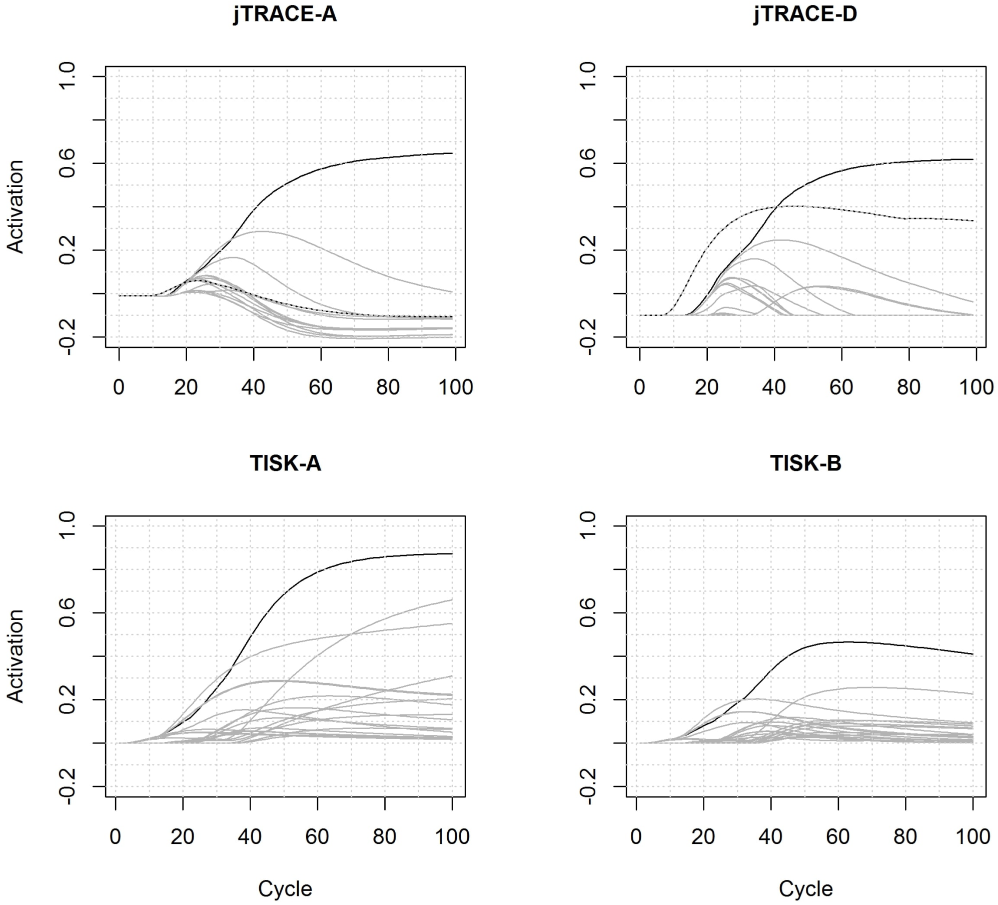

Figure 1. An example of the competition process in the jTRACE and TISK models based on activation values for the word "shield" given in black and 19 closest competitors given in gray. Activation is given on the y-axis, and the time cycle is given on the x-axis. Parameter sets used are given as titles for each plot. The silence phoneme as a word competitor in jTRACE is presented with a dotted black line.

(y-axis) as more of the signal is presented (x-axis). After a while, most competitors will decrease in activation and a small group will continue to rise. In the example we provide, the black line represents the target word "shield", and it is apparent that it stands out in comparison to other competitors in later time cycles. We did not name each competitor in Figure 1, but in all simulations the competitors that received higher activations seemed sensible. In the case of jTRACE, they were words like "she", "sheet", "sheep", and "dull" (there are no words other than "shield" ending with /ld/ in *slex*, so there were no rhyme competitors available). In TISK simulations, high activation is also reached by, e.g., "lid" and "blood", since a match in unordered diphone combinations is important in gaining activation in this model.

Another detail noticeable in Figure 1 is that even though the target word has the highest activation, it does not reach the threshold of .75 except when TISK-A was used. Indeed, free word recognition accuracy across models and model parameters was low when the absolute criterion accuracy was used. For jTRACE, the accuracy was only 10% for parameter sets B and D, and 12% for parameter sets A and C, regardless of the alignment used. For TISK-A accuracy was 86%, but for TISK-B and TISK-C none of the words in the lexicon were correctly recognized. Most often it was the case that the activation level of .75 was never reached by any of the competitors (see the bottom right plot in Figure 1), as higher activation levels under those settings can only be obtained with longer words. Therefore, we lowered the absolute criterion to .4, enabling nearly all of the words to reach this activation level and increasing word recognition accuracy (see below).

Additionally, jTRACE has an entry in the lexicon for silence, and this competitor would often qualify as the winner when the relative and the time-dependent criteria are used — before any other word could become the leading candidate. This was often the case when the MAX-ADHOC alignment was used, as can be seen in the top right plot in Figure 1, for jTRACE-D. Besides the target word given in black, the silence is represented by a dotted black line. It is visible that silence is the leading candidate between cycles 10 and 40, and by more than .05 activation value, qualifying it as the winner using both the relative and the time-dependent criterion. Therefore, we further adjusted our criteria to simply ignore the silence as a potential winner (although we kept it as a competitor and calculated its activation level).

Table 2 shows accuracies for different combinations of models, parameter sets, and decision criteria when simulations are run with the changes noted above. We see that the absolute criterion achieved high accuracies with the change in the jTRACE model. With TISK, accuracies improved, but were still not very high, so perhaps a further reduction in the threshold may be required. The other two criteria performed very well, except when the specified alignment was used in jTRACE (parameter sets A and B).

Table 2. Free word recognition accuracies of *slex* words present in MALD1 for the different parameter sets and winner selection criteria used in the two models.

Model	Parameter set	Criterion accuracy(%)		
		absolute	relative	time-dependent
jTRACE	A	95	82	64
	B	92	94	76
	C	97	98	99
	D	92	99	99
TISK	A	79	97	99
	B	42	97	99
	C	61	97	98

Correlations between response latency estimates in simulations were likewise varied, ranging from no correlation to $r = 1$. High correlations were noted between estimates generated by the same model (jTRACE or TISK) and, in the case of jTRACE, using the same alignment (especially between parameter sets C and D). Correlations between jTRACE and TISK estimates were the highest when jTRACE-A and B were used with the relative and the time-dependent criteria. In that case, certain high correlations were between approximately $r = .7$ and $r = .85$, depending on the TISK parameter set. The calculated correlations are too numerous for all of them to be presented here, but are available in the supplementary material.

The correlation between any of the model estimates and participant responses is much lower. When mRT is used, the correlation ranges between $r = -.07$ and $r = .09$. With dRT, we see some small improvement as all of the correlations increase slightly and three of the model estimates have a correlation above the .1 value (Figure 2).

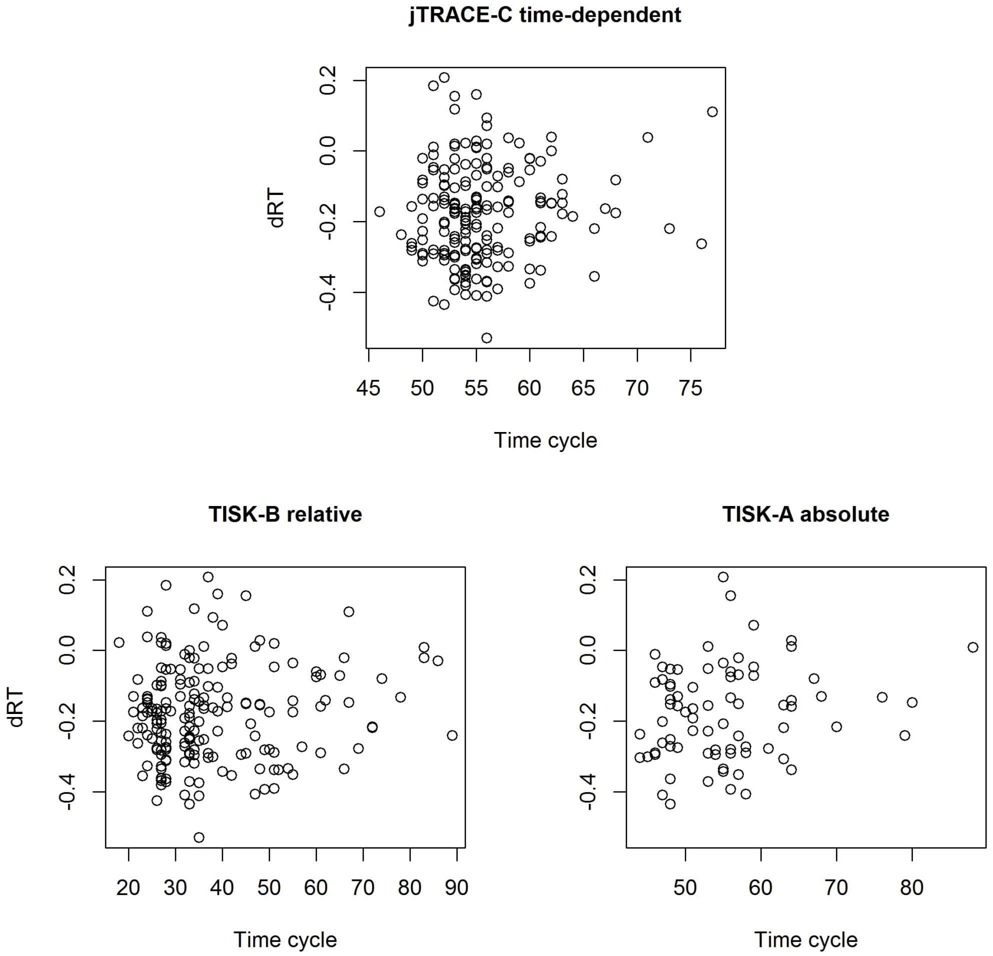

Figure 2. The highest correlations between participant performance and computational model estimates of the time cycle when the winner should be selected recorded in Simulation 1. The time cycle when the model selected the winning word is given on the x-axis, and dRT is given on the y-axis.

jTRACE-C with the time-dependent criterion and TISK-A with the relative criterion used have a correlation of $r = .1$ with dRT. TISK-B with the absolute criterion used has the highest correlation with dRT ($r = .17$), but it should be noted that this setup has a low accuracy rate in free word recognition.

Finally, we fitted separate linear models with dRT as the dependent variable and each of the model estimates as the predictor. We included logged frequency as a predictor, as our simulations did not take it into account. None of the model estimates were significant predictors of dRT. We also noted that in these models the effect of word logged frequency was often non-significant as well. The models are available in our supplementary material.

4.3. Discussion

The initial simulation provided us with important information about implementing jTRACE and TISK to model responses to words in an auditory lexical decision task.

The basic expectations of model performance were met as the activation-competition plots exhibited all of the expected properties of the activation-competition process, with most competitors decreasing in activation, and a singular winner emerging from a smaller group of more persistent competitors later on. Furthermore, we achieved acceptable and even very high accuracy in free word recognition for some of the parameter settings that we used, although we must have had certain parameters different to the simulation reported in Hannagan et al. (2013) and You and Magnuson (2018), given that we had to, for example, reduce the absolute criterion threshold. We also noted a high correlation between jTRACE and TISK estimates in at least some of the setups we used. Together, these results seemed encouraging as we successfully matched previous model simulations.

