Computational Modelling of Spoken Word Recognition in the Auditory Lexical Decision Task

by

Filip Nenadić

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Linguistics
University of Alberta

Examining committee:

Benjamin V. Tucker, Supervisor
Petar Milin, Supervisor
Michael Kiefte, Supervisory Committee
Chris F. Westbury, Examiner
Juhani Järvikivi, Examiner
Odette Scharenborg, External examiner

Abstract

The process of spoken word recognition has been an important topic in the field of psycholinguistics for decades. Numerous models have been created, many of which received their own computational implementation. However, large-scale simulations using these models performed on the same dataset by an independent researcher are rare at best. In the present dissertation, three models of spoken word recognition (TRACE, DIANA, and the discriminative lexicon approach) are tested in their ability to simulate the spoken word recognition process as captured by the auditory lexical decision task. The simulated data comes from the Massive Auditory Lexical Decision project, a large-scale study that enables us to estimate model performance on thousands of English words and compare it with performance of hundreds of human listeners. The main goals of the present work are threefold. The first goal is to assess models' performance in simulating the auditory lexical decision task. The second goal is to learn about the process of spoken word recognition through differences in models and model setups. The third goal is to provide suggestions for model improvement or future model development. The dissertation begins by outlining the history of development and the current state of computational models of spoken word recognition, motivating the conducted research. The central part of the dissertation is split into three separate sections. The first section describes the TRACE model in more detail and the simulations of MALD data performed using TRACE's re-implementations called jTRACE and TISK. The second section describes an implementation of an end-to-end model of spoken word recognition called DIANA and simulations performed using that model. The third section presents the simulations performed using the discriminative lexicon approach to spoken word recognition. Each of these sections includes a separate discussion of the results, focusing predominantly on the model in question. A joint conclusion brings together the findings from these three separate studies and also includes a suggestion to creating a hybrid model using strong aspects of the tested computational models of spoken word recognition.

Preface

The three chapters are intended to be published as separate research articles. Benjamin V. Tucker was the supervisory author and was involved with concept formation and manuscript composition in all three chapters.

Chapter 2 has been published as:

Nenadić, F. & Tucker, B. V. (2020). Computational modelling of an auditory lexical decision experiment using jTRACE and TISK. *Language, Cognition and Neuroscience*. Advance online publication. https://doi.org/10.1080/23273798.2020.1764600.

Chapter 3 was completed with help from Louis ten Bosch, the author of DIANA, and has been submitted as:

Nenadić, F., ten Bosch, L., & Tucker, B. V. (2020). Computational modelling of an auditory lexical decision experiment using DIANA. Manuscript submitted for publication.

Chapter 4 was completed with help from a part of the team developing the discriminative lexicon: Elnaz Shafaei-Bajestan, Yu-Ying Chuang, and R. Harald Baayen, following a two-week visit to their laboratory in Tübingen.

Additional support regarding Chapter 3 came from Petar Milin during a three-month visit to the University of Birmingham, funded by the Mitacs Globalink Research Award. During this stay, the outlines of a hybrid model described in this dissertation's discussion were also created.

Acknowledgements

My work on this dissertation was funded by the Social Sciences and Humanities Insight Grant #435-2014-0678 (The spoken English lexicon: Massive auditory lexical decision project). I was also funded during my visit to Eberhard Karls University of Tübingen by the host, funded by the Mitacs Globalink Research Award during my visit to the University of Birmingham, and supported by a University of Alberta travel award.

I thank my supervisors and all of the members of the examining committee. I thank the researchers that were part of the collaborative projects presented in the dissertation. I also thank all fellow researchers that I had the chance to talk about the work related to this dissertation, whether we ever met in person or not.

Thank you to friends and family for being friends and family.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Human communication using spoken words can be said to have been “solved”, but not “explained”. Numerous technical solutions enable high-performing automatic speech recognition, yet our understanding of how the process of spoken word recognition (SWR; ‘spoken word recognition’ and ‘SWR’ are used interchangeably throughout the text) unfolds still harbors many unknowns or, at the very least, uncertainties. The process of recognizing even isolated spoken words has proven to be complex and special even in comparison to reading. Given the number of factors and moving parameters playing a role in the process of SWR, researchers interested in explaining SWR have made use of computers and technology. However, instead of merely trying to produce an output matching human performance, they attempted to do so while exerting strict control over the computational process. In the case of this cognitive computational modelling, the ultimate goal is to create a computational model with an architecture that is a plausible representation of what might happen when a human listener hears and understands a word. Knowledge of human physiology and cognition as well as the results from behavioral experiments in which real listeners perform various tasks are invaluable in navigating the meanders of decision-making involved in the creation of a computational model of SWR. At the same time, behavioral data are the best tests that we can put before a model — the model needs to perform similarly to the human listener in numerous and varied tasks.

The present dissertation is one such test. Three models of spoken word recognition are tested for their ability to match human performance in the auditory lexical decision task, i.e., the task in which listeners decide whether an isolated signal is a word existing in their language or not. The purpose of this investigation is to tell us more about both the models of spoken word recognition we employ and the process of spoken word recognition itself. In the remainder of the introduction, I first present a brief overview of models of spoken word recognition. I then focus on the goals of the present dissertation, describing its layout in more detail. The following Chapters present a series of simulations performed with other researchers using an assorted selection of models of SWR and the discussions of obtained results.

1.1 Overview of models of spoken word recognition

The early models of spoken word recognition introduced principles that still hold for many current models. These models are often referred to as "first-generation" models of SWR. The logogen model (Morton, 1969) introduced the logogen as a stored unit of meaning that accepts information and then responds when sufficient information has been gathered. In other words, when enough supporting evidence has been collected, a word is provided as a response. Research conducted as part of the frequency ordered bin search model investigations (Forster & Bednall, 1976) stressed the importance of word frequency and also set the focus on the process of searching for the right item from a set of candidates. Additional aspects of the process of SWR that these studies touched on include, e.g., morphological aspects of storing units of meaning (Taft & Forster, 1975). These models were in time replaced and are not used in present-day simulations, not least because they assumed lexical access was the same for written and auditory stimuli. Still, the metaphor of word activation and search — the notion that a signal stretch "activates" a number of candidates based on their matching characteristics during the search of the lexicon — still serves as a baseline for most models of spoken word recognition.

The introduction of the “second-generation models” marked the beginning of a proliferation in the number and variety of models of spoken word recognition. Each of the models was special in its own way and introduced features that were not considered in other models, often making direct model comparison very difficult. The most influential models from this period, which are also still relevant in the literature, were the Cohort model (Marslen-Wilson & Tyler, 1980; Marslen-Wilson & Welsh, 1978, note that the Cohort model never received a computational implementation), the TRACE model (McClelland & Elman, 1986), and the mathematical model called Neighborhood Activation Model (NAM; Luce, 1986; Luce & Pisoni, 1998). All three models were subsequently adapted or improved, as Cohort II was developed (Marslen-Wilson, 1987, see, e.g.,), TRACE was moved to a Java platform and re-implemented as jTRACE (Strauss et al., 2007), and NAM gained a connectionist instantiation in the form of PARSYN (Luce et al., 2000). Since then, many more models were developed, most notable of which is the Shortlist model (Norris, 1994), and to it connected Merge (Norris et al., 2000b). These models were updated as well in the form of Shortlist B and Merge B (Norris & McQueen, 2008). Additionally, the Time-Invariant String Kernel model (TISK; Hannagan et al., 2013; You & Magnuson, 2018) was recently developed as a reimplementation of TRACE that changes some of the architecture but retains similar performance.

Despite the general agreement with regard to the activation-competition metaphor in these notable models, the specifics of the process also raised many points of disagreement (see Weber & Scharenborg, 2012, for an extended discussion). The models differ in the assumed input units (e.g., whether they are phonemes or subphonemic features), the number of prelexical abstract layers and their characteristics, the possibility of information flow from higher to lower levels (that is, feedback from higher levels affecting what is recognized in lower levels), existence of competitor inhibition, and others. I will use another big point of contention — the selection and retention of candidates during the activation-competition process — to illustrate these differences.

Cohort assumes that after some initial information is made available (matching the first couple of phonemes of input), a cohort of competitors would be retained as all the other competitors that do not have the same starting phonemes are excluded from further activation and competition, with this shedding of candidates continuing as more information becomes available during the unfolding of the signal. In TRACE, close candidates are all words that share the first two phonemes with the target word (cohorts), all words that share the final two phonemes with the target word (rhymes), and all the words that are fully embedded in the target word (embeddings). The Neighborhood Activation Model (NAM) captures phonological neighbors, which are usually operationalized as all words that are one phoneme edit away from the target word (e.g., neighbors of /bit/ are /pit/, /big/, /bɑt/, and /sit/). In Shortlist, a smaller set of 30 most viable candidates is observed at each time step as the signal unfolds, and only those items are considered. Still, two important characteristics are also shared between these models (albeit with slight differences). All of these models use some form of an abstract, prelexical unit as input to the model. Additionally, all of these models represent the mental lexicon as an unconnected list of units (words) that are essentially strings of sublexical units (phonemes).

An important addition in some of the most recent models was inclusion of actual sound signal as input to the model. The central benefit of such models is that they capture the vast variability present in the speech signal and take into account fine-phonetic detail. The Speech-based Model (SpeM; Scharenborg et al., 2005) used acoustic models that are developed in automatic speech recognition to recognize phonemes present in the signal and, subsequently, calculate most likely lexical paths and word activation. SpeM's successor Fine-Tracker (Scharenborg, 2008, 2009) operated in a similar fashion, but extracted prelexical units were articulatory-acoustic features, not phonemes. The representation at the lexical level was necessarily changed as well, and each word was represented by articulatory-acoustic feature vectors. The most recent addition to this type of model is DIANA (ten Bosch, Boves, & Ernestus, 2015). DIANA also uses automatic speech recognition techniques to create acoustic models. These are applied to the acoustic signal to derive estimates of segment (phoneme) activation and word activation. Importantly, DIANA is one of the only models of SWR to explicitly define a decision component. This allows DIANA to state whether the input matches a word in the mental lexicon (well enough) or not, which word is the winning candidate, and also estimate the time when the decision is made, making it an end-to-end model of SWR.

Besides the previously described models of spoken word recognition which are referred to as abstract models, exemplar or episodic models were also created. The principal difference between the two is that episodic models do not postulate the existence of one abstract representation for each stored unit. Instead, such models assume that the storage features multiple representations, each related to some previous occurrence (example or episode) of that unit being encountered. The process of SWR would then aim to "land" somewhere in the vicinity of a group of exemplars of essentially the same unit of meaning. Similarly, prelexical levels also need not include abstract units, as an exemplar model should incorporate idiosyncratic aspects of speech. A prime example of an episodic model is presented by Goldinger (1998) in a series of word-shadowing experiments followed by computational simulations using the Minerva2 model (Goldinger, 1998; Hitzman, 1986). Although the question of representation of lexical and prelexical units as abstract or episodic is central to the theory of SWR and yet unresolved, episodic models are far less prevalent than abstract models of SWR and their number is smaller. These models are not featured in the present dissertation. For additional information about episodic models of SWR, see Scharenborg and Boves (2010) and Weber and Scharenborg (2012).

Lastly, yet another separate group of models is often referred to as the learning models (see Magnuson et al., 2012). Learning models are most often connectionist models using networks trained to make connections between input and outcome, changing the way they perform with additional training material. Major representatives of such models are Simple recurrent networks (Elman, 1990), the Distributed Cohort Model (Gaskell & Marslen-Wilson, 1997), the PK99 (Plaut & Kello, 1999), and the Adaptive Resonance Theory, with ARTword in particular (Grossberg & Myers, 2000). The most recent addition to this set of models of SWR is the discriminative lexicon approach (Baayen, Chuang, Shafaei-Bajestan, et al., 2019). The discriminative lexicon is an extension of the naive discriminative learning approach which was used to model various language phenomena (see, e.g., Baayen et al., 2011; Milin, Divjak, et al., 2017; Milin, Feldman, et al., 2017; Ramscar et al., 2014). The authors have also made a point of using the acoustic signal to create more realistic input for the model (see Arnold et al., 2017).

1.2 The present dissertation

With such an abundance of models of spoken word recognition (and there are others not mentioned in the previous section) impartial testing is necessary for their fair assessment and guided further development. However, although often discussed, models of SWR are rarely implemented by those that did not create them — except if these are not creators of their own model of SWR comparing the models. When independent researchers do use models of SWR, the simulations are mostly designed to support a conducted behavioral experiment or to generate hypotheses for a behavioral experiment, rather than to assess model performance. Although this kind of research is exactly the reason why models of SWR are developed, it uses models of SWR as a tool and/or as a theoretical framework, not as the focus of investigation.

Additionally, models of spoken word recognition are rarely tested on large and varying datasets. The testing (with notable exceptions) is ordinarily performed on small datasets or even minimal, toy problems. The reason for this is that toy problems and proofs of concept are computationally feasible and more easily presented and replicated. In other cases, the computational model is developed on a larger dataset, but never re-implemented or tested on another similar set of data. This is because large datasets are rare, and there rarely exists another large dataset with similar features. The vast majority of reports referencing models of spoken word recognition do not actually implement the computational model. Instead, they develop predictions or discuss the results of empirical investigation based on previously performed and reported computational simulations, or speculate as to how the model would perform in the context of that study.

Using small datasets and toy simulations is also in collision with simulating participant performance in tasks where effects necessarily span hundreds of stimuli and the set of plausible competitors (i.e., the size of the mental lexicon) is not explicitly limited. An example of such a task is the auditory lexical decision task (see, e.g., Goldinger, 1996). The task for the participant in the auditory lexical decision task is to decide whether the brief speech signal they just heard is a word of the language in question or not. Usually half of such stimuli are actual words, while the other half merely sound like they could be words in the given language, but actually have no assigned meaning. These stimuli are called pseudowords. Participants are ordinarily instructed to perform as quickly and as accurately as possible, and measures obtained from this task include participant response latency and accuracy. Although computational models of SWR did attempt to simulate particular phenomena captured using the auditory lexical decision task, such as the effects of subcategorical mismatch described by Marslen-Wilson and Warren (1994), to the best of my knowledge, the only computational model of SWR that tried to match participant response latency and accuracy from a behavioral dataset of responses in the auditory lexical decision task is DIANA (see, e.g., ten Bosch, Boves, Tucker, et al., 2015). However, a model of spoken word recognition should be able to simulate a large variety of experimental tasks investigating spoken word recognition, including the auditory lexical decision task — such simulations are necessary if we wish to claim that the model matches the processes occurring in a listener when they are presented with auditory stimuli and in order to make predictions about new experiments.

Although these issues (lack of third party testing of model performance, using large datasets instead of small toy problems, and simulating a task such as the auditory lexical decision task which cannot be easily framed in a small sample simulation) are only prevalent in the field and not present in every model and every simulation performed, very few (if any) models have avoided all three of them simultaneously. From this stem three closely related goals of the present dissertation. The first goal is to test a number of models of SWR in their success in simulating the auditory lexical decision task. The central question is whether models of SWR can capture the process as it unfolds in the human listener. Furthermore, by comparing ease of use and estimate-to-data fit we can drive forward the discussion on model adequacy and ultimately make an informed decision about which (type of) model and structure should be favored. Note that in some cases it may be clear how a certain model of SWR can be improved upon even before any simulations are performed. The first goal, however, is not to reach peak performance by altering the models — the goal is rather to subject them to impartial testing in their current state.

The second goal is to use the simulations to draw conclusions about the spoken word recognition process as captured by the auditory lexical decision task. By switching models or model parameters, we are also changing our assumptions about the process of spoken word recognition. Therefore, some of the questions which will be discussed are how input should be presented and processed, what the nature of the abstract sub-word units is (if they exist), how competition unfolds, how and when does the model (or human) reach their decision that a stimulus is a word or a pseudoword, and how the storage of units of meaning (the mental lexicon) should be organized.

The third goal of this dissertation is different for each particular model used and ties to their particular characteristics. We can use the results of the simulations to provide suggestions for improvements and further model development. These suggestions, however, remain informed by certain characteristics of other models of spoken word recognition. I also hope that the presented research will prompt future studies to use large datasets and set them as targets for the computational simulation instead of using toy problems and proofs of concept.

The present simulations are conducted as part of the ongoing Massive Auditory Lexical Decision project (MALD; Tucker et al., 2019). MALD includes datasets with a large number of responses from numerous participants to a large number of English words. The advantage of using data from a large-scale study in comparison to a smaller, target auditory lexical decision experiment lies in the generalizibilty of the findings as any patterns captured in a dataset are in part due to idiosyncrasies present in the participant or stimulus sample. Given that models of spoken word recognition aim to match general regularities in the process, having a larger set of words and human listeners set as a benchmark for the model makes the simulation outcomes more reliable.

Two MALD datasets are used in the current dissertation. The first one is MALD1, i.e., the first published MALD dataset. MALD1 tested 231 monolingual native Canadian English listeners (180 female, 51 male, aged 17 to 29) as they responded to approximately 26,000 English words. The participants were allowed to participate in up to three sessions, never listening to the same items. A single session contained 400 words and 400 pseudowords. The total number of recorded MALD sessions was 284, and the total number of recorded participant responses was 227,179. The second dataset we refer to as MALD_semrich (for 'semantic richness'). The database was created as a replication of the Goh et al. (2016) study. In this case, 27 participants were presented with a single list of 442 MALD nouns and 442 randomly selected MALD pseudowords. MALD_semrich has a smaller pool of words, but offers significantly stronger item power in comparison to MALD1. Both of these datasets are described in more detail in the central part of this dissertation.

The models of spoken word recognition, as we have seen, are both numerous and varied. Since a single dissertation cannot implement all of them, the models tested in the present dissertation were selected to cover a variety of model characteristics.

The selected models are implementations of the TRACE model, DIANA, and the discriminative lexicon approach. TRACE was selected as a representative of the second-generation models of spoken word recognition. TRACE is a textbook model (see, e.g., Traxler & Gernsbacher, 2006) that was at the center of many discussions about the process of spoken word recognition. The model continues to be relevant today (see Chawla & Chillcock, 2019), in part due to its reimplementation jTRACE, and it is certainly one of the most cited models of spoken word recognition (if not the most cited). Simulations are performed using jTRACE and TISK. DIANA was selected as a representative of models that take a step further from the somewhat older models by featuring input based on the actual acoustic signal. DIANA also defines a decision component that provides an explicit estimate of when the model recognized the input as a certain word. It is also interesting to note that, in theory, DIANA's candidate selection during the competition process matches that of the Cohort model (Marslen-Wilson & Tyler, 1980; Marslen-Wilson & Welsh, 1978), providing a complement to TRACE. The discriminative lexicon was selected primarily because it represents a new and different approach to simulating spoken word recognition. As will be described in more detail in Chapter 4, this approach does not share the same background as most abstract models of SWR. Furthermore, unlike most models of SWR, the discriminative lexicon offers a semantically rich, interconnected representation of meaning storage as a replacement for the standard representation of the mental lexicon.

Availability and convenience also had a role in model selection. For example, TRACE was selected over Shortlist B because it had a developed software with a user manual (Strauss et al., 2007), whereas Shortlist B has no user-friendly code and is restricted to the Dutch language. PARSYN and Cohort also do not have an instantiation that enable ready third-party use. DIANA shares many traits with Fine-Tracker and SpeM, but was selected because it was more recently used to simulate the auditory lexical decision task from another large project, BALDEY (Ernestus & Cutler, 2015), and has even previously been applied to a pilot MALD dataset (ten Bosch, Boves, Tucker, et al., 2015). Both the Distributed Cohort Model and the discriminative lexicon approach offer a learning perspective to the process of spoken word recognition as well as a semantically organized mental lexicon. The discriminative lexicon approach, however, is a more recent development and has been implemented through available R packages (R Core Team, 2018), making its implementation significantly easier.

The selection of the models is therefore not reflective of their quality or importance only. Although I also kept in mind (potential) model impact, aiming to assess some of the most influential models, there are many other notable models that could have been used instead of the three models that have been selected. Indeed, I hope other models will be implemented to simulate larger sets of auditory lexical decision data in the future.

Three separate chapters present the simulations conducted with the selected models of spoken word recognition. The chapters are intended to be accessible as separate research articles, so there is some inevitable overlap in the information they include. This overlap mostly considers broad overviews or comments on computational models of spoken word recognition and the Massive Auditory Lexical Decision project. Barring these similarities, each chapter focuses on a different model — introducing in more detail (j)TRACE and TISK in Chapter 2, DIANA in Chapter 3, and the discriminative lexicon in Chapter 4. The central part of the articles, as well as the dissertation as a whole, are the simulations performed using these models. Each article also includes its own discussion of the simulation results, focusing on the implemented model.

The last chapter is a joint discussion of the simulations performed. In it, the findings from the three central chapters are combined to offer general conclusions about the process of spoken word recognition as captured by the lexical decision task, as well as suggestions for future development of models of spoken word recognition. Additionally, I offer a broad description (a sketch) for a hybrid model of spoken word recognition that brings together features from models tested in the current dissertation.

Chapter 2

jTRACE and TISK

Chapter 2 has been published as:

Nenadić, F. & Tucker, B. V. (2020). Computational modelling of an auditory lexical decision experiment using jTRACE and TISK. *Language, Cognition and Neuroscience.* Advance online publication. https://doi.org/10.1080/23273798.2020.1764600.

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 this work.

2.1 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 ordinarily 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 et al., 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 et al., 2007) and the more recently developed TISK model (Hannagan et al., 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.

2.1.1 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 et al., 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 et al., 2000b), or at least reported findings that the model does not fully account for (see, e.g., Chan & Vitevitch, 2009; Frauenfelder & Content, 2000; Gaskell et al., 2008; McMurray et al., 2009; Smith et al., 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 et al., 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).

2.1.2 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 that they claimed match the findings of previous empirical studies. Furthermore, the authors introduced changes to the parameter set values that did not significantly affect the relationship between jTRACE and TISK simulations.

2.1.3 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 et al., 2018; Ventura et al., 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 et al., 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 et al., 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 et al., 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.

2.2 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 et al. (2016) study. The full datasets are also available at this link: mald.artsrn.ualberta.ca.

2.2.1 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.

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.

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).

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 et al., 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.

2.2.2 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.

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.

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.

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.

2.3 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 et al., 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 et al., 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 et al. (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 2.1). 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.

$$
\begin{aligned}
predRT_1 &= RT_1 \\
\forall r > 1 : predRT_r &= \alpha \cdot RT_{r-1} + (1-\alpha) \cdot predRT_{r-1} \qquad (2.1) \\
dRT_r &= RT_r - predRT_r
\end{aligned}
$$

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 et al., 2014), valence (Warriner et al., 2013), and the number of semantic features (McRae et al., 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.

2.4 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.

2.4.1 Simulation setup

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 2.1 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, et al. (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".

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

Table 2.1: 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

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.

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 Table 2.2.

Table 2.2: The three parameter sets used for TISK simulations.

Parameter	Set A	Set B	Set C
step	10	10	10
nPhone threshold	0.91	0.91	0.91
phoneme decay	0.001	0.01	0.001
diphone decay	0.001	0.01	0.1
single phone decay	0.001	0.01	0.1
word decay	0.01	0.05	0.05
input to phoneme weight	1.0	1.0	1.0
phoneme to phone weight	0.1	0.1	0.1
diphone to word weight	0.05	0.05	0.05
single phone to word weight	0.01	0.01	0.01
word to word weight	-0.005	-0.05	-0.010
word to diphone activation feedback	0	0	0.1
word to single phone activation feedback	0	0	0.1
word to diphone inhibition feedback	0	0	-0.05
word to single phone inhibition feedback	0	0	-0.05

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.

2.4.2 Results

We first performed a visual inspection of model performance by creating activation-competition plots for both jTRACE and TISK simulations. Figure 2.1 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 (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 2.1, 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 2.1 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

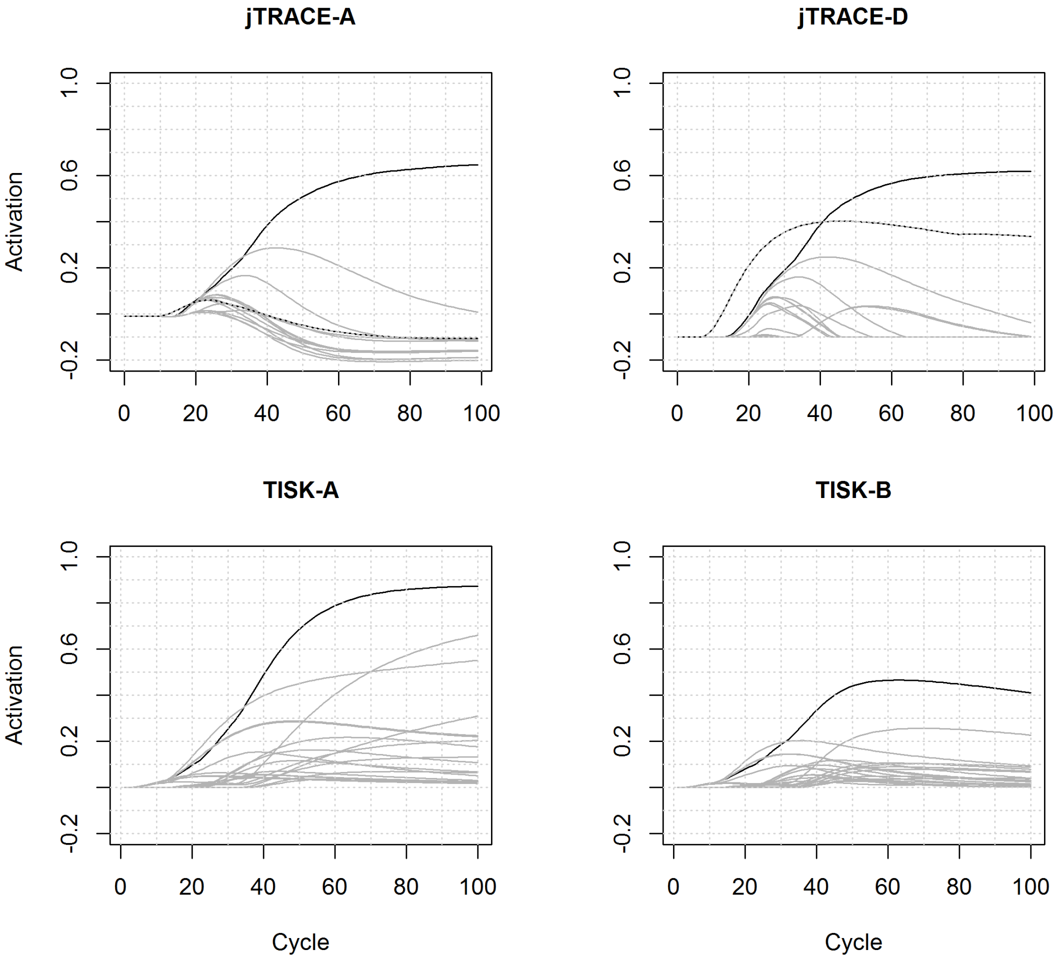


Figure 2.1: An example of the competition process in the jTRACE and TISK models based on activation values for the word "shield" given in black and 19 closest competitors given in gray. Activation is given on the y-axis, and the time cycle is given on the x-axis. Parameter sets used are given as titles for each plot. The silence phoneme as a word competitor in jTRACE is presented with a dotted black line.

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 2.1), 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 2.1, 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 2.3 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).

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

Table 2.3: 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

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 2.2). 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

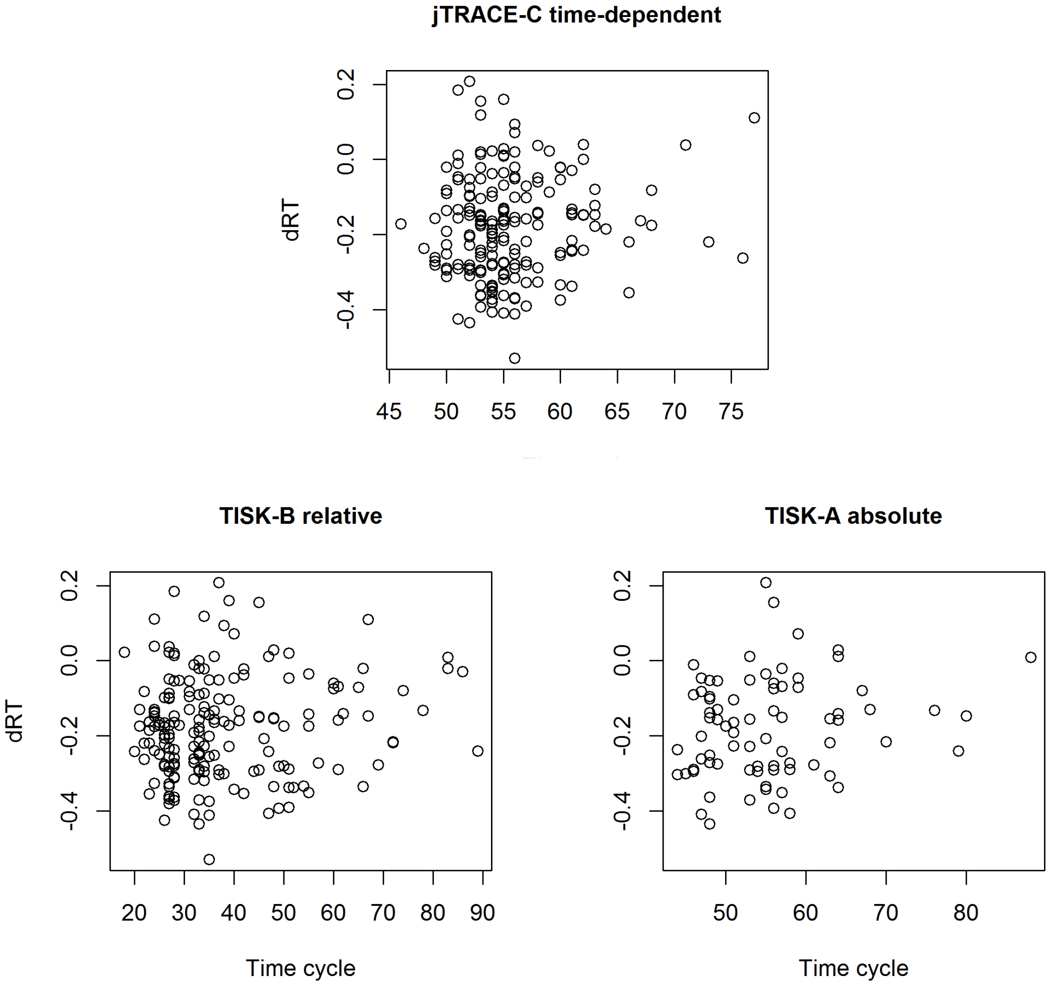


Figure 2.2: The highest correlations between participant performance and computational model estimates of the time cycle when the winner should be selected recorded in Simulation 1. The time cycle when the model selected the winning word is given on the x-axis, and dRT is given on the y-axis.

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.

2.4.3 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.

2.5 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.

2.5.1 Simulation setup

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 2.4: 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

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.

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.

2.5.2 Results

Table 2.5 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 2.5: 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 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 2.3). 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 $r = .65$,

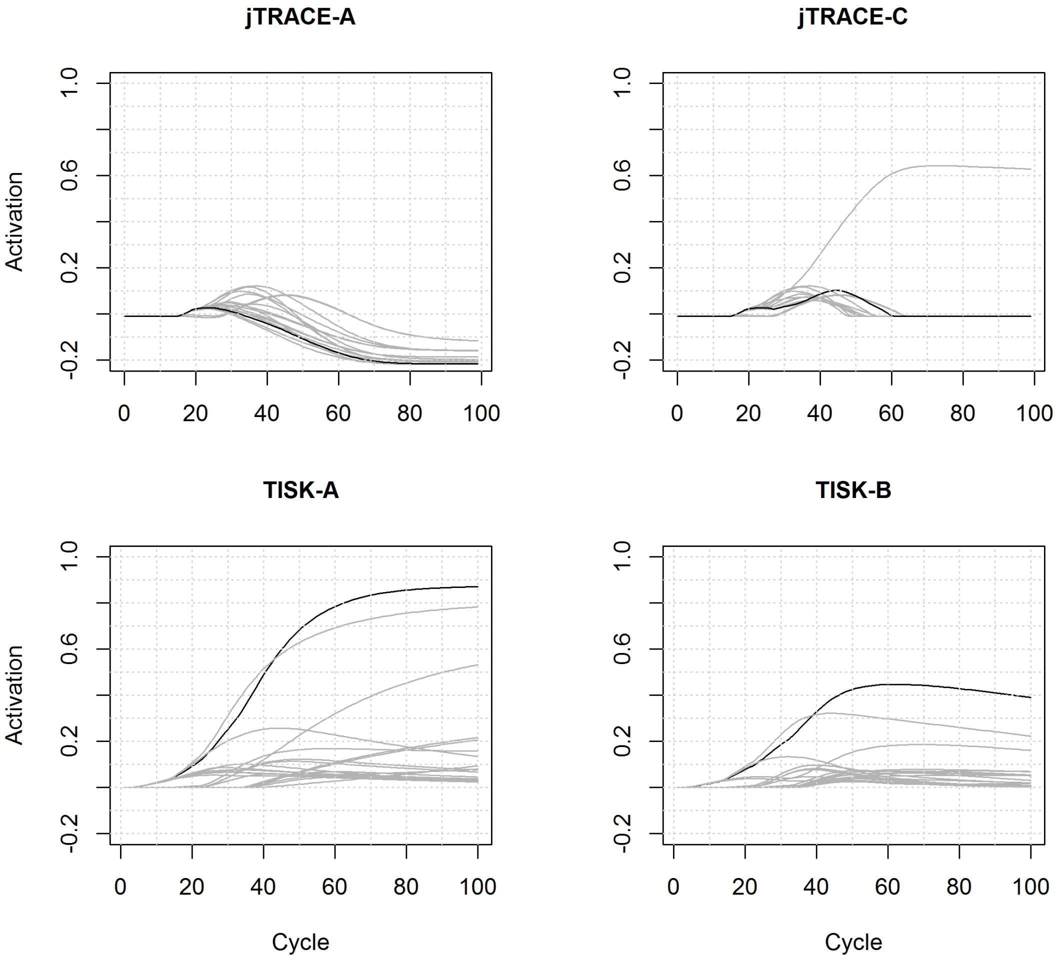


Figure 2.3: An example of the competition process in the jTRACE and TISK models when MALD_semrich words are used. The figure presents activation values for the word “cherry” given in black and 19 closest competitors given in gray. Activation is given on the y-axis, and the time cycle is given on the x-axis. Parameter sets used are given as titles for each plot.

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 2.6 as an example.

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 $r = .83$.

2.5.3 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,

Table 2.6: 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(>$\lvert t \rvert$)
(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

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.

2.6 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.

2.6.1 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, Luthra, et al., 2018), so we assumed that results obtained in TISK should translate well to jTRACE should it be able to represent these phonemes.

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.

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.

RT comparison

The same approach as in Simulation 2 was used.

2.6.2 Results

After creating custom competitor lists for each word, free word recognition accuracy dropped severely (Figure 2.4). 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 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 2.5a, 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

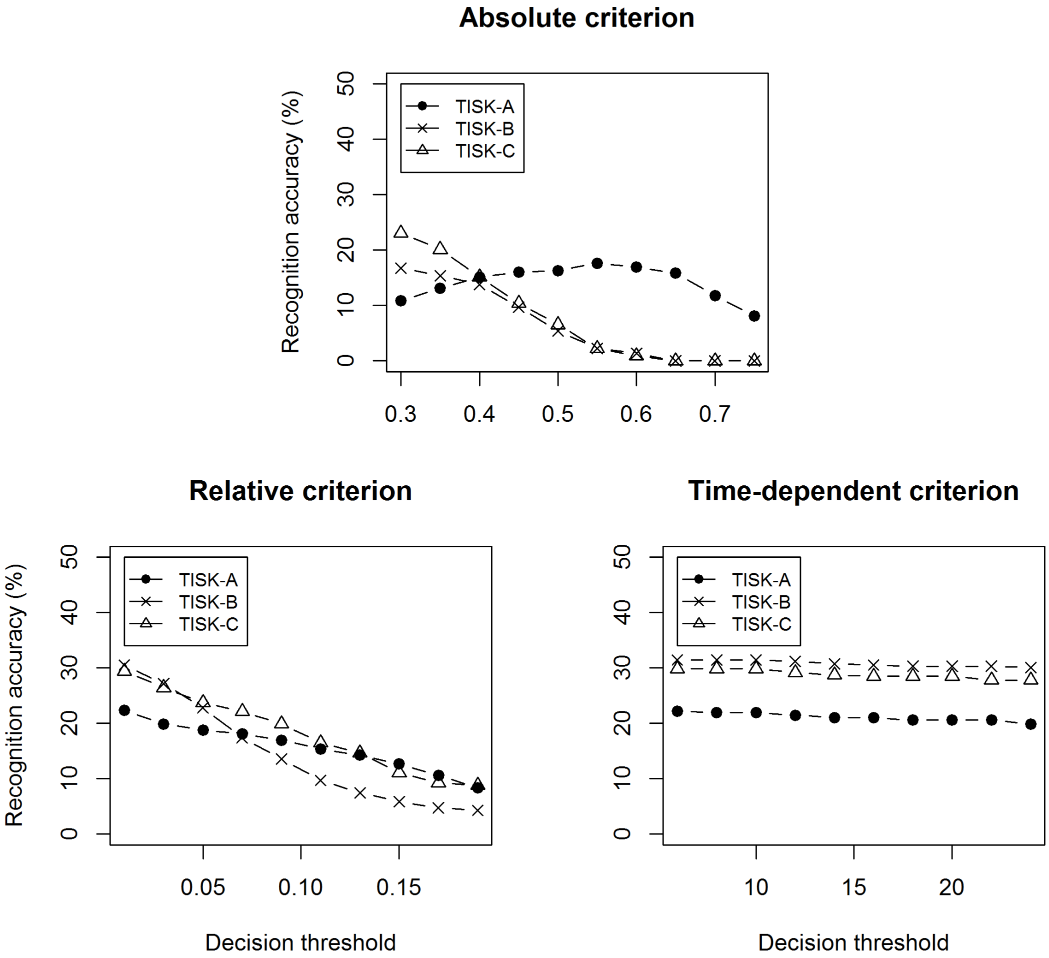


Figure 2.4: TISK model accuracies in free word recognition per decision criterion (separate figures) and parameter set (separate lines) used. The percent of correctly recognized words is given on the y-axis, while the decision threshold is given on the x-axis.

words that were correctly recognized by the model, regardless of the decision criterion or the parameter set used. As an example, Figure 2.5b shows a box-plot when the time-dependent criterion with the decision threshold of 10 was used in TISK-B.

However, Figure 2.5b 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 2.5c 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 2.6). 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 2.6b). 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 2.6c). 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 2.6d). (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

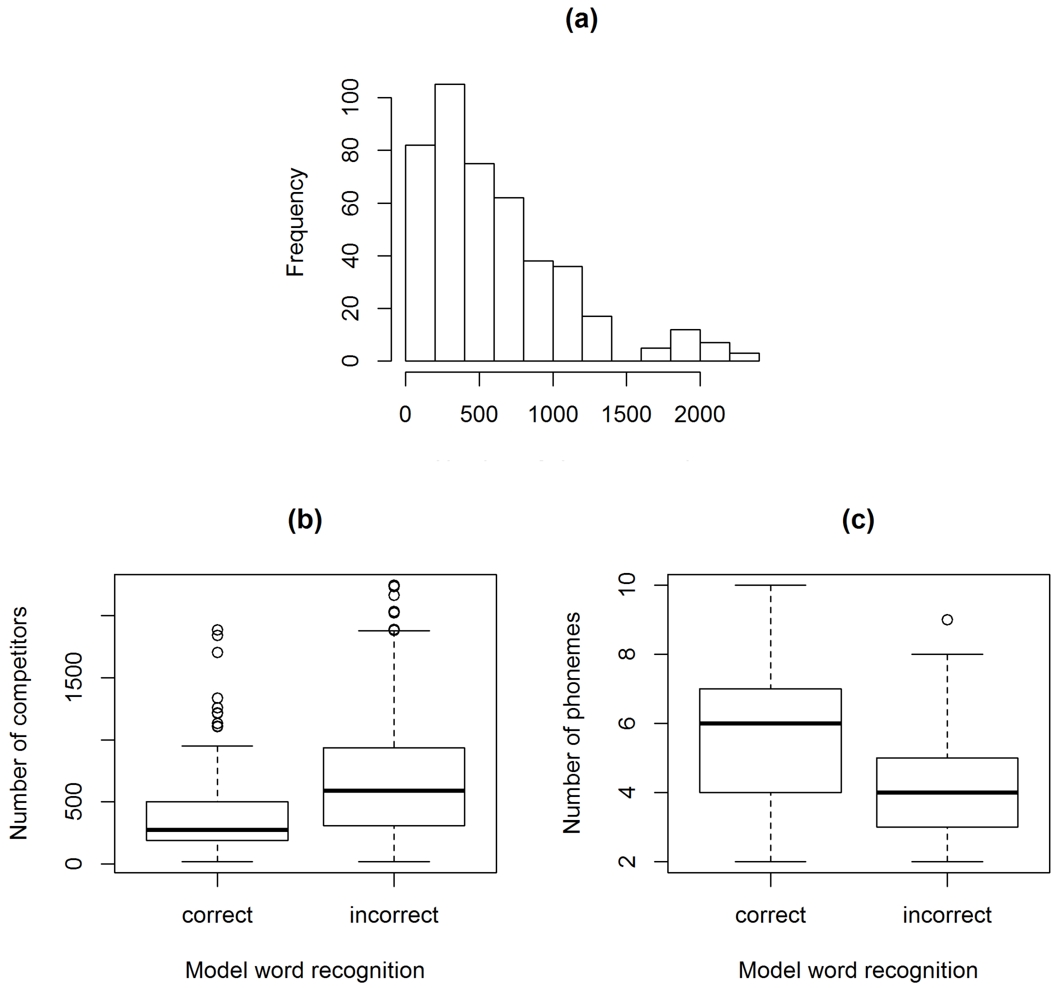


Figure 2.5: Figure (a) is the histogram of the number of close competitors (cohorts, rhymes, and embeddings) extracted from the CMU dictionary for the 442 MALD_-semrich words, with many words having hundreds of close competitors. Bottom figures show that the model more often correctly selected the target 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.

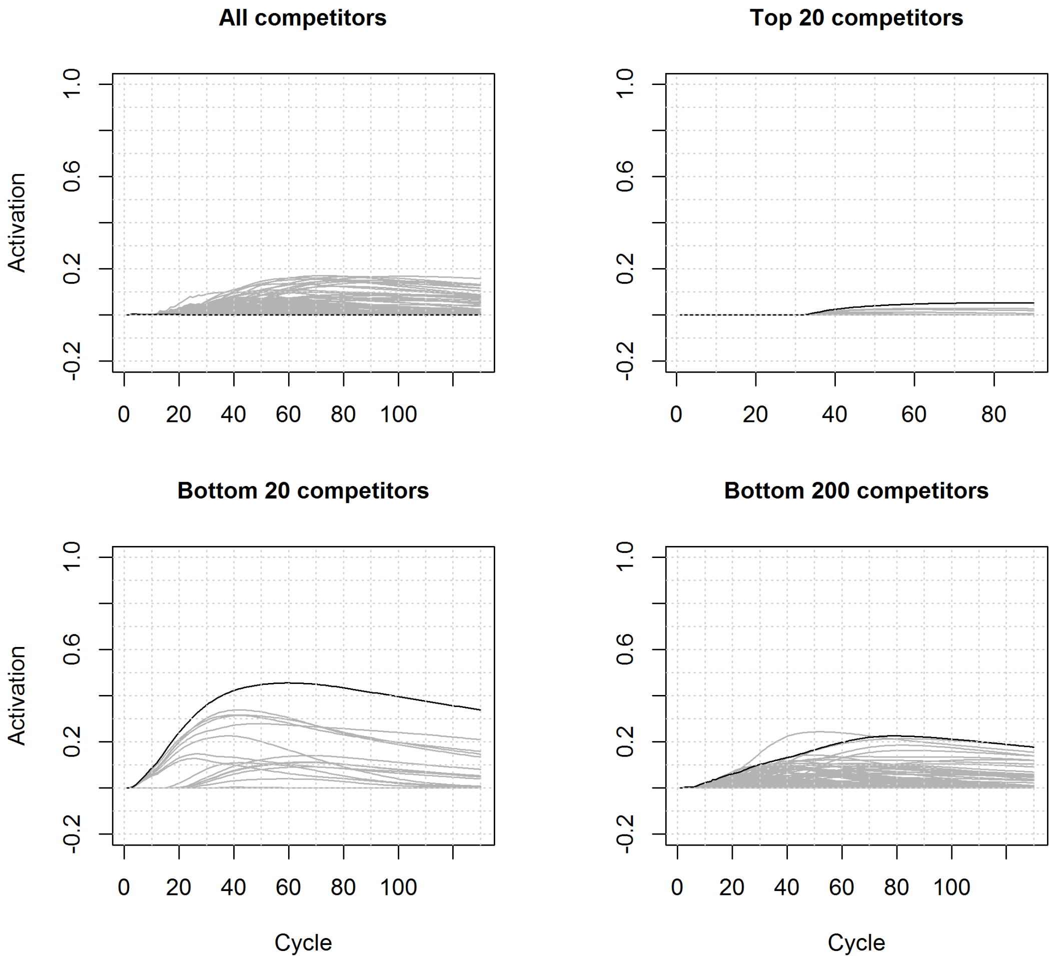


Figure 2.6: An example of the competition process for the target word “belt” in the TISK-B model when different close competitors are included. Activation is given on the y-axis, and the time cycle is given on the x-axis. The target word is given in black and other competitors are given in gray. The 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.

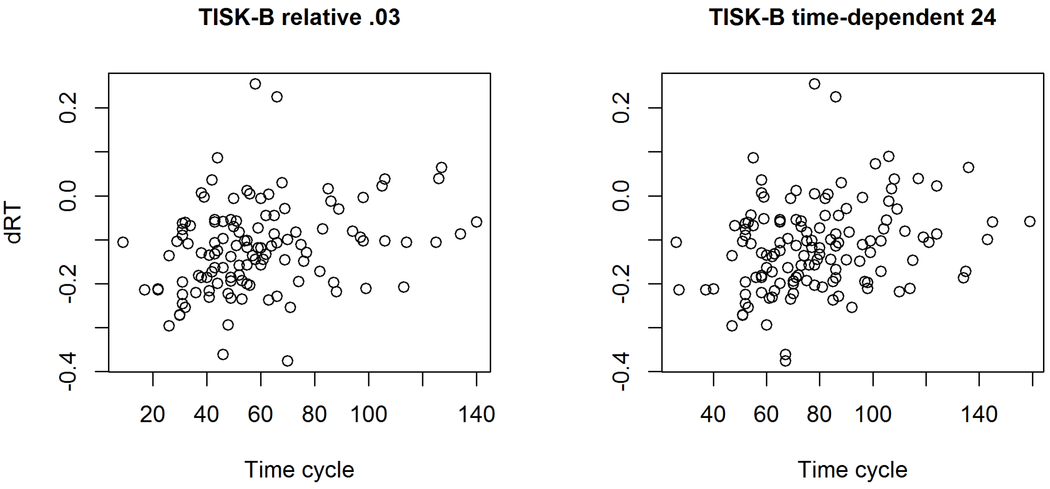
when the relative decision criterion is used, and maximally 30 time cycles with TISK-B when the time-dependent criterion is used. Importantly, not only did the time cycle simply increase, it also changed differently for different words. The correlations between the time cycles when the winner is selected 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 2.7). Unfortunately, as in Simulation 2, both of these model estimates 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.

2.6.3 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_sem-


TISK-B relative .03

TISK-B time-dependent 24

Figure 2.7: 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.

rich 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.

2.7 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.

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 pseudowords 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.

2.7.2 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 2.7 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 2.7: 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

2.7.3 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.

2.8 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 et al., 1994; Salverda et al., 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 & Warner, 2011; Ernestus & Baayen, 2007; 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 et al., 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 combining /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 et al., 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 et al., 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, et al., 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 et al., 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, Tucker, et al., 2015; ten Bosch et al., 2014; ten Bosch, Boves, & Ernestus, 2015). 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, Chuang, Shafaei-Bajestan, 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 et al., 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 et al. (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, Chuang, Shafaei-Bajestan, 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 et al., 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, Chuang, Shafaei-Bajestan, et al., 2019) the acoustic input is represented using the so-called frequency band summary features (Arnold et al., 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.

Chapter 3

DIANA

Chapter 3 was completed with help from Louis ten Bosch, the author of DIANA, and has been submitted as:

Nenadić, F., ten Bosch, L., & Tucker, B. V. (2020). Computational modelling of an auditory lexical decision experiment using DIANA. Manuscript submitted for publication.

We also thank Terrance M. Nearey and Matthew C. Kelley for their help in shaping this study.

3.1 Introduction

The question of how a listener understands the meaning of what is being said is central to the field of speech perception and spoken word recognition. Isolating characteristics of the speech signal that act as reliable cues of its content has proven difficult due to lack of invariance, leading to a long debate and numerous explanations of how this process unfolds. Still, most models of spoken word recognition (SWR) sidestep the problem of analyzing the acoustic speech signal and "instead use an artificial, often-hand crafted, idealised discrete (prelexical) representation of the acoustic signal as input" (Scharenborg & Boves, 2010, pp. 144).

The main reason for eschewing the acoustic signal were technical limitations that all first and second-generation models of SWR faced, not lack of understanding of its importance. Topics ranging from acoustic-phonetic invariance to prosodic cues were central in the development of the Lexical Access From Spectra (LAFS) model proposed by Klatt (1979). The acoustic-phonetic representation in bottom-up approaches to SWR is also discussed by Pisoni and Luce (1987) as they overview what are mostly considered first-generation models of SWR, but also the Cohort model (Marslen-Wilson & Tyler, 1980; Marslen-Wilson & Welsh, 1978) and TRACE (McClelland & Elman, 1986). Where the more recent, second generation models of SWR are concerned, TRACE I is built around the acoustic signal being used as input, but TRACE II, the model that was actually implemented, employed acoustic pseudofeatures instead (McClelland & Elman, 1986; Strauss et al., 2007). Similarly, Shortlist (Norris, 1994) used phoneme strings as input, but Shortlist-B (Norris & McQueen, 2008) made a step towards representing their variability better by using sequences of phoneme probabilities calculated over time slices, which were obtained from a diphone gating study with human listeners. One consequence of using pseudo-acoustic input is difficulty of establishing a direct comparison between model output and human performance; that is, forming what Magnuson et al. (2012) refer to as a linking hypothesis. Very few models of SWR offer a clear-cut estimate of when the model recognized the input as a certain word.

However, although many of the technical limitations have since been alleviated, most models of SWR continue to settle for pseudo-acoustic input. Even the more recently developed Time-Invariant String Kernel model (TISK; Hannagan et al., 2013; You & Magnuson, 2018) uses phonemes as input, assuming that the process of phoneme recognition had already been successfully completed. Notable exceptions to this trend are SpeM and Fine-Tracker (Scharenborg, 2008, 2009; Scharenborg et al., 2005) and the discriminative lexicon approach to spoken word recognition (Baayen, Chuang, Shafaei-Bajestan, et al., 2019). Fine-Tracker maps the acoustic signal to a set of articulatory features, and it is capable of simulating durational and fine-phonetic detail effects captured in behavioral experiments (see, e.g., Andruski et al., 1994; Salverda et al., 2003). The discriminative lexicon approach extracts frequency band summary features (Arnold et al., 2017) which are strings describing a particular frequency band of a temporal chunk of the signal in terms of its initial, final, median, and maximum amplitude. Both approaches are promising, but may require additional fine-tuning, as is the case with Fine-Tracker (see Scharenborg & Merkx, 2018), or additional testing, in case of frequency band summary features.

To the best of our knowledge, the only other model to also implement acoustic signal as input is DIANA (ten Bosch, Boves, & Ernestus, 2015). Similarly to other notable models of SWR, DIANA is an activation and competition model. The current setup uses speech corpora to develop acoustic models, which are then used to calculate phone activations in novel acoustic signals presented to DIANA. What sets DIANA apart even in comparison to models like Fine-Tracker and the discriminative lexicon approach is that it acts as an end-to-end model of SWR. Models of SWR generally do not include a decision rule of when the target word should be selected; DIANA, in turn, allows for generating estimates of response latencies and simulates word/pseudoword decisions that can be directly compared to those obtained in behavioral experiments.

In the present paper, we first give a more detailed description of DIANA and previous simulations performed using this model. We then motivate the present simulations, presenting the goal of the current study. The central part of the paper describes the simulations we performed and discusses the findings to both develop DIANA and further scrutinize the process of spoken word recognition.

3.1.1 DIANA

A coarse visual representation of DIANA's components and parameters is given in Figure 3.1 (adapted from ten Bosch, Boves, Tucker, et al., 2015). The model takes the speech signal as input and uses existing automatic speech recognition acoustic models (described in detail in later sections) to activate subword units, which are phones in the current setup, and words in the mental lexicon represented as phone strings. The model can also weigh the activations using top-down information such as word frequency. The impact of top-down information is controlled by changing the parameter γ. Besides the activation component, DIANA also includes a decision and an execution component.

The decision component operates in parallel with the activation component, reassessing whether a winner can be found at each time step. As words in the lexicon gain activation, they compete without lateral inhibition until a winner is selected based on the difference in activation between the leading candidate and the runner-up. This difference is determined by an adjustable threshold θ_{wc}. If no winner is clear before signal offset, parameter β determines the added time required to make a decision depending on remaining competition. DIANA also makes a word/pseudoword decision by examining whether word activation is similar enough to the string of phones that the model activates independently of the lexicon. If a string of phones not present in the mental lexicon (i.e., a possible pseudoword that happens not to be a lexical item) is activated much better than any phone string that is present in the mental lexicon (i.e., any word), then the input signal is categorized as a pseudoword. The difference in activation required to make a "pseudoword" decision is governed by an adjustable threshold θ_{lb}.

Finally, the third major component of the model is the execution component. The execution happens after the decision has been made and represents the time taken to actually respond (e.g., press a button). Usually, this time is set to 200 ms in DIANA. This estimate of the execution time is based on existing measures of response times in different tasks (see Kosinski, 2008, for a review). Note, however, that increasing or reducing the execution time is a linear transformation that would not impact the correlation between model estimates and some existing behavioral measure. The more important question is whether an approximation that is a fixed number can represent the variability in human reaction time well — not all humans react equally fast. Since the intent of the model is to represent general tendencies in human behavior and since it is unclear how a distribution of execution times can be modeled, we employ the standard approach and use a fixed number. We provide more technical detail about the current implementation of DIANA when we describe the setup of our simulations.

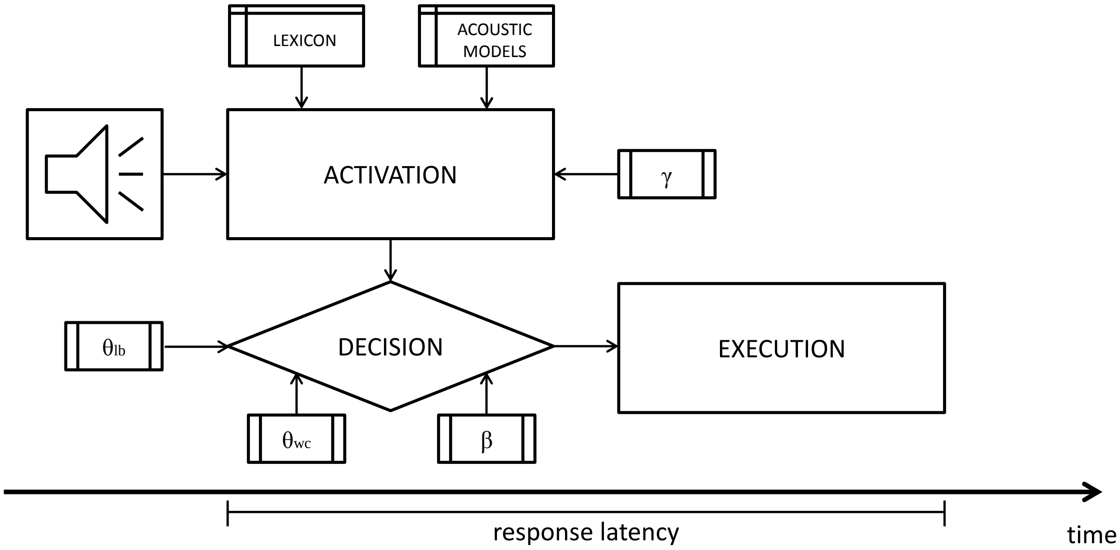


Figure 3.1: DIANA takes acoustic signal as input and has three components (activation, decision, and execution). Word activation depends on the input signal, the acoustic models, and the impact of top-down information adjustable by changing the parameter γ. The decision component is affected by two thresholds (θ_{wc} for the decision that a word wins in comparison to all other word competitors and θ_{lb} for the word vs. pseudoword decision). If no decision is made prior to signal offset, parameter β determines the added decision time. The execution component represents the time needed to execute the decision.

DIANA was predominantly used to simulate the process of isolated spoken word recognition in word repetition and auditory lexical decision tasks (but see ten Bosch, Giezenaar, et al., 2016, for a simulation of L2 listener errors in comprehension of reduced word forms in a sentence dictation task). These simulations were performed almost exclusively in Dutch. The first such instance (ten Bosch et al., 2013) modeled auditory lexical decision responses to 613 disyllabic monomorphemic Dutch words made by 20 participants. The model showed comparable error rates to human participants as its accuracy was 96% for the “word” (participant average: 94%) and 93%

for the “not a word” (participant average: 95%) response. Model estimates of when the decision should be made also correlated quite well with tendencies in participant response latency. The average correlation between the model estimate and human participant performance was $r = .47$. In comparison, the correlations in response latencies between any two participants never exceeded $r = .30$. Although these results imply that DIANA is able to simulate general tendencies in participant responses, low correlations between participants themselves raise additional questions.

In a subsequent report describing a simulation of participant performance in a word repetition task using the same stimuli, ten Bosch et al. (2014) explain the lack of between-participant correlation by using the notion of “local speed effects” (Ernestus & Baayen, 2007). Local speed effects explain the tendency of response latencies to a certain stimulus to correlate with response latencies to a number of previous stimuli. It is assumed that these correlations are a product of, for example, learning, fatigue, or shifts in attention. Since the variation induced by these factors can be treated as noise in comparison to long term effects such as general cognitive abilities, their effect should be attenuated, especially considering that a computational model of SWR is not susceptible to similar effects. Therefore, ten Bosch et al. (2014) took into account response latencies to five preceding stimuli when estimating the “true” response latency to a stimulus, similarly to taking into account previous RT in statistical modelling (see also ten Bosch et al., 2018). The number of relevant preceding stimuli was selected to achieve maximum participant-to-participant correlation in response latency. The correlation between participants and the correlation of DIANA to the average participant response latency both increased after the local speed effects have been attenuated, with the latter being $r = .41$.

The same procedure that removed local speed effects was used in a later study (ten Bosch, Boves, & Ernestus, 2015) that again simulated participant responses to 613 Dutch words in the word repetition task. This time, however, different model parameters were also varied, showing that word frequency plays an important role in approximating participant response latencies, that the model should not just take the word with the highest activation score as the winner but should add extra choice time if there is a close competitor at word offset, and that a word should in general have a substantial advantage to be selected as the winner. Applying these rules increased the average correlation between DIANA's estimates and actual participant response latencies to $r = .76$.

DIANA was also implemented outside of Dutch, albeit only once (ten Bosch, Boves, Tucker, et al., 2015). The dataset included responses from 10 to 12 native and non-native speakers of English to 1,200 words. The results of the simulation still showed satisfactory performance of DIANA, with the correlation between model estimates and average participant response latency in an auditory lexical decision task being $r = .45$. In general, it seems that DIANA achieved higher correlations with participant performance when simulating word repetition than auditory lexical decision data.

3.1.2 The present study

One of many advantages of studies with very large item and/or participant sample sizes, often called megastudies (see Balota et al., 2012), is that they enable testing how well model estimates correspond to human performance by providing a behavioral database for comparison. The results of these comparisons are necessary for further model development. An extensive overview of existing megastudies is given in Keuleers and Balota (2015), while a more recent list is available at http://crr.ugent.be/archives/2141.

Most large studies investigate responses to visually presented, written stimuli. The number of existing databases and their sizes are smaller for the auditory domain. Still, such databases are instrumental in the development of many models of SWR. One of the first larger databases was created by Luce and Pisoni (1998) and the data gathered in this study was used in the development of the Neighborhood Activation Model (see also Luce, 1986). Another example is the study conducted by Smits et al. (2003), as the collected data was used in the development of Shortlist-B Norris and McQueen, 2008. Biggest Auditory Lexical Decision Experiment Yet (BALDEY; Ernestus & Cutler, 2015) collected responses to 5,541 Dutch content words and pseudowords from 20 young native Dutch speakers and was instrumental in testing DIANA (ten Bosch, Boves, & Ernestus, 2016; ten Bosch, Boves, & Ernestus, 2015).

Massive Auditory Lexical Decision (MALD; Tucker et al., 2019) is a still ongoing project designed to provide an even larger database of responses to isolated words presented in the auditory modality, with the goal of complementing the existing databases in the visual domain such as the English Lexicon Project (Balota et al., 2007). One of the purposes of building a large database of MALD responses is to test existing computational models of SWR. The goal of the present study is to implement DIANA in English and test how well it matches participant performance in an auditory lexical decision task using MALD data. Although correspondence to actual participant behavior is only one of the criteria for estimating adequacy of models of SWR (see Scharenborg & Boves, 2010, for an extended discussion), an acceptable fit is still necessary for a model to be considered credible.

DIANA aims to be a language-independent model of SWR and in our simulations we want to investigate the challenges of implementing DIANA for the first time. Therefore, although DIANA was already tested in English on a smaller scale (ten Bosch, Boves, Tucker, et al., 2015), we develop new acoustic models, completing the entire process a researcher in any language would have to undertake to implement DIANA for their own purposes. Once the models are created, we test DIANA's performance in recognizing words in novel speech signals by calculating between-word competition as a function of time, and, most importantly, by simulating the lexical decision task. Crucially, we compare model estimates to actual participant performance in MALD on a large scale and test model adequacy in that way. Original data accompanied with DIANA and statistical analysis scripts are available as supplementary material at https://doi.org/10.7939/r3-jdpa-dn72.

3.2 Behavioral experiment

As we noted in the introduction, we compare DIANA model estimates to human performance in the Massive Auditory Lexical Decision (MALD) project database (Tucker et al., 2019). We use the first version of the dataset (MALD1) available at mald.artsrn.ualberta.ca. In the present paper, we provide only the necessary information about the MALD experiment and the word and pseudoword recordings. More details about the stimuli and procedure are available in Tucker et al. (2019).