However, the computational model estimates for the most part failed to match participant performance in the auditory lexical decision task. There were no notable correlations between any of the computational model estimates and mean logged participant response latency per word, de-trended or not. Linear models with frequency included as a predictor showed that the computational model estimates are not significant predictors of participant response latency. Apparently, the model failed to capture and match the same difficulties participants have when responding to words in the experiment.

At the same time, we saw that word frequency also failed to predict de-trended response latencies, even though its effects are well-documented. Given these results, we wanted to compare model estimates to a larger set of more reliable estimates of central tendencies in participant responses. A set of only 189 words, some of which are excluded when the model selects the wrong winner, may be a poor benchmark for the computational model. Furthermore, these 189 words were not all responded to by the same participants, introducing between-participant variability in the central tendency estimate.

5. Simulation 2

Simulation 1 showed that the even though the high performance and similarities between jTRACE and TISK were somewhat replicated, the computational model estimates did not correlate with general tendencies in participant responses from the MALD1 dataset. However, MALD1 includes only a small number of responses per item, and it could be that the calculated general tendencies were less reliable due to between-participant variability. In Simulation 2 we provide a similarly small dataset of words to which we have more than a few participant responses per item. We used MALD_semrich which provides us with up to 27 responses for each of the 442 nouns in the stimulus set.

Importantly, MALD_semrich also includes words containing phonemes other than the original 14 phonemes in the TRACE model. Such a list of words forced us to expand the phoneme set for both models, and allowed us to inspect the performance of jTRACE/TISK under these new conditions. Although jTRACE and TISK seem to perform similarly in Simulation 1, we decided to continue using both models in Simulation 2 as they do not represent their input in the same manner. jTRACE uses pseudofeatures and TISK uses phonemes, so the inclusion of additional phonemes may influence the performance of these two computational models differently.

5.1. Simulation setup

5.1.1. jTRACE setup

The target words and the lexicon of competitors were replaced in comparison to Simulation 1. Instead of using *slex*, we use the 442 MALD_semrich words as target words and as lexicons of competitors. The computational model parameter sets and decision criteria were the same as those used in Simulation 1. Given our initial observations of the absolute criterion threshold being too high and the silence sometimes emerging as the winner, we again reduced the absolute criterion value to .4 and excluded silence as the potential winner in jTRACE.

An important issue that arose in Simulation 2 was how to represent the set of phonemes that are not described in the default phoneme set available in jTRACE. Both the original TRACE model and jTRACE have only 14 phonemes and the silent phoneme described in terms of their seven pseudofeature values. Mayor and Plunkett (2014) expanded this set to include additional phonemes of English and we adopted their phoneme pseudofeature values for our simulations. The only exceptions were diphthongs, affricates, and the r-colored vowel which cannot be represented directly in the TRACE model. The reason for this is that pseudofeatures used in TRACE must have constant values assigned throughout the phoneme duration. In turn, diphthongs, affricates, and the r-colored vowel require for certain characteristics (such as burst or diffuseness for affricates) to change as the phoneme unfolds in time. We decided to represent these phonemes the same way Mayor and Plunkett (2014) did — as combinations of two phonemes with their duration reduced to six time slices, i.e., half of the standard phoneme duration (see Table 1). We hoped that this setup would at least to a degree maintain the relationship between particular speech sounds and their acoustic (pseudo)features. With the new phonemes included, we could now represent all 39 phonemes occurring in the CMU dictionary (Weide, 2005) and therefore in MALD as well. The symbols used to represent all new phonemes and pseudofeature values assigned to them can be found in our supplementary material.

Table 3. Affricates, diphthongs, and the r-colored vowel as operationalized in jTRACE. The duration of component phonemes was halved.

ARPAbet	IPA	Components
CH	ʧ	t+ʃ
JH	ʤ	d+ʒ
AW	aʊ	a+ʊ
AY	aɪ	a+ɪ
EY	eɪ	e+ɪ
OW	oʊ	ɔ+ʊ
OY	ɔɪ	ɔ+ɪ
ER	ɝ	e+r

5.1.2. TISK setup

The same parameter sets and decision criteria were used as in Simulation 1. In the case of TISK, any singular symbol present in the lexicon is considered a separate phoneme. We therefore simply used 1-letter ARPAbet notation for the TISK simulations. Since MALD_semrich includes words longer than the longest word in *slex*, the number of time cycles used in TISK simulations was not limited to 100. Instead, this parameter was left blank, which by default automatically sets it to fit the longest competing word.

5.1.3. RT comparison

Model estimates of the time cycle when the winner should be selected were again compared to mRT, dRT, and used as predictors in the previously described statistical linear models. However, in contrast to Simulation 1, in Simulation 2 we used behavioral data from the MALD_semrich dataset, rather than from MALD1 dataset. All other aspects of this analysis were identical to those from Simulation 1.

5.2. Results

Table 4 shows the accuracies in free word recognition for all model parameters and decision criteria used. jTRACE accuracies are all lower than in Simulation 1, while TISK accuracies are all higher than in Simulation 1.

Table 4. Free word recognition accuracies of MALD_semrich words for the different parameter sets and winner selection criteria used in the two models.

Model	Parameter set	Criterion accuracy(%)		
		absolute	relative	time-dependent
jTRACE	A	77	72	51
	B	77	82	68
	C	79	70	52
	D	83	83	65
TISK	A	82	100	99
	B	74	100	100
	C	90	99	100

We first examined the potential causes for jTRACE to perform worse. The explanation that the lexicon now includes a larger number of words and phonemes did not seem sufficient, as TISK performed better using the same lexicon. The actual cause of lower accuracy in jTRACE simulations were probably diphthongs, affricates, and the r-colored vowel. The average accuracy across all parameter sets and decision criteria was 86% for the words that do not contain these phonemes, and only 55% for the words that do contain them. This discrepancy is sufficient to lower overall accuracies significantly because as many as 46% of MALD_semrich words contain at least one of these phonemes. Affricates, diphthongs, and the r-colored vowel are even more frequent in the CMU dictionary, as at least one of these speech sounds is found in as many as 53% of approximately 116 thousand unique pronunciations.

In TISK, all phonemes are represented merely as symbols, so it is not surprising that accuracy remained high even for words containing phonemes jTRACE struggled with. What is interesting, however, is that we see a further increase in accuracy in comparison to model performance when *slex* was used, even though the sheer number of words and the number of phonemes in the lexicon increased. Although strange at first, this result makes sense when we count the number of close competitors each of the words had in *slex* versus in the MALD_semrich dataset. Using a TISK command, we extracted the number of cohort competitors and rhymes, i.e., items that share the first two or the last two speech sounds with the target, and the number of words in the lexicon that are embedded in their entirety in the target word. On average, *slex* words

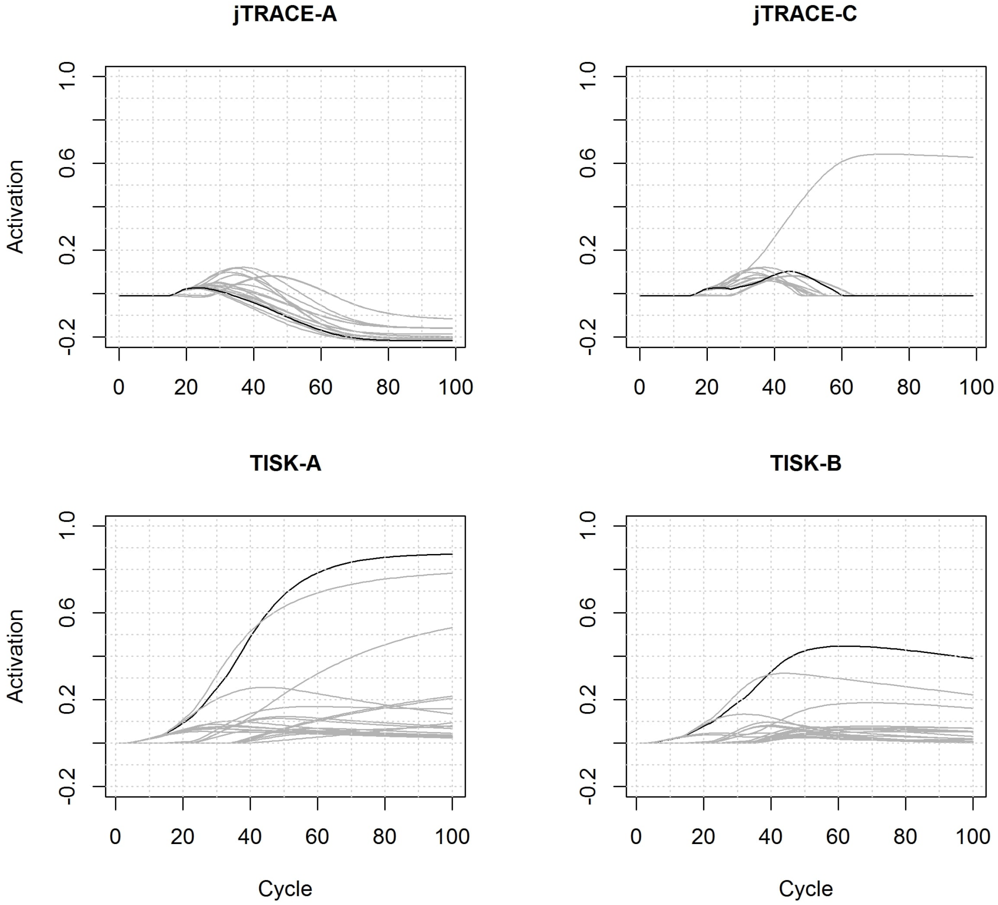

Figure 3. An example of the competition process in the jTRACE and TISK models when MALD_semrich words are used. The figure presents activation values for the word "cherry" given in black and 19 closest competitors given in gray. Activation is given on the y-axis, and the time cycle is given on the x-axis. Parameter sets used are given as titles for each plot.

have seven such close competitors, as the authors designed *slex* to include at least some level of competition. When the MALD_semrich words (which were not designed to investigate competition) are used, the average number of close competitors is less than three, making for easier competition.

This is exemplified in the activation-competition process for the word "cherry" (Figure 3). We see that when jTRACE-A is used all words have very low activations (best competitors were "tent", "telephone", "toy", "pear", "hair", and only towards the last of the 20 were "chair" and "cherry"). When jTRACE-C was used, we see a winner emerge other than the word "cherry", and it was an unlikely winner "stereo". The lower two plots show that TISK had no issue assigning high activation to some competitors, and they were the winning target word "cherry", "chair", and "cheese".

Estimates generated by jTRACE mostly correlated well with each other, and the correlation between TISK estimates were even higher than in Simulation 1 (all higher than $r = .8$ and often close to $r = 1$, except for the absolute criterion in TISK-C, which acted differently in comparison to other setups). However, correlations between estimates coming from the two models further show the discrepancies in jTRACE and TISK simulations. The highest correlations were again obtained when the relative and the time-dependent criterion were used with any jTRACE parameter set. The values of notable correlations ranged between approximately $r = .4$ and $r = .65$, depending on the particular setup. Other correlations were lower, and sometimes even non-existent.

The correlations between mRT and dRT on one side and computational model estimates on the other did not differ much. They also increased in comparison to Simulation 1, ranging from $r = -.08$ and $r = .2$. Ten correlation coefficients were higher than $r = .1$, whereas in Simulation 1 only three setups had such a high value. The highest correlations were recorded using the relative criterion in TISK-A and TISK-B (for detailed information regarding correlations, please consult the supplementary material).