3.2.1 Sample

The MALD1 dataset includes responses from 231 native monolingual English listeners (180 females, 51 males; age M = 20.11, SD = 2.39). All participants were recruited from the University of Alberta, receiving partial course credit for their participation.

3.2.2 Stimuli

Stimuli recordings were made by one 28-year-old male speaker of western Canadian English. The speaker was recorded reading isolated words and pseudowords on a computer monitor. He was instructed to produce the words written in their standard spelling as naturally as possible. Pseudowords were presented in their IPA phonemic transcription and the speaker was instructed to read them as if they were words. All word and pseudoword recordings are available as separate wave files and have been aligned using the Penn Forced Aligner (Yuan & Liberman, 2008).

The recording procedure and post-processing of the stimuli yielded 26,800 words and 9,600 pseudowords used in the experiment. The words were split into 67 sets, and the pseudowords were split into 24 sets. Each word and pseudoword set contains 400 unique items. A total of 134 pairings of one word and one pseudoword set were made (i.e., each word list was paired separately with two different pseudoword lists), creating 134 balanced 800-item lists used in the behavioral experiment.

The simulations described in the following sections have many steps and there was small word/pseudoword loss between these steps for various reasons. In the interest of clarity and brevity, we do not document all of these losses in the paper because they are minor and because we always maintain a high standard of hundreds or thousands of items used. We do provide the exact number of items used for critical simulations and comparisons to MALD data. Detailed information about the simulation process, including item loss, can be found in our supplementary materials.

3.2.3 Procedure

The participants were seated in a sound-attenuated booth for the experiment. A single 800-item list of stimuli was presented using the E-Prime experimental software (Schneider et al., 2012). Stimuli order was randomized. After a visual fixation cross lasting 500 ms, a word or a pseudoword was presented over headphones and the task for the participants was to decide whether the signal was a word of English or not by pressing the "yes" or "no" button on the button box. Responding during stimulus presentation would interrupt it and the experiment would proceed to the next trial. If no response was made within three seconds, the following trial was presented. The participants had the option of returning for another session and a new experimental list up to three times. Some participants therefore completed more than one list (but never the same word or pseudoword set), and a total of 284 responses to experimental lists were recorded.

3.3 Simulation 1 – Acoustic models

The first goal of Simulation 1 was to follow the process of setting up DIANA from scratch. We developed our own acoustic models and compared their performance with the performance of existing acoustic models for English in a free word recognition test. We do not compare model estimates to participant data in this simulation.

3.3.1 Simulation setup

Acoustic models can be trained using careful (read) speech corpora such as TIMIT (Garofolo et al., 1993) or LibriSpeech (Panayotov et al., 2015), which was used in the development of Montreal Forced Aligner (McAuliffe et al., 2017). Acoustic models can be also trained using spontaneous speech corpora such as SCOTUS (Yuan & Liberman, 2008), which was used in the development of FAVE (Rosenfelder et al., 2014). We used two unpublished spontaneous speech corpora as a baseline for creating acoustic models. The Western Canadian English spontaneous speech corpus (WCE) contains telephone call recordings made by 11 speakers, while the Corpus of Spontaneous Multimodal Interactive Language (CoSMIL) contains conversation recordings of 8 pairs of speakers. We decided to use WCE and CoSMIL to train acoustic models for three reasons. First, many languages do not have extensive support in terms of previously available speech corpora. By using our own corpora, we show that an independent researcher could create a spontaneous speech corpus for their language of interest and use it to create acoustic models for DIANA. Second, the speakers in WCE and CoSMIL speak the western Canadian variety of English, same as the MALD speaker. Third, human listeners are more often exposed to spontaneous, conversational speech than to careful enunciations. It is best when a model of SWR can be presented with the same input as the human listener; in our study both are presented with MALD items in the test phase. However, we also wanted to represent the kind of "practice" human listeners receive as faithfully as possible, so we used spontaneous speech in the training phase.

In our implementation of DIANA, similar to previous implementations, we trained the acoustic models using automatic speech recognition training in the Hidden Markov Model Toolkit (HTK; Young et al., 2006). WCE and CoSMIL recordings were separated into brief speech intervals, and we further split the longer transcribed intervals to create speech chunks shorter than 10 seconds. We excluded speech chunks that entirely consisted of silent pauses, laughter, or other non-speech noise. In total, just over nine hours of speech were isolated and split into 20,086 speech chunks each shorter than 10 seconds. We downsampled the speech chunks to 16 kHz, and excluded 31 speech chunks due to potential sound clipping.

The first step in the training procedure takes the speech chunk input and creates estimates for all transcribed units (in this case, phones) as three-state hidden Markov models (HMMs), while the acoustic characteristics of phones are represented by Gaussian mixture models (GMMs). Speech chunks from conversational speech often included two or more connected words. Therefore, we expanded the acoustic models to also include estimates for short pauses in speech, that is, we created the so-called "sp models".

Increasing the number of GMMs per state may reliably reduce error rate in word recognition (Vertanen, 2006), so in the second step of creating the acoustic models we increased the number of GMMs per HMM state from 1 to 2, then 4, 8, 16, and finally to the usually recommended 32 GMMs. The currently employed monophone system assumes that phones are context independent. In reality, they are not, so with larger training material triphone models can be created to take into account phonetic context. One drawback of such an approach is that it is even more substantially based on automatic speech recognition, and thereby less likely to serve as a genuine proxy with regard to the representation of human processing. We kept our models simple due to our limited training material, but also because HTK is just a technical mechanism to bridge audio on the one hand and activations of words as items in a dictionary on the other.

The third and final step in creating the acoustic models was speaker adaptation. In this step, recordings from the MALD speaker (the speaker that the model will be tested on) are introduced to realign acoustic model estimates. Using a portion of speakers' recordings for training purposes limits the amount of material remaining for the test phase. Considering that the amount of material from the same speaker used in a behavioral experiment may be small to begin with, we wanted to test how much material is required to create adequate acoustic models. Starting with the “sp model” described above, we created separate speaker-adapted models differing in the number of MALD word recordings used for adaptation. Pseudoword recordings were not used in training. The smallest adaptation set included only three MALD word lists with a total of 1,200 words. Larger adaptation sets were created in increments of three (6 lists, 9 lists, 12 lists, etc.) up to 45 MALD word lists with a total of 18,000 words. Each list includes approximately just under 4 minutes of speech.

We compared speaker-adapted models in their ability to recognize the input signal from a list of competitors comprised of all 26,000 MALD words. We used 6 MALD lists (46 to 51) as test material. In the current implementation of DIANA, the activation component analyzes the acoustic input by converting it into MFCC vectors, while the acoustic characteristics of every phone in the lexicon, as we stated above, are represented by GMMs specifying the distribution of MFCC vectors for the three states in a hidden Markov model that each phone has. The matching is performed using a Bayesian framework and calculated for every ten milliseconds of input, as per the HTK default settings. Since the goal was to assess the quality of the acoustic models, activation values were not weighted by word frequency ($\gamma = 0$). Furthermore, we did not use the decision component of DIANA; we simply observed whether the correct word has the highest activation value. We also compared our acoustic models based on spontaneous speech corpora with FAVE acoustic models (Rosenfelder et al., 2014), likewise adapted for the MALD speaker.

Finally, we created n-best lists to show the top competitors and their activations at word offset. These lists allow us to see whether the competitors considered alongside the word with top activation are sensible, and also inspect the cases in which the wrong winner is selected. We created 20-best lists, that is, observed top 20 competitors for every target word. The choice of the number of competitors was arbitrary and made to ascertain that no important competitors will be omitted, but also to allow feasible computation and data manipulation. The number of retained candidates is comparable to those used in established measures such as orthographic Levenshtein distance 20 and phonological Levenshtein distance 20 (OLD20 and PLD20; Yap & Balota, 2009; Yarkoni et al., 2008).

3.3.2 Results

Free word recognition accuracy

Free word recognition results are presented in Figure 3.2. We can see that free word recognition accuracy is relatively low when models unadapted to the MALD speaker are used. In this initial step, the FAVE model performs slightly better than our own model (although the line fit is favoring our model, the actual accuracy presented by separate dots shows higher FAVE accuracy). Adapting the acoustic models on more MALD words leads to a large improvement in free word recognition at first, but this effect is reduced for adaptations performed on more than 9 MALD word lists. Acoustic models created based on WCE and CoSMIL slightly outperform the FAVE acoustic model after speaker adaptation. This difference also becomes smaller as more words are added and disappears when the adaptation is performed on 40 MALD word lists or more. Free word recognition accuracy never reaches 90%. As another point of comparison, the acoustic models used by ten Bosch, Boves, Tucker, et al. (2015) in the pilot DIANA simulations of MALD data had an accuracy of 82% when 500 words were tested with a lexicon of 36,000 word competitors.

We selected the model adapted on 30 MALD word lists, henceforth referred to as AM30, for all subsequent simulations. The difference in average accuracy between AM30 and the acoustic model adapted on 45 MALD lists is only 1%. The model adapted on 33 MALD lists is the first model where we see a slight decline rather than an increase in free word recognition accuracy, indicating that any additional realigning may be volatile. The model still offers a bit more (1.4%) than the model adapted on 15 MALD word lists, as well as a smaller difference in accuracy across

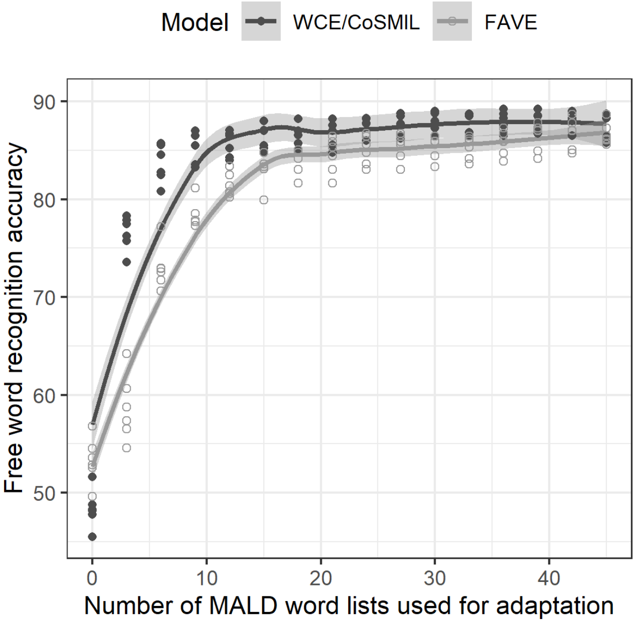


Figure 3.2: Accuracy in free word recognition of 2,403 MALD words. Average accuracy per list (46 to 51) is presented by separate points. The number of MALD word lists used for speaker adaptation is given on the x-axis. Models based on WCE and CoSMIL are given in dark gray, while the models adapted from the FAVE acoustic model are given in light gray.

the six test lists. Choosing AM30 as the model to be used leaves 37 MALD word lists available for testing purposes.

Inspecting top competitors and incorrectly recognized words

We also used AM30 to extract 20-best competitors for the target words in the six MALD test lists. We noted sensible competitors in all cases, regardless of whether the correct word was selected as the winner or not. Table 3.1 shows the top four competitors for target words *tales* and *proceed*. For the first word, the string of phones was correctly recognized although the target word shares the same activation level as its heterographic homophone *tails*. The correct word was selected as the winner because it appears earlier in an alphabetized list of words. (Note that weighing activation using word frequency would change activation values of the homophones so the more frequent homophone would be selected as the winner; regardless, in later simulations we treat a win by any of the homophones to the target word as correct.) High activations of rhyme competitors *pales* and *hails* indicate that the model is considering candidates with initial phone mismatches. The word *proceed* was incorrectly recognized as *precede* by a very small difference in activation, indicating that small differences in vowel characteristics may be difficult for the model to tease apart. Other close competitors include words that have the same lemma as the target word.

Table 3.1: Activation of top four competitors at word offset for two example words. Higher values indicate higher activation (e.g., -100 is better than -200). Activation level is also dependent on signal length, with longer words reaching lower negatives than shorter words. For the word *tales* the correct phone string was detected and selected as the winner. For the word *proceed*, *precede* was incorrectly detected as the winner, with the target word being a close second.

Target word	Competitor	Activation
TALES (correct winner)	TALES	-2,861
	TAILS	-2,861
	PALES	-2,870
	HAILS	-2,880
PROCEED (incorrect winner)	PRECEDE	-5,093
	PROCEED	-5.095
	PROCEEDS	-5,148
	PROCEEDED	-5,153

Out of 2,403 words considered, only 14 were not one of the top 20 competitors for their signal: *bow*, *curb*, *dear*, *tongues*, *desirous*, *boors*, *brazier*, *juggle*, *bairn*, *beer*, *betrothed*, *croquette*, *mowing*, and *priority*. We found no errors in the recordings of these words, and no commonalities between them. The 20 closest competitors for these words were still sensible. In all other cases, even when the correct word is not selected as the winner, it is at least a close competitor. In 59% of the remaining cases the correct word is the runner-up and in 88% it is within the top five competitors.

A closer (manual) inspection of errors showed that some of them were made because there is a heterographic homophone among the competitors, such as in *urns* and *earns* or *genes* and *jeans*. Other errors mainly occur due to uncertainty whether there should be an initial stop or not (e.g., *breast* winning instead of *arrest* and *aiding* winning instead of *bathing*), due to omitting the final stop (*individualize* instead of *individualized*), or due to the wrong vowel being activated (*cake* instead of *kick*). Complete information on 20-best lists can be found in our supplementary material.

3.3.3 Discussion

Setting up the HTK acoustic models required for DIANA simulations was relatively successful. It seems that approximately nine hours of transcribed and labeled spontaneous speech is sufficient to create acoustic models that will, after speaker adaptation, perform on par with certain existing acoustic models. Where speaker adaptation itself is concerned, we selected the model trained on 30 MALD word lists (AM30), that is, we used slightly less than two hours of careful speech from the MALD speaker. It also seems that similar results in free word recognition can be obtained with the equivalent of 40 minutes of speech (approximately 10 MALD word lists).

Free word recognition never reached 90% which is a result that could be improved. However, the competition process included 26,000 competitors for every word and even when a mistake was made the target word was often among the closest competitors. Therefore, we decided to use AM30 in subsequent DIANA simulations of the auditory lexical decision task and simply exclude cases in which the model is making a mistake.

3.4 Simulation 2 – Lexical decision

When simulating the auditory lexical decision task, DIANA treats the task as two distinct decisions — (1) the decision of whether a signal is a word or a pseudoword and (2) the decision of which word is the winning candidate and when it is selected as the winner. In Simulation 2, we use DIANA to simulate the first decision: whether a signal is a word or a pseudoword. We also compare DIANA errors to MALD1 participant errors.

3.4.1 Simulation setup

DIANA decides whether a signal is a word or a pseudoword by comparing the best possible activation of a word competitor present in the mental lexicon to the best possible activation achieved if any phone sequence is allowed. We will refer to the first activation as *word activation* and to the second as *free phone activation*. Word activation is the same activation presented in free word recognition in Simulation 1. In free phone activation, the language model does not include a mental lexicon (i.e., a list of word competitors). Instead, it only contains phones, and, optionally, probabilities of moving from one phone to the other. In our simulation, we treated all possible phone transitions as equally probable.

Word activation can never exceed free phone activation because words form a subset of the set of all word-like, phonotactically licensed phone sequences. Free phone activation adjusts itself as best it can to the audio signal and generates a string of phones; the activation of a phone string already present in the lexicon (i.e., word activation) can at best perfectly match free phone activation. Sometimes, however, the signal does not have a perfect match with any of the words in the lexicon and the activation pattern is coerced to adapt to an existing entry it fits best, leading to imperfect matching and therefore lower word activation levels. The larger the difference in word activation and free phone activation, the less the signal resembles the given word. Finally, it is also possible for free phone activation and word activation to match very well, even though the wrong word has the highest activation and is winning instead of the target word.

When a pseudoword is presented to the model, free phone activation should deviate significantly from word activation for any word in the mental lexicon, simply because phone strings comprising pseudowords are not present in the mental lexicon. DIANA differentiates between word and pseudoword signals using the extent to which free phone activation matches the activation of a word candidate from the mental lexicon — words should have similar free phone and word activations, while pseudowords should have much higher activation of a free phone string than any word activation. This should yield two distinct distributions of differences between free phone and word activation, forming a group in which the difference is 0 or close to 0 (words) and a group in which the difference is larger (pseudowords).

Ideally, there would be no overlap between these two groups of stimuli, allowing the model to perfectly distinguish between them. However, this would require acoustic models that perform perfectly, in addition to all word and pseudoword recordings having very careful enunciation of every phone in the word that align well with the acoustic model. Instead, DIANA employs a threshold θ_{lb} that specifies the difference between free phone activation and word activation that is small enough for a signal to be considered a word. This threshold is adjustable and we investigate what value leads to best accuracies in word and pseudoword classifications. At the same time, we are careful to select a "strategy" for the model that will at least to a degree match participant word vs. pseudoword response rates.

Besides calculating free phone activation and introducing pseudowords, we made additional changes in comparison to Simulation 1. We performed the simulation on all MALD words from lists 31 to 67 (i.e., all lists that were not used in adapting the model AM30, a total of 14,800 words) and on all MALD pseudowords. Instead of using all of the MALD words as the lexicon of competitors, we created separate lexicons for every word and pseudoword. Since DIANA endorses a Cohort-like competition (Marslen-Wilson & Tyler, 1980; Marslen-Wilson & Welsh, 1978), the lexicon included all short words (three phones or fewer) and all words that share the first three phones with the target word. The competitors were selected from the CMU dictionary (Weide, 2005). This procedure yielded lexicons of approximately 25 thousand words mostly comprised of words with three phones or fewer. In other words, the intention was to limit the competitor list under the assumption that the first three phones would be correctly recognized as the signal unfolds, but expand the number of close competitors by including more similarly-sounding words.

We noticed in Simulation 1 that sometimes the wrong word is selected as the winner because the target word had a heterographic homophone in the lexicon (as in the *urns* vs. *earns* example mentioned previously). Using the entire CMU dictionary to create separate lexicons of competitors for each target word introduced many such heterographic homophones. To avoid this issue, we treated all cases in which a homophone of the target word was selected as the winner to be accurate, given that homophones have identical activations in DIANA and that in an auditory task where single words are recognized both are technically correct.

The activation scores for words were still left unaffected by word frequency weighted by the parameter γ. The decision whether a signal is a word or a pseudoword compares the activation of a single top word candidate and the best free phone activation, meaning that relative positioning of word competitors is insignificant. Furthermore, participants in an auditory lexical decision experiment quickly learn that the number of words is roughly equal to the number of pseudowords: any given signal has an equal chance of being a word or a pseudoword and these prior probabilities cancel each other out.

3.4.2 Results

Differences in word activation and free phone activation for words and pseudowords

For word recordings, the maximum difference between word and free phone activations was 371.87, recorded for the word *depopulation*. Mean difference in activation was 29.64, while the median was 16.93. A total of 3,303 words (22.26%) had the activation difference of 0, meaning that free phone activation perfectly matched word activation.

For pseudoword recordings, the differences between word and free phone activations were expectedly higher. The highest difference was 1,482.45 and it was recorded for the pseudoword /ɛkmɪsieɪskləɹoʊsiz/ which the model fitted as *Izzy's* — competitors with more than three phones beginning with /ɛkm/ were rare and not similar to this pseudoword. The mean difference in activations for pseudoword recordings was 222.34 and the median was 184.38. Still, 337 pseudowords (3.5%) had the activation difference value of 0, meaning that the model incorrectly interpreted them in a way that perfectly matched with a word in the lexicon.

Figure 3.3 shows the distribution of word and free phone activation differences for word and pseudoword recordings. We set the x-axis limit to activation difference of 500 to make the distribution in the lower values more visible, but the long tail of differences continues for pseudoword recordings up to 1,482.45. In the case of words, as the activation difference increases, the number of words with that difference between word and free phone activation decreases; most words tend to have a small difference between word and free phone activation. For pseudowords, this trend can also be noted, but with a much smaller slope, as the distribution is, especially towards the lower hundreds, nearly uniform.

Adjusting the ratio of "word" vs. "pseudoword" responses and inspecting DIANA's accuracy in lexical decision

We then examined the ratio of "word" vs. "pseudoword" responses and model accuracy in predicting whether the input signal is a word or a pseudoword as a function of θ_{lb}. To make a balanced list of words and pseudowords for our computational simulation, we randomly selected a subset of words that matched the number of MALD pseudowords retained at this point in the simulation process (9,591). The activation difference was equal in words that were selected and words that were not selected for the subset, confirmed by a Wilcoxon rank sum test with continuity correction ($W = 25096000$, $p = .87$). We varied θ_{lb} in increments of 10, starting from 0, mean-

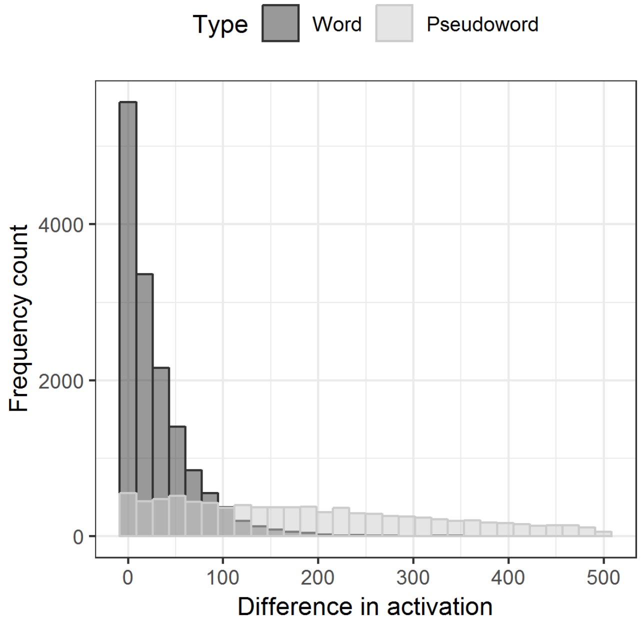


Figure 3.3: Histogram of differences in word activation and free phone activation for MALD words (lists 31 to 67) given in dark gray and MALD pseudowords given in light gray. The x-axis is limited to activation difference of 500, but the same trend continues to the maximum activation difference recorded which is 1,482.45.

ing that only a perfect match between word activation and free phone activation yields a "word" response, and ending with 200, leaving out only 66 words (0.004%) with an activation difference higher than this number.

The percent of "word" responses increases as the θ_{lb} increases. When θ_{lb} is 0 only 13.01% of the stimuli are selected as real words. With θ_{lb} of 150 the percent of "word" responses in all stimuli rises to 70.44%. Figure 3.4a shows this relationship and also includes three points of special interest on the curve. These points mark the quartiles of the percent of "word" responses in all experimental sessions in MALD1. The middle half of MALD1 sessions (the interquartile range) are found between points Q1 (48.12% "word" responses) and Q3 (54.83% "word" responses). These results indicate that in most sessions MALD1 participants had a fairly balanced response regime, making roughly an equal number of "word" vs. "pseudoword" responses. Since DIANA aims to simulate general tendencies in participant behavior, it seems that θ_{lb} should be set in the range between approximately 50 and 70. This threshold is dependent on the choice of features (in our case, MFCCs), the quality of the speech recordings, and the speaker — θ_{lb} needs to be adjusted for each new experiment.

DIANA's accuracy in classifying words versus pseudowords is also dependent on θ_{lb}. As the threshold rises, so does the percent of word signals correctly recognized as words. At the same time, the number of false alarms increases, as more and more pseudowords are mistakenly taken for words. Figure 3.4b shows how accuracy for words and pseudowords changes as a function of θ_{lb} and again includes quartiles from MALD1 sessions for comparison. We see that in our current setup DIANA inevitably performs worse than an average MALD1 participant, as it cannot have a performance higher than the value of Q1 for both words and pseudowords. Since the focus of our simulations were responses to words and since we wanted to match the participants' balanced response regime, we settled for θ_{lb} value of 70. With this threshold value the model made 55.65% of "word" responses and had an accuracy of 87.92% when responding to words and 76.44% when responding to pseudowords.

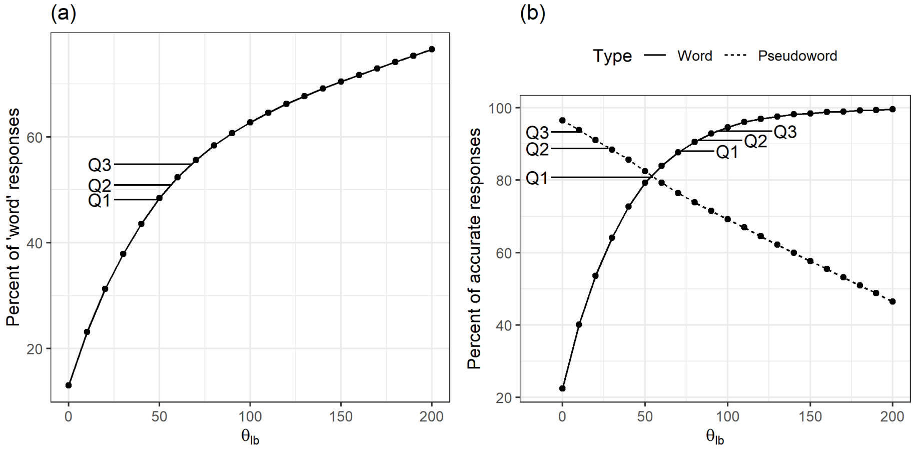


Figure 3.4: Figure (a) presents the relationship between threshold θ_{lb} and the percent of "word" responses DIANA makes. Points Q1, Q2, and Q3 are added for comparison and represent quartiles of the percent of "word" responses in MALD1 sessions. Figure (b) shows DIANA's accuracy in lexical decision for words and pseudowords as a function of threshold θ_{lb}. Points Q1, Q2, and Q3 connected to word and pseudoword curves represent quartiles from MALD1 session data.

Cross-tabulation of lexical decision and winner selection accuracy at word offset showed that 78.46% of word items were both selected as a word by the lexical decision process and the right competitor won at signal offset. In 10.57% of the cases, the correct word was the winner based on word activation, but the word activation was still smaller than free phone activation by more than 70, so these signals were incorrectly flagged as pseudowords. In 9.45% of the cases, the signal was accurately selected as a word (the difference between word and free phone activation was less than 70), but the wrong word had the highest activation at word offset. Only 1.52% of word recordings were both mistakenly marked as pseudowords and the wrong word had the highest activation at word offset. Words tend to be misinterpreted as another word rather than a pseudoword if they are shorter in duration (Welch's unequal variances t-test: $t_{(2736.2)} = -47.62$, $p < .001$) and have fewer phones (Welch's unequal variances t-test: $t_{(2471.3)} = -56.18$, $p < .001$), as these words have more close competitors.

Inspecting the relationship between lexical decision accuracy and stimulus duration

As noted in Simulation 1, longer recordings reach lower negative activation values. One concern that we had is whether higher differences between word activation and free phone activation would simply be a product of longer signals and a higher opportunity of mismatch between the two. Figure 3.5 shows that this is indeed the case, but mostly for pseudowords ($r = .76$), not words ($r = .33$). Words of any length can have an activation difference of less than $\theta_{lb} = 70$ and therefore be recognized as a word by DIANA. Only 8.16% of pseudowords split into more than 50 frames (approximately 520 ms in duration or longer) are incorrectly recognized as words, and this percent drops further to only 0.58% for pseudowords with more than 70 frames (720 ms).

We do not see such a strong correlation between duration and accuracy in MALD1 participants. There is no correlation between word duration and the proportion of correct responses to that word ($r = .03$). For pseudowords, the correlation between

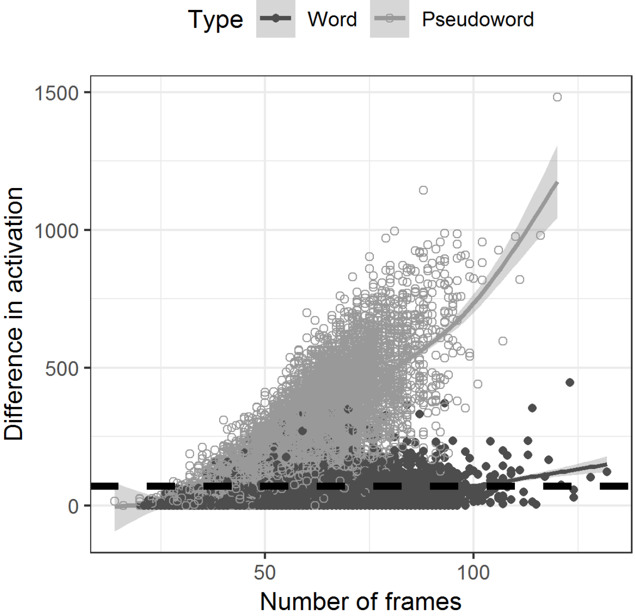


Figure 3.5: The relationship between the number of frames (duration) of a stimulus and the difference between word and free phone activation for that stimulus presented separately for words and pseudowords. The black dashed lines marks the selected difference value of $\theta_{lb} = 70$. The eventual score a hypothesis (phone string) receives is based on the summation of local scores that are associated to individual MFCC frames — in principle, the longer the stimulus, the larger the deviations will be between the scores of competing candidates.

pseudoword recording duration and the proportion of correct responses to that pseudoword is much lower than the one recorded in DIANA (being $r = .27$ in MALD1 data). The relationship between the proportion of correct responses to words and the activation difference in DIANA is also practically non-existent ($r = -.08$), but pseudowords with higher activation differences are also recognized as pseudowords by participants more often ($r = .29$).

3.4.3 Discussion

The goal of Simulation 2 was to test DIANA's approach to modeling lexical decision, that is, the word/pseudoword decision all human participants make in an auditory lexical decision experiment. Specifically, we wanted to establish the best value of the threshold θ_{lb} which determines whether a stimulus will be recognized as a word by

DIANA. We found that the approach can distinguish between the two types of stimuli fairly well, although the model in our current setup does perform somewhat worse than an average MALD1 participant. It is important to note that DIANA's response accuracy could be increased by selecting an unrealistic response strategy — in our case, by increasing the number of "word" responses the model makes. However, this is a poor approach if the goal of the simulation is matching participant behavior. The goal of cognitive simulation is to explain a process such as spoken word recognition in humans using plausible solutions, not to maximize model performance.

The reasons for making mistakes are only partly shared between DIANA and human participants. Both can "mishear" the signal, taking a pseudoword for a word, a word for a pseudoword, or mistaking the word for some other word. However, participants also make mistakes because they do not know a word, whereas DIANA has all the MALD/CMU words stored in its lexicon. Additionally, a human participant can simply lose attention and press the wrong button (ten Bosch et al., 2019), whereas DIANA always performs on the same level. In the current simulation, DIANA's performance fully depends on the quality of the acoustic models, the characteristics of the incoming novel acoustic signal, and the available competitors in the mental lexicon.

In the case of pseudowords, we note a trend in which longer pseudowords are more accurately categorized by DIANA. We explain this finding in terms of cumulative activation and lexicon structure. There are more opportunities for longer pseudowords to mismatch with an existing word. Additionally, the number of plausible word candidates is smaller for longer pseudowords and with that so are the odds of the pseudoword signal being mistakenly taken for an existing word. Still, due to imperfect acoustic models, we see that certain short pseudowords are mistaken for words using the threshold $\theta_{lb} = 70$.

Although a similar relationship between pseudoword duration and accuracy exists in the MALD1 data, it is much less pronounced. But why are not MALD1 participants benefiting (as much) from more opportunities for mismatch and fewer plausible word candidates when listening to longer pseudowords? We argue that, unlike our current DIANA setup, MALD1 participants are aware of the morphological and even semantic characteristics of pseudowords, making certain long pseudowords more word-like to a human listener. MALD pseudowords were created from actual words of English by replacing a third of their subsyllabic constituents with another phonotactically licit and probable segment, yielding pseudowords with some apparent morphological complexity (Tucker et al., 2019). One example is the pseudoword /ɛnspeɪzd/ that was correctly classified in only 36% of occurrences in MALD1 sessions. Morphologically, this word may resemble a combination of *en* plus *spaced* (although we do note that the final sounds are voiced in the pseudoword), as in, for example, *encircled.* Another example is the pseudoword /trænzvɑɹmɪŋ/. Although there are differences in comparison to existing words such as *transforming*, or a potential "word" *transwarming*, the prefix *trans* and the suffix *ing* in conjunction with the central part of the word that sounds like existing words are likely the reason why this pseudoword was correctly responded to in only 27% of its MALD1 trials. DIANA is not sensitive to this kind of similarity and the top word competitor to /ɛnspeɪzd/ is the word *inspires*, while the top competitor for the pseudoword /trænzvɑɹmɪŋ/ is *tensiometer.* Furthermore, recent research shows that processing written pseudowords is not free of frequency or semantic effects (Cassani et al., 2020; Hendrix & Sun, 2020), as pseudowords do not necessarily have a frequency of 0, and as form-meaning patterns learned from words can extend to pseudoword processing.

Where words are concerned, although we changed the lexicon of competitors, tailoring them for every target stimulus, word recognition accuracy remained as high as in Simulation 1. However, in a portion of cases a pseudoword has higher activation than the target word. Additionally, among correct lexical decisions there are cases in which the wrong word had the highest activation. Both of these kinds of errors in DIANA's word recognition stem from the same root cause — relatively low activation of the target word. Since the goal of our simulations is to give DIANA its best possible chance at simulating the process of spoken word recognition, we will only use those words in which both the target word is activated the highest and the signal is correctly classified as a word in the following simulation of response latency.

3.5 Simulation 3 – Response latency

The goal of Simulation 3 was to test how well DIANA's estimates of when a word is recognized and selected as the winner match general tendencies in participant response latency from MALD1 data.

3.5.1 Simulation setup

In Simulation 3, we only considered the 11,592 words that were both correctly recognized at word offset and treated as words (not pseudowords) by DIANA in Simulation 2. We used the same lexicons of competitors as in Simulation 2. However, in Simulation 3 we calculated word competitor activation using a gating procedure. We split all word recordings into 20 ms frames. Model estimates were made upon addition of every new frame. Since the process is computationally demanding and since the initial stages of word competition are uninformative, we only observed the activation of top 20 competitors in the last 300 ms of the sound signal. In effect, the gating procedure allows us to estimate competitor activation and observe the activation-competition process as the signal unfolds. Additionally, DIANA's decision component can make a decision at every selected point in time during the signal presentation.

The activation at the final phase of the gating procedure (word offset) is identical to the activation used in the lexical decision simulation from Simulation 2. We already determined the value of the lexical decision threshold θ_{lb} based on the difference in free phone and word activation when the entirety of the signal was available to the model. The majority of responses in auditory lexical decision experiments are made after signal offset, and our reasoning was that one viable strategy for the listener would be to make the best possible decisions when all of the information is available.

Additionally, varying all parameters in DIANA at the same time would create too many combinations for feasible computation and analysis of results, so we determined θ_{lb} independently from γ, θ_{wc}, and β.

We followed similar reasoning when determining plausible values for parameter γ that controls the contribution of top-down (frequency) effects; the value of γ needs to be determined experimentally because it depends on the type of word material used. Since we only selected words that were correctly recognized in Simulation 2, accuracy in selecting the right word at word offset is 100% with no contribution of word frequency. However, modifying acoustic activation using word frequency may change the order of top competitors if a runner-up has a much higher frequency than the top competitor and a high weight is assigned to the top-down effect, that is, a high γ is used. As shown in Equation 3.1, a competitor's *total activation* (TA) was calculated as a sum of its *acoustic activation* from the acoustic model (AM) and logged frequency count (f) from the Corpus of Contemporary American English (COCA; Davies, 2009) weighted by parameter γ.

$$TA = AM + \gamma * log(f) \tag{3.1}$$

We assessed which values of parameter γ are acceptable as weights for logged frequency so that word recognition accuracy is not severely reduced. The word recognition process in the auditory lexical decision task is primarily guided by acoustic information, not prior probabilities or context; as Norris and McQueen (2008, pp. 371) state: "Once the perceptual evidence becomes completely unambiguous, frequency should never override it". In effect, we opted for an approach that increases the difference between the top competitor and other competitors if the top competitor is a high frequency word, and reduces this difference if the top competitor is a low frequency word, but ultimately does not determine which word is heard. This should yield results in which high frequency words are isolated and recognized sooner, while low frequency words are more difficult to isolate and are recognized later. It should also be stressed that the word frequency effect in the current setup is further limited because it only modifies the activations of up to the top 20 acoustic competitors.

The decision of which word is the winning candidate in DIANA is regulated by a threshold θ_{wc} determining the required difference in activation between the leading candidate and the runner-up. Since there are many heterographic homophones in the dictionary that will have identical activation (e.g., *tails* and *tales*), we only considered non-homophone competitors when we determined the difference between the leading candidate and the runner-up. We calculated this difference at every step in the gating procedure. When determining the range of acceptable values for threshold θ_{wc}, we again used MALD1 responses as a benchmark. Increasing θ_{wc} increases the required difference between the top competitor and the runner-up for a winner to be selected, and therefore increases the number of word signals which do not have a clear winner before word offset. A very low value of θ_{wc} will in turn yield many winners before word offset — which can also lead to many wrong competitors being selected as winners based on early activation. We decided to adjust the value of θ_{wc} so that the percent of words that win before word offset is roughly equal to the percent of word responses that happen before word offset in MALD1 data. When determining this percent for MALD1, we added 200 ms to word duration to take into account the time required to execute the response, as assumed by DIANA.

We only selected words that were correctly recognized at signal offset in Simulation 2 to be used in Simulation 3. However, a wrong word may be the leading candidate prior to signal offset, especially considering that top-down information now affected competitor activation. Therefore, we also tested which word is the leading candidate at the time frame when the winner is selected.

When a winner is selected prior to word offset, DIANA takes the time at which it was selected and adds the aforementioned 200 ms for execution. In the case when the required difference between the top-competitor and the runner-up (controlled by threshold θ_{wc}) is not attained at stimulus offset, a controllable parameter β estimates the added time for the final winner decision. The time needed to decide on the final winner depends on the number of remaining plausible competitors, that is, all the words with an activation difference of less than θ_{wc} from the top competitor. However, when simulating the lexical decision task, DIANA assumes that the listener is at this stage also considering viable phone strings which are not present in the mental lexicon. In other words, pseudowords are also competing with real words, increasing the perplexity of the decision at signal offset. Unlike for highly activated word competitors, we cannot obtain the activation values for all potential pseudowords. The number of pseudoword competitors at word offset is approximated by raising 3 to the power of the number of phones of the target word. This is a crude estimation in itself that also assumes that these non-word competitors are still plausible competitors at stimulus offset. The formula for estimating choice reaction time then follows the Hick-Hyman law (Hick, 1952; Hyman, 1953) by calculating the logarithm of the total number of remaining word and pseudoword competitors weighted by parameter β (Equation 3.2). Choice reaction time is finally added to the total duration of the signal, in addition to the 200 ms required for execution.

$$RT_{choice} = \beta * log(N_{words} + 3^{N_{pseudowords}}) \tag{3.2}$$

With acceptable ranges for parameter γ and threshold θ_{wc} determined, we adjusted the value of parameter β to maximize the match in mean response latency between DIANA and MALD1. We then observed the correlation between logged DIANA's response latency estimates per word calculated using the selected values of γ, θ_{wc}, and β and mean logged MALD1 response latency per word. We followed the procedure from ten Bosch et al. (2018) to de-trend MALD1 response latencies, limiting the degree of local speed effects (Ernestus & Baayen, 2007). Maximum between-participant correlation on the entirety of MALD1 data ($r = .19$) was achieved when ten previous responses were taken into account to determine the "true" current response latency. The code for MALD1 data de-trending is available alongside all other data and scripts in our supplementary material.

3.5.2 Results

Frequency effects and word recognition accuracy

We first tested how word recognition accuracy at signal offset changes when top-down frequency effects are introduced to the model. We tested γ values from 0 (no frequency effect) to 20, in steps of 2. With $\gamma = 20$, word recognition accuracy dropped to 90%, meaning that in 10% of the cases a more frequent competitor won instead of the less frequent target word. Word recognition accuracy decreased by approximately 0.5% at each step of γ increase. We decided to allow less than 5% error rate at word offset due to frequency effects and only considered values of parameter γ up to 10 in our comparisons to participant response latency.

Adjusting the required difference between top competitor and runner-up activation

We then assessed plausible ranges for threshold θ_{wc} by comparing the percent of decisions made before word offset in DIANA and MALD1. The correlation between the percent of responses made before word offset and the percent of correct lexical decisions was very low in MALD1 sessions ($r = -.10$). In our simulations, we only considered words correctly recognized by DIANA and compared their estimates to response latency in correct trials from MALD1. Therefore, we decided to only observe the percent of correct responses made before word offset per MALD1 session.

Figure 3.6 shows how the number of winner selections that happen before word offset decreases as the required difference in activation between the top competitor and the runner-up (θ_{wc}) increases. This relationship is nearly identical for all considered levels of γ (0 to 10). MALD1 data includes a wide distribution of percents of responses made before word offset when 200 ms are deducted from the response latency to account for execution time. This indicates a wide range of participant strategies: while some opt to make practically no (correct) responses before they heard the entirety of the signal, certain other participants make up to 80% of their responses at least slightly before the signal ended. A portion of this variability may be attributed to simple differences in speed, as it is probable that not all participants take exactly 200 ms to execute a response to every stimulus. The mean percent of correct responses made before word offset in MALD1 sessions was 26%, while the median was 24%. Since our goal was to match general tendencies in participant performance, it would be reasonable to opt for θ_{wc} values that would yield 16% (Q1) to 35% (Q3) of responses made before word offset. As can be seen in Figure 3.6, θ_{wc} values between 150 and 220 fit that range.

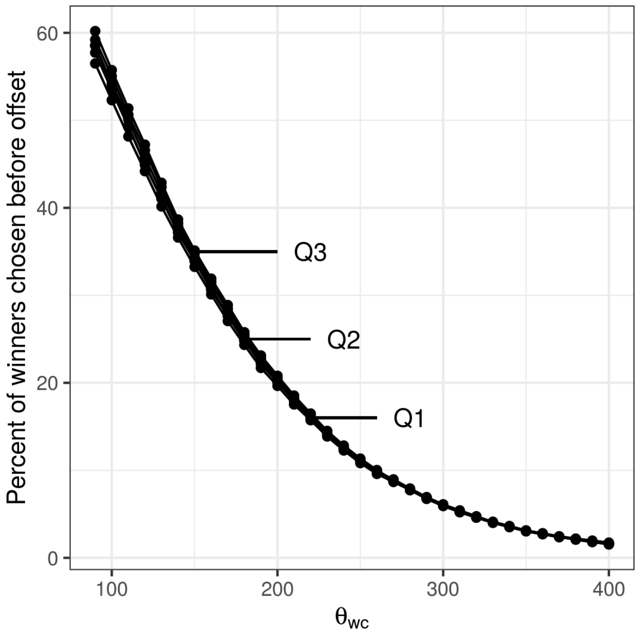


Figure 3.6: DIANA's percent of decisions made prior to word offset as a function of threshold θ_{wc}. Separate lines are drawn for different parameter γ values (0 to 10), and these do not seem to affect the results. Points Q1, Q2, and Q3 represent quartiles from MALD1 session data.

However, making a decision prior to word offset also introduces the risk of choosing the wrong word as the winner: at some point during the activation-competition process, a candidate may get highly activated and win, even though the remainder of the signal would reduce its activation. (Remember that we previously excluded all the words that were incorrectly recognized in Simulation 2 due to imperfect alignment of acoustic models and word recordings, so all words are correctly recognized at word offset.) Therefore, we tested how accuracy in selecting the right word as the winner changes as a function of θ_{wc}. Figure 3.7 shows that the number of wrong selections for responses prior to word offset decreases as θ_{wc} increases. When the model is more conservative in selecting the winning word and fewer words are recognized before word offset, there is less of a chance that the wrong word will be selected as the winner. Frequency again plays only a minor role, especially in the more favored, higher values of θ_{wc}. The vertical dashed lines in Figure 3.7 represent the margins within which the average MALD1 session operates (16 and 35% of word responses before signal offset). According to DIANA, that would indicate that for 10 to 15% of the responses before word offset the participants actually heard the wrong word. We cannot know whether this is true as the standard auditory lexical decision task (unlike, e.g., word repetition task) does not require the participant to state which word they heard. Additionally, this is not entirely implausible: considering the number of word stimuli in a MALD session, this would mean that around 10 to 15 word responses were actually made prematurely, thinking of a different word than the one presented. We also found that target words for which DIANA selects the winner before word offset have relatively earlier phonological uniqueness points (when the total number of phones in the word are taken into account), confirmed by a Wilcoxon rank sum test with continuity correction ($W = 21033820$, $p < .001$). This finding indicated that the selection of words for which a response was made before offset by DIANA is plausible.

We decided to be more conservative and consider θ_{wc} values between 200 and 400 for estimating response latency in DIANA, primarily to decrease the number of wrong word recognitions. This range of θ_{wc} values yields a percent of responses made before word offset that is not out of range of MALD1 sessions. We excluded very early DIANA estimates of a word winning (before 420 ms pass with 200 ms for response included) as unrealistic. We also excluded words that were RT outliers in

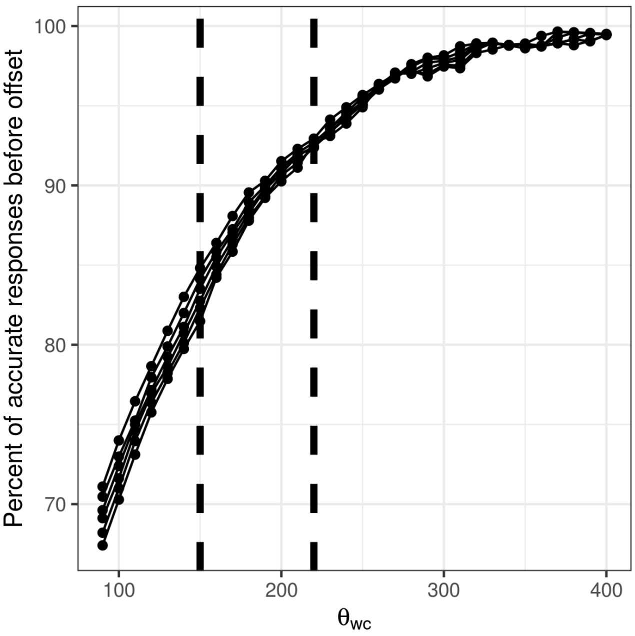


Figure 3.7: DIANA's percent of accurate responses prior to word offset as a function of threshold θ_{wc}. Separate lines are drawn for different parameter γ values (0 to 10), and these do not seem to affect the results, especially in higher values of θ_{wc}. The two vertical dashed lines delineate the value range of θ_{wc} that yields a percent of responses made before offset equal to the values between the first and third quartile in MALD1 session data.

MALD1 data. The final number of words used to correlate DIANA's response latency estimates with response latency from MALD1 data was 11,488. DIANA estimates of when the target word should be selected were compared to de-trended MALD1 data from sessions 31 to 67. We used γ values of 0 to 10 in steps of 2 and θ_{wc} values between 200 and 400 in steps of 20. If the decision was made prior to word offset, we took the time when the slice ends and added 200ms to account for execution time. In all other cases, we adjusted the value of β to maximize the similarity of estimated response latency between the model and behavioral data.

Adjusting the calculation of decision time past word offset and predicting MALD dRT using DIANA estimates

To adjust plausible levels of β, we assessed the final response latency estimate in milliseconds provided by DIANA and observed whether this duration fits within the general time frame of responses in MALD1. Figure 3.8 shows how the average response latency estimate by DIANA changes as a function of β (ranging from 0 to 100 in steps of 10) for different levels of θ_{wc}. Word frequency effect controlled by γ is not presented as it had little impact on the overall trends. We see that the estimated RT increases as β increases, which is expected given that β governs how much additional time will be spent on each remaining competitor past word offset. Different lines represent different values of threshold θ_{wc}. DIANA's estimated RT also increases as θ_{wc} increases because more words are not responded to until word offset and have choice reaction time added to their estimate. The dashed line represents the mean MALD1 response latency for the 11,488 MALD1 words (949 ms). Similarly, the mean RT of correct response to word stimuli calculated per MALD1 session is 941 ms. Therefore, a plausible value of parameter β ranges between 25 and 37.

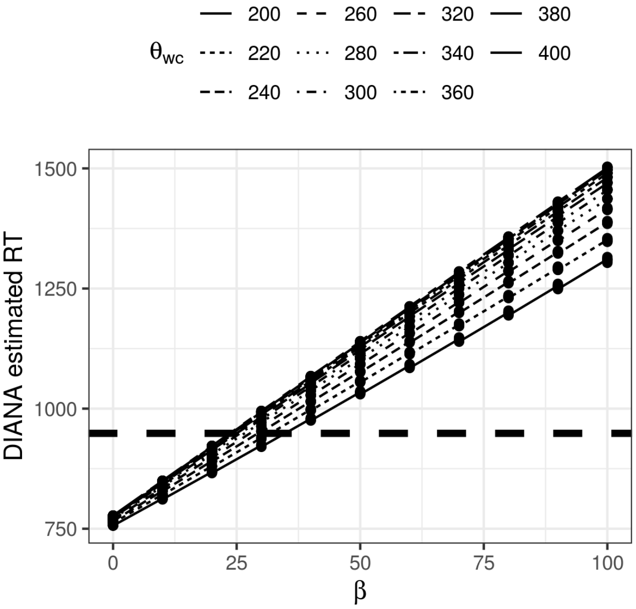


Figure 3.8: DIANA's estimated RT as a function of parameter β for different values of θ_{wc}. The horizontal dashed line is set at 949 ms and is equal to the mean MALD1 RT for the same 11,488 words considered by DIANA.

However, for the range of β values between 25 and 37 we noticed that the correlation between logged DIANA's estimated RT and participant de-trended RT (dRT) decreases as β increases. We decided to use a broader range of β values to better explore this trend. Figure 3.9 represents the change in the correlation between DIANA and MALD1 response latency for values of β ranging from -50 to 100 in steps of 10. Different lines stand for different θ_{wc} values (different γ values were again not considered as word frequency had a very small effect on the overall results). The highest correlation between logged DIANA's estimates and MALD1 dRT was $r = .43$ and it was obtained in negative values of β, specifically, for any value of γ (0 to 10), θ_{wc} of 400, and β of -10. In other words, the best result was obtained when RT was subtracted from word offset rather than added to it. When no choice reaction time was added to word duration and execution time ($\beta = 0$), the recorded correlation between DIANA's estimated response latency and MALD1 data was only slightly smaller ($r = .42$).

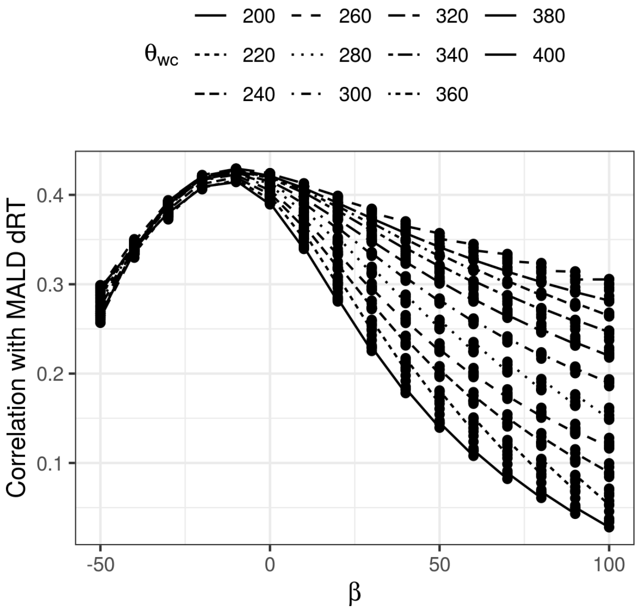


Figure 3.9: Correlation between DIANA estimates and MALD1 dRT as a function of parameter β for different values of θ_{wc} when both words and pseudowords are considered as plausible competitors at word offset.

3.5.3 Discussion

The goal of Simulation 3 was to simulate participant response latency when responding to words in the auditory lexical decision task. We used words that were correctly recognized in Simulation 2 and adjusted the values of parameters γ and β and threshold θ_{wc} to calculate DIANA estimates of participant response latency. Crucially, we developed parameter and threshold values that lead to plausible model behavior by comparing model performance to human performance.

Our results show relatively negligible effects of word frequency on simulation outcome. High values of γ, which would increase the impact word frequency has on final DIANA estimates, lead to a large number of wrong words winning instead of the target word. To prevent this kind of "hallucination", low values of γ need to be used. Furthermore, once the ranges of θ_{wc} and β values are adjusted and their values are varied within those ranges, the variation of γ within its acceptable range has a very small relative impact. In other words, DIANA estimates of participant response latency in the current setup perform best when they are almost exclusively based on bottom-up, acoustic information.

Varying the threshold θ_{wc} leads to conceptually similar results. Although MALD1 participants seem to be making at least a portion of their correct word responses before the end of the signal (when 200 ms is calculated as execution time), DIANA favors a very conservative approach in which the response is made after the entirety of the signal has already been presented. A lower threshold leads to an increased percent of wrong word selections prior to word offset, where some of the decisions are made implausibly early. Additionally, all correlations of MALD1 participant response latency with DIANA estimates favored the highest θ_{wc} we used (400), which indicates that virtually all DIANA's decisions should be made only after the entirety of the word signal had been presented to the model.

DIANA also includes a parameter β which weighs the formula accounting for choice reaction time — once the signal has reached its end, the model calculates added time needed to make a decision between the remaining plausible candidates. DIANA assumes that in the lexical decision task (but not word repetition task) the set of plausible candidates at word offset includes the remaining word competitors whose activation is within the threshold θ_{wc} from the activation of the leading candidate. The correlations obtained with MALD1 response latency were moderate, reaching the value of $r = .43$. Although this correlation with participant data is not by any means low, the contribution of choice reaction time to the overall correlation between model estimates and MALD1 data was very modest. More importantly, the highest correlation between DIANA's estimates of response latency and MALD1 response latency was registered when using negative values of β. DIANA's assumptions is that an increase in the number of remaining competitors should lead to longer choice reaction times. Instead, given that negative β values were optimal, a higher number of remaining competitors was connected to shorter response latency. Furthermore, negative values of β lead to *removing time* from the total duration of the signal in order to estimate the duration of the decision process after that signal had already completed. It is clear that such a procedure is in collision with the physical reality in which human listeners operate.

Simulating MALD1 response latency data shows a shortcoming of DIANA in the sense that the transformation from choice entropy to choice response time is not precise enough. This could indicate that either Hick-Hyman's law is not applicable in its full form, or that the computation of the entropy is not precise enough — e.g., due to the quite rough estimation of the number of pseudoword competitors at stimulus offset. We offer a more thorough discussion of the theoretical implications of these findings in the following section.

3.6 General discussion

In this study, we used DIANA (ten Bosch, Boves, & Ernestus, 2015) to simulate participant performance in an auditory lexical decision task. In three simulations, we (1) created new acoustic models for western Canadian English, (2) simulated the lexical, that is, word/pseudoword decision, and (3) correlated DIANA's estimates of when the winning word is selected with general tendencies in participant responses from the MALD project (Tucker et al., 2019). The results of these simulations can be used to guide future development of models of spoken word recognition including DIANA and, at the same time, inform the theory regarding the process of spoken word recognition.

In Simulation 1, we show that setting up DIANA in a new language is possible even without existing acoustic models: we used our own, relatively small, in-house spontaneous speech corpora to make new acoustic models. This process is labor-intensive, as it requires recording and annotating a speech corpus, training acoustic models, and recording enough additional material by the speaker whose recordings are used in experiments to adapt these acoustic models. It would be time-consuming for an independent researcher to take DIANA as an off-the-shelf model even with existing acoustic models, given that speaker adaptation must be performed regardless. We provide the acoustic models we developed and adapted for the MALD speaker as part of our supplementary materials. These adapted acoustic models should allow researchers to perform DIANA simulations using MALD recordings as material.

DIANA is not isolated with regards to model setup complexity. SpeM and Fine-Tracker (Scharenborg, 2008; Scharenborg et al., 2005) require similar preparatory work. Shortlist B (Norris & McQueen, 2008) depends on a large database of responses to gated diphones, which is likely the reason this model has only been implemented in Dutch. For comparison, using instantiations of the TRACE model, jTRACE (Strauss et al., 2007) and TISK (You & Magnuson, 2018), requires installations that can be completed in a matter of hours. However, the additional work yields a crucial advantage: DIANA deals with actual acoustic input (but see Norris & McQueen, 2008, for criticism of SWR models based on automatic speech recognition).

One consequence of a good representation of the variability in the acoustic signal is that DIANA performs well in free word recognition. Accuracy in selecting the correct word as the winner from a corpus of approximately 26,000 words was slightly under 90%. This level of word recognition accuracy is much higher than those we obtained using TISK, where lexicons with close competition never yielded word recognition accuracy higher than approximately 30% (Nenadić & Tucker, 2020, but note that the competitor structure was different in TISK simulations). Additionally, DIANA seems to perform better in free word recognition than Shortlist A and SpeM, as these models never exceeded 75% recognition accuracy although the lexicon of competitors was smaller than in our simulations see Scharenborg et al., 2005. Word recognition accuracy using the discriminative lexicon approach yielded accuracy of up to 25%. This simulations analyzed word recordings isolated from spontaneous speech and human participants generally did not perform better on the same material (Arnold et al., 2017). Lastly, our acoustic models work on par with the FAVE acoustic models (Rosenfelder et al., 2014). Even higher accuracy may be obtained with improved base acoustic models or extended model adaptation. A high standard of model performance in terms of input (free word) recognition is crucial for simulations that involve large word sets — that is, for any simulation that aims to be more than a proof of concept using a toy example.

Another important advantage of using acoustic signal as input is that competitor activation is dependent on the characteristics of the sound signal, not on preconceptions about which words should sound similar (see, e.g., Hawkins, 2003, for an extended discussion). For example, TRACE (McClelland & Elman, 1986; Strauss et al., 2007) relies on acoustic pseudofeatures to determine phoneme identity. These pseudofeatures always have the same values for a particular phoneme, meaning that every occurrence of a phoneme is always the same (barring some pseudofeature overlap of neighboring phonemes that accounts for coarticulation). DIANA, in turn, can analyze any number of unique recordings of the same word, each time generating a different activation-competition pattern. Besides providing much better estimates of what words the signal actually resembles most, this also allows researchers to explore and simulate phenomena that were not part of our simulation, such as subphonemic, acoustic effects (e.g., Andruski et al., 1994; Marslen-Wilson & Warren, 1994), effects of prosody (e.g., Kemps et al., 2005; Salverda et al., 2003), or effects related to processing reduced variants of a word (Ernestus & Warner, 2011; Ernestus & Baayen, 2007; Tucker, 2011; Tucker & Ernestus, 2016).

As the signal unfolds in time, activations of different word competitors rise and fall. A highly activated candidate at an early point in the signal may lose activation later, as more suitable candidates gain prominence. However, even towards the end of a signal, many competitors had high activation despite initial mismatch with the target word. Exploring the list of top competitors in Simulation 1 provides us with an example for this phenomenon: *pales* and *hails* are the highest activated competitors for the recording of word *tales*. This model performance is in line with human performance. A recent MEG experiment supports the claim that subsequent contextual information influences the perception of preceding segments as subphonemic detail is preserved in the auditory cortex and reanalyzed as additional signal becomes available (Gwilliams et al., 2018). Models of SWR in general attempt to include this kind of flexibility in word recognition and not discard a competitor based on differences in early phonemes, as was done in the original Cohort model (Marslen-Wilson & Tyler, 1980; Marslen-Wilson & Welsh, 1978). Cohort II (Marslen-Wilson, 1987; Marslen-Wilson et al., 1988) was specifically adapted to be able to recognize the correct word despite initial mismatch (see also Weber & Scharenborg, 2012). TRACE (McClelland & Elman, 1986) also retains competitors even if there is some initial mismatch, as besides cohorts (words sharing the first two phonemes with the target word), the model also considers rhymes (words sharing the last two phonemes with the target word) and embeddings (words that are fully embedded in the target word) to be close competitors to a target word. The authors of Shortlist B (Norris & McQueen, 2008) also make a point of that model successfully recognizing words despite some (initial) mismatch with the target.

Although DIANA's current implementation in HTK allows creation of large lexicons including as many as 36,000 words, this implementation still cannot deal with lexicons the size of, e.g., the entire CMU dictionary (approximately 135,000 words). Therefore, the initial set of plausible competitors needed to be decided by the researcher. DIANA endorses a Cohort-like competition so in our subsequent simulations (Simulations 2 and 3) we created separate lexicons to include all words with up to three phones and all words that share the first three phones with the target word. The assumption behind this procedure was that DIANA will resolve initial inconsistencies for longer words and that only the closest of competitors will matter towards word offset. If the acoustic models and the enunciations of every segment in MALD recordings were perfect, this would indeed be the case, but we have seen that DIANA made mistakes in recognizing the correct word due to, i.a., initial stop elision. Therefore, this approach seems to be faulty as it uses initial mismatch to disqualify words that could have won instead of the target word. Additionally, even if the model successfully resolves initial inconsistencies, that does not mean that competitors with initial mismatch are not some of the closest competitors to the target word. As stated above, one such example is *pales* being one of the closest competitors for the recording of word *tales*. By pre-excluding competitors based on initial phone mismatch we inevitably affected both the potential winner of the activation-competition process and the structure of close competitors (which may be relevant for response latency estimation).

Then how do we select which competitors should be included in the limited lexicon of competitors created for every target word? Of course, competitor selection also depends on the task the model of SWR is simulating. A clear example is the visual world paradigm where the lexicon may be limited to include only the competitors that are visually presented on the screen (see, e.g., Allopenna et al., 1998). In many other tasks, such as the word repetition or the auditory lexical decision task (and in everyday communication), any word in the lexicon could be activated if fitting acoustic signal is presented. If we continue to think of close competitors to words in terms of the phonemes they share, using competitor selection criteria from TRACE (McClelland & Elman, 1986) seems like a better approach. We already described above that competitors in TRACE include cohorts, rhymes, and embeddings to the target word. Note that this approach encompasses word neighbors from NAM (Luce & Pisoni, 1998) and word cohorts from the Cohort model (Marslen-Wilson & Welsh, 1978).

Still, it is possible (although not too probable) for a word competitor to be highly activated and not belong to any of these three groups of TRACE competitors, especially prior to word offset. This issue may be solved through brute force, that is, by the sheer size of the lexicon that the current implementation of DIANA can handle. The number of TRACE close competitors extracted from the CMU dictionary for 442 English words ranges from 17 to 2,243, with the average of 605 close competitors (Nenadić & Tucker, 2020). DIANA, in turn, can handle quite sizable lexicons. Therefore, we propose using the competitor selection approach from TRACE but also capitalizing on DIANA's capacity for large lexicons by selecting 30,000 words that have the lowest edit distances from the target word. This approach is yet to be tested, but all cohorts, rhymes, and embeddings should be present in these 30,000 selected competitors — in fact, it should be true that most of the words that are not within 30,000 most similar competitors to the target word (based on phonemic transcription) are indeed not very similar to the target word. One downside of that approach is that it is more computationally demanding than building smaller lexicons.