We then fitted a statistical linear model with dRT as the dependent variable and frequency, concreteness, valence, and number of semantic features as predictors. All of these variables acted as significant predictors of dRT, with the semantic predictors contributing approximately 6% to the variance explained. Then we created separate linear models in which we added one of the jTRACE/TISK computational model estimates. In the case of jTRACE, none of the computational model estimates were significant predictors of dRT. In the case of TISK, all of them were, again except when the absolute criterion was used in TISK-C. The linear model summary for the time-dependent criterion using TISK-C is given in Table 5 as an example.

Table 5. Summary of a linear model predicting dRT with a number of standard predictors and TISK-C estimates of the cycle when the winner is selected using a time-dependent criterion.

Coefficients:	Estimate	Std. Error	t value	Pr(>$\|t\|$)
(Intercept)	0.500	0.118	4.233	2.88e-05
Frequency	-1.66e-06	4.15e-07	-4.012	7.22e-05
Concreteness	-0.109	0.025	-4.411	1.33e-05
Valence	-0.013	0.005	-2.426	0.016
N semantic features	-0.005	0.001	-3.285	0.001
time-dependent TISK-C	0.001	2.96E-04	3.044	0.002

Multiple R-squared: 0.17, Adjusted R-Squared: 0.16
F-statistic: 16.31 on 5 and 390 df, p-value: 1.287e-14

However, once we introduced the number of phonemes into the linear models already containing the predictors mentioned previously and TISK model estimate, the number of phonemes was a significant predictor of dRT and the effects of TISK model estimates ceased to be significant. TISK estimates of the time cycle when the winner should be selected correlated highly with the number of phonemes in the word (excluding the absolute criterion in TISK-C), ranging from $r = .77$ to $r = .83$.

5.3. Discussion

The second simulation presented a much richer environment for jTRACE and TISK simulations. We used a novel set of words in the lexicon, expanded the phoneme set, provided more reliable estimates of central tendencies in participant responses, and introduced the number of morphemes and semantic richness measures as additional predictors of participant response latency alongside frequency and computational model estimates.

jTRACE did not fare well under these new conditions. Free word recognition accuracies were lower than in the first simulation and estimates obtained from correctly recognized words did not predict participant response latencies. jTRACE also deviated from TISK model estimates, with the correlations between the two models being noticeably lower. A large portion of errors occurred for words that include affricates, diphthongs, and the r-colored vowel. We see no fitting way of representing these phonemes in jTRACE under its current framework. At the same time, approximately half of English words contain them. Simply put, it does not seem possible for jTRACE to correctly represent word competition when the lexicon is not limited to a small set of preselected word candidates containing only certain phonemes, while the auditory lexical decision task (and many other tasks and everyday scenarios) does not incur such preselecting.

Unlike jTRACE, TISK performed better in free word recognition than in the first simulation, even with a larger lexicon and more phonemes. This is likely due to fewer close competitors present in the MALD_semrich lexicon than in *slex*. We also registered very high correlations between TISK estimates, indicating that changes in parameter values do not affect winner selection under the selection criteria used.

However, we do notice that once again the absolute criterion was a poorer approach to selecting the winning candidate — a certain activation level may never be reached for very short words, and for longer words there is a risk that a plausible candidate may reach the threshold before the target word has made itself distinct. This finding supports a general notion that the overall activation level is not sufficient for selecting a candidate as the winner. Rather, the selection should favor a relative approach, either in terms of relative difference between the winner and the runner-up, or in terms of a candidate leading in activation for long enough. Another potential approach not utilized in our simulations would rely on entropy of top candidates' activation levels, where the winner is selected if the entropy is low.

TISK estimates also correlated with mRT and dRT better than in Simulation 1. This result may be attributed to a more reliable estimate of participant response latencies than when MALD1 data was used, but also may be due to the reduced number of close competitors considered or due to a larger, new set of words being used. Crucially, TISK estimates seemed predictive of participant response latencies, but only until the number of phonemes a word has was introduced into the linear models. It is entirely expected for the TISK model estimates to be related to length characteristics of words such as duration or the number of syllables, and especially the number of phonemes, as the phoneme is the basic unit used in TISK. Still, TISK is an activation-competition model, and it should also be expected that it offers more than what a simple number of phonemes in a word tells us when estimating the process of activation and competition.

We saw that accuracy in TISK increased, while the generated model estimates did not reflect competition, but rather the number of phonemes in a word. Both of these findings may be reflecting of low ecological validity of the competitor set used, as the number of close competitors per word in the MALD_semrich set is very small. Using a larger set of close competitors for every word may allow us to assess free word recognition accuracy in TISK in the more realistic circumstances of difficult, close competition. A more ecologically valid competition could also allow TISK to better represent the activation-competition process in the human listener, which in turn could yield computational estimates that are more in line with participant response latencies.

6. Simulation 3

In Simulation 3 we attempted to represent a more ecologically plausible competition scenario. Previous simulations had limited lexicons and the target words only competed against other words in those lexicons. The results of Simulation 2 showed that a larger number of close competitors may influence free word recognition accuracy. More importantly, computational model estimates did not predict general tendencies in participant response latency well — the contribution of the computational model estimates, where significant, could be completely replaced with the sheer number of phonemes in the word. Such poor performance in predicting human response latencies may also have been caused by lax competition. Therefore, the main goals of Simulation 3 were to provide a challenging word set for the computational model to better test its accuracy, and to allow the model to calculate estimates from a dataset that better represents actual competition, in turn potentially making it a better predictor of participant response latencies.

6.1. Simulation setup

Our initial intention was to use both jTRACE and TISK with a variety of established parameter value sets to simulate the activation-competition process in all 26,800 MALD1 words. However, given the outcomes of Simulation 1 and 2, in Simulation 3 we only used TISK, not jTRACE, and focused solely on MALD_semrich words. The main reason for using only TISK is that jTRACE was unable to correctly recognize many words in Simulation 2 due to poor representation of diphthongs, affricates, and the r-colored vowel. There was no reason to assume jTRACE would perform better with closer competition than it did using only MALD_semrich words as competitors. Additionally, we have seen in Simulation 1 that jTRACE and TISK produce comparable estimates, and this was demonstrated by the authors of TISK as well (Magnuson, Mirman, et al., 2018), so we assumed that results obtained in TISK should translate well to jTRACE should it be able to represent these phonemes.

6.1.1. TISK setup

TISK has been tested on lexicons of up to 20 thousand words and the processing time per word remained very short, being less than a second (You & Magnuson, 2018). However, the CMU dictionary contains a bit over 116 thousand unique pronunciations and our computer was unable to successfully initialize a TISK model when all of them were included. Instead, we used the same TISK command mentioned in Simulation 2 to extract close competitors from the CMU dictionary according to the TRACE model (cohorts, rhymes, and embeddings, that is, items that share the first two or the last two speech sounds with the target, and the number of words in the lexicon that are embedded in their entirety in the target word) for each of the 442 MALD_semrich words. In other words, each MALD_semrich word had its own unique lexicon — the only words in the lexicon for each input word were the close competitors of that word. We then created separate TISK models for each of the target words and its close competitors using the same three TISK parameter sets (A, B, and C).

As in the previous simulations, we extracted the winning candidate and the time cycle in which the winner was detected using three decision criteria (absolute, relative, and time-dependent). However, since we saw in Simulation 1 that the absolute criterion set at 0.75 was too high for most words to reach, we were concerned that the model may perform poorly not due to close competition, but due to the wrong decision threshold being used. To circumvent that potential issue, we calculated the time-cycle when the winner should be selected using ten different decision thresholds for each of the three decision criteria. In the case of the absolute criterion, the thresholds used ranged from .3 to .75, increasing by .05; relative criterion thresholds ranged from .01 to .19, increasing by .02; time-dependent criterion thresholds ranged from 6 to 24, increasing by 2 as well.

6.1.2. Exploration of competitor structure effects

If close competition impedes word recognition in TISK, an additional question of interest arises concerning the number and the structure of close competitors needed for the model to make a mistake. To test this, we also conducted separate toy simulations on three arbitrarily selected words ("sofa", "belt", and "clarinet"). We observed how the activation-competition process and winner selection change as the number of close competitors increases and as the considered competitors are closer competitors to the target word. Regarding competitor "closeness", neither jTRACE nor TISK, to the best of our knowledge, have a definition of which competitor is "closer" to the target (all cohorts, rhymes, and embeddings are treated equally, as close competitors). Therefore, we estimated how close of a competitor a certain word is to the target word based on how highly the competitor was activated at word offset when all of the close competitors were included. We then created three subsets of the close competitor lists for "sofa", "belt", and "clarinet" based on these simulations. The first subset included the 200 least activated close competitors (i.e., contained a high number of competitors, but no closest competitors). The second subset included the 20 most activated close competitors (i.e., contained a low number of competitors but all of them were top competitors to the target word). The third subset included the 20 least activated competitors (i.e., contained a low number of the least activated close competitors to the target word). In other words, in this part of Simulation 3 we coarsely vary and explore the effects of the number and the closeness of a word's close competitors on word recognition accuracy.

6.1.3. RT comparison

The same approach as in Simulation 2 was used.

6.2. Results

After creating custom competitor lists for each word, free word recognition accuracy dropped severely (Figure 4). Changing the decision criteria thresholds, for the most part, did not improve model accuracy. Absolute decision criterion remained the least successful of the three decision criteria, never reaching 30% accuracy in any of the parameter sets and thresholds used. Higher threshold values further decreased accuracy. Relative decision criterion threshold increase likewise only decreased model accuracy, and highest accuracies were recorded when a difference between the leading candidate and the runner-up was merely .01. Finally, changing the decision threshold for the time-dependent criterion yielded no differences in free word recognition accuracy, indicating that if the correct word becomes the leading candidate, it will remain the winner indefinitely.