However, relying on categories such as the phoneme, as we have previously noted, misses a lot of variability present in the fine-phonetic detail and stemming from reduction or other pronunciation variants in words. Ideally, the close competitors would be determined using similarity in the acoustic signal, rather than generalized categories such as phonemes. For example, with our MALD word set, we could calculate acoustic distances between word recordings and use those to form sets of competitors for each word (Kelley, 2018). An even better alternative would be to calculate acoustic distances between many recordings of many words — although this would require a very large (truly, representative) set of word recordings — and use those as a benchmark.

We also note that the necessity of preselecting competitors in a model of SWR has been at least as much a question of its technical implementation as it has been of its theory. If a model always considers all the options stored in the mental lexicon, there would be no need to determine whether it, for example, calculates activation for rhymes to the target word or not (even more so since the model cannot know what word the signal is supposed to represent beforehand). Design choices such as the one whether there is lateral inhibition between candidates would still be important, but there would be no need to discard a candidate during the activation-competition process, and especially no need to prevent a word candidate from competing before the activation-competition process has even started. Our implementation of DIANA relied on HTK and the lexicon limitations we had were a matter of technique: the model could be implemented using, for example, KALDI (Povey et al., 2011), allowing for better performance and a much more fine-grained view of unfolding activations. The lexicon size could also be dramatically increased to hundreds of thousands of words, removing the issue of candidate pre-selection. Technical limitations and novel advances will certainly continue to shape models of SWR and in part determine which questions regarding their architecture are considered relevant.

DIANA's lexical decision accuracy was also fairly high. The model uses a simple but powerful solution of comparing the best possible activation of a word in the mental lexicon with the best possible activation of any phone string. The activation of a word signal should remain high even if it is coerced to fit an existing string of phones (i.e., a word in the lexicon), while the activation of a pseudoword signal should be much lower if it is forced to match a string of phones that it does not correspond to. There may be one objection to the current approach. Inevitably, DIANA and the human participant have different causes of errors, both in free word recognition and lexical decision. For the computational model, the only cause of error is a poor match between the acoustic signal and the existing acoustic models, leading to a misinterpretation of the input. Listener errors may have other causes besides issues in interpreting the acoustic signal. For example, a human may not have the target word stored in their mental lexicon (i.e., the person may not know a word), may not be able to retrieve the target word at that particular time, or may miss portions of the signal or press the wrong button.

This leads to two specific issues. The first issue is that pseudoword accuracy highly depends on signal length. We will address this finding in more detail below, when we consider the representation of the mental lexicon in DIANA. The second issue is that word frequency does not affect the outcome of the lexical decision, while MALD1 and other lexical decision data generally show that word frequency predicts response accuracy. As we said above, some of the correlation between accuracy and word frequency in behavioral experiments is certainly due to the fact that lower frequency words are known by fewer participants. Additionally, perhaps signals of low frequency words require a higher threshold of attention due to less practice with that signal; it is easier to get confused and make a mistake for a word one does know if that word is encountered rarely. Given high-performing acoustic models, future simulations could include a parameter that would estimate the probability of a word being responded to as a pseudoword based on that word's frequency (or other characteristics that prove relevant). Generally, however, we do not want the model to purposely "throw away" data, so the goal is to get lexical decision accuracy to be as high as possible.

The central aim of our simulations was to simulate the time needed to make a response from the onset of the signal. Effect of word frequency, regularly registered in statistical analyses of behavioral responses, was found to be negligible. Higher values of γ that would increase the impact frequency has on activation could not be used as they would lead to a large number of incorrect words winning simply due to their higher frequency. It is important to note that the current implementation of the frequency effect in DIANA is not as straightforward as it may appear. In statistical modelling of auditory lexical decision data, word frequency is ordinarily included as a predictor of response latency to that word. In DIANA, the effect of word frequency is not as direct. Instead, the impact of frequency is best described as an interaction between a word's frequency and the frequency of its close acoustic competitors. If a high frequency target word has a high frequency competitor, then the activation difference between the two will remain dependent on acoustic activation alone and the winner may be selected rather late. In contrast, a high frequency target word that has no high frequency competitors will become the sole plausible competitor much sooner. Statistical analyses of participant responses should investigate whether this sort of frequency relationship between top acoustic competitors is a better predictor of human response latency than using solely the frequency of the target word.

Another reason for low impact of word frequency is due to the model estimating that the optimal strategy is to wait until word offset. This behavior is unsurprising in an auditory lexical decision task, as a signal can become a pseudoword at any point before signal offset. At that point, the number of pseudoword competitors far outweighs the number of word competitors, making the values of θ_{wc} and γ less important for final response latency estimation. If it is confirmed that direct impact of word frequency is a better predictor of participant response latency than taking into account the relationship between frequency values of the target word and its close competitors, and if we consider response after word offset to be optimal, than perhaps word frequency should not affect its acoustic activation levels. Instead, we would argue for word frequency being included in the choice reaction time formula and used to facilitate sifting through remaining candidates in search for the target word. In such a setup, a high frequency word would stand out from its other plausible acoustic competitors better than a low frequency word when the acoustic signal before word offset was insufficient to make an early decision.

Once signal offset is reached, DIANA assumes that the task is to choose the correct winner from the number of remaining competitors, with the decision being weighed by parameter β. The list of competitors includes all words that have their activation within the value of θ_{wc} in comparison to the top competitors and all potential pseudowords. The number of pseudowords is approximated by raising 3 to the power of the number of phones in the signal word. A six-phone word would therefore have as many as 729 potential pseudoword competitors at word offset, and a word with seven phones would have 2,187. It is clear that when using this estimation pseudoword competitors far outnumber remaining plausible word competitors. In effect, the number of plausible word competitors and the distribution of their activations become insignificant in comparison. In turn, this means that the more phones a word has the longer choice reaction time will be for that word (as more potential pseudoword competitors are registered at word offset). In contrast, MALD1 data shows that longer words are responded to faster when response latency is calculated from word offset, that is, from the point when the signal of the word has ended. Simply put, while DIANA assumes that the time needed to select the winner will be longer in longer words due to many pseudoword competitors remaining at word offset, behavioral data shows an opposite trend in which participants respond faster to longer words relative to word offset.

In line with participant data, we found that optimal values of parameter β controlling for choice response time are negative, indicating that time should be deducted from word offset rather than added to it, and deducted more for longer words. Even with such a setup that would make the raw response latency estimates much shorter than those observed in the behavioral experiment, the added benefit of choice response time to the correlation with MALD data is very limited. The highest correlation between DIANA response latency estimate and mean de-trended logged participant response latency was $r = .43$. Although this correlation is moderate and higher than any correlation we managed to obtain using TISK and jTRACE (we never exceeded $r = .2$ using these models; Nenadić & Tucker, 2020), it is almost exclusively due to the fact that DIANA and the human participants were presented with the same sound recordings. The correlation when β is set to 0, that is, when no choice response time is added and word duration alone is used, was nearly as high ($r = .42$). These results indicate that a different way of representing the decision process is needed. The main issue seems to lie in the way choice response latency is calculated, especially with regards to estimating the number of plausible pseudoword competitors at word offset in longer words. A possible improvement for DIANA would be to adapt the current estimations of pseudoword competitors at stimulus offset, as the estimations are likely too high. The current number of pseudoword competitors is based on the full phonetic length, but perhaps only the most recent changes (those made towards signal end) matter — as all earlier ones are already knocked out as implausible.

It is also clear that many words have heterographic homophones in the CMU dictionary. These words have identical acoustic activation, although their activation may change after frequency effects controlled by parameter γ are introduced. Our approach was to simply treat any win by a homophone of the target word as correct. This is in line with the general status in the auditory lexical decision task as participants cannot know which exact meaning of the word was intended, nor can the researcher analyzing the data know which meaning the participant thought of when responding to a word stimulus. Still, this approach ignores an important characteristic of words — their plurality of meaning and contexts in which they are used, which ties directly to our interpretation of how the mental lexicon is organized and accessed.

Currently, DIANA represents the mental lexicon as a list of unconnected strings of phones (words), focusing on form alone. Under this setup, recognizing a word is in no way affected by the word's meaning beyond its frequency of occurrence. It is unclear how the model should treat word frequency of heterographic homophones. Would these words have a joint, sum frequency? Would only the most frequent word be considered? Or are these words separate lexical entries, each with their own frequency, the way they were represented in our simulations? Additionally, effects of word meaning in spoken word recognition extend beyond frequency of occurrence. We have seen in Simulation 2 that a representation of the mental lexicon that stores information on form and frequency alone leads to lexical decisions to pseudowords being mostly guided by direct acoustic mismatch, making long pseudowords very easy to discard for DIANA. Human participants, however, do not have this sort of certainty when responding to long pseudowords, given that these pseudowords share, for example, morphological characteristics with existing English words.

We argue that not just DIANA, but any model of SWR would benefit from a representation of the mental lexicon that does not consider word form (and frequency) only, especially given mounting recent findings that semantic characteristics of words are predictive of participant performance even when context is very scarce, such as in the auditory lexical decision task (Goh et al., 2016; Tucker et al., 2019). Certain models of SWR attempt to expand on the representation of the mental lexicon. The Distributed Cohort Model (Gaskell & Marslen-Wilson, 1997, 1999, 2002) and the discriminative lexicon approach (Baayen, Chuang, Shafaei-Bajestan, et al., 2019) represent units in the mental lexicon as semantic vectors. These vectors are correlated, creating a network of word meaning. Similar solutions could be implemented in DIANA in the future.

Another important aspect of simulating the auditory lexical decision task is estimating the time when the model (and therefore the human listener) should make a "pseudoword" decision. In Simulation 2, we do not estimate the time point when the "pseudoword" decision is made and we make no comparisons between general tendencies in pseudoword response latency in MALD1 data and a model estimate. Given DIANA's architecture, two different options for simulating pseudoword response latency seem plausible. First, the decision could be made as soon as the difference between free phone activation and word activation exceeds the value of θ_{lb}. This estimation regime would lead to many pseudoword decisions happening before pseudoword signal offset. However, responses to pseudowords tend to be slower than responses to word stimuli and this was true in MALD1 data as well (Tucker et al., 2019). Given that "word" decisions mostly happen after word offset (both in participant data and optimal model setups), this approach would likely lead to poor results.

For this reason we favor the second option in which pseudoword decisions are always made after signal offset, when it becomes clear that the signal will not match any word after all. At this point, an elegant solution would be to calculate the added choice reaction time using the same formula as in words, if that formula yields longer response latency estimates for pseudowords than for words. An exception to this rule could be speech signals that break the phonotactic rules of the language — *nonwords* (cf. Ziegler et al., 1997). In the case that the signal is a nonword, an early decision may be warranted, though these do not occur in MALD. One possibility would be to treat nonwords the same as pseudowords, expecting choice reaction time to be shorter because the number of plausible word competitors remaining at signal offset is minimal if not zero. Another possibility would be to have two thresholds of difference between free phone activation and word activation. The first threshold is θ_{lb} and it would still determine whether the signal is a word or not at signal offset. The second threshold would be higher than θ_{lb} and when met it would discard the signal as nonsense (i.e., a nonword) even if the signal is not yet completed. Future simulations are needed to empirically test these solutions by comparing DIANA simulations to existing participant responses to phonotactically licit (pseudoword) and illicit (nonword) speech signals not present in the mental lexicon. MALD data could not be used in this case as MALD, alongside words, only contains phonotactically licit pseudowords.

Despite many challenges that the current implementation of DIANA faces, we believe that this model may be the most promising model of spoken word recognition yet. The primary reason for this is that DIANA successfully uses the acoustic signal as input and has no binding limitations in terms of language it can be used for (as long as acoustic models exist or can be created), lexicon size that can be implemented, or behavioral tasks that it can or cannot be used to simulate. All of the very pressing issues that we discuss are best viewed as venues for model improvement, rather than be considered crippling to the model. The development of the field of spoken word recognition depends on its models being tested against various behavioral data and improved based on the findings. We argue that the primary frontier for current models of spoken word recognition is to simulate spoken word recognition phenomena using realistic conditions (e.g., realistic input and realistic competitor sets) and be adaptable enough to simulate data from a plethora of different behavioral tasks. In our view, DIANA has the potential to do both.

Chapter 4

Discriminative Lexicon

Chapter 4 was completed with help from a part of the team developing the discriminative lexicon: Elnaz Shafaei-Bajestan, Yu-Ying Chuang, and R. Harald Baayen. Additional support was received from Petar Milin.

We also thank Terrance M. Nearey, Matthew C. Kelley, and Ryan Podlubny for fruitful discussions about the discriminative lexicon.

4.1 Introduction

Spoken word recognition has been an important topic of investigation within the field of psycholinguistics for decades and numerous explanations of how this process unfolds have been offered. Some good overviews of models of spoken word recognition are given in Protopapas (1999), McQueen (2007), Scharenborg and Boves (2010), Weber and Scharenborg (2012), and Magnuson et al. (2012). Although there are notable differences between them, a common thread connecting virtually all contemporary models is the activation-competition metaphor — with the input presented to the model, various units (e.g., features, phonemes, and words) are activated and compete to be selected as the winner.

In this study, we test a new computational model called the discriminative lexicon (Baayen, Chuang, Shafaei-Bajestan, et al., 2019). Specifically, we test its ability to simulate the process of spoken word recognition in the auditory lexical decision task.

What sets the discriminative lexicon apart from other notable models of spoken word recognition is that it does not stem from the activation-competition tradition of the first-generation models such as the logogen model (Morton, 1969) or the frequency ordered bin search model (Forster & Bednall, 1976; Taft & Forster, 1975). Instead, the discriminative lexicon is grounded in a learning rule independently proposed by Rescorla and Wagner (1972) and Widrow and Hoff (1960). The discriminative lexicon was designed to use shallow and wide linear networks with no hidden layers or layers of abstract representations (such as phonemes or morphemes) to map the input directly onto meaning. The model is therefore conceptually far removed from most notable second-generation models of spoken word recognition such as TRACE (McClelland & Elman, 1986) or Shortlist (Norris, 1994; Norris & McQueen, 2008). At the same time, the discriminative lexicon, as will be clear when we describe its architecture in more detail, shares many traits with connectionist models present in the field. Crucially, the discriminative lexicon allows the user to calculate matches between the input and potential outcomes and to select the winning outcome.

At first, the Naive Discriminative Learning (NDL) approach was used to model findings from studies investigating reading. Within this approach the so-called *naive discriminative reader* model was developed and it simulated reading performance with special focus on morphological processing (see Baayen et al., 2011). However, NDL was also successfully implemented in simulating a wide range of other language and psycholinguistic phenomena (see, e.g., Baayen, 2010, 2011; Baayen et al., 2013; Baayen, Milin, & Ramscar, 2016; Baayen, Shaoul, et al., 2016; Milin et al., 2009; Ramscar et al., 2014; Tomaschek et al., 2019). These simulations include predicting responses in the visual lexical decision task (Milin, Feldman, et al., 2017). More recently, Baayen, Chuang, Shafaei-Bajestan, et al. (2019) replaced NDL with Linear Discriminative Learning (LDL) in which the outcomes (e.g., stored units of meaning; in most models of spoken word recognition, these would be words stored in the mental lexicon) do not have to be binary — with the binary indicator stating whether a particular outcome was present/predicted or not — and therefore orthogonal. In LDL, outcomes can be numbers and can be correlated, making an interconnected semantic vector space. LDL, which we describe in more detail later, is the basis of the discriminative lexicon, which in turn represents a unified model of processing and production of written and spoken language.

4.1.1 The discriminative lexicon and spoken word recognition

The foundation of using discriminative learning to represent spoken word recognition is presented in Baayen, Shaoul, et al. (2016). These initial NDL simulations primarily served as a proof of concept that triphones (or any such larger units) would perform better than a single phoneme layer, while also matching various findings from experimental studies that argue in favor of the phoneme's existence. The model was at this time still using phoneme strings to represent units of meaning, forming a list of unconnected items. Although the authors do assume that the process of spoken word recognition unfolds in time, the simulations, except for marking phoneme order within triphones, was atemporal.

However, the theoretical assumptions behind the discriminative lexicon were quite distinct to these initial simulations. In stark contrast to most (abstract) models of spoken word recognition which form a mental lexicon as a list of isolated abstract units (words), the discriminative lexicon instead stores lexical dimensions which are connected in a system of world knowledge — *lexomes*. In this regard, the discriminative lexicon is reminiscent of the semantic element of the Distributed Cohort Model (Gaskell & Marslen-Wilson, 1997, 2002) which included binary semantic features to represent the positioning of an element in a semantic space (albeit these elements were still words). The discriminative lexicon currently represents a lexome's position in its semantic space in terms of lexome co-occurrence in a language corpus. This procedure is again similar to the one employed by Gaskell and Marslen-Wilson (1999) when they performed actual computational simulations with the Distributed Cohort

Model; semantic features were abandoned because they were difficult to select and define. The key difference is that the discriminative lexicon, besides simply being larger than the attempt of Gaskell and Marslen-Wilson (1999), is built on the basis of discriminative learning, with a network learning whether certain lexomes occur and, importantly, do not occur together. The learning is performed using information from the TASA corpus (Ivens & Koslin, 1991; Landauer et al., 1998). An additional characteristic of lexomes is that they are assumed to be continually shaped and changed by the flow of experience, as the network itself would also change shape with more of the TASA sentences being presented to the learning algorithm. Finally, lexomes serve as discriminative units of meaning, so besides content and function words they can also be, for example, markings of tense or number, or anything else that creates lexicalized contrast.

As noted above, Baayen, Shaoul, et al. (2016) also posited that adequate recognition of lexomes can be obtained in a model without feature, phoneme, morpheme, etc., layers or any other intermediary abstract layers that are present in most models of spoken word recognition (cf. Luce, 1986; Luce & Pisoni, 1998; Marslen-Wilson & Tyler, 1980; Marslen-Wilson & Welsh, 1978; McClelland & Elman, 1986; Norris & McQueen, 2008; Taft & Forster, 1975). Instead, the discriminative lexicon assumes a direct connection between sound and meaning and uses a wide yet shallow network implementing a relatively simple biologically plausible learning algorithm (Rescorla & Wagner, 1972) to learn the connections between acoustic cues and lexomes. Therefore, the model has no hidden layers that act as a black box, as do certain models of spoken word recognition see, e.g., Elman, 2004; Gaskell and Marslen-Wilson, 1997; Scharenborg, 2009, and deep learning approaches in general.

Attempts following the Baayen, Shaoul, et al. (2016) study made changes in the modelling procedure that aimed at better representation of these core principles of the discriminative lexicon. Arnold et al. (2017) made the first step by using cues developed from the actual acoustic signal instead of triphones. Using acoustic input is a trait shared by only a few models of spoken word recognition, such as Fine-Tracker (Scharenborg, 2008, 2009) and DIANA (ten Bosch, Boves, & Ernestus, 2015). Arnold et al. (2017) trained a model on 20 hours of speech from 13 female speakers using the aforementioned Rescorla-Wagner learning algorithm (Rescorla & Wagner, 1972). In this modelling procedure, the cues used were newly developed *Frequency Band Summary Features* (FBSFs). FBSFs are unique identifiers (opaque to the learning algorithm) that describe a time-chunk of a word in terms of its minimum, maximum, median, first, and last pseudo-intensity value for 21 mel-frequency spectrum bands into which the sound is split. The model remained unconcerned with the incremental aspect of the spoken word recognition process, as all of these features were presented to the model at the same time. In addition, the authors conducted a behavioral experiment in which listeners responded to words sampled from the same speech corpus used for training. The results showed that the model was accurate in selecting the right lexome in 20 to 25% percent of the cases. Even though this accuracy rate may seem low, the selection was made among thousands of competitors, and the model still performed better than at least some of the human participants, whose maximum individual accuracy was just a bit above 40%. This also indicates that recognizing words isolated from spontaneous speech is not an easy task for a human listener, and it was likely made even more difficult by the fact that almost half of the words used in the experiment were function words.

In a simulation with even more challenging material, Shafaei-Bajestan and Baayen (2018) trained and tested a model on the spontaneous and noisy speech from TV broadcasts. As could be expected, model accuracy was lower than in Arnold et al. (2017) with word recognition accuracy never exceeding 14%. However, the model still outperformed the Mozilla DeepSpeech speech-to-text engine (Mozilla Organization, 2013) to which it was compared. No human responses were collected as part of that investigation.

The second major step towards better representing the principles of the discriminative lexicon was introduced in Baayen, Chuang, Shafaei-Bajestan, et al. (2019). The authors tackled the same dataset as Shafaei-Bajestan and Baayen (2018), but instead of NDL they used the newly-developed Linear Discriminative Learning (LDL). The principal change was that instead of having strings of phonemes representing the units of meaning, each unit (lexome) was represented in terms of its semantic vector space. The same Rescorla-Wagner equations were used to generate connections between a target word and other words that occurred frequently enough in the TASA corpus (Ivens & Koslin, 1991; Landauer et al., 1998) within the same sentence as the target word. At the same time, morphological characteristics were also registered as lexomes (e.g., PL for plural forms or PAST for past tense verbs). In other words, a matrix was created in which every lexome was described in terms of its learned connection to other lexomes. Therefore, instead of matching to a binary outcome, the model attempted to connect the (lack of) occurrence of a particular FBSF to a vector of a lexome's semantic space. In this setup, model accuracy for more than 130 thousand words from a spontaneous speech corpus was 33%.

The discriminative lexicon approach was implemented once more for spoken word recognition, this time to model responses to pseudowords from the Massive Auditory Lexical Decision project (Tucker et al., 2019). Chuang et al. (2020) showed that model estimates are predictive of response latency to pseudowords. Importantly, the model was augmented again to re-introduce a triphone layer as an indirect route connecting FSBFs and semantic vectors, as this was observed to improve model performance in the visual modality (Baayen, Chuang, Shafaei-Bajestan, et al., 2019). Therefore, the current setup of the part of the discriminative lexicon model dealing with spoken word recognition includes FBSFs derived from the acoustic signal as input and represents stored units of meaning in terms of a semantic vector space. Between the two are a direct connection and also an indirect route bridging input and the semantic space using smaller bundles of speech operationalized in terms of triphones.

4.1.2 The present study

One of the central drives for this study stems from the observation that although models of spoken word recognition are usually developed based on relatively rich empirical datasets, many of them have rarely been tested outside of the process in which they were created (in part due to inaccessible programming). When they have, the testing, with notable exceptions, was ordinarily performed on small datasets or even toy problems. Given that the discriminative lexicon approach seems to offer a promising yet simple way of explaining how listeners make connections between the speech sounds they hear and meanings they form, we wish to further test it on an independent dataset of audio recordings and participant responses. To that end, we use word stimuli from the Massive Auditory Lexical Decision project MALD; Tucker et al., 2019 to perform an LDL simulation and compare model estimates to actual participant performance in MALD behavioral experiments.

The MALD1 dataset includes recordings of more than 26 thousand words and responses from over 250 participants in an auditory lexical decision task, offering a large source of information both for model creation and comparison to actual participant behavior. More importantly, testing performance using MALD provides a new challenge for the discriminative lexicon approach. Previous simulations show that the discriminative lexicon approach performs on par with human participants and better than certain speech-to-text models with words sampled from spontaneous speech recordings (Arnold et al., 2017; Shafaei-Bajestan & Baayen, 2018). However, model performance was not yet compared to participant responses to carefully produced words recorded in a laboratory setting. As could be expected, participant response accuracy to MALD words far exceeds that from the Arnold et al. (2017) study, being 90% on average (Tucker et al., 2019). It is an empirical question whether an LDL-trained model can at least somewhat keep up with such an increase in accuracy.

On the other hand, there are circumstances to the specific setup we employ that may help the model achieve greater accuracy than its earlier implementations. We train the model specifically on the recordings made by the MALD speaker, so between-speaker variability is not an issue. All the recordings were made in a laboratory setting, and the speech is much more careful than the conversational speech used in previous simulations, significantly reducing the noise in the signal. Every word is only represented by a single rendition, removing within-word variability, whereas both Arnold et al. (2017) and Shafaei-Bajestan and Baayen (2018) dealt with multiple recordings of words from multiple speakers in spontaneous, noisy speech recordings. This last characteristic is also a limitation of our study — the model is trained and tested on the same material, as we do not have more than one recording for every MALD word. Our stance is that this simulation is still a good investigation of the potential the discriminative lexicon approach has: before facing the model with a more daunting task of being trained on certain material and tested on another (perhaps of mixed speakers as well), we want to see whether linear discriminative learning is at all capable of learning the connections between input and outcome in a closed set of careful speech.

There are two additional questions of interest aside from model accuracy in word recognition. First, we want to inspect whether the model isolates plausible candidates as highly "activated". In other words, disregarding whether the actual target was the winner, we still may wonder whether the candidates that were in close consideration are sensible. Second, we want to use the information gained from the model to predict participant response latency in the auditory lexical decision task, i.e., test whether or not model estimates are at least to a degree reflective of the organization of meaning in humans and the process of accessing units of meaning.

4.2 Behavioral experiment

Discriminative lexicon model estimates are compared to participant responses from the Massive Auditory Lexical Decision (MALD) project. In this study, we use the

MALD1 dataset, which is available online at <mald.artsrn.ualberta.ca> and described in detail in Tucker et al. (2019).

4.2.1 Sample

The MALD1 dataset includes responses from 231 monolingual English listeners. All participants were recruited from the University of Alberta (180 females, 51 males; age M = 20.11, SD = 2.39) and were awarded partial course credit for participation.

4.2.2 Stimuli

MALD stimuli were presented in separate experimental lists. Each list contained 800 stimuli, combining one of 67 sets of 400 words and one of 24 sets of 400 pseudowords. Each of the 67 word sets was coupled with two different pseudoword sets to create the total number of 134 experimental lists. Selection of words that can be a part of MALD was not restrictive. For example, the database included various parts of speech, even function words and interjections. The final selection includes participant responses to 26,793 words with up to nine syllables, 17 phones, and 6 morphemes. All stimuli were recorded by a 28-year-old male speaker of Western Canadian English. The recordings are available online in wave format accompanied with Praat text grids (Boersma & Weenink, 2011) as part of the MALD dataset.

4.2.3 Procedure

The experiment was built using the E-Prime experimental software (Schneider et al., 2012). The task for the participants was to listen to the stimulus presented over the headphones and make a lexical decision. They used their dominant hand to make a “word” response, and their non-dominant hand to make a “pseudoword” response. Each stimulus was preceded by a fixation cross lasting 500 ms, and the participants had up to 3000 ms to respond before the experiment would proceed to the following stimulus. The same participant could complete up to three separate experimental lists (with different word and pseudoword sets) and the total number of recorded sessions was 284.

4.3 Simulation setup

The simulations were performed in the R statistical environment (R Core Team, 2018), using packages WpmWithLdl (Baayen, Chuang, & Heitmeier, 2019), tuneR (Ligges et al., 2018), seewave (Sueur et al., 2008), and AcousticNDLCodeR (Arnold, 2018). Where possible, we followed the same procedures as the ones described in Baayen, Chuang, Shafaei-Bajestan, et al. (2019) and (Chuang et al., 2020). The code used is available as part of our supplementary material at: https://doi.org/10.7939/r3-8qvn-em70.

4.3.1 Cue and outcome matrices

Not all of the words recorded in the MALD project were frequent enough in the TASA corpus to have their semantic vectors calculated, as the minimum frequency required was 8. The final number of MALD words used in all analyses was 19,410. A total of 8,606 unique triphones were registered within these words. We generated Frequency Band Summary Features (FBSFs) developed by Arnold et al. (2017) for the selected MALD word recordings. FBSFs represent a particular recording in terms of sound intensity level broken into frequency bands. First, a word recording is split into one or more temporal chunks. The chunk borders are set at minima of the Hilbert-transformed envelope for that particular sound file. Note that a word may be split in the middle of what would otherwise be considered a single segment, such as a phone or syllable. Then, in non-overlapping 5 ms windows, power spectra are registered and mapped onto 21 mel-frequency bands. Intensity within each band is represented as a discrete variable with five levels and further simplified by registering only the initial, final, minimum, maximum, and median intensity within the given mel-frequency band. To summarize, a word recording is described using strings containing information about the principal discretized intensities within 21 mel-frequency bands for each of its temporal chunks. For example, one of the generated FBSFs for the MALD recording of the word *zucchini* was *band1-start1-median2-min1-max4-end2-part1*. This particular FBSF describes the intensities in the first mel-frequency band of the first part (chunk) of the word recording and we see that the initial intensity in that part of the audio signal is 1, median intensity is 2, minimum intensity is 1, maximum intensity is 4, and final intensity is 2. If a word has only one chunk, the number of FBSFs used to summarize its acoustic signal is 21, that is, equal to the number of mel-frequency bands. The recording of *zucchini* was split into two chunks, so it is described in 42 FBSFs.

We then created the cue (C) and the outcome (S) matrices. The C matrix serves as the input for the model and marks for every MALD recording whether it contains a particular FBSF (noted as 1) or not (0). The S matrix describes lexomes in terms of their semantic vector space — it effectively serves as the representation of stored units of meaning, i.e., what would ordinarily be referred to as a mental lexicon. We used the same semantic vector values calculated based on co-occurrence within sentences of the TASA corpus (Ivens & Koslin, 1991; Landauer et al., 1998) as the ones used in Baayen, Chuang, Shafaei-Bajestan, et al. (2019). It is important to note that semantic vectors for inflected words are created by adding the semantic vectors of word stem and affix. Since some of the considered 19,410 MALD words also have a degree of inflection, we took into account number (i.e., whether a word can be interpreted as a PLURAL form), comparison (COMPARATIVE and SUPERLATIVE), tense (CONTINUOUS, PAST, and PERFECTIVE), and whether a verb is in third person (PERSON3) when generating semantic vectors. For example, the word *abandoning* is a sum of semantic vectors for the stem *abandon* and the CONTINUOUS tense affix *ing*.

Some of the retained 19,410 MALD words could be decomposed in different ways, that is, the same MALD word can have different meanings depending on the context in which it is used. Therefore, some MALD words are represented by multiple rows in the S matrix. For example, the word *confused* is represented by three separate entries: one for the adjective *confused*, and two for the PERFECTIVE and PAST verb forms of *confuse*. The total number of rows (lexomes) in matrix S was 22,075. Since matrices C and S need to have the same number of rows, matrix C was expanded to include multiple instances of the same set of FBSFs where a certain MALD word yielded more than one lexome entry in matrix S. For example, *confused* was represented in three rows in both matrices C and S.

The columns of matrix S are lexomes as well, including word stems (e.g., *magnet*) but also characteristics that lead to a lexical contrast, even broader than the ones we used to decompose MALD words. Examples include lexical characteristics of words that mark whether they denote FUTURE, whether they are a PATIENT in the sentence, whether the number is ORDINAL and so on. Not all of the columns in matrix S are equally good at distinguishing between different rows of matrix S (i.e., the 22,075 MALD lexomes). We therefore reduced the outcome matrix to include only the 5,030 columns that had the highest variance, meaning that they change their value the most for different lexomes presented in rows (see Baayen, Chuang, Shafaei-Bajestan, et al., 2019). Each of the 22,075 MALD lexomes was described in terms of a semantic vector space operationalized as their relation to 5,030 S-matrix columns that seemed to discriminate between them the best.