We then investigated why the word recognition accuracy using TISK dropped so

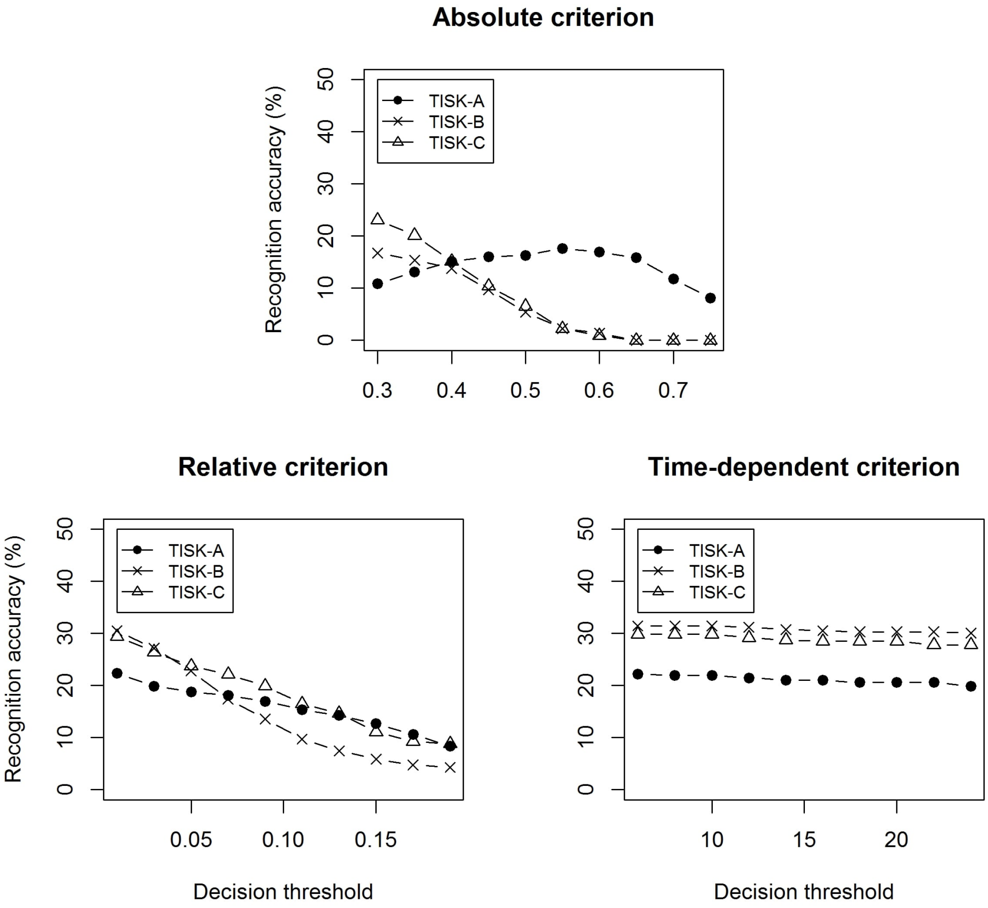

Figure 4. TISK model accuracies in free word recognition per decision criterion (separate figures) and parameter set (separate lines) used. The percent of correctly recognized words is given on the y-axis, while the decision threshold is given on the x-axis.

significantly in comparison to the perfect or near-perfect accuracy rates recorded in certain Simulation 2 setups. The number of close competitors per target word increased dramatically in comparison to previous simulations, as can be seen in Figure 5a, ranging from 17 (for the word "owl") to 2,243 ("deer"). The average number of close competitors was 605, which is close to a hundred times more than in *slex*. In total, as many as 83,122 (71%) out of the 116,726 unique pronunciations in the CMU dictionary acted as a close competitor to at least one of the 442 MALD_semrich words. Furthermore, the number of competitors was significantly lower in those words that were correctly recognized by the model, regardless of the decision criterion or the parameter set used. As an example, Figure 5b shows a box-plot when the time-dependent criterion with the decision threshold of 10 was used in TISK-B.

However, Figure 5b also shows that certain words with more than 500 close competitors are still recognized correctly by the model, while others with few competitors, like the word "owl" with only 17, were not. Figure 5c indicates that the model struggled to correctly recognize shorter words, which have a higher probability of (full) phoneme overlap with other English words. Therefore, we wanted to explicitly test whether it is the number of close competitors, or their composition, that causes inaccuracies in selecting the winning candidate in TISK. We present the activation plots for the word "belt" as an example (Figure 6). When only a small number of very close competitors are included in the model, even though the target word "belt" wins according to the relative and the time-dependent criterion, overall activations remain very low for all competitors (Figure 6b). A similar pattern was observed for "clarinet" (although with a bit higher and less equal activation values) and "sofa". On the other hand, a model created using only the worst close competitors shows a pattern of activation that better resembles the expected ideal, while still allowing the target words to be selected as winners (Figure 6c). Increasing the number of close competitors to 200 leads to another flatlining of activations — even if the close competitors are not the best possible competitors to the target word, their number can weigh down the activation for all considered words (Figure 6d). (It should be noted that it is entirely possible to have a large competitor pool that is very dissimilar to the target word, as in Simulations 1 and 2, in which case the activations are shaped as expected). Again, "clarinet" and "sofa" show similar activation patterns.

Next we tested whether the change in the competitor list also affected the time cycle when the winner is selected by comparing model estimates to those obtained in Simulation 2. In all setups, the time cycle when the winner is selected increased between simulations, with the increase being minimally 17 time cycles for TISK-A when the relative decision criterion is used, and maximally 30 time cycles with TISK-B when the time-dependent criterion is used. Importantly, not only did the time cycle simply increase, it also changed differently for different words. The correlations between the time cycles when the winner is selected for the same setups in Simulations 2 and 3 ranged from $r = .28$ (absolute criterion in TISK-A) and $r = .61$ (relative criterion in TISK-B).

Finally, we observed how model estimates from Simulation 3 correlate with participant response latency. We only considered those setups that had an accuracy rate of at least 20%. The results showed that the two setups that correlated the highest with participant responses were TISK-B with the relative decision criterion threshold set at .03 and TISK-B with the time-dependent criterion threshold set at 24. The accuracy rates for these two setups were 27% and 30%, respectively, and they correlated with dRT somewhat higher than what we observed in Simulation 2 — $r = .27$ and $r = .26$ (Figure 7). Unfortunately, as in Simulation 2, both of these model estimates

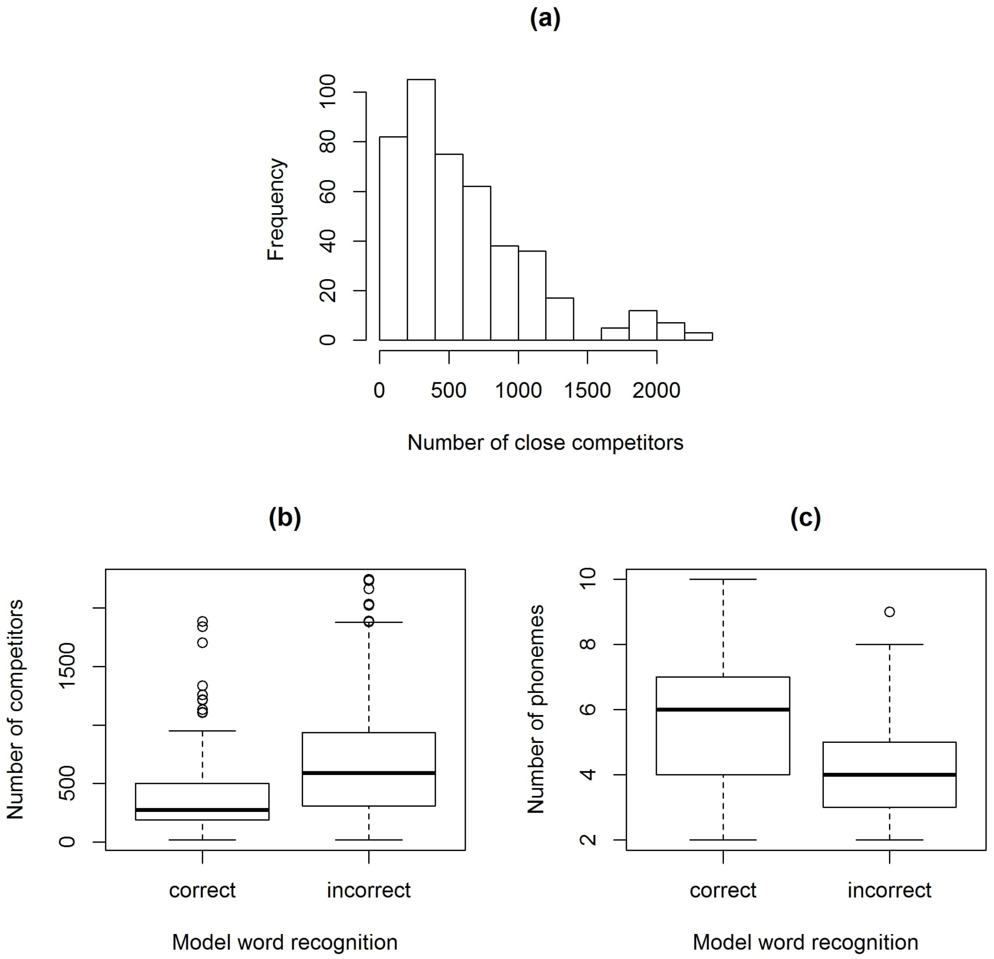

Figure 5. Figure (a) is the histogram of the number of close competitors (cohorts, rhymes, and embeddings) extracted from the CMU dictionary for the 442 MALD_semrich words, with many words having hundreds of close competitors. Bottom figures show that the model more often correctly selected the target word as the winner when there were fewer close competitors (b) and when the number of phonemes in the target word was smaller (c). The accuracies were taken from the simulation using TISK-B parameter set and the time-dependent criterion with the threshold equaling 10.

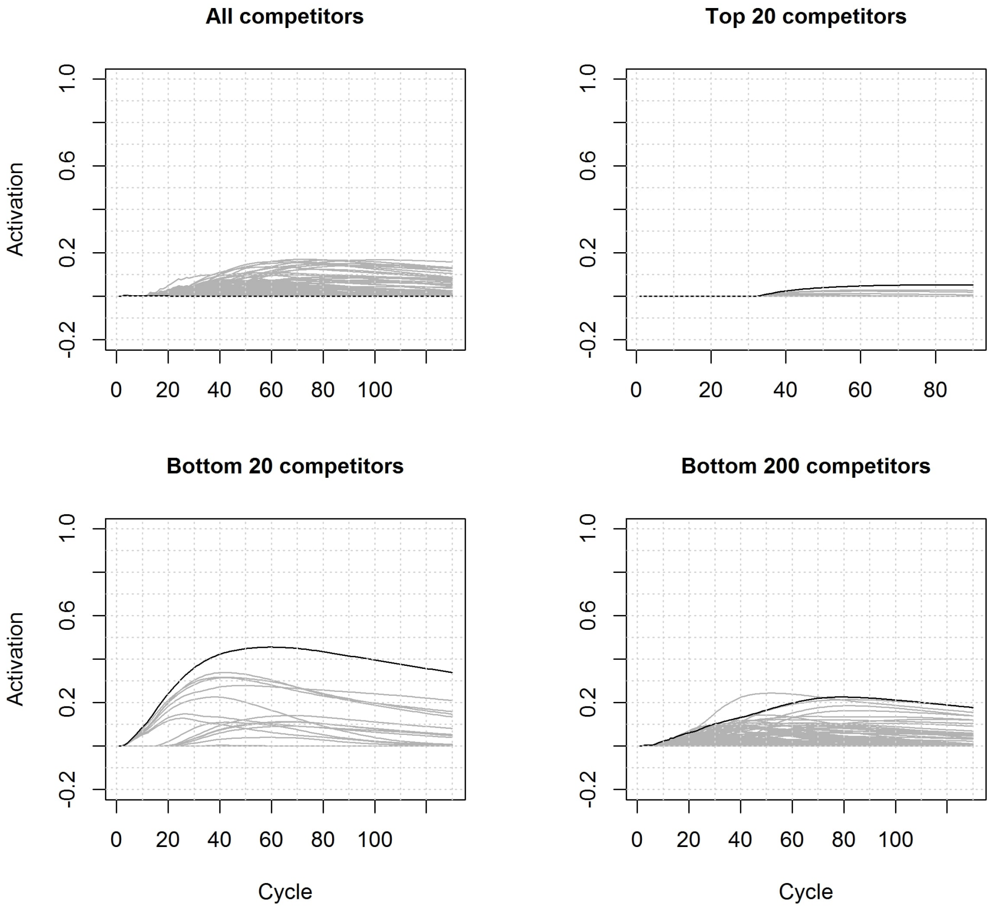

Figure 6. An example of the competition process for the target word "belt" in the TISK-B model when different close competitors are included. Activation is given on the y-axis, and the time cycle is given on the x-axis. The target word is given in black and other competitors are given in gray. The upper left figure labeled "All competitors" presents activation values for all 750 close competitors to the target word "belt", with no particular peaking competitor. Upper right figure labeled "Top 20 competitors" shows activations from a model which only included the top 19 competitors to the target word and the target word, again with very low activation values. The bottom left figure labeled "Bottom 20 competitors" shows the activation values from a model which only included the bottom (worst) 19 competitors to the target word and the target word, showing expected activation patterns and a clear win from the target word. The bottom right figure labeled "Bottom 200 competitors" similarly included bottom (worst) 199 competitors to the target word "belt" and the target word, and in this case there are again no distinct peaking words, similarly to the figure in the upper left corner when all competitors were used.

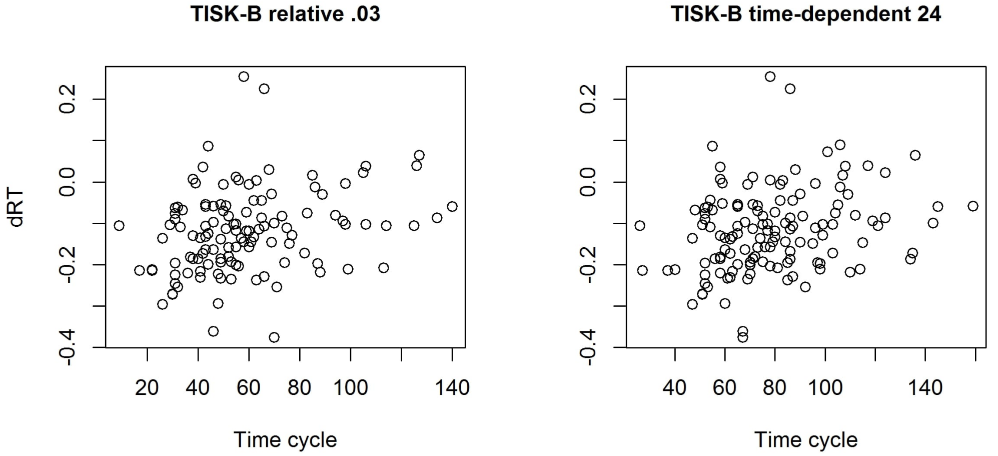

Figure 7. The highest correlations between participant performance and computational model estimates of the time cycle when the winner should be selected recorded in Simulation 3. The time cycle when the model selected the winning word is given on the x-axis, and dRT is given on the y-axis.

only act as significant predictors of dRT in a statistical linear model until the number of phonemes in the word is introduced as a predictor.

6.3. Discussion

The goal of Simulation 3 was to provide the TISK model with a plausible competition scenario, both to test its accuracy, and to allow it to better match participant performance. We created separate lexicons of close competitors for every MALD_semrich word and found that English words have many close competitors, far more than the instantiations of TRACE usually account for. With such an increase in the number of close competitors in the lexicon, word recognition accuracy drops significantly, making the model practically unable to successfully recognize the input, regardless of the decision criterion and threshold.

The decline in accuracy is not caused solely by the number of competitors, as we have seen that the model is successful with, for example, the 442-word MALD_semrich lexicon in Simulation 2, and also with certain words that have many competitors in Simulation 3. Additionally, the result that shorter words are more difficult to be correctly recognized implies that the potential for greater overlap with other words in the lexicon may impede the selection of the target word as the winner. Our targeted simulations showed that even if the correct word is selected as the winner using the relative and the time-dependent criterion, the activation-competition ceases to resemble its standard depictions when a small number of close competitors are selected. On the other hand, 200 competitors, even if they are the least close of the close competitors, altered the activation-competition process in our example words. It seems that both the number and the composition of the close competitors (and especially a combination of the two) may provide insurmountable challenges to the model under the current setup.

Changing the list of competitors for every word affected not just model accuracy, but the time cycle in which the winner is selected. Closer competition forced the model to select the winning word later than in Simulation 2 regardless of the setup. This is not surprising, as the decision criteria require one word (the target word) to make itself distinct from other competitors, and that is more difficult if multiple words share many of the phonemes with the target word. Furthermore, the increase was different for different words, and the correlations between Simulation 2 and Simulation 3 estimates were rarely strong.

This change in model estimates did not translate into much better correlation with participant response latency. Although the correlation between model estimates and dRT somewhat increased, it remained of a low degree. Most importantly, as in Simulation 2, model estimates were unable to predict participant response latency better than the number of phonemes in the word. In other words, we see a clear impact of realistic, close competition on free word recognition accuracy in TISK and on the model estimates of when the winner should be selected. However, these model estimates, when the correct word is selected, remain mostly related to the number of phonemes in the word, and do not seem to be able to predict how long the word recognition process should be in the human listener.

7. Simulation 4

In Simulation 4 we investigate how TISK performs when presented with a word that is not present in the lexicon, that is, when the model is presented with a pseudoword. The decision criteria employed by a model of SWR may be successful in picking a certain target word as the winner, but at the same time may lead to many pseudowords being wrongly recognized as existing words. Previous simulations have shown that TISK performs very well in free word recognition, regardless of the phonemes used, when the competitor set does not include too many close competitors to the target word (i.e., in Simulation 2). We once again give the model its best chance at high performance, and observe whether the decision criteria can discard pseudowords as not present in the lexicon under those same, lax competition conditions used in Simulation 2.

7.1. Simulation setup

Simulation 4 is effectively a repetition of simulations performed using TISK described in Simulation 2, but instead of MALD_semrich words, we presented the model with MALD_semrich pseudowords. The lexicon of competitors was still the same set of 442 MALD_semrich words and we used the same parameter sets and the same decision criteria as in Simulation 2.

In Simulation 4, we did not estimate the time cycle when the decision should be made that the input is a pseudoword. The reason for this was simply because there are no guidelines made by either TRACE or TISK stating how this decision should be made. Therefore, we also made no comparisons between TISK model estimates and participant response latencies to pseudowords in MALD_semrich. The purpose of Simulation 4 was to test whether using the decision criteria that yielded high word recognition accuracy would also cause the model to incorrectly flag pseudoword input as a word.

To make the simulation as comparable to Simulation 2 as possible, we excluded all pseudwords that were longer than the longest MALD_semrich word. We also excluded all the pseudowords that contained phonemes that were not present in the word list. The total number of retained pseudowords was 416.

7.2. Results

MALD_semrich words had on average three close competitors (cohorts, rhymes, and embeddings) present within the other 442 words; MALD_semrich pseudowords on average had 2.56 close competitors within those words. Among these, 101 (24%) pseudowords had no close competitors. In other words, it seemed that it should be fairly easy for TISK to recognize that the input does not match any of the words present in the lexicon.

The results presented in Table 6 show that the relative and the time-dependent criterion perform poorly regardless of the parameter set used. When parameter sets B and C are used with the relative decision criterion we do see a bit of an increase in the number of cases when no word has been selected as the winner, but 4 out of 5 pseudowords still activate a word in the lexicon highly enough for the input signal to be recognized as that word. The best results are obtained using the absolute criterion and parameter sets B and C. The accuracy obtained in these conditions (88 and 93%) might even match participant performance in the auditory lexical decision task fairly well.

Table 6. Accuracy in discarding MALD_semrich pseudowords when MALD_words are used as the lexicon of competitors for different parameter sets and decision criteria in TISK.

Model	Parameter set	Criterion accuracy(%)		
		absolute	relative	time-dependent
TISK	A	1	3	0
	B	88	16	0
	C	93	19	0

7.3. Discussion

Simulation 2 showed that, with lax competition, using TISK parameter set C and the standard decision criteria leads to very high free word recognition. In Simulation 4, we used the same parameter sets and decision criteria and presented TISK with pseudoword input. Our results show that the current setup would lead to an unacceptably large number of mistakes when the relative and the time-dependent decision criteria are used. These errors happen even in those pseudowords that have practically no close competitors in the lexicon that could confuse the model. The absolute criterion performed significantly better (at least when parameter sets B and C were employed). However, previous simulations have shown that the absolute criterion performs the worst in free word recognition with word input, while also being highly dependent on word length and the number of time cycles in the simulation. We discuss these findings in more detail in the following section.

8. General discussion

In the first simulation, we showed that both jTRACE and TISK perform with high accuracy in free word recognition when the default lexicon of 212 words and 14 phonemes is used. The two models also performed quite similarly, especially in certain setups. However, the model estimates did not correlate well with participant response latency. In the second simulation, we expanded the phoneme set to 39 phonemes.

jTRACE was unable to successfully represent diphthongs, affricates, and the r-colored vowel (as combinations of two shorter phonemes), and word recognition accuracies dropped significantly. In contrast, word recognition in TISK was even higher than in the first simulation. The correlations between TISK estimates when the winner should be selected and participant response latency increased slightly. Still, TISK model estimates could completely be replaced by the number of phonemes in the word when predicting response latency. In the third simulation, using TISK only, words competed only against their close competitors. Word recognition accuracies decreased severely and TISK model estimates could again be replaced by the number of phonemes in the word when predicting participant response latency. In the fourth simulation, we show that the decision criteria which yielded very high free word recognition results in Simulation 2 also lead to a large number of false positive responses when TISK is presented with a pseudoword. In short, we found that jTRACE simulations were impeded by poor phoneme representation, that TISK simulations were impeded by close competition, and that neither model provided estimates of when the winning word should be selected that contributed to better prediction of participant response latency, regardless of the setup used. Furthermore, it seems that the decision criteria are not fitting for making a lexical decision task, that is, choosing whether the input signal is present in the lexicon or not. Although we were relatively unsuccessful in simulating participant performance in the auditory lexical decision task, the simulations presented in this paper provided several important insights into the direction in which contemporary models of spoken word recognition should develop, as well as some hypotheses about the spoken word recognition process.

The fact that jTRACE and TISK estimates did not predict participant response latency in our simulations is not an immediate proof of model failure. Magnuson et al. (2012) discuss heuristics for model evaluation, differentiating between issues with the linking hypothesis, parameters used, model implementation, and the theory itself. We believe that the computational model and the participants were presented with similar tasks and, as much as possible for these two computational models, similar input, i.e., that the time cycle in which a winner is selected based on the activation-competition process in jTRACE and TISK should roughly correspond to the average time it takes participants to respond to the word stimulus in the auditory lexical decision task, especially if word characteristics such as frequency or concreteness are accounted for. However, we must also address model implementation and the parameter sets used before assessing the theories behind jTRACE and TISK.

A model of spoken word recognition should be able to represent as much of the variability present in the actual acoustic speech signals as possible. Not only does this allow the model to simulate more of the speech perception phenomena, it also makes it more plausible as it is presented with the same information a human listener is presented with. The best way to achieve this is to use the acoustic signal as input for the model. In jTRACE (TRACE II), the signal is instead represented using a number of acoustic pseudofeatures, and their values define the phoneme set. Although this solution is well-founded and allowed the model to be used in simulations that propelled the field forward, three decades since the model was created and more than a decade since jTRACE was developed as its reimplementation, arguably the biggest issue with using jTRACE is its input representation (i.e., input implementation).