4.3.2 Learning through linear mappings

With matrices C and S ready, the spoken word recognition process can be simulated using linear mappings. Linear mappings are used as a computationally convenient alternative to the discriminative learning rules — they replace the need to incrementally develop the model and provide an output matching the one that could be obtained with many learning events see Baayen, Chuang, Shafaei-Bajestan, et al., 2019; Chuang et al., 2020; Milin, Feldman, et al., 2017, for additional detail. Figure 4.1 presents the portion of the discriminative lexicon model explaining and simulating spoken word recognition. This figure is an adaptation of an existing figure depicting the entirety of the discriminative lexicon approach that also includes speech production, reading, and writing presented in Baayen, Chuang, Shafaei-Bajestan, et al. (2019). In the upper, direct route, matrices C and S are connected via a transformation matrix F. Matrix F transforms matrix C into matrix S so that $CF = S$. To that end, the generalized inverse of matrix C is created (C^{-}) so that $F = C^{-}S$. Effectively, rows of matrix F are FBSFs, while columns, as in matrix S, are the 5,030 semantic vectors from the TASA corpus. Using the obtained F matrix and a set of cues given in the C matrix, we can generate model estimates $\hat{S}$ and compare these values with the original matrix S. Note that unlike in matrix S, all instances of the same lexome (e.g., *confused*) will have the same estimated semantic vector.

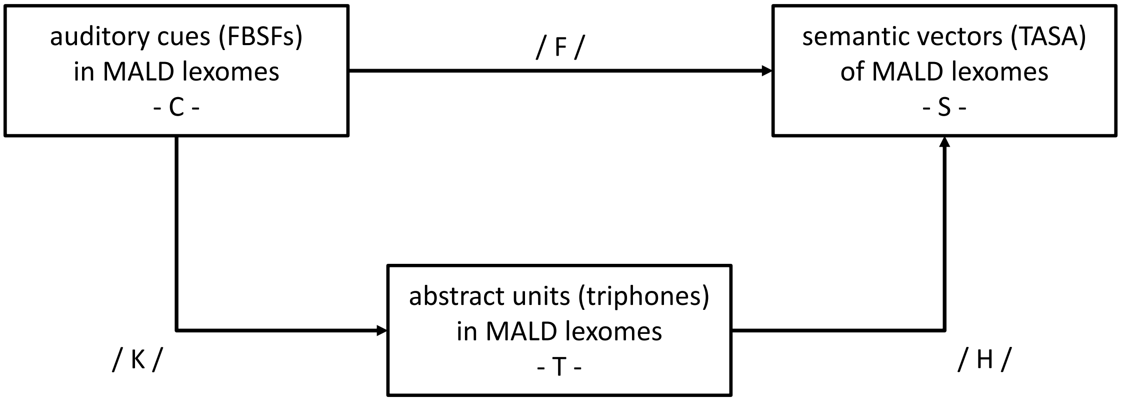


Figure 4.1: The discriminative lexicon approach to spoken word recognition. Figure adapted from Figure 11 in Baayen, Chuang, Shafaei-Bajestan, et al. (2019).

We calculated the correlation between semantic vectors from the original S matrix and the obtained values in $\hat{S}$ to examine model predictions. For each lexome, a winner is selected from matrix S for having the highest correlation with that particular lexome's row in matrix $\hat{S}$. Ideally, each row (lexome) in the estimated matrix $\hat{S}$ will correlate the highest with that same lexome's row in matrix S, leading to a perfect match and 100% accuracy. A model cannot perform perfectly, however, and we inspect potential causes for model errors as part of our analysis. The current approach also allows us to observe the top candidates for every input, again by observing the strength of correlation between the target row in matrix Ŝ and rows in matrix S. We observed the top 20 correlations between each estimated and actual semantic vector values to see whether or not plausible candidates are extracted, regardless if the correct lexome was selected as the winner or not. We compared the set of top 20 "activated" candidates to the sets of top 20 closest semantic vectors in matrix S and top 20 lexomes with the highest number of shared FBSFs for every target lexome. Note that these sets of 20 lexomes always include the target lexome as well, technically making the set contain the target lexome plus 19 other close lexomes.

In the lower, indirect route, matrices C and S are bridged by matrix T. Instead of FBSFs used in matrix C, matrix T has 8,606 triphones as its columns. The procedure is otherwise the same as in the direct route. A transformation matrix K connects matrices C and T and is calculated as $K = C^{-}T$. The rows of matrix K are FBSFs and the columns are the triphones. Using the calculated transformation matrix K, we can create an estimated matrix T̂, that is, generate predictions about which triphones are present in each MALD word signal based on FBSFs. The second step in the indirect route is governed by transformation matrix H which connects matrices T̂ and S so that $H = T^{-}S$. The rows of matrix H are the triphones, while columns (as in matrix S) are the 5,030 semantic vectors from the TASA corpus. Although this time we are predicting the semantic vectors from triphones, not FBSFs, a matrix Ŝ can be generated for the indirect route as well. Winner selection is then performed in the same manner as in the direct route — by comparing the strength of correlation between semantic vectors in matrices S and Ŝ.

4.3.3 Predicting response latency

Besides observing model accuracy in the two routes, we also wanted to predict participant response latency in MALD1 data. In this part of the analysis, we only considered correctly recognized lexomes. We used three measures derived from the correspondence between matrices Ŝ and S as indicators of how "accessible" the semantic vector was for the model. The first estimation is the strength of correlation between the lexome's semantic vector in matrices Ŝ and S. If the estimated semantic vector better matches the observed semantic vector, the time for the process to ascertain lexome identity should be shorter. We refer to this measure as CorStr, short for correlation strength. However, sometimes the semantic vectors of other lexomes can also be plausible candidates for the winner and delay the decision due to higher uncertainty. The second measure takes this into account by calculating the difference in correlation strength between the winning lexome and the second best candidate. We call this measure CorDiff, short for correlation difference. The last, third measure aims to take into account more than just the second most plausible candidate. We calculated Shannon entropy of the 20 highest correlations for every lexome as a representation of "competitor" density. We refer to this estimate as CorEnt, standing for correlation entropy.

The three measures derived from the simulation process — CorStr, CorDiff, and CorEnt — were included as predictors to participant response latency in multiple linear regression models. As the discriminative lexicon approach is currently atemporal and does not directly register frequency of occurrence, we also included MALD word recording duration and logged COCA word frequency increased by 1 (Davies, 2009) as predictors in these models. Other phonemic, morphological, and semantic predictors that have proven relevant in statistical analyses of auditory lexical decision task data were not included. We assumed that, at least in part, FBSFs capture relevant acoustic characteristics of the recordings and that the structure of the S matrix captures the morphological and semantic characteristics of lexomes.

The dependent variable in this statistical model is mean logged response latency to each of the MALD1 word recordings. Before the response latencies were averaged, a de-trending procedure was applied to account for the so-called "local effects" that impact response latency (ten Bosch et al., 2018). Local effects include getting tired and having one's mind wander. These factors impact consecutive responses to groups of stimuli in human participants, whereas the computational model always performs on the same level. The de-trending procedure described in ten Bosch et al. (2018) entails adjusting the response latency based on the response latency to a number of preceding stimuli, which is not unlike the well-established approach of taking into account the preceding response latency when analyzing response latency data. The number of preceding stimuli that are considered is set to maximize average pairwise correlations between participant response latencies to the same stimuli. In the case of MALD1 data, this number was 10. We will refer to this estimated response latency as dRT. A more detailed description of MALD1 data de-trending is given in Nenadić and Tucker (2020).

4.4 Results

The results section is split into three subsections: we first describe word chunking and generated FBSFs, then we investigate the direct route results, and finally we present the results obtained through the indirect route of the modelling procedure.

4.4.1 Chunks and FBSFs

Descriptive statistics for chunks and the relationship between the number of chunks and other measures of word length

We first explored word recording chunking, necessary for describing a word signal in terms of FBSFs. MALD words were split in up to 9 chunks. More than 98% of word recordings were split into 2, 3, or 4 chunks, or were considered to be a single chunk in their entirety (Figure 4.2a). In order to offer a comparison of the chunking process to ordinarily used measures of word length or complexity, we also correlated the number of chunks in a word with number of phonemes, number of syllables, and number of morphemes. We observed a moderate correlation between the number of chunks and number of phonemes a word has (r = .66), followed closely by the correlation with word duration ($r = .63$), and number of syllables ($r = .57$). A low correlation was observed between the number of chunks and number of morphemes in a word ($r = .27$). Despite the number of chunks and word duration in milliseconds being moderately correlated, it is visible in Figure 4.2b that words longer than 800 ms were sometimes considered to have a single chunk, while words shorter than 500 ms were sometimes split into 4 or 5 chunks. Similarly, there is no direct correspondence between the number of chunks in a word and the number of phonemes, triphones, or syllables it has.

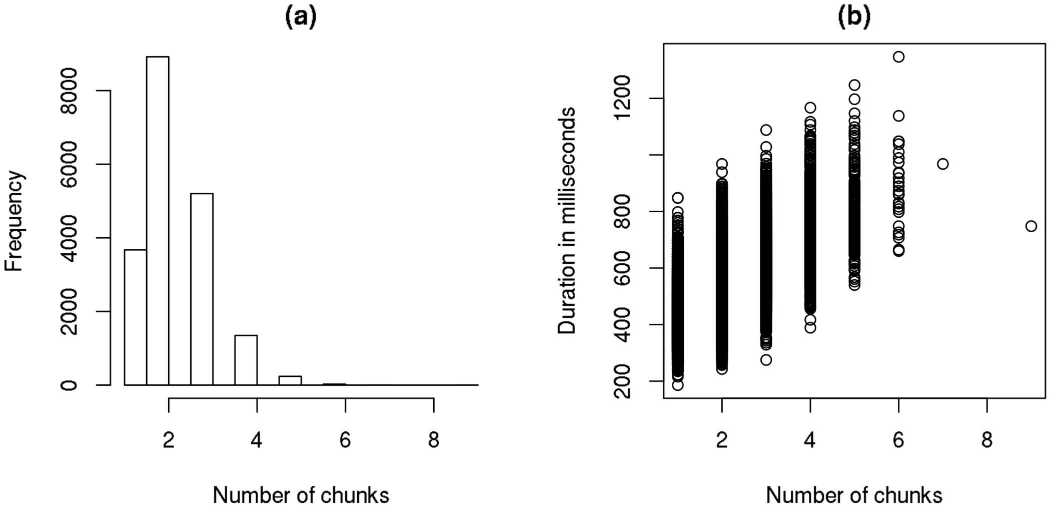


Figure 4.2: Figure (a) is the histogram of the number of chunks present in MALD word recordings. Most words are split into four or fewer chunks. Figure (b) presents a scatter-plot of how the number of chunks (x-axis) changes as MALD word recording duration in milliseconds (y-axis) changes.

Descriptive statistics for FBSFs

The distribution of the extracted FBSFs was even more skewed towards lower values than the number of chunks in MALD word recordings. A total of 26,336 unique FBSFs were registered. The most frequent FBSF was *band18-start1-median2-min1-max3-end2-part1* and it occurred in 2,303 MALD word recordings. However, most FBSFs occur rarely. The mean number of occurrences was 35 and the median was

4. As many as 7,409 FBSFs (28.13%) occurred only once (Figure 4.3a). FBSFs that occur only once are found in 4,050 MALD words. Accordingly, certain MALD words have more than one FBSF that does not occur in any other MALD word. This is likely because the chunk (part) number is embedded in the FBSF, so words with a large number of chunks have a higher chance of not sharing some of their FBSFs with any other word.

The number of FBSFs exceeds the number of MALD word recordings. Therefore, we tested whether or not this number of FBSFs would likely continue to grow if additional word recordings were to be included. We created 8 subsamples of MALD words using a stepwise procedure. We started with 1,000 MALD words and then added 3,000 new words at each step, up to a total of 19,000 MALD words. We observed how the number of unique FBSFs increases as the number of considered MALD word recordings increases. We then repeated this process 99 more times, randomly selecting new subsets of MALD words in each iteration. Figure 4.3b presents these 100 observations. It appears that the number of unique FBSFs does not reach a plateau before 19,000 MALD word recordings are considered — adding additional novel words would likely introduce new FBSFs.

The relationship between the number of shared FBSFs and word-form similarity

Since FBSFs are representations of the acoustic signal, we investigated whether similarly sounding words share more FBSFs. We arbitrarily selected groups of arguably similarly sounding words and counted the number of FBSFs they share with one another. We also noted the MALD word with which each of these target words shares the most FBSFs. Table 4.1 presents an example set of five similarly sounding words and shows how many FBSFs these words share. Additionally, the table shows the maximum number of shared FBSFs each of these five words has with any other word in the MALD lexicon. We can see that the number of shared FBSFs between simi-

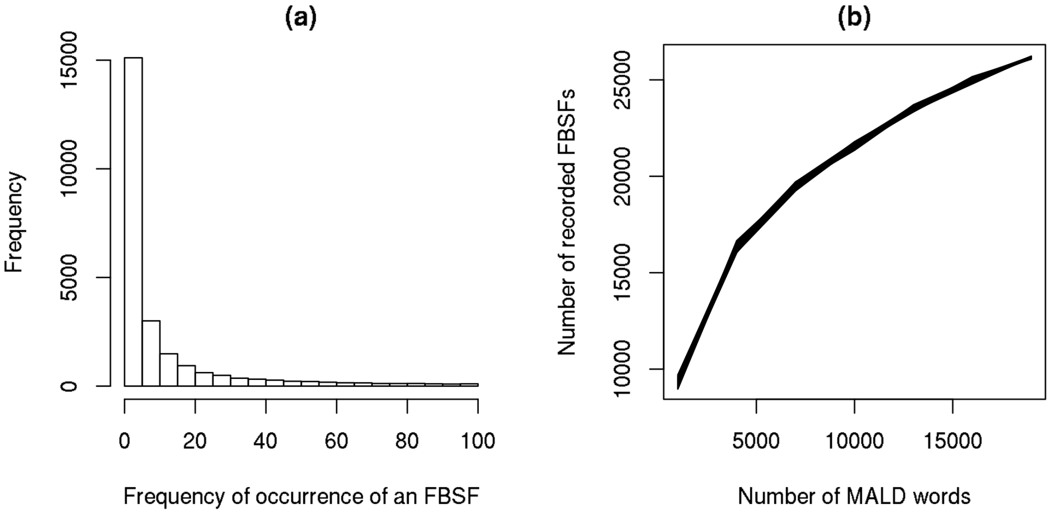


Figure 4.3: Figure (a) is a histogram of FBSFs found in MALD word recordings. Most FBSFs occur very rarely so this histogram shows only the frequency of occurrence of up to 100. A long tail continues to the maximum number of repetitions observed for any FBSF (2,303). Figure (b) shows how the number of unique FBSFs increases as the number of MALD word recordings increases. These estimates are based on 100 samplings of MALD words.

larly sounding words is rather small, especially in comparison to the number of FBSFs they have (given in the diagonal). Furthermore, these words apparently share many more FBSFs with some seemingly acoustically unrelated words. Similar trends were observed in multiple subsets of similarly sounding words.

In addition, a more detailed investigation of how FBSFs capture acoustic similarity between MALD words was performed by correlating the number of shared FBSFs with a recently developed measure of acoustic distance between MALD word recordings (for detail, see Kelley, 2018). The acoustic distance measure is a replacement for measures such as phonological neighborhood density and it takes into account fine phonetic detail rather than differences in assumed abstract sub-word units when estimating word recording similarity. Acoustic distance is also a more fitting match to the FBSFs, since FBSFs are likewise based on the acoustic signal and do not rely on a predetermined abstract unit such as the phoneme. This correlation was calculated separately for every word. As can be seen in Figure 4.4, the correlation tended to be

Table 4.1: Number of shared FBSFs for a small set of five similarly sounding words. The diagonal represents the total number of FBSFs in the word. This number divided by 21 is also the number of chunks into which the word recording is split. The bottom part of the table shows the maximum number of FBSFs each of the target words shares with any MALD word and names that word.

	comprehend	comprehensive	compression	compressor	comprise
comprehend	63	2	2	1	1
comprehensive	2	63	2	3	0
compression	2	2	84	6	0
compressor	1	3	6	63	1
comprise	1	0	0	1	42
max shared	17	14	18	11	12
word	unbecoming	unavoidable	ammunition	aggressiveness	metropolis

very weak. There is some tendency for a negative correlation to be registered, which is expected, as higher acoustic distance should mean fewer shared FBSFs. However, this correlation never exceeds $r = -.02$.

Comparing the number of shared FBSFs could only be performed between recordings of different words, as each word is represented by a single audio recording in MALD. However, we also wanted to test whether different recordings of the same word share many FBSFs. To test this, we randomly sampled five words (*deemed*, *flowering*, *tabby*, *warship*, and *presentation*) and invited the same speaker that recorded the original MALD stimuli to record four new renditions of each of these words. Word order was randomized (the randomization process had the same word be produced twice in a row only once) and an additional four words were included at the beginning and end of the list as fillers. We found that the number of chunks into which these word recordings were split was inconsistent in all words except the word *deemed*. We also found that the recordings of the same word shared but a small number of FBSFs with one another. The highest number of shared FBSFs was 19, but in the case of two renditions of the word *presentation* which were split into 4 chunks, meaning they had

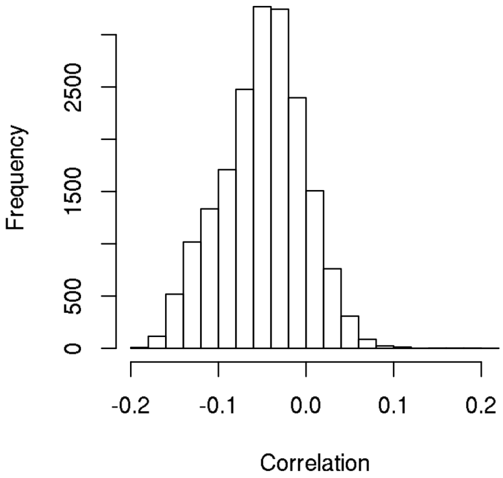


Figure 4.4: Histogram of correlations between the number of shared FBSFs and acoustic distance calculated for every word recording.

84 FBSFs each. In relative terms, the highest number of shared FBSFs was recorded for the word *tabby*, where three renditions that only had a single chunk shared 10 out of 21 FBSFs. Additionally, it was not unusual for a certain word recording to share more FBSFs with a recording of some other word in this small set than any other recording of the same word. For example, two renditions of *deemed* had more in common with one of the renditions of *tabby* than with other renditions of *deemed.*

Table 4.2: Number of shared FBSFs in four recordings of the same five words. The word is given in the first column. The number of chunks its renditions were split into is given the second column. The third column has the mean number of shared FBSFs between the four renditions of the word. The fourth column notes the maximum number of shared FBSFs between any two renditions of the word.

Word	Number of chunks	Mean shared FBSFs	Max shared FBSFs
deemed	1	6.5	9 (43%)
flowering	1 or 2	1.67	5 (24%)
tabby	1 or 2	4.67	10 (48%)
warship	2, 3, or 4	2	8 (10%)
presentation	3 or 4	11.17	19 (40%)

4.4.2 Direct route

Model accuracy and the relative contribution FBSFs have to model accuracy

Model accuracy in mapping the set of FBSFs onto the right unit (lexome) in the discriminative lexicon was 92% when the direct route was used. Not all FBSFs contribute equally to this accuracy, as can be seen by observing the values in the transformation matrix F. These values can be either positive or negative, but in either case their absolute value is an indicator of FBSF importance. We measured the maximum and the mean absolute value per FBSF in matrix F as indicators of cue importance. A scatter-plot combining these two estimates is given in Figure 4.5a. Most FBSFs that are highly important for a single column in matrix S also tend to be more important on average ($r = .83$). Importantly, however, many FBSFs arguably contribute little to model success — the gray dotted square in the bottom-left part of Figure 4.5a delineates an area in which 49% of FBSFs reside. Low-frequency FBSFs seem to carry higher weights in matrix F, as evidenced in both Figure 4.5b and Figure 4.5c. However, there are low-frequency FBSFs that also do not have substantial weight; occurring rarely seems to be only one of the requirements for an FBSF to have a large impact as a cue. The other factors are likely related to (1) whether or not such a rare FBSF is paired up with another rare FBSF in recordings in which it is found and (2) whether the word recording in which the FBSF is found has a semantic vector similar to semantic vectors of many other lexomes. High-frequency FBSFs have low values in matrix F.

Inspection of incorrectly recognized lexomes

Most of the incorrect responses (83%) happen in lexomes whose recordings where split into a single chunk. The remaining incorrect responses are registered in recordings split into two chunks (17%), with a handful of cases in lexomes represented by recordings split into three (9 cases) or four (2 cases) chunks. The model selects the correct

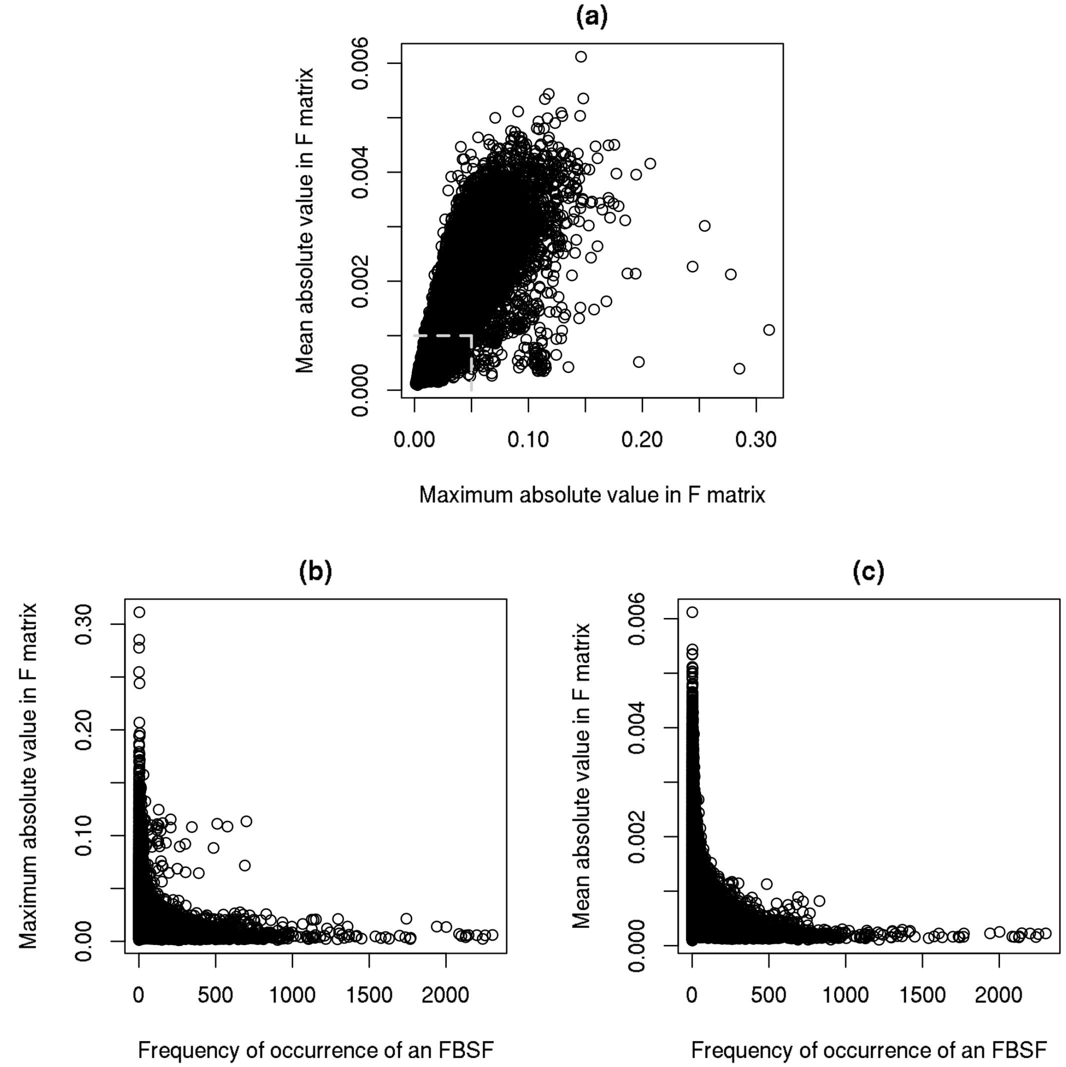


Figure 4.5: Figure (a) is a scatter-plot showing the maximum absolute value for an FBSF by its mean absolute value in matrix F. The dotted gray line delineates a square area in which 49% of FBSFs are placed. Figure (b) shows the relationship between the frequency of occurrence of an FBSF (x-axis) and its maximum absolute value in matrix F (y-axis). Figure (c) shows the relationship between the frequency of occurrence of an FBSF (x-axis) and its mean absolute value in matrix F (y-axis).

lexome as the winner in only 49% of the cases if the input for that lexome is in the form of a single chunk. Accuracy is 97% if the input for a lexome has two chunks and even higher if a recording was split into more than two chunks. Additionally, a lexome will more likely be recognized correctly if it is represented by FBSFs that are on average less frequent (Wilcoxon rank sum test with continuity correction $W = 29737476$, $p < .01$) or if the minimum frequency of its FBSFs is lower ($W = 30210026$, $p < .01$). In other words, having a large number of chunks increases the chance of having rare or even unique FBSFs (as an FBSF carries the chunk number as part of its identity), whereas low frequency FBSFs are related to higher chance of correct recognition. All words that have unique FBSFs were correctly recognized.

In comparison, having a semantic vector that is very similar to another semantic vector does not negatively impact the chances of a lexome to be correctly recognized. Correlation with the nearest semantic neighbor in matrix S was actually higher in the set of lexomes that were correctly recognized ($W = 13727174$, $p < .01$). Note, however, that the overall accuracy of the model still depends on the relatively complex interplay between, on the one hand, the distribution of FBSFs in the target lexome and other lexomes, and, on the other hand, on the semantic vectors of the target lexome and its close neighbors. This makes a straightforward prediction of which items will be difficult for the model to correctly recognize difficult. Still, the items that the computational simulation struggled with and provided the wrong answer to were also on average more difficult for the human listener. Mean accuracy for the items that were correctly recognized in the simulation was 94% in the MALD1 dataset, while the mean accuracy for the set of items that the model incorrectly recognized was 89% ($W = 15646348$, $p < .01$).

Inspection of top candidates to each target lexome

Besides selecting the correct lexome as the winner, the model should also consider plausible candidates in addition to the target. We observed which lexomes in S matrix were top candidates to be selected as the winner for every semantic vector obtained in matrix Ŝ. Top candidates are generally connected through their similar semantic vectors in matrix S, not similarity in their acoustics (i.e., shared FBSFs). For example, the semantic vector in matrix Ŝ for lexome *warship* correlated the highest with the semantic vector in matrix S for that same lexome, indicating that the model chose the correct lexome as the winner. Other semantic vectors in matrix S to which the semantic vector for *warship* in matrix Ŝ correlated highly were *warships*, *naval*, *fleet*, *troop*, and so on. These top candidates were at the same time semantically most correlated to *warship* in matrix S. The same was observed in the case of lexome *presentation* which was also correctly recognized by the model. The semantic vector for this lexome in matrix S is most closely related to semantic vectors of lexomes such as *presentations*, *critique*, *communicative*, *discussion*, and *customer*; these same semantic vectors were also most highly correlated with the semantic vector for *presentation* in matrix Ŝ.

This relationship between top candidates and lexomes with semantic vectors closest to the target lexome is not maintained in every case. Figure 4.6a shows that there are cases in which not all of the top 20 semantic neighbors are also the top 20 candidates, but also cases in which only the target word was within the top 20 candidates, although it was still selected as the winner. Figure 4.6b shows that when an incorrect response is made, most often none of the semantic neighbors are included in the top 20 candidates.

In comparison, sharing a high number of FBSFs was not sufficient for a lexome to be a close candidate to the target lexome. For example, words with which *warship* shares most of its FBSFs are *snip*, *worship*, *frantic*, and *inescapable*; *presentation* shares most FBSFs with recordings of *organizations*, *rehabilitations*, *urbanization*, and *stimulate.* None of these options were close candidates to these two target lexomes and this trend is visible in all lexomes. When correct decisions are observed, the mean number of top 20 candidates that are also within the 20 lexomes that share most FBSFs with

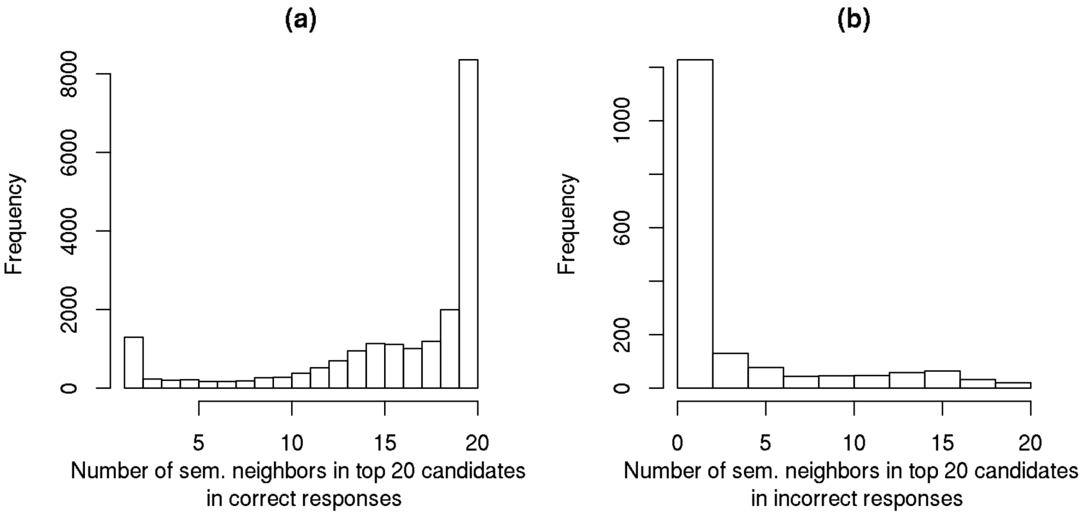


Figure 4.6: Figure (a) is a histogram representing the number of semantic neighbors (i.e., 19 lexomes in S matrix whose semantic vectors correlate the highest with the semantic vector of the target lexome plus the target lexome) that are also within the top 20 candidates for the target lexome, observing correct responses only. Figure (b) is the same histogram generated for incorrect responses. Note that the number of incorrect cases is much smaller so that the y-axes are not the same in the two figures.

the target lexome was 1.1 and the maximum was 7. Remember that if the answer is correct, one of these lexomes must be the target lexome — and in 91% of the cases it is the only one. This result does not change even in incorrect responses, where we have seen that semantic neighbors are not as prominent within the top 20 candidates, making more room for other lexomes. When the incorrect winner was selected, the mean number of close candidates that were also among the top 20 lexomes that share most FBSFs with the target lexome was 0.61, while the maximum was 4.

Predicting MALD dRT using model estimates

Finally, we used the semantic vectors obtained in matrix Ŝ to predict mean logged detrended response latency (dRT) in MALD1 data. All three multiple linear regression models showed the expected facilitatory effect of frequency and inhibitory effect of word duration. The model including CorStr showed an overall significant effect of the three predictors ($F(3, 17724) = 1654$, $p < .001$, $R^2 = .22$), but the effect of the critical predictor CorStr was not significant ($\beta = -0.01$, $p = .12$). The second model was also significant ($F(3, 17724) = 1666$, $p < .001$, $R^2 = .22$) and included a significant effect of CorDiff ($\beta = 0.04$, $p < .001$). Higher CorDiff was associated with longer response latency. The third multiple regression model was significant as well ($F(3, 17724) = 1669$, $p < .001$, $R^2 = .22$), while the effect of CorEnt was facilitatory ($\beta = -0.04$, $p < .001$). We note that the contribution of the two critical predictors that were significant is modest in comparison to the effects recorded by frequency and duration, which were approximately four and ten times larger, respectively.

4.4.3 Indirect route

Model accuracy with focus on triphone recognition and the relative contribution triphones have to model accuracy in lexome recognition

Indirect route accuracy was significantly lower in comparison to the direct route (57%). Again the group of items that was recognized correctly also had higher mean accuracy in MALD1 data ($t(17443) = -9.00$, $p < .001$). This difference was smaller in the indirect route than in the direct route; the mean MALD1 accuracy for the items that were correctly recognized in the indirect route was once more over 94%, but the mean accuracy for items that were not recognized by the model was approaching 93%.

The bulk of the reduction in recognition accuracy in the indirect route does not seem to happen at the step between the C matrix and the T matrix. The T matrix registers whether a triphone is present in a particular MALD word or not (1 or 0). Although most of the estimates obtained in $\hat{\mathrm{T}}$ were very close to 1 or 0, they are not integers. Therefore, we approximated which triphones are detected in $\hat{\mathrm{T}}$ by treating all generated estimates higher than 0.5 as 1 and all lower as 0. After applying this procedure, we noted a perfect match between estimated and observed triphones in 90% of the cases. Not all of FBSFs were contributing equally to this accuracy level — we noted identical patterns in the transformation matrix K as in the previously described transformation matrix F. To repeat, most FBSFs have small weights and high weights are by rule registered in low-frequency FBSFs. Additional figures showing this trend for matrix K are available in our supplementary materials.

At this stage, the computational model attempts to learn the connection between triphones present in the signal and the semantic vectors stored in matrix S. A total of 8,606 different triphones were recorded and their distribution, shown in Figure 4.7, is not unlike the distribution of FBSFs. The most frequent triphone /iŋ#/, where # stands for the end of the word, occurred 1,184 times. The mean number of occurrences of a triphone was 14.4 and the median, as in FBSFs, was 4. Our of 8,606 triphones recorded, 2,036 (23.66%) occurred only once, which is a slightly smaller percent than in FBSFs. Unlike in the case of FBSFs, unique triphones are found in a relatively small set of 1,667 words.

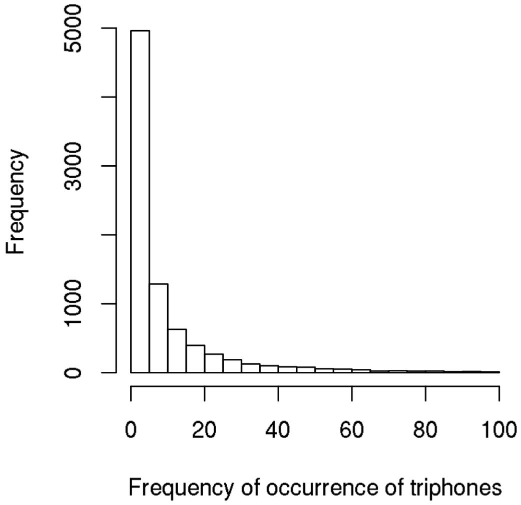


Figure 4.7: Histogram of the triphones recorded in MALD words. Most triphones occur very rarely so this histogram shows only the frequency of occurrence of up to 100. A long tail continues to the maximum number of repetitions observed for any triphone (1,814).

Transformation matrix H connects the triphone layer to the S matrix. We again assessed the contribution of each cue (in this case, triphone) by measuring its maximum and mean value in matrix H. The correlation of the two measures was lower than in matrix F, estimated at $r = .45$ (Figure 4.8a). Additionally, although the maximum weight remained high for low-frequency triphones (Figure 4.8b), the mean weight was not negligible for high-frequency triphones (Figure 4.8c). We take this result as an indication that the model cannot rely on a small number of very distinct cues alone in order to successfully generate model estimates.

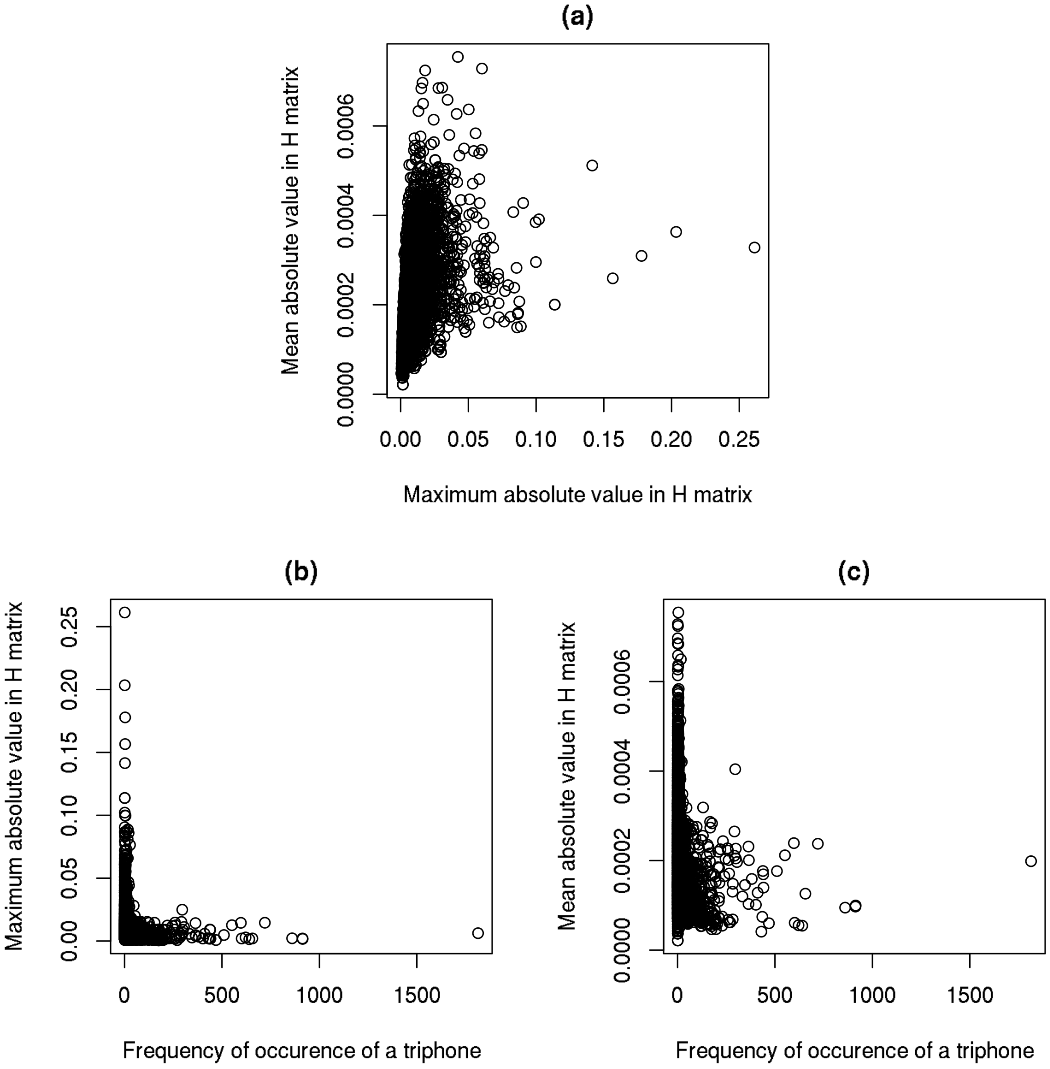


Figure 4.8: Figure (a) is a scatter-plot showing the maximum absolute value for a triphone by its mean absolute value in matrix H. Figure (b) shows the relationship between the frequency of occurrence of a triphone (x-axis) and its maximum absolute value in matrix H (y-axis). Figure (c) shows the relationship between the frequency of occurrence of a triphone (x-axis) and its mean absolute value in matrix H (y-axis).

Inspection of incorrectly recognized lexomes

For the most part, incorrect answers in the direct route remained incorrect in the indirect route as well (92%). The mistakes made in the indirect route are, unlike in the direct route, more evenly distributed across lexomes with different numbers of chunks. Error rates for lexomes with 1 to 6 chunks were 59, 36, 40, 45, 55, and 67%, respectively. The one lexome with seven chunks was recognized correctly, and the one lexome with nine chunks was recognized incorrectly. This lack of correlation between the number of chunks and recognition accuracy in the indirect route is likely a consequence of the cues used — the frequency of an FBSF largely depends on the chunk it is from, whereas the number of chunks has a much smaller impact on the frequency of the triphones present in a lexome.

In the indirect route, accuracy seems to depend on triphone frequency. Having a lower mean frequency of triphones is beneficial ($t(20817) = 8.75$, $p < .001$). The group of lexomes that was correctly recognized had a mean triphone frequency of 132, whereas the group of lexomes that were incorrectly recognized had a mean triphone frequency of 144. However, having at least one very low frequency triphone seems to be even more important ($t(14161) = 42.13$, $p < .001$). The minimum triphone frequency in the group of correctly recognized lexomes was 9, whereas it was 17 in the group of incorrectly recognized lexomes. We also recorded larger closeness of the nearest semantic neighbor for lexomes that were correctly recognized versus those that were not ($t(18145) = -44.79$, $p < .001$), just like in the direct route. The correlation between the target lexome's semantic vector and its nearest semantic neighbor's semantic vector was .71 in correctly recognized lexomes and .54 in incorrectly recognized lexomes.

Inspection of top candidates to each target lexome

Accuracy decreased in the indirect route, but even when the correct lexome is selected as the winner, the top candidates that are isolated do not always match those from the direct route. Whereas most top candidates in the direct route's correct recognitions were also the top semantic neighbors to a lexome in matrix S, we see a much less skewed distribution in the indirect route (Figure 4.9a). In many cases, not all of the top semantic neighbors are also highly activated in the simulation, meaning that the obtained matrix Ŝ does not match matrix S so well. In incorrect responses (Figure 4.9b), the results remained the same as most often none of the semantic neighbors are included in the top 20 candidates. Another result that remained the same is that top candidates were again not isolated based on their similarity with the target lexome in the cues (triphones) they share.

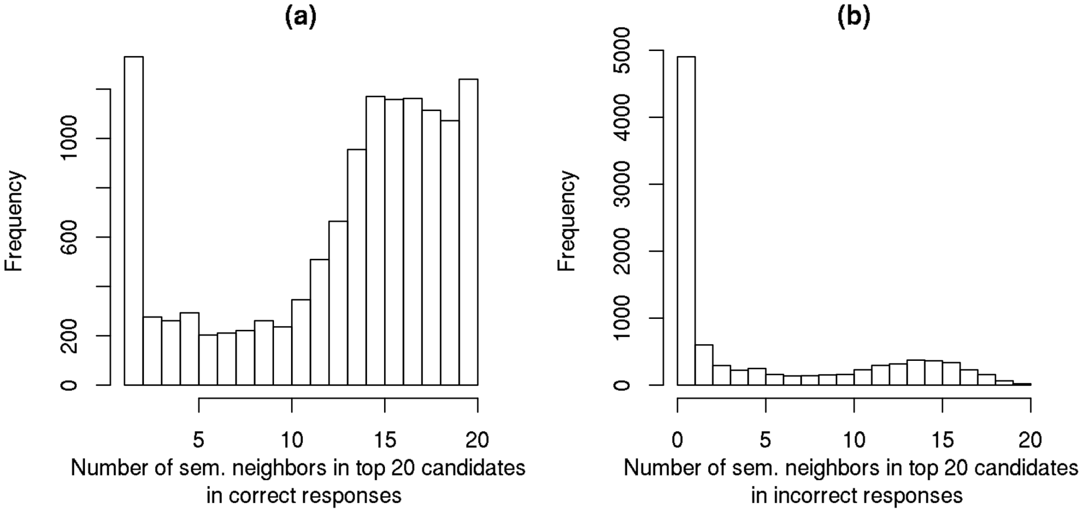


Figure 4.9: Figure (a) is a histogram representing the number of semantic neighbors (i.e., 19 lexomes in S matrix whose semantic vectors correlate the highest with the semantic vector of the target lexome plus the target lexome) that are also within the top 20 candidates for the target lexome, observing correct responses only. Figure (b) is the same histogram generated for incorrect responses. Note that the values on the y-axis are different for the two figures.

Predicting MALD dRT using model estimates

Lastly, we used CorStr, CorDiff, and CorEnt calculated from the correlations between the matrix Ŝ obtained in the indirect route and matrix S to predict dRT. Again all three multiple linear regression models showed the expected facilitatory effect of frequency and inhibitory effect of word duration, as well as more humble effects of the estimates obtained from the computational simulation. The multiple linear regression model that included CorDiff was significant ($F(3, 10489) = 969$, $p < .001$, $R^2 = .22$), with a significant effect of the critical predictor ($\beta = 0.02$, $p = .038$). CorEnt also had a significant effect ($\beta = -0.03$, $p < .001$) in its own linear regression model ($F(3, 10489) = 971.9$, $p < .001$, $R^2 = .22$). Unlike in the direct route, in the indirect route CorStr was also a significant predictor of response latency ($\beta = -0.03$, $p < .001$) in the multiple linear regression that included this predictor ($F(3, 10489) = 973.3$, $p < .001$, $R^2 = .22$).

4.5 Discussion

The goal of the present study was to implement the discriminative lexicon approach (Baayen, Chuang, Shafaei-Bajestan, et al., 2019) and to simulate participant behavior in the Massive Auditory Lexical Decision project (Tucker et al., 2019). This is the first implementation of the discriminative lexicon that compares model performance to a large set of lexical decisions made as a response to a large set of words produced in isolation (but see Chuang et al., 2020, for a simulation focusing on MALD pseudowords). The results in the direct route showed a lexome recognition accuracy that was similar to the percent of correct responses made by MALD participants in the MALD1 dataset. In other words, the computational model seemed capable of learning direct connections between a set of features describing the acoustic signal and semantic vectors of lexomes using a wide and shallow network with no hidden layers. Moreover, lexomes which were recognized incorrectly were associated with words with higher error rates in MALD1 data and measures extracted from model output were predictive of participant response latency. These results indicate that the modelling procedure partly mimics the difficulties faced by human listeners when responding to auditory stimuli. This is yet another implementation of the discriminative lexicon that shows high model performance, either in comparison to human listeners or other computational solutions.

Although we consider the simulations to be successful, a concern for the discriminative lexicon approach to simulating spoken word recognition seems to be its input representation. The model does make an important step towards using actual acoustic signal as input by creating a measure that aims to be faithful to the human hearing system and the auditory cortex (see Arnold et al., 2017; Baayen, Chuang, Shafaei-Bajestan, et al., 2019). However, the chunking process produces an outcome in which most word recordings are split into up to four chunks and many are treated as one single chunk. In turn, this means that the FBSFs often cover a very lengthy period of time. During this time, many important changes in the amplitude of a frequency band could be happening and the chunking process would engulf them (i.e, multiple segments) into a single chunk, while the only information retained would be the initial, final, median, minimum, and maximum intensity. Simply put, two very different stretches of acoustic signal recorded within a mel-frequency band of a chunk could have the same or very similar FBSF assigned to describe them.

At the same time, even a small change in the acoustic signal means that an entirely different FBSF should be assigned. Two signals that only differ in their initial (start) amplitude value by one degree would be represented by two different FBSFs. The learning algorithm does not have the information that certain FBSFs are more similar to others, as these are the smallest, indivisible, opaque units; all FBSFs are equally distant from one another. Since FBSFs capture multiple levels of multiple estimates of amplitude, the number of possible different FBSFs for any given chunk is measured in tens of thousands, making FBSFs very recording-specific and unable to relate to generalizable characteristics of the acoustic signal. The consequences of these characteristics of input representation were visible in our results: FBSFs were very numerous and therefore often of low frequency of occurrence, introduction of new recordings kept introducing new FBSFs, there was little similarity in FBSFs used to describe acoustically related words, repeated recording did not produce reliable

FBSFs for the same word, and low-frequency FBSF were instrumental in correct lexome recognition as they were rare if not unique cues.

The importance of distributional characteristics of the cues is evidenced in the current simulation as well. In the first step of the indirect route, FBSFs were connected to a set of triphones present in lexomes. Recognition accuracy at this stage was comparable to the one found in the direct route. Accuracy in lexome recognition dropped significantly in the second step of the indirect route, when a set of triphones was connected to the semantic vectors. In comparison to FBSFs, there is a significantly smaller overall number of different triphones, there are slightly fewer unique triphones, unique triphones are found in a smaller set of lexomes, and each lexome is represented by a smaller set of triphones than FBSFs. Together, this means that there is a higher chance that cues will overlap for different lexomes, making lexome recognition much more difficult. We also note that most lexomes that were recognized incorrectly in the direct route are also recognized incorrectly in the indirect route, meaning that this portion of the simulation process cannot act as a "safety net" for incorrect recognitions in the direct route.

The specific benefit of including the indirect route should be more carefully examined in future simulations. However, we see three ways in which the issues we had with implementing FBSFs can be addressed. The first is to notice that these issues may be a consequence of the particular setup employed in the current simulation. The discriminative lexicon assumes that the connections between input and meaning are learned after a large number of events. In our case, the set of events was limited as every lexome was only presented once and therefore represented by a single set of FBSFs. Our small additional analysis of five MALD words being produced four times (*tabby*, *flowering*, *presentation*, *deemed*, and *warship*) did show that there is significant variability in the FBSFs extracted for different renditions of the same word, which is somewhat concerning, but it is still an open question whether or not numerous repetitions of the same word would ultimately yield some sort of a "central tendency" in FBSFs extracted for a particular word. With a larger number of recordings, perhaps similarities could even be found between groups of similarly sounding words. This is an empirical question that could be the focus of future research. The drawback of this solution, of course, is that it requires very large datasets, limiting the number and variety of simulations that can be performed.

The second option is to adapt FBSFs. In our view, such an adaptation should aim to reduce the number of potential FBSFs, as that should increase the possibility of overlap between similarly sounding words or different recordings of the same word. In turn, this would make the representation of the acoustic signal more generalizable and less recording-specific. Another potential step would be to make the duration of a single chunk smaller, as a lot of important information may be lost with an overly long chunk. A special concern are word recordings that are represented by a single chunk. The signal could still be split into mel-frequency bands, to match the assumptions of how acoustic signal is analyzed in the cochlea. For example, the chunking procedure could be replaced by relatively lengthy moving windows with moderate overlap, and a single amplitude value could be recorded for that stretch. The new string would contain information about the window, the mel-frequency band, and the amplitude recorded (e.g., *window1-band1-amp3*). If 30 ms windows with 15 ms overlap are chosen, a 600 ms signal would have 40 windows with 21 bands and 5 amplitude levels, making for a total of only 3,000 different strings. Whatever the adaptation, it would have to be empirically tested to see whether it actually outperforms FBSFs while also being faithful to the theory at the core of the discriminative lexicon.

The third option is to replace FBSFs with another, more generalizable representation of the acoustic signal. Both DIANA (ten Bosch, Boves, & Ernestus, 2015) and Fine-Tracker (Scharenborg, 2009; Scharenborg & Boves, 2010) have had some success using mel-frequency cepstral coefficients, in effect outsourcing the analysis of the acoustic input to an automatic speech recognition system. A similar matrix to matrix C could be devised in which FBSFs are replaced with units extracted from the automatic speech recognition software, where a 1 would be assigned to a certain number of most highly activated units. These units would not have to correspond to phonemes, as they currently do in DIANA. (These units could correspond to triphones, however, replacing the first part of the indirect route.) Instead, the recognition process could be designed to create its own set of units based on the acoustic data. Although there are a number of decisions a user can make to control the creation of such acoustic models, the obvious downside to this approach is that it largely relinquishes control over the process of analyzing the acoustic signal, potentially making this aspect of the model incomparable to the process occurring in human listeners. The goal behind creation of FBSFs was exactly opposite.

Arguably the most significant contribution of the discriminative lexicon approach to the field of spoken word recognition is the representation of the discriminative lexicon itself. Most models of spoken word recognition end their process at what is usually referred to as "lexical access", that is, recognition of the input in terms of units of form stored in the mental lexicon see, e.g., Gaskell and Marslen-Wilson, 2002; Weber and Scharenborg, 2012. These units are usually not connected in any way except in the units of form that they share, making the mental lexicon a list of unconnected strings. However, the goal of listening to words is understanding their meaning and form recognition is often impossible to tease apart from the meaning it carries. Furthermore, the mental lexicon should be organized in a manner that testifies to the history of its use, making it adapted to the requirements and outcomes of previous access. Recent behavioral studies also show that semantic richness effects can be captured even in isolated word recognition using the auditory lexical decision task (Goh et al., 2016; Tucker et al., 2019).

The current simulation is yet another argument in favor of including a semantic element to models of spoken word recognition. Measures derived from "activation" estimates, that is, correlations between rows of the obtained matrix $\hat{\mathrm{S}}$ and matrix S, were predictive of response latency. Interestingly, these results do not follow the usual notion that more competition and closer competition means longer decision time (although this notion is preserved in the finding that lexomes represented by less frequent cues, i.e., that compete with fewer other lexomes, are recognized correctly more often). We found that the smaller the difference in activation between the top candidate and the runner-up, the shorter the response latency. Similarly, shorter response latency was associated with higher entropy in the activations of the top 20 candidates. We interpret these results by noting that top candidates in correctly recognized lexomes are usually the top semantic neighbors of the target lexome. As we stated before, this is unsurprising as the goal of the simulation is to replicate the semantic vectors stored in matrix S. In incorrectly recognized lexomes, the candidates are usually not semantic neighbors. Therefore, the final output of the model is closer to a spread of semantic activations given the correctly recognized lexome, rather than a set of close competitors. Participant response latency is then predicted primarily by the characteristics of the semantic neighborhood of a lexome (with higher semantic neighborhood density being beneficial), rather than competition between formal characteristics of lexomes.

There are two additional important aspects of the auditory lexical decision task that the current architecture of the discriminative lexicon does not support. First, the auditory lexical decision task presents participants with both words and pseudowords. We do not present simulations investigating how the discriminative lexicon predicts response latency to pseudowords, as these simulations were the focus of extensive and detailed simulation performed by Chuang et al. (2020). However, one of the crucial elements of the lexical decision task is the very lexical decision, that is, the listener's decision whether the input signal represents a lexical item (word, lexome) in the given language or not. The discriminative lexicon currently does not include a decision component that would categorize the input as fitting a certain existing lexome or not, the way DIANA does for words (ten Bosch, Boves, & Ernestus, 2015). We also note that the discriminative lexicon is not an exception in this regard, as most models of spoken word recognition do not have an explicit decision rule for making a lexical decision. For example, TRACE does not have a built-in moment of recognition (Strauss et al., 2007), so Hannagan et al. (2013) had to devise their own solution that seemed appropriate for estimating the time cycle in which the model recognizes the signal as a certain word. The decision component is necessary to fully simulate participant behavior in the auditory lexical decision task, while models of spoken word recognition need to be able to simulate a variety of different experimental tasks in order to properly assess their performance and make them comparable to one another.

Second, the current implementation of the discriminative lexicon approach to modelling spoken word recognition is atemporal. This directly relates to our previous point as it means that the model currently cannot simulate some of the phenomena that have been of central interest to the field of spoken word recognition and litmus tests for model performance. Notable examples include subphonemic effects captured by using sound splicing (Marslen-Wilson Warren 1994), mismatch effects see, e.g., Marslen-Wilson and Zwitserlood, 1989; McQueen, 2007, and the time-course of spoken word recognition as investigated using the visual world paradigm e.g., Allopenna et al., 1998; Dahan, Magnuson, Tanenhaus, and Hogan, 2001. One way in which the signal can be incrementally presented to the model is by using temporal chunks into which the word is split. Instead of having FBSFs from all chunks of a word in a single row of matrix C, we could represent each word in multiple rows. FBSFs of one additional chunk could be added with each new row, with semantic vectors effectively being estimated after each chunk. This approach, however, does not offer much detail. We have seen that most MALD word recordings were split into four chunks or fewer, while many were parsed into a single chunk in their entirety. Additionally, there is a risk that additional repetition of FBSFs belonging to the early chunks (especially the first one) would further reduce model accuracy. Therefore, an alternative to the current input may be necessary. We have already discussed the topic of the input representation in the discriminative lexicon approach and there is, at this point, no one clear direction in which it may develop. It seems that including some option to make the input incrementally presented to the model would be useful in expanding the set of experimental tasks that the model can simulate.

Finally, it is only fair to once more point out that the computational model used in the current study was trained and tested on the same set of data. Although the results are promising and some important insight into how the model could be improved were gained, additional simulations that would use an expanded set of recordings (part of which can be used for training and part for testing) are necessary. Future iterations of the MALD project will include recordings of two additional speakers besides the speaker whose recordings were used in the MALD1 sessions. It would be interesting to see the extent of speaker variability as captured by FBSFs (or whatever other input representation) and whether the model is capable or learning from recordings of one speaker and perform well when tested using the recordings of another speaker. The Auditory English Lexicon Project (Goh et al., 2020) may be another good choice, as it includes recordings of six speakers from three dialects of English producing isolated words. The learning models (or transformation matrices) could also be trained on spontaneous speech corpora and then applied to one of these large auditory lexical decision task databases, even though the circumstances in which these recordings were made would not be the same. Whatever the direction of future investigation, simulations are necessary for any further model development and, at the same time, development of the theory explaining the process of spoken word recognition. This holds not only for the discriminative lexicon approach, but also for other models of spoken word recognition, very few of which have been tested in their ability to simulate human performance in the auditory lexical decision task on larger sets of data.

Chapter 5

General Discussion and Conclusion

The bulk of the discussion of simulation results is found in the separate discussions present in each central Chapter of the present dissertation. In this final Chapter, I first give an overview of the results of the three separate studies and provide a comparison of model performance. I then focus on what the findings in these studies, taken together, bring to our understanding of the process of spoken word recognition as captured by the auditory lexical decision task and how this process of SWR is implemented in the tested models. The final portion of this Chapter is reserved for a sketch for a new hybrid model of spoken word recognition, combining features from two models of spoken word recognition.

5.1 Overview of simulation results

The first goal of the dissertation was to perform computational simulations of the auditory lexical decision task and compare model adequacy in matching the process as it unfolds in the human listener. The three models we employed do not seem to perform very well when simulating the auditory lexical decision task and I will summarize the issues noted when implementing them. TRACE's (McClelland & Elman, 1986) reimplementations jTRACE (Strauss et al., 2007) and TISK (Hannagan et al., 2013; You & Magnuson, 2018) arguably performed the worst. The current setup of jTRACE does not allow for a proper representation of certain phonemes of English, specifically affricates and the r-colored vowel, as phonemes that have dynamic, changing characteristics. Our solution that included merging two shorter versions of existing phonemes (previously used by Mayor & Plunkett, 2014) lead to a very significant decrease in phoneme and hence word recognition accuracy in jTRACE. I argue that there is still a possibility to define these phonemes as "steady-state" (similarly to how TRACE does not represent diphthongs as vowels with dynamic formants). For example, the diffuseness and the burst pseudofeatures could have high values at the same time in an affricate. Although suboptimal, this solution might yield higher recognition accuracy and allow jTRACE to cross the hurdle of having low phoneme and word recognition accuracy when all phonemes of English are included.

However, what lays on the other side of that hurdle could likely be bleak. Ultimately, TISK and jTRACE perform quite similarly (Hannagan et al., 2013) and TISK sidesteps the issue of phoneme recognition altogether. Still, we have seen that TISK recognition accuracy plummeted with any form of close competition. Furthermore, in correctly recognized words, estimates of response latency were not predictive of human response latency, or at least not any more predictive than using a simple measure of the number of phonemes in a word. The decision criteria had an additional flaw: applying them on pseudoword input (i.e., on a phoneme string not present in the lexicon) did not disqualify the input as a pseudoword, instead generating a large number of false positives. The decision criteria we used were created to perform model-to-model (jTRACE to TISK) comparisons, and, as far as I know, were never used to make model-to-human comparisons.

DIANA (ten Bosch, Boves, & Ernestus, 2015) is a model that was developed with auditory lexical decision and word repetition tasks in mind. The model was also previously tested on human data from those experimental tasks. However, the simulations we performed had their issues. One difficulty with applying DIANA is that it seems unlikely that a researcher could use it "off the shelf", as jTRACE and TISK, or even the discriminative lexicon. DIANA requires acoustic models that are adapted to the speaker, that is, approximately at least forty minutes of transcribed recording to adapt an existing acoustic model. Additionally, the simulation process we used requires an at least somewhat skilled user of the Hidden Markov Toolkit (Young et al., 2006). This obstacle is not insurmountable, and we also offer our own acoustic models that have been adapted to the MALD speaker to those researchers that would use MALD stimuli in their experiments as well.

A more important concern is that DIANA's current setup showed best performance when the decision component of the model was effectively neutralized. The highest correlation between model estimates and participant response latency estimates was obtained when the model estimate was essentially equal to the duration of the word recording. In a way, this result is reminiscent of the TISK simulation result — whereas the contribution of TISK to predicting human response latency could be reduced to the number of phonemes in a word, the contribution of DIANA could be reduced to recording length. This result shows that the decision process taking place as the signal unfolds and, importantly, after the signal has ended needs to be changed in future simulations using DIANA.

The discriminative lexicon approach (Baayen, Chuang, Shafaei-Bajestan, et al., 2019), with its current architecture, does not estimate response latency. Instead, we used measures obtained from model output that relate to candidate density (or closeness) as predictors of human response latency. Of course, a procedure could be devised similar to the one used in DIANA: the model could assume that all responses are made after signal offset and then add an estimate of decision time past signal offset to the duration of the signal. The duration of the signal would need to be stored outside of the cue matrix. However, I am uncertain how the decision time should be estimated in the discriminative lexicon approach. Candidate set density, that is, the closeness of semantic neighbors, had a negative correlation with human response latency, as more competitors meant shorter response latency. In DIANA and in most approaches to modelling SWR more candidates mean more time needed to select the correct winner due to higher competition or uncertainty. Adding negative time to signal duration to account for the opposite trend noted in the discriminative lexicon would not yield realistic estimates of response latency, making them shorter than the duration of the signal. This question of the missing decision component is closely related to the characteristic of the model being atemporal and should be addressed in future model development.

The direction of the correlation between response latency and the closeness of candidates is opposite to the one expected at least partly because candidate "activation" in the discriminative lexicon reflects the structure of the outcome matrix (the lexicon itself). To repeat, TISK yields response latency estimates dependent on the number of phonemes, DIANA on signal duration, while the discriminative lexicon's estimates seem to relate to what we may call semantic neighborhood density (i.e., similarity in semantic vectors between the target lexome and close candidates in the S matrix). At the same time, the candidates extracted in the discriminative lexicon simulation do not seem to reflect similarities in input (the acoustic signal). It seems to me that, in a successful simulation, the outcome matrix aiming to match the S matrix should never inherit similarities in the acoustic characteristics present in the C matrix, as that is not the goal of the simulation process. Therefore, I am uncertain what kind of input could be used in the discriminative lexicon approach that would retain acoustic similarity in simulation outcomes. Our analysis of the characteristics of frequency-band summary features (Arnold et al., 2017) also showed that they are very recording-specific, which also means that they are not similar for similar sounds. The first step in further development of the discriminative lexicon's approach to simulating SWR should still be making significant changes in the input representation.