Limits imposed on the lexicon size can be remedied by creating subsets of close competitors, as we did using TISK. A limitation in the set of phonemes that can be represented in the model, however, is not as easily sidestepped. In jTRACE, every occurrence of a phoneme is necessarily equal to every other occurrence of that same phoneme. Phoneme overlap introduced to account for coarticulation slightly alters the signal depending on the preceding and the following phoneme, but it is not uncommon for a phoneme to find itself in the same immediate environment in multiple words, and, regardless, the central part of the phoneme as represented in jTRACE always remains the same. This makes jTRACE unable to account for the fine changes in the acoustic characteristics of speech sounds that affect spoken word recognition (e.g., Andruski, Blumstein, & Burton, 1994; Salverda, Dahan, & McQueen, 2003), and makes the model miss the variability created by various other speaker and contextual factors, phenomena that Fine-Tracker (Scharenborg, 2008, 2009) was specifically developed to simulate. Certain targeted phoneme changes can be made in jTRACE explicitly, but these are made for the purposes of simulating effects on the phoneme level, and cannot reasonably be a part of a large-scale simulation at the word level. Therefore, some of the important topics of investigation in the field of spoken word recognition, such as representing reduction in speech (Ernestus & Baayen, 2007; Ernestus & Warner, 2011; Tucker, 2011; Tucker & Ernestus, 2016) or accounting for other fine variation in the acoustic signal remain outside the realm of abilities of jTRACE, as they can only be presented via coarse changes in pseudofeature values or phoneme splicing.

Additionally, all of the phonemes are practically steady-state phonemes in jTRACE. The pseudofeature values do rise at the beginning and decrease towards the end of a phoneme's presentation, but this change is a fixed value, and the number of time cycles assigned to ramping on and ramping off are necessarily identical for all pseudofeatures. Besides this representation not fitting the reality of the acoustic signal, as, for example, even monophthongs often have a degree of formant value change throughout their production (Hillenbrand, 2013; Hillenbrand, Getty, Clark, & Wheeler, 1995; Nearey & Assmann, 1986), it also disables the model from representing diphthongs or affricates, which are defined by the change in their acoustic (pseudo)features as they unfold. The solution we adopted, the one also used by Mayor and Plunkett (2014), was to create phonemes of half duration and put them together, e.g., create /ʧ/ by combinging /t/ and /ʃ/. This solution is not perfect and, more importantly, it does not seem to allow jTRACE to correctly recognize the target word in our simulations. There may be other solutions. One is, of course, to develop a system in which pseudofeature values rise and fall independently from one another as the phoneme unfolds. Another solution would be to treat a phoneme in jTRACE as internally a-temporal. For example, /ʧ/ would have a relatively high value for both burst and frication at the same time, and these values would ramp on and ramp off together, even though the "burst" part of /ʧ/ happens before the "frication" part. Regardless of the approach taken, jTRACE needs to be able to represent all of the speech sounds in a language or it can only be used to run simulations on limited toy word sets. If this limitation is not also present in the experimental task (as it can be, for example, in the visual world paradigm), any comparison between model estimates and experiment data can only be conceptual, not direct.

TISK does not have this problem as all of the phonemes of English (or any other language) can be represented in it. Our simulations have shown that there is no decrease in word recognition accuracy between TISK Simulation 1 and Simulation 2, even though the number of phonemes increased from 14 to 39. TISK does this by assuming that the phoneme recognition process is already complete, and uses phoneme strings as input. This approach does come at a cost, and we are unsure whether disposing with acoustic pseudofeatures is a step in the right direction. All of the acoustic variability within speech sounds (phonemes) is obliterated with this approach. This approach, however, conflicts with studies which show the importance of sub-phonemic differences (e.g., Andruski et al., 1994; Marslen-Wilson & Warren, 1994) and prosodic cues (e.g., Kemps, Wurm, Ernestus, Schreuder, & Baayen, 2005; Salverda et al., 2003) for speech recognition. Furthermore, the competition process treats all phonemes as equally probable competitors, as in the Neighborhood Activation Model (Luce & Pisoni, 1998), making /sæt/ an equal competitor to /bæt/ as /pæt/, even though /bæt/ and /pæt/ should sound much more similar. Finally, the process which leads to a successful recognition of constituent speech sounds in a word is by no means trivial or easily solved for, and therefore needs to be explained.

Fine-Tracker (Scharenborg, 2008, 2009), SpeM's (Scharenborg, Norris, Bosch, & McQueen, 2005) more contemporary successor, already uses the acoustic signal as input. The most recent additions to the group of models that simulate spoken word recognition, DIANA (ten Bosch, Boves, & Ernestus, 2015) and the discriminative lexicon model based on linear discriminative learning (Baayen, Chuang, Shafaei-Bajestan, & Blevins, 2019), do the same. The authors of jTRACE and TISK themselves notice the issue of the field not moving away from what was intended to be a temporary solution, i.e., using pseudofeatures or phonemes as intermediary layers and assuming these were already successfully recognized, and have already started developing their own solution (EARSHOT) that also relies on actual acoustics (Magnuson, You, et al., 2018).

Combining TISK with a system that recognizes phonemes from the acoustic signal, as in DIANA or Fine-Tracker, could greatly enrich the model and perhaps make its estimates more similar to participant responses. Pilot simulations using DIANA with MALD data indicate that the model can be quite successful in recognizing novel speech input (Nenadić, ten Bosch, & Tucker, 2018). The highest accuracy DIANA attained in free word recognition was approximately 95% with a lexicon of competitors of close to 25,000 words (ten Bosch, Boves, & Ernestus, 2015). Additionally, in comparison to jTRACE and TISK, DIANA shows significantly higher correlations to participant behavior in the auditory lexical decision and word repetition tasks, ranging from $r = .4$ to $r = .76$ (Nenadić et al., 2018; ten Bosch, Boves, & Ernestus, 2015; ten Bosch, Boves, Tucker, & Ernestus, 2015; ten Bosch, Ernestus, & Boves, 2014). However, we are currently investigating the contribution of signal duration to this correlation (similarly to how the number of phonemes highly corresponds to TISK estimates).

A model of spoken word recognition should attempt to provide a representation of the structure of the mental lexicon. There is now a growing body of research showing that semantic variables play a role even in isolated spoken word recognition (Goh et al., 2016; Sajin & Connine, 2014; Tucker et al., 2019). We noted the same albeit modest contribution in the statistical models predicting MALD response latency in this study. However, the mental lexicon is ordinarily presented as a simple list of unconnected units in models of SWR — "lexical access" is treated separately from "meaning access" (Gaskell & Marslen-Wilson, 2002). Currently, jTRACE and TISK can at best include top-down frequency effects to modulate activation.

Rare exceptions to this sort of representation of the mental lexicon are the Distributed Cohort Model (DCM; Gaskell & Marslen-Wilson, 1997) and an approach to modeling spoken word recognition based on discriminative learning (Baayen et al., 2019). DCM and the discriminative lexicon describe units in the lexicon as semantic vectors. In these models, the semantic vectors can be correlated, allowing maps of meaning to be formed. In turn, the competition process and final competitor activations are in part shaped by item characteristics other than frequency. Including a well-developed representation of the mental lexicon is not a primary concern for jTRACE and TISK, but it will be beneficial to future development of models of SWR.

A model of spoken word recognition should provide guidelines to model parameter values. We already mentioned in the introduction and when describing the simulation setup that the parameters in jTRACE and TISK are rarely changed and considered robust. We could only add another comment that the parameters are in "delicate equilibrium" and that their change may unpredictably affect the outcome of the simulation, recounted by the authors of jTRACE (Strauss et al., 2007, pp. 30). Therefore, in all our simulations, we relied on suggested (established and used) parameter sets for both jTRACE and TISK. Our results show that, in general and regardless of the setup, within-model estimates of when the winner should be selected tend to be high, and word recognition accuracies tend to be comparable. For example, nearly all TISK setups in Simulation 2 produce very similar results (even if the activation-competition process, when plotted, does not look the same). This lead us to think that perhaps altering TISK parameter values does not greatly impact the qualitative result (selection of the winning word); the parameter values may indeed be very robust. However, certain setups sometimes performed strikingly worse. For example, TISK-B with the absolute criterion in Simulation 2 and jTRACE-A with the time-dependent criterion in Simulation 1 had notably lower accuracies then other parameter sets with the same decision criterion used, indicating that some changes may greatly impact the results.

At this point, we can only state that the ordinarily used jTRACE and TISK parameter sets are not successful in simulating the auditory lexical decision task. Indeed, Magnuson et al. (2012) mention that the necessity to change model parameters for every simulation can be used as an argument against a model. However, it stands to reason that different parameter sets may be required for simulating different experimental tasks — similarly to how participants (probably) adopt different strategies when confronted with different experimental tasks. Since this is the first time, to the best of our knowledge, that simulations of the auditory lexical decision task were performed by comparing model estimates of the time cycle when the winner is selected to participant responses, it may simply be that different, new parameter values are required.

Therefore, the parameter space of jTRACE and TISK still needs to be further explored. We did not test all possible (plausible) parameter values, and there may yet be a setup that will lead to both higher word recognition accuracy and perhaps better correlation with participant response latency. jTRACE and TISK have a large number of parameters, and each can be fine-tuned using value continua. This makes the number of parameter value combinations exceedingly, unfeasibly large to be tackled using some sort of a hypothesis-driven manual system — considering possible combinations even when we wish to test merely five different values for each parameter would require thousands of simulations. (We attempted various informed manual searches, not reported in this paper, in order to improve the activation-competition process as visible in the figures and word recognition accuracy, but we were unsuccessful.) We believe that contemporary computational power and machine learning approaches may allow researchers to test the full breadth of parameter combinations in search for the optimal solution. Only then would we be able to state with some certainty whether jTRACE and TISK can produce estimates that match participant response latency in the auditory lexical decision task, and then further investigate whether that parameter set can be successfully applied to other comparable datasets without significant changes.

A model of spoken word recognition should incorporate a decision component. DIANA (ten Bosch, Boves, & Ernestus, 2015) is a good example of an end-to-end model of spoken word recognition, as the model defines the decision-making process as well, allowing the researchers to explicitly test whether that aspect of the model is fitting experimental data. Currently, jTRACE and TISK have no built-in function or recommendation as to how the winning word should be selected. In our simulations, we used the three decision criteria used by the developers of jTRACE and TISK to compare the two models (Hannagan et al., 2013), and we even tried modifying the decision thresholds. However, there are many other ways in which the winner could have been selected — and none of these choices, including the ones we used, can be said to be an integral part of jTRACE or TISK. Our simulations have shown that the relative and the time-dependent criterion seem to be better than the absolute criterion in selecting the target word as the winner. This is an important finding, as it seems that the sheer activation level should not be sufficient in spoken word recognition; words and competing candidates differ in their length (number of phonemes) and the density of the competitor pool, and some reach activation levels others do not. In other words, it seems that the decision should be made on the principle that a certain word is simply the best candidate there is (for long enough), regardless of the level of activation generally registered for different words.

However, relative approaches come with a risk. In the auditory lexical decision task, participants are presented with both words and pseudowords. A decision process that correctly selects the target word as the best (winning) competitor from a group of competitors may still select a certain word as the winner even if the input is not in the lexicon (a pseudoword). Simulations with pseudowords act as another test for the decision criterion employed — if a decision criterion recognizes words successfully, but at the same time leads to a word being selected as the winner even though the input is not present in the lexicon, then the decision process needs to be altered. This concern was well represented in our results from Simulation 4: the absolute criterion performed fairly well in lax competition, while the relative and the time-dependent criteria yielded an unacceptably large number of false positive responses.

One option to circumvent the issue of a word being selected with pseudoword input could be to combine decision criteria and select a competitor as the winning word only if it is the best candidate for sufficiently long, but has also reached an absolute activation level that marks it as "sufficiently word-like". Another option is to treat the "word/not word" and "which word?" as two separate decisions (as is currently done in DIANA): perhaps the decision when the leading candidate should be selected as the winning candidate is not necessarily the same decision as the one stating whether the input is present in the lexicon or not. With the decision criteria used in our report, it remains unclear when the "not a word" decision should be made if no candidates ever reach the threshold, as should happen if the input is not a word in the lexicon.

A model of spoken word recognition should be easily accessible and allow even users with lower proficiency in programming to conduct simulations. Assessing model performance and further model development fully depend on conducting simulations and matching the findings with data from behavioral experiments. A model that is accessible to the wider research public can be tested on numerous and varying datasets, where simulations can be replicated. Furthermore, the model can then be tested in its ability to match findings from a wide variety of experimental tasks investigating spoken word recognition, subject to the interest of a particular research group. As we have seen, a certain model with certain parameters may be successful in simulating data from one task, with the same model and setup failing to match participant data from another task.

The jTRACE reimplementation of the TRACE II model allowed many researchers to conduct their own simulations, leading to a significant increase in the number of studies that report computational simulations (Chawla & Chillcock, 2019). Choosing jTRACE and TISK for our simulations was in part governed by the fact that there are not many models of spoken word recognition that a researcher can independently delve into, without the assistance of the developer. However, as the authors of jTRACE note, we still found using jTRACE scripts to be “unfortunately cumbersome” (You & Magnuson, 2018, pp. 876), and its graphic user interface to have numerous errors. In turn, TISK is arguably the most approachable model of spoken word recognition at this time. We have tested and confirmed the claim made by the authors that a user with some experience using platforms such as R (R Core Team, 2018) can successfully navigate TISK simulations in Python, even if they have no experience with that programming language (You & Magnuson, 2018). There are certain features which would be useful to have as part of the standard TISK code, but an advanced (or a persistent) user can expand the code on their own for other purposes. This makes jTRACE and TISK, with their faults, invaluable assets to the field of computational modelling of spoken word recognition.

Finally, it may be that no changes in the model implementation or parameter values would yield high word recognition accuracy and results that fit participant responses. In our attempts to simulate the auditory lexical decision task, the most striking observation was how close the competition between words actually was. The number of very similar competitors that are extracted for every target word using the criteria from notable models such as TRACE (McClelland & Elman, 1986), COHORT (Marslen-Wilson & Tyler, 1980; Marslen-Wilson & Welsh, 1978), or NAM/PARSYN (Luce, 1986; Luce, Goldinger, Auer, & Vitevitch, 2000; Luce & Pisoni, 1998) seems to be creating extensive subsets.

COHORT reduces the list of competitors after the initial two or three phonemes and keeps excluding competitors upon mismatch (but see COHORT II, where slight initial mismatch is allowed; Marslen-Wilson, 1987; Marslen-Wilson, Brown, & Tyler, 1988). However, it may be that the cohort size is unfeasibly large at the very beginning of reducing the list. The mean number of close competitors extracted from the CMU dictionary using the TISK command for the 442 MALD_semrich words is approximately 650, and ranges between 17 and 2,243 close competitors per target word. In comparison, mean number of phonological neighbors (including all the words that are one phoneme edit away from the target word, also based on the CMU dictionary) in all MALD words is only 13, and ranges between 0 and 240 phonological neighbors. Using only NAM neighbors as competitors may therefore benefit model accuracy, and perhaps even the correlation to participant response latency. On the other hand, phonological neighborhoods may be quite extensive in highly inflective languages like Finnish.

Certain suggestions were already made to remedy the issue of competition including too many words. In Shortlist A and B (Norris, 1994; Norris & McQueen, 2008) a smaller number of candidates is selected at each time step and they are the only ones considered in the competition process. In Shortlist B, simulations include over 20,000 competitors to every word, while the focus remains on the small number of “shortlisted” closest competitors only. TISK can also accommodate large lexicons, but still not the entirety of the CMU dictionary, at least with our computational resources. Since we wanted to investigate close competition, in Simulation 3 we manually preselected the lexicon of only close competitors from the CMU dictionary for every target word, effectively “shortlisting” candidates following TRACE’s approach to what should comprise close competitors. The results of the simulation showed that the activation-competition process ceases to resemble the expected distribution when only 20 of the closest competitors are used in the TISK lexicon. Therefore, an application of a manual "shortlisting" solution based on TRACE categorization of close competitors in TISK would still require additional parameter changes to those currently employed in order to obtain acceptable results.

The finding that having 20 closest competitors in the lexicon of competitors prevents TISK from properly performing may have implications to other models of SWR as well. Both Shortlist B and DIANA (ten Bosch, Boves, & Ernestus, 2015), similarly to TISK, allow for large lexicons of 20 to 30 thousand words to be employed. However, we have seen that having a sizable lexicon of 20 or 30 thousand words does not guarantee that all (or most) of the close competitors to the target word are included — 71% of 116 thousand CMU words were a close competitor to at least one of only 442 MALD_semrich words. This indicates that some close competitors would be missing if 20 or 30 thousand word lexicons are used (note: based on TRACE criteria of what comprises close competitors). Ideally, in models of SWR there would be no need to preselect competitors or to create "shortlists" of competitors, but it seems that technical limitations and computational feasibility would likely force researchers to make certain assumptions and adapt their lexicons, at least for now. Furthermore, the question of competitor selection is at the core of many models of SWR. Future simulations should compare multiple competitor selection approaches (e.g., TRACE vs. COHORT vs. Shortlist vs. NAM, etc.) and increase the number of close competitors for every word based on these criteria. It would be very interesting to see how not just large, but also close competition affects model performance in cases of Shortlist B and DIANA, as we have seen it have substantial impact on TISK performance under the current parameter setups used.

Another approach is to assume that the decision is only made once the entirety of the signal is present, which is in line with behavioral data — especially in the case of the auditory lexical decision task (Ernestus & Cutler, 2015; Tucker et al., 2019), where a presumed "word" stimulus could become a pseudoword at any point and less than 3% of all responses are made before signal offset. If we take into account some time for the response to be made — e.g., 200 ms, which is the amount assumed by DIANA (ten Bosch, Boves, & Ernestus, 2015) — 20% of all responses to words are made before this time elapses in the MALD1 dataset. Perhaps the "entire signal" should instead refer to the uniqueness point of the word, and we find in MALD1 that practically no responses are made before the temporal uniqueness point of the word, even when 200 ms are added to represent time needed for executing the response. Although additional investigation is needed to better describe the cause for early responses and what is considered "sufficient information" (e.g., the entire signal or the uniqueness point), it is apparent that certain models of SWR are shifting their focus from the early activation-competition process towards the word offset. The current implementation of the discriminative lexicon (Baayen et al., 2019) abandons the incremental aspect of the process of spoken word recognition. In DIANA (ten Bosch, Boves, & Ernestus, 2015) the simulations include estimates for the time it takes to make the decision which word is the winning word after the signal offset, as it is assumed that in many cases the decision cannot be made until that point. However, we have seen that the solution that disregards the temporality (incrementality) of the signal in TISK, at least with the current simulation setup, was not successful in simulating MALD data.

Yet another possibility is for the listener (and therefore the model) to consider larger chunks of the continuous acoustic signal (more than what would correspond to, e.g., the first two phonemes in the TRACE model). This would reduce the number of plausible competitors, and the model could assess whether a winning word is found, again, at larger time steps than those currently employed. Once it is clear that the signal is complete (i.e., past the signal offset), the decision-making process would pick the best match from the list of remaining (hopefully few) competitors. In other words, the incrementality of the process of spoken word recognition is maintained, but the estimates of competitor activation are based on longer (larger) chunks of the acoustic signal. In a way, TISK already does this by taking into account all of the possible diphone combinations in the word. We also see this inclination in certain learning models of SWR (see Magnuson et al., 2012). Adaptive Resonance Theory (ART; Goldinger & Azuma, 2003; Grossberg, Boardman, & Cohen, 1997; Grossberg & Myers, 2000) stores chunks that can be phonemes, syllables, or even entire words if they co-occurred often enough in the learning process. In the discriminative lexicon approach, (Baayen et al., 2019) the acoustic input is represented using the so-called frequency band summary features (Arnold, Tomaschek, Sering, Lopez, & Baayen, 2017) that are calculated for larger portions of the acoustic signal of a word, e.g., in two or three chunks for a three-syllable word.

Many more simulations, alongside behavioral study findings, are required to test these assumptions and solutions. It is clear that the field of computational modelling of spoken word recognition cannot advance without actual simulations that will adapt model parameters and the models themselves, which in turn is fully dependent on the models being accessible. The most pressing changes that need to be made, especially considering jTRACE and TISK, would include using actual acoustic signal as input, a detailed investigation of how parameter values and decision criteria impact simulation outcomes, and simulations of various experimental tasks and datasets.

Acknowledgements

We thank Matthew C. Kelley for programming assistance and fruitful discussions regarding the paper topic. We thank Julien Mayor for sharing his simulation material and insights. We thank one anonymous reviewer and James McQueen for their helpful comments that improved the quality of the manuscript.

Funding

This project was funded by the Social Sciences and Humanities Research Council: Grant #435-2014-0678.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(4), 419–439.

Andruski, J. E., Blumstein, S. E., & Burton, M. (1994). The effect of subphonetic differences on lexical access. *Cognition*, *52*(3), 163–187.

Arnold, D., Tomaschek, F., Sering, K., Lopez, F., & Baayen, R. H. (2017). Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit. *PloS One*, *12*(4), e0174623.

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., & Blevins, J. P. (2019). 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*, *4895481*.

Baayen, R. H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, *94*, 206–234.

Balling, L., & Baayen, R. H. (2008). Morphological effects in auditory word recognition: Evidence from danish. *Language and Cognitive Processes*, *23*(7-8), 1159–1190.

Balota, D. A., Yap, M. J., Hutchison, K. A., & Cortese, M. J. (2012). Megastudies: What do millions (or so) of trials tell us about lexical processing? In J. S. Adelman (Ed.), *Visual word recognition volume 1: Models and methods, orthography and phonology* (p. 90-115). Hove, England: Psychology Press.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... Treiman, R. (2007). The english lexicon project. *Behavior Research Methods*, *39*(3), 445–459.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, *46*(3), 904–911.

Chan, K. Y., & Vitevitch, M. S. (2009). The influence of the phonological neighborhood clustering coefficient on spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6), 1934.

Chawla, M., & Chillcock, R. (2019). What is the role of computational models in cognitive science? A quantitative and qualitative analysis of the history of the trace model of speech segmentation. *PsyArXiv*. Retrieved from `https://doi.org/10.31234/osf.io/m79fw`

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.

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes*, *16*(5-6), 507–534.

Davies, M. (2009). The 385+ million word corpus of contemporary american english (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*(2), 159–190.

Ernestus, M., & Baayen, R. (2007). The comprehension of acoustically reduced morphologically complex words: The roles of deletion, duration, and frequency of occurrence. In *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 773–776).

Ernestus, M., & Cutler, A. (2015). Baldey: A database of auditory lexical decisions. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1469–1488.

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

Ferrand, L., Méot, A., Spinelli, E., New, B., Pallier, C., Bonin, P., ... Grainger, J. (2018). Megalex: A megastudy of visual and auditory word recognition. *Behavior Research Methods*, *50*(3), 1285–1307.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., ... Pallier, C. (2010). The french lexicon project: Lexical decision data for 38,840 french words and 38,840 pseudowords. *Behavior Research Methods*, *42*(2), 488–496.

Forster, K. I., & Bednall, E. S. (1976). Terminating and exhaustive search in lexical access. *Memory & Cognition*, *4*(1), 53–61.