With these limitations in mind, I turn to the successes of the present series of simulations. In my opinion, of the three models used, DIANA seems best fit to simulate the process of spoken word recognition as captured by the auditory lexical decision task. The model is fairly accurate in selecting the correct word from thousands of competitors using actual acoustic signal as input even with in-house acoustic models like the ones we created, which may be unprecedented in models of spoken word recognition. The model is also relatively successful in distinguishing between words and pseudowords. The correlation between model estimates and human responses is the highest in DIANA, even if only because the model "listens" to the same acoustic signal as the human listener and does not lose count of its duration. Crucially, the model is built so that it can explicitly perform the three required tasks — it defines how it selects whether the input signal is a word or not, which word it is, and when the decision is made. The decision element of the model is necessary for calculating model estimates and I do not see it as an inherent part of other models of SWR; it is certainly not a core element of TRACE or the discriminative lexicon. In the case of DIANA, the decision process is assumed to be different for the auditory lexical decision task versus the word repetition task (which we did not investigate in the present dissertation), and it can be adapted to fit the task at hand. Although additional work may be needed to define the decision process or at least adjust parameter values, as we have seen that the current setup does not fit MALD data well, DIANA is an important step forward and a good starting point for future simulations.

The discriminative lexicon approach also showed high recognition accuracy, even higher than DIANA, and in line with human accuracy. The current setup does so without necessitating any sort of prelexical abstract units, given that the indirect triphone route seemed to perform worse than the direct route. Furthermore, estimates derived from the simulations were predictive of human response latency in the MALD1 dataset — estimates that seem to mostly be related to distributional semantics of lexomes. As we have stated in Chapter 4, our simulations are yet another confirmation that the discriminative lexicon approach merits further testing and development. Importantly, this model puts under question the somewhat ingrained ideas about the necessity of prelexical abstract units and the mental lexicon being represented as a list of unconnected items.

Lastly, there does not seem to be much favorable to be said about the mostly unsuccessful simulations conducted using jTRACE and TISK. However, I wish to stress one important thing about using jTRACE and especially TISK to simulate the auditory lexical decision task: of the three models employed in the present dissertation (but also certain other models that I have investigated), jTRACE and TISK have proven to be the easiest to set up and use. I will say more about this towards the end of the following section.

5.2 Implications for the theory of spoken word recognition

The second major goal of the dissertation was to learn more about the process of spoken word recognition by simulating this process as captured by the auditory lexical decision task using a set of computational models. This goal was only partly met. The present simulations scarcely provide researchers in the field with novel important aspects to be aware of in the process of spoken word recognition; rather, they point to how the known important aspects of the process are technically implemented (or omitted) in the tested models. In other words, it seems that the majority of issues noted in the present series of computational simulations were not related to the theory behind the employed models of SWR, but to the technical implementations (or lack thereof) of crucial elements in any model of SWR.

In this section, I return to these important elements — input representation, abstract units, competition, and the representation of the mental lexicon (i.e., storage) — to discuss how they are implemented in the tested models. Further, I discuss how these elements should be treated in models of SWR, basing my arguments with regards to the implementations in the tested models and their performance. Still, the implications to our understanding of the process of spoken word recognition and to how we operationalize this process in computational simulations that I present in this section should be taken with some reservation for two reasons. First, these conclusions are based on simulations of a single experimental task (the auditory lexical decision task) and they may not extend equally to all other tasks or situations that require the human listener to recognize spoken words. Second, the conclusions drawn here are partly speculative and require further testing because of the various technical limitations or other hindrances we faced using the three models of spoken word recognition.

Models of SWR need to abandon pseudo-acoustic input, as it does not properly represent the variability present in the speech signal. At the same time, we cannot assume that the process of analyzing the acoustic signal is trivial or somehow already solved, as is done in TISK. Using higher level abstract prelexical units that are assumed to be already recognized is not fitting because in that case the model does not prove that those units can be extracted without error from the acoustic signal, nor that those abstract units could not be replaced by some other, better abstract representation of the acoustic signal. Putting input representation under careful scrutiny through new challenges may reveal that it is not fitting. We have seen in the case of jTRACE that expanding the phoneme set impeded recognition even though pseudo-acoustic features were used. We have also shown that not all representations of the acoustic signal do it justice by noting issues with frequency-band summary features (Arnold et al., 2017).

The need to use better input representation (i.e., one based on the acoustic signal) has been acknowledged for more than thirty years at the point when the present dissertation is written. What changed is that we now have models that make valid attempts at deriving input from the acoustic signal, like DIANA, the discriminative lexicon, or Fine-Tracker (Scharenborg, 2008, 2009; Scharenborg & Merkx, 2018). Both the successes and the failures of such attempts provide important information about how the acoustic signal should be treated to derive model input from it. Given that DIANA and the discriminative lexicon both had to offer more than instantiations of

TRACE in the present dissertation, I see no reason why a contemporary model of spoken word recognition would not be expected to provide a solution that is based on analysis of the acoustic signal.

Another important question in the field of SWR is the number and nature of abstract sublexical units. The three models we employed all use progressively fewer layers of abstract sublexical units and larger units. jTRACE uses pseudofeatures from which it derives phonemes, TISK uses already recognized phonemes as input, DIANA calculates phoneme activation directly from the speech signal, and the discriminative lexicon has a direct route that connects frequency-band summary features (that is, acoustics) directly to units of meaning (lexomes) without any prelexical layers. Even in the indirect route, the discriminative lexicon uses a larger abstract unit, triphones. Our results indicate that multiple abstract layers may not be needed, as both DIANA and the discriminative lexicon approach perform fairly well without them, and at the same time better than jTRACE. Furthermore, DIANA's accuracy would likely be higher if triphones were used instead of the flat monophone acoustic models we created (see Young et al., 2006).

However, it is difficult to make claims about which approach is better. Other model characteristics between TRACE instantiations, DIANA, and the discriminative lexicon were also different. In order to test what kind of abstract prelexical units are needed (if any), other model characteristics need to be kept constant, or at least systematically varied to cover major possible combinations/interactions of model characteristics. DIANA could be a good model to test such assumptions on, as it could replace phonemes with triphones, syllables, or any other abstract prelexical unit. DIANA could also use an option to have the automatic speech recognition process extract its own categories based on the speech signal, rather than extract imposed man-made sublexical units. Regardless whether DIANA is used as the framework model to test these assumptions, future studies and simulations are necessary to investigate whether larger, chunked abstract sublexical units should be favored over multiple layers of sublexical units increasing in size and whether the answer to this question is task-dependent.

The question of how competitors are selected and removed from contention as the process of spoken word recognition unfolds was one of the central points of discussion in second-generation models of SWR. This issue was solved differently in TRACE, (Marslen-Wilson & Tyler, 1980; Marslen-Wilson & Welsh, 1978), Neighborhood Activation Model (NAM; Luce, 1986; Luce & Pisoni, 1998), and Shortlist A and B (Norris, 1994; Norris & McQueen, 2008). Based on the simulation results in DIANA, it seems that it is better not to entirely remove candidates at the first point when they show mismatch with the input. In the initial simulations where we adapted acoustic models we saw close competitors that had initial mismatch with the target word. As DIANA follows a Cohort-like competition, such candidates were excluded from shorter lexicons created for every target word in later DIANA simulations; we already discussed in Chapter 3 that this may have had some impact on the simulation results. Therefore, it seems that the candidate retention process should be rather TRACE-like than Cohort-like, as rhymes (in TRACE terms) to the target word are often retained as closest competitors in DIANA if they are not previously excluded.

Although likely an improvement, even the TRACE approach determines similarity between candidates based on similarities in the abstract units they contain. It would be better yet to estimate candidate similarity based on acoustic signal similarity (Kelley, 2018), as it is imaginable to have two words that do not share their first nor last phonemes while still being acoustically similar. For example, the difference between /pɑp/ and /bɑb/ is only in the voicing of the bilabial stops. Neither TRACE nor NAM would consider these two words to be close candidates to one another, but renditions of these words may be acoustically quite similar (or treated as a viable candidate by the human listener). With high variability present in different renditions of the same word, it is also possible to have a certain recording be more ambiguous than another recording of the same word. Preselections made based on similarities of abstract units between words do not take into account this variability.

Furthermore, as discussed in Chapter 3, the issue of candidate retention or preselection may be a thing of the past, present only because of the technical limitations that disallow the entire lexicon participating in the activation-competition process. DIANA and the discriminative lexicon approach (and possibly TISK) have the potential to include the entire lexicon in every simulation and estimate candidate activation based on similarities between the acoustic input and, currently, phonemes present in words stored in the mental lexicon. The optimal approach seems to be for the model to define how the activation is calculated and have candidate plausibility or prominence be estimated as a direct consequence of that decision, without necessitating candidate preselection.

The representation of the mental lexicon is quite reduced in most models of SWR. The units in the mental lexicon are most often unconnected strings of phonemes. Our simulation using the discriminative lexicon and behavioral studies (e.g., Goh et al., 2016) clearly show that there is something to the organization of the mental lexicon or the characteristics of its units that plays a role in the process of SWR even when the task is to recognize isolated items, as it is in the auditory lexical decision task. These simulation and experimental results seem to counter an approach that is entirely feed-forward (see, e.g., Norris et al., 2000a) and that entirely disregards the structure of the mental lexicon and how that structure shapes accessing units stored within. As mentioned in Chapter 4, it is of course clear that with a perfectly unambiguous signal and a model that always performs on the same level there would be no need for any other information for correct recognition. However, it is also clear that the goal of cognitive simulation is not to solve the spoken word recognition problem, but to solve it the way a human does. Humans listen for meaning and have adapted to do so in an often noisy environment. We necessarily generate predictions based on our previous experiences, as that facilitates signal disambiguation and allows a fast reaction when needed — depending on the (supposed) content of the message. This characteristic of human behavior is now more readily being recognized as an important element of spoken word recognition (see, e.g., Norris et al., 2016), but how we could simulate this prediction process and create an interwoven, meaning-rich representation of the mental lexicon remains an interesting and open question. I think that the semantic vector approach somewhat shared between the Distributed Cohort Model (Gaskell & Marslen-Wilson, 1997) and the discriminative lexicon is a good place to start, although this solution would likely not be able to capture effects that are difficult to extract from distributional characteristics of words, such as the recorded effect word valence has on participant response latency in the auditory lexical decision task (see Goh et al., 2016; Tucker et al., 2019).

As I mentioned previously, it is crucial to note that models of spoken word recognition can only simulate the process of spoken word recognition as captured by a certain experimental task. Humans adapt their strategies when performing different tasks, and the model needs to be able to match them. DIANA explicitly defines the decision process and makes it different for the auditory lexical decision versus the word repetition task (the latter was not investigated in the present dissertation). The simulations of the process of spoken word recognition should likely be different if the task for the participant is to, for example, make a lexical decision, perform a phoneme monitoring task (see, e.g., Connine & Titone, 1996), or participate in the visual world paradigm experiment (see, e.g., Allopenna et al., 1998). A model of spoken word recognition needs to be adaptable and applicable to many different experimental tasks. If a model can only simulate a smaller subset of experimental tasks that does not overlap with the simulations performed with another model, the two models would be very difficult to compare — a model cannot be considered an improvement in comparison to another model if they simply model a different set of experimental tasks or findings (Coltheart et al., 2010).

There is one more point that I want to make although it does not concern any of the important elements of models of SWR (input, abstract layers, competition, and the representation of the mental lexicon) that I discussed in the present section: The series of simulations performed also made clear that the models of spoken word recognition are very inaccessible. Although this issue is not directly related to our conception of spoken word recognition and the theories, computational models, and hypotheses we form, it is without a doubt having a direct impact on the research we perform and disseminate. It seemed to me that implementing DIANA and the discriminative lexicon without aid from their creators would be close to impossible, as no freely available code (before what we offer in this dissertation) was available and previous publications (understandably) do not offer all the details. Additionally, many decisions that needed to be made in the simulation process are not explicitly codified and are in the hands of the researcher performing the simulation. In those cases the opinions and advice from model creators were very important. Model flexibility expressed through many simulation choices is, of course, necessary, but so are guidelines, manuals, tutorials, and sample code.

The lack of manuals or guidelines is not unexpected in DIANA and the discriminative lexicon approach, as both models are recent developments that are yet to receive additional publications and user support. However, there are numerous models of SWR that have been prominent for years, if not decades, that are still quite inaccessible. The only models I consider fairly well-documented are jTRACE (with all the faults the scripting environment and the GUI have) and especially TISK. Model accessibility is, in my opinion, one of the biggest contributions jTRACE and TISK offer to the field of computational modelling of spoken word recognition. These two models show that models of SWR can and should be more than dense under-commented code available per request with a warning that some things may not work properly.

The development of scientific fields depends on accessible knowledge and replicable procedures. Experimental psychology (and psycholinguistics) was made easy and widespread with experiment-building software. Even undergraduate students in some of their higher-level courses can be expected to create their own experiments. A study that cannot show what stimuli and experimental procedures it used (without a good reason) would likely be considered unacceptable for journal publication. Psychology of individual differences as a paradigm of personality psychology partly owes its dominance to computers and software that can easily analyze patterns in hundreds of thousands of questionnaire responses in minutes or hours. Factor analyses were once performed by the select few researchers that had the skills and the computational power (and the time) to do so, yet now they are a part of the standard basic training in statistics in social sciences. The field would not have developed as much if applying its procedures took weeks or months. The significance of software like Praat (Boersma & Weenink, 2011) for the field of phonetics can hardly be overstated, even with scripting difficulties encountered with some more advanced options. One may argue that computational models are not the same as experimental, statistical, or sound-analysis software, and that is somewhat true. In that case, I suggest we take the example of life sciences, where one can find a curated database of hundreds of mathematical and computational models that has been active for fifteen years (see Malik-Sheriff et al., 2020).

Most publications that regard models of SWR do not actually implement them, and I believe that this is largely because the time required to understand, set up, and then adapt a model of SWR to a particular research experiment or question is simply too long. This significantly impedes the process of model testing, scrutinizing, and, consequently, model development. Furthermore, this limits the number and variety of experimental tasks and datasets considered for model simulation, as the models mostly remain confined within a select few groups of researchers. In my opinion, this issue is so important that if we envision a “third generation” of models of SWR, they should not be marked by their implementation of the actual acoustic signal as input, better representation of the mental lexicon that takes into account word meaning, nor their ability to perform large-scale simulations with tens of thousands of words in the mental lexicon. Instead, they should be marked by their accessibility and applicability to a large number of different experimental tasks by different research groups which would enable their continuous and gradual testing and improvement.

5.3 GRAFT: A sketch for a hybrid model of SWR

The third goal of the dissertation was to provide suggestions for improvements and further model development. This goal was in part met in the central Chapters of the dissertation, as we discussed each of the implemented models separately. Here, I present a sketch for a hybrid model of spoken word recognition that is a direct result of the investigations performed using TRACE instantiations, DIANA, and the discriminative lexicon. The following ideas were developed together with Petar Milin and Benjamin V. Tucker.

In many cases, computational models of spoken word recognition have been developed as a response to a characteristic of a previous model that was either computationally unfeasible or unable to simulate certain findings from experimental studies (see, e.g., Magnuson et al., 2012; McQueen, 2007; Protopapas, 1999; Weber & Scharenborg, 2012). At the same time, the newly developed model was often radically different to the predecessor it criticised, making direct comparison between models difficult; a difficulty that affected the current dissertation as well. We argue for a different approach called the nested incremental approach, succinctly discussed by Coltheart et al. (2010). Instead of developing a new model altogether, we would attempt to create a combination of two existing models of SWR. When constructing this grafted model (hence, GRAFT), our focus is placed on getting the most out of two arguably irremovable elements of any model of SWR (alongside the algorithm connecting them) — the representation of the input signal and the representation of stored meaning.

We notice two tendencies in models of SWR. First, although the goal of human speech perception is to understand the meaning of what is being said, most models of SWR are designed to simulate the process taking the listener from the input —

in everyday speech this is the acoustic signal — to the form of the word stored in the mental lexicon (we use the terms 'word' for a unit of meaning and the term 'mental lexicon' for the storage and organization of units of meaning for simplicity, although any other assumed units of meaning and conceptualizations of storage could be used). The goal of such a computational model is therefore not to access the meaning behind the word, but rather the face of its form. The textbook statement (see Traxler & Gernsbacher, 2006, pp. 97) that "most models of lexical access do not actually deal with activation of meaning" (Gaskell & Marslen-Wilson, 2002, pp. 261) remains unfortunately true, despite occasional efforts to change this. The mental lexicon is usually presented as an unconnected list of words which are in turn strings of phonemes. This is the case in TRACE (McClelland & Elman, 1986; Strauss et al., 2007), TISK (Hannagan et al., 2013; You & Magnuson, 2018), Cohort (Marslen-Wilson & Tyler, 1980; Marslen-Wilson & Welsh, 1978), Shortlist A and B (Norris, 1994; Norris & McQueen, 2008), NAM/PARSYN (Luce, 1986; Luce et al., 2000; Luce & Pisoni, 1998), and the current implementation of DIANA (ten Bosch, Boves, & Ernestus, 2015). Fine-Tracker (Scharenborg, 2008; Scharenborg & Boves, 2010), for example, describes words using articulatory-acoustic feature vectors, but these vectors are generated as representations of phonemes contained in the word, and, more importantly, the words in the lexicon remain unconnected with regards to their meaning. Certain models, like DIANA or TRACE, do allow for a weight representing word frequency to be applied to the activation levels that are based on the input. However, a frequency weight still does not make the mental lexicon interconnected, as it only places certain words metaphorically higher in the imaginary bin from which they are taken, regardless of their meaning. Other models, following the same notion that lexical access is form access, postulate that the process of spoken word recognition is entirely feed-forward, which in turn obscures the need for a realistic representation of meaning storage (see Norris & McQueen, 2008; Norris et al., 2000a, 2000b). In effect, competition between words and word recognition in most models of SWR is driven primarily if not exclusively by the similarity in the representation of a words' (pseudo-)acoustic representation.

One exception to this conceptualization is the Distributed Cohort Model (DCM; Gaskell & Marslen-Wilson, 1997, 2002). In DCM, the organization of words in the mental lexicon is dependent both on their phonological features and semantic features. Initially, as a proof of concept, the semantic features were represented by eleven arbitrarily chosen (or more precisely, random) binary vectors. In their subsequent simulations, Gaskell and Marslen-Wilson (1999) replace the idea of binary semantic feature vectors as they are difficult to choose/define and use word co-occurrence statistics instead. Word activation in the DCM is therefore a product of both its phonological and meaning characteristics. As we have seen in the present dissertation, a similar approach is used in a more recent addition to the group of models that simulate SWR, the discriminative lexicon (Baayen, Chuang, Shafaei-Bajestan, et al., 2019).

Second, although models of SWR seem to focus on the input-to-form process, very few of the models use actual acoustic signal to create the input. This is primarily a consequence of technological limitations present at the time when these models were created. Pseudo-acoustic input was implemented mostly as a temporary stand-in until better representation can be obtained (see, e.g., McClelland & Elman, 1986; Scharenborg & Boves, 2010; Weber & Scharenborg, 2012). The models often use strings of phonemes as input, as in TISK (Hannagan et al., 2013; You & Magnuson, 2018) or Shorlist A and B (Norris, 1994; Norris & McQueen, 2008), or perhaps some form of phonetic/phonological pseudofeatures as in TRACE (McClelland & Elman, 1986) and the DCM (Gaskell & Marslen-Wilson, 1997).

Besides the aforementioned FBSFs implemented in the discriminative lexicon, one exception to this tendency is Fine-Tracker, a model that was specifically designed to deal with acoustic input and fine-phonetic detail (Scharenborg, 2008; Scharenborg & Boves, 2010). In Fine-Tracker, time steps in the acoustic signal are converted into vector representations of articulatory-acoustic features. Subsequently, these feature vectors are input for another module that connects them to words in the lexicon that are also represented in vectors of articulatory-acoustic features. Although promising, Fine-Tracker's performance may still need to be honed, as the model still performed somewhat worse than the human listener in more recent tests where new and improved articulatory-acoustic representations were used (see Scharenborg & Merkx, 2018). We have implemented a more recent addition to the group of models of SWR that use actual acoustic input, DIANA (ten Bosch, Boves, & Ernestus, 2015), described and tested in the Chapter 3 of the present dissertation.

The fact that the correct form, i.e., a word can be accessed solely based on the (pseudo-)acoustic input, with the mental lexicon being organized as a list, does not mean that this is what happens when a human listener tries to understand spoken words. Frequency effects are well-documented and are also recorded in the most recent large auditory lexical decision experiments (Goh et al., 2020; Tucker et al., 2019). There is evidence that other meaning-related, semantic richness predictors, also help predict the time it takes a listener to process isolated spoken words (Goh et al., 2020; Goh et al., 2016; Tucker et al., 2019). Similarly, the fact that certain models can select the proper target based on pseudo-acoustic input or a phoneme string does not mean that fine acoustic detail does not affect the human listener. Findings showing the importance of subphonemic differences (Salverda et al., 2003) and prosodic cues (Andruski et al., 1994; Kemps et al., 2005) in fact propelled the creation of Fine-Tracker as a model using actual acoustic signal as input (Scharenborg, 2008).

Our intention is to combine elements of DIANA and the discriminative lexicon approach. We take 'the best of both worlds' to create a combination that better represents both the fine detail of the acoustic signal in the process of form recognition, and conceptualizes what is usually referred to as the mental lexicon as an interconnected network of meaning. This new structure would use word activations as they are generated in DIANA. At every time step, however, it would weigh these activations using various estimates derived from the discriminative lexicon matrix representing stored units of meaning. The goal of this simulation would be to test whether computational model estimates generated in this fashion can outperform the standard DIANA and discriminative lexicon implementations in predicting participant processing (response) time in an auditory lexical decision task, again using MALD as the benchmark (Tucker et al., 2019). If the test is successful, this study would lend further support to the necessity of properly representing the acoustic signal and fleshing out the mental lexicon as more than a list of unconnected items in models of SWR. In effect, we are hoping to enhance the weak sides of DIANA (the representation of the mental lexicon) and the discriminative lexicon approach (the representation of the acoustic signal). Additionally, it is important to make the model implementation accessible. A model cannot be used nor properly scrutinized by the community of researchers if its implementation is not well-documented and relatively easy to set up. Ideally, the implementation of GRAFT would learn from TISK and feature a package in Python, or use a set of packages in R like those developed for easier use of the discriminative lexicon approach.

Bibliography

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. (2018). *Acousticndlcoder: Coding sound files for use with ndl* [R package version 1.0.2]. R package version 1.0.2. https://CRAN.R-project.org/package=AcousticNDLCodeR

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., & Heitmeier, M. (2019). *Wpmwithldl: Implementation of word and paradigm morphology with linear discriminative learning* [R package version 1.3.18]. R package version 1.3.18.

Baayen, R. H. (2010). Assessing the processing consequences of segment reduction in dutch with naive discriminative learning. *Lingue e linguaggio*, *9*(2), 95–112.

Baayen, R. H. (2011). Corpus linguistics and naive discriminative learning. *Revista Brasileira de Linguística Aplicada*, *11*(2), 295–328.

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., Hendrix, P., & Ramscar, M. (2013). Sidestepping the combinatorial explosion: An explanation of n-gram frequency effects based on naive discriminative learning. *Language and Speech*, *56*(3), 329–347.

Baayen, R. H., Milin, P., Đur đević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*(3), 438–481.

Baayen, R. H., Milin, P., & Ramscar, M. (2016). Frequency in lexical processing. *Aphasiology*, *30*(11), 1174–1220.

Baayen, R. H., Shaoul, C., Willits, J., & Ramscar, M. (2016). Comprehension without segmentation: A proof of concept with naive discriminative learning. *Language, Cognition and Neuroscience*, *31*(1), 106–128.

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? (J. S. Adelman, Ed.). In J. S. Adelman (Ed.), *Visual word recognition volume 1: Models and methods, orthography and phonology*. Hove, England: Psychology Press.

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

Boersma, P., & Weenink, D. (2011). Praat, a system for doing phonetics by computer. www.praat.org

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.

Cassani, G., Chuang, Y.-Y., & Baayen, R. H. (2020). On the semantics of nonwords and their lexical category. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(4), 621–637.

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*. https://doi.org/10.31234/osf.io/m79fw

Chuang, Y.-Y., Vollmer, M. L., Shafaei-Bajestan, E., Gahl, S., Hendrix, P., & Baayen, R. H. (2020). The processing of pseudoword form and meaning in production and comprehension: A computational modeling approach using linear discriminative learning. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01356-w

Coltheart, M., Tree, J. J., & Saunders, S. J. (2010). Computational modeling of reading in semantic dementia: Comment on woollams, lambon ralph, plaut, and patterson (2007).

Connine, C. M., & Titone, D. (1996). Phoneme monitoring. *Language and Cognitive Processes*, *11*(6), https://doi.org/10.1080/016909696387042, 635–646. https://doi.org/10.1080/016909696387042

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.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Elman, J. L. (2004). An alternative view of the mental lexicon. *Trends in cognitive sciences*, *8*(7), 301–306.

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.

Ernestus, M., & Baayen, R. H. (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*.

Ferrand, L., Méot, A., Spinelli, E., New, B., Pallier, C., Bonin, P., Dufau, S., Mathôt, S., & 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., Augustinova, M., & 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*, Nijmegen, The Nethelands: Max-Planck Institute for Psycholinguistics.

Frauenfelder, U. H., & Peeters, G. (1990). Lexical segmentation in trace: An exercise in simulation (G. T. M. Altmann, Ed.). In G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives*. 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 (J. Grainger & A. M. Jacobs, Eds.). In J. Grainger & A. M. Jacobs (Eds.), *Localist connectionist approaches to human cognition*. Mahwah, NJ: Lawrence Erlbaum.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., & Pallett, D. S. (1993). DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon technical report*, *93*.

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. (1999). Ambiguity, competition, and blending in spoken word recognition. *Cognitive Science*, *23*(4), 439–462.

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., & Chee, Q. W. (2020). The Auditory English Lexicon Project: A multi-talker, multi-region psycholinguistic database of 10,170 spoken words and nonwords. *Behavior Research Methods*, 1–30.

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. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251.

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.

Gwilliams, L., Linzen, T., Poeppel, D., & Marantz, A. (2018). In spoken word recognition, the future predicts the past. *Journal of Neuroscience*, *38*(35), 7585–7599.

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

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, *31*(3-4), 373–405.

Hendrix, P., & Sun, C. (2020). A word or two about nonwords: Frequency, semantic neighborhood density, and orthography-to-semantics consistency effects for nonwords in the lexical decision task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, *4*(1), 11–26.

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

Hitzman, D. (1986). Schema abstraction in a multiple-trace memory model. *Psychological Review*, *93*(4), 411–428.

Hyman, R. (1953). Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology*, *45*(3), 188.

Ivens, S. H., & Koslin, B. L. (1991). *Demands for reading literacy require new accountability methods*. Touchstone Applied Science Associates.

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

Kelley, M. C. *How acoustic distinctiveness affects spoken word recognition: A pilot study*. Presented at the 11th International Conference on the Mental Lexicon (Edmonton, AB). 2018. https://doi.org/https://doi.org/10.7939/R39G5GV9Q.

Kemps, R. J., Wurm, L. H., Ernestus, M., Schreuder, R., & Baayen, R. 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.

Klatt, D. H. (1979). Speech perception: A model of acoustic–phonetic analysis and lexical access. *Journal of Phonetics*, *7*(3), 279–312.

Kosinski, R. J. (2008). A literature review on reaction time. *Clemson University*, *10*, 337–344.

Kuperman, V. (2015). Virtual experiments in megastudies: A case study of language and emotion [PMID: 25406972]. *Quarterly Journal of Experimental Psychology*, *68*(8), 1693–1710. https://doi.org/10.1080/17470218.2014.989865

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25*(2-3), 259–284.

Ligges, U., Krey, S., Mersmann, O., & Schnackenberg, S. (2018). *tuneR: Analysis of music and speech*. https://CRAN.R-project.org/package=tuneR

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 (M. Spivey, M. Joanisse, & KenMcRae, Eds.). In M. Spivey, M. Joanisse, & KenMcRae (Eds.), *The Cambridge Handbook of Psycholinguistics*. 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. https://doi.org/10.3389/fpsyg.2018.00369

Magnuson, J. S., & You, H. (2018). Feedback in the time-invariant string kernel model of spoken word recognition (C. Kalish, M. Rau, J. Zhu, & T. Rogers, Eds.). *Proceedings of the Cognitive Science Society*, 732–737. http://par.nsf.gov/biblio/10097512

Magnuson, J. S., You, H., Nam, H., Allopenna, P., Brown, K., Escabi, M., Theodore, R. M., Luthra, S., Li, M. Y.-C., & Rueckl, J. (2018). EARSHOT: A minimal neural network model of incremental human speech recognition. *PsyArXiv*. https://doi.org/10.31234/osf.io/m79fw

Malik-Sheriff, R. S., Glont, M., Nguyen, T. V., Tiwari, K., Roberts, M. G., Xavier, A., Vu, M. T., Men, J., Maire, M., Kananathan, S., Et al. (2020). Biomodels—15 years of sharing computational models in life science. *Nucleic Acids Research*, *48*(D1), D407–D415.

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

Marslen-Wilson, W. D., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(3), 576.

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

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi, In *Proc. interspeech 2017*.

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 (S.-A. Rueschemeyer & G. Gaskell, Eds.). In S.-A. Rueschemeyer & G. Gaskell (Eds.), *The Oxford Handbook of Psycholinguistics*. 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.

Milin, P., Divjak, D., & Baayen, R. H. (2017). A learning perspective on individual differences in skilled reading: Exploring and exploiting orthographic and semantic discrimination cues. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(11), 1730.

Milin, P., Đur đević, D., & del Prado Martín, F. M. (2009). The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from serbian. *Journal of Memory and Language*, *60*(1), 50–64.

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., & Baayen, R. H. (2017). Discrimination in lexical decision. *PloS one*, *12*(2), e0171935.

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.

Mozilla Organization. (2013). Project DeepSpeech. https://github.com/mozilla/DeepSpeech

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*. https://doi.org/10.21437/Interspeech.2018-2081

Nenadić, F., & Tucker, B. V. (2020). Computational modelling of an auditory lexical decision experiment using jTRACE and TISK. *Language, Cognition and Neuroscience*, 1–29.

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. (2000a). Feedback on feedback on feedback: It's feedforward. *Behavioral and Brain Sciences*, *23*(3), 352–363.

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

Norris, D., McQueen, J. M., & Cutler, A. (2016). Prediction, bayesian inference and feedback in speech recognition. *Language, Cognition and Neuroscience*, *31*(1), 4–18.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books, In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

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

Plaut, D. C., & Kello, C. T. (1999). The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist approach (B. MacWhinney, Ed.). In B. MacWhinney (Ed.), *The emergence of language*.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Et al. (2011). The Kaldi speech recognition toolkit, In *IEEE 2011 Automatic Speech Recognition and Understanding Workshop (ASRU 2011)*, Hawaii, USA. IEEE Signal Processing Society.

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*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, R. H. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in Cognitive Science*, *6*(1), 5–42.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement (A. H. Black & W. F. Prokasy, Eds.). In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory*. New York: Appleton Century Crofts.

Rosenfelder, I., Fruehwald, J., Evanini, K., Seyfarth, S., Gorman, K., Prichard, H., & Yuan, J. (2014). FAVE (Forced Alignment and Vowel Extraction) suite version 1.1.3. https://doi.org/10.5281/zenodo.9846

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*. ISCA Archive.

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*. ISCA Archive.

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

Scharenborg, O., & Merkx, D. (2018). The role of articulatory feature representation quality in a computational model of human spoken-word recognition, In *Proceedings of the Machine Learning in Speech and Language Processing Workshop*, Hyderabad, India.

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*. Psychology Software Tools Inc. Pittsburgh.

Shafaei-Bajestan, E., & Baayen, R. H. (2018). Wide learning for auditory comprehension. *Proc. Interspeech 2018*, 966–970.

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.

Smits, R., Warner, N., McQueen, J. M., & Cutler