Frauenfelder, U. H., & Content, A. (2000). Activation flow in models of spoken word recognition. In *Proceedings of the Workshop on Spoken Word Recognition* (p. 79-82). Nijmegen, The Nethelands: Max-Planck Institute for Psycholinguistics.

Frauenfelder, U. H., & Peeters, G. (1990). Lexical segmentation in trace: An exercise in simulation. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (p. 50-86). Cambridge, MA: MIT Press.

Frauenfelder, U. H., & Peeters, G. (1998). Simulating the time course of spoken word recognition: An analysis of lexical competition in trace. In J. Grainger & A. M. Jacobs (Eds.), *Localist connectionist approaches to human cognition* (p. 101-146). Mahwah, NJ: Lawrence

Erlbaum.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes*, *12*(5-6), 613–656.

Gaskell, M. G., & Marslen-Wilson, W. D. (2002). Representation and competition in the perception of spoken words. *Cognitive Psychology*, *45*(2), 220–266.

Gaskell, M. G., Quinlan, P. T., Tamminen, J., & Cleland, A. A. (2008). The nature of phoneme representation in spoken word recognition. *Journal of Experimental Psychology: General*, *137*(2), 282.

Goh, W. D., Yap, M. J., Lau, M. C., Ng, M. M., & Tan, L.-C. (2016). Semantic richness effects in spoken word recognition: A lexical decision and semantic categorization megastudy. *Frontiers in Psychology*, *7*, 976.

Goldinger, S. D. (1996). Auditory lexical decision. *Language and Cognitive Processes*, *11*(6), 559–568.

Goldinger, S. D., & Azuma, T. (2003). Puzzle-solving science: The quixotic quest for units in speech perception. *Journal of Phonetics*, *31*(3-4), 305–320.

Goldstein, R., & Vitevitch, M. S. (2017). The influence of closeness centrality on lexical processing. *Frontiers in Psychology*, *8*, 1683.

Grossberg, S., Boardman, I., & Cohen, M. (1997). Neural dynamics of variable-rate speech categorization. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(2), 481.

Grossberg, S., & Myers, C. W. (2000). The resonant dynamics of speech perception: interword integration and duration-dependent backward effects. *Psychological Review*, *107*(4), 735.

Hannagan, T., Magnuson, J. S., & Grainger, J. (2013). Spoken word recognition without a TRACE. *Frontiers in Psychology*, *4*, 563.

Hillenbrand, J. M. (2013). Static and dynamic approaches to vowel perception. Modern acoustics and signal processing. In G. Morrison & P. Assmann (Eds.), *Vowel inherent spectral change* (pp. 9–30). Springer, Berlin, Heidelberg.

Hillenbrand, J. M., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of american english vowels. *The Journal of the Acoustical society of America*, *97*(5), 3099–3111.

Jusczyk, P. W., & Luce, P. A. (2002). Speech perception and spoken word recognition: Past and present. *Ear and Hearing*, *23*(1), 2–40.

Kemps, R. J., Wurm, L. H., Ernestus, M., Schreuder, R., & Baayen, H. (2005). Prosodic cues for morphological complexity in dutch and english. *Language and Cognitive Processes*, *20*(1-2), 43–73.

Keuleers, E., & Balota, D. A. (2015). Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments. *Journal of Experimental Psychology*, *68*(8), 1457–1468.

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The british lexicon project: Lexical decision data for 28,730 monosyllabic and disyllabic english words. *Behavior Research Methods*, *44*(1), 287–304.

Kuperman, V. (2015). Virtual experiments in megastudies: A case study of language and emotion. *Quarterly Journal of Experimental Psychology*, *68*(8), 1693-1710. (PMID: 25406972)

Luce, P. A. (1986). A computational analysis of uniqueness points in auditory word recognition. *Perception & Psychophysics*, *39*(3), 155–158.

Luce, P. A., Goldinger, S. D., Auer, E. T., & Vitevitch, M. S. (2000). Phonetic priming, neighborhood activation, and parsyn. *Attention, Perception, & Psychophysics*, *62*(3), 615–625.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*(1), 1.

Magnuson, J. S., Mirman, D., & Harris, H. D. (2012). Computational models of spoken word recognition. In M. Spivey, M. Joanisse, & KenMcRae (Eds.), *The Cambridge Handbook of Psycholinguistics* (pp. 76–103). Cambridge University Press Cambridge, UK.

Magnuson, J. S., Mirman, D., Luthra, S., Strauss, T., & Harris, H. D. (2018). Interaction in spoken word recognition models: Feedback helps. *Frontiers in Psychology*, *9*, 369. Retrieved from `https://www.frontiersin.org/article/10.3389/fpsyg.2018.00369`

Magnuson, J. S., & You, H. (2018). Feedback in the time-invariant string kernel model of spoken word recognition. *Proceedings of the Cognitive Science Society*, 732–737. Retrieved from `http://par.nsf.gov/biblio/10097512`

Magnuson, J. S., You, H., Nam, H., Allopenna, P., Brown, K., Escabi, M., … Rueckl, J. (2018). EARSHOT: A minimal neural network model of incremental human speech recognition. *PsyArXiv*. Retrieved from `https://doi.org/10.31234/osf.io/m79fw`

Marslen-Wilson, W. D., & Warren, P. (1994). Levels of perceptual representation and process in lexical access: Words, phonemes, and features. *Psychological Review*, *101*(4), 653-675.

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

Marslen-Wilson, W. D., Brown, C. M., & Tyler, L. K. (1988). Lexical representations in spoken language comprehension. *Language and Cognitive Processes*, *3*(1), 1–16.

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

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*(1), 29–63.

Mayor, J., & Plunkett, K. (2014). Infant word recognition: Insights from trace simulations. *Journal of Memory and Language*, *71*(1), 89–123.

McClelland, J. L., & Elman, J. L. (1986). The trace model of speech perception. *Cognitive Psychology*, *18*(1), 1–86.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2009). Within-category vot affects recovery from lexical garden-paths: Evidence against phoneme-level inhibition. *Journal of Memory and Language*, *60*(1), 65–91.

McQueen, J. M. (2007). Eight questions about spoken-word recognition. In S.-A. Rueschemeyer & G. Gaskell (Eds.), *The Oxford Handbook of Psycholinguistics* (pp. 37–53). Oxford University Press, USA.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547–559.

Mirman, D., McClelland, J. L., Holt, L. L., & Magnuson, J. S. (2008). Effects of attention on the strength of lexical influences on speech perception: Behavioral experiments and computational mechanisms. *Cognitive Science*, *32*(2), 398–417.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, *76*(2), 165.

Nearey, T. M., & Assmann, P. F. (1986). Modeling the role of inherent spectral change in vowel identification. *The Journal of the Acoustical Society of America*, *80*(5), 1297–1308.

Nenadić, F., ten Bosch, L., & Tucker, B. V. (2018). Implementing diana to model isolated auditory word recognition in english. In *Proc. interspeech 2018* (pp. 3772–3776).

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.

Norris, D., McQueen, J. M., & Cutler, A. (1995). Competition and segmentation in spoken-word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(5), 1209.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, *23*(3), 299–325.

Protopapas, A. (1999). Connectionist modeling of speech perception. *Psychological Bulletin*, *125*(4), 410-436.

R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Sajin, S. M., & Connine, C. M. (2014). Semantic richness: The role of semantic features in processing spoken words. *Journal of Memory and Language*, *70*, 13–35.

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition*, *90*(1), 51–89.

Sauval, K., Perre, L., & Casalis, S. (2018). Phonemic feature involvement in lexical access in grades 3 and 5: Evidence from visual and auditory lexical decision tasks. *Acta Psychologica*, *182*, 212–219.

Scharenborg, O. (2008). Modelling fine-phonetic detail in a computational model of word recognition. In *The 9th Annual Conference of the International Speech Communication Association* (pp. 1473–1476).

Scharenborg, O. (2009). Using durational cues in a computational model of spoken-word recognition. In *The 10th Annual Conference of the International Speech Communication Association* (pp. 1675–1678).

Scharenborg, O., & Boves, L. (2010). Computational modelling of spoken-word recognition processes: Design choices and evaluation. *Pragmatics & Cognition*, *18*(1), 136–164.

Scharenborg, O., Norris, D., Bosch, L., & McQueen, J. M. (2005). How should a speech recognizer work? *Cognitive Science*, *29*(6), 867–918.

Schneider, W., Eschman, A., & Zuccolotto, A. (2012). E-prime reference guide [Computer software manual]. Pittsburgh.

Shuai, L., & Malins, J. G. (2017). Encoding lexical tones in jtrace: a simulation of monosyllabic spoken word recognition in mandarin chinese. *Behavior Research Methods*, *49*(1), 230–241.

Smith, A. C., Monaghan, P., & Huettig, F. (2017). The multimodal nature of spoken word processing in the visual world: Testing the predictions of alternative models of multimodal integration. *Journal of Memory and Language*, *93*, 276–303.

Strauss, T. J., Harris, H. D., & Magnuson, J. S. (2007). jTRACE: A reimplementation and extension of the trace model of speech perception and spoken word recognition. *Behavior Research Methods*, *39*(1), 19–30.

Taft, M., & Forster, K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior*, *14*(6), 638–647.

ten Bosch, L., Boves, L., & Ernestus, M. (2015). DIANA, an end-to-end computational model of human word comprehension. In *The 18th International Congress of Phonetic Sciences (ICPhS 2015).* Retrieved from `https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0480.`

ten Bosch, L., Boves, L., Tucker, B., & Ernestus, M. (2015). DIANA: Towards computational modeling reaction times in lexical decision in North American English. In *Interspeech 2015: The 16th Annual Conference of the International Speech Communication Association* (p. 1576-1580).

ten Bosch, L., Ernestus, M., & Boves, L. (2014). Comparing reaction time sequences from human participants and computational models. In *Interspeech 2014: The 15th Annual Conference of the International Speech Communication Association* (pp. 462–466).

ten Bosch, L., Ernestus, M., & Boves, L. (2018). Analyzing reaction time sequences from human participants in auditory experiments. In *The 19th annual conference of the international speech communication association* (p. 971-975). Hyderabad, India: ISCA.

Tucker, B. V. (2011). The effect of reduction on the processing of flaps and/g/in isolated words. *Journal of Phonetics*, *39*(3), 312–318.

Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., & Sims, M. (2019). The Massive Auditory Lexical Decision (MALD) database. *Behavior Research Methods*, *51*(3), 1187–1204.

Tucker, B. V., & Ernestus, M. (2016). Why we need to investigate casual speech to truly understand language production, processing and the mental lexicon. *The Mental Lexicon*, *11*(3), 375–400.

Ventura, P., Morais, J., Pattamadilok, C., & Kolinsky, R. (2004). The locus of the orthographic consistency effect in auditory word recognition. *Language and Cognitive Processes*, *19*(1), 57–95.

Vitevitch, M. S., Siew, C. S. Q., & Castro, N. (2018, 09). *Spoken word recognition.* Oxford University Press.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, *45*(4), 1191–1207.

Weber, A., & Scharenborg, O. (2012). Models of spoken-word recognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*(3), 387–401.

Weide, R. (2005). *The Carnegie Mellon pronouncing dictionary [cmudict. 0.6].* Carnegie Mellon University. Retrieved from `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`

You, H., & Magnuson, J. S. (2018). Tisk 1.0: An easy-to-use python implementation of the time-invariant string kernel model of spoken word recognition. *Behavior Research Methods*, *50*(3), 871–889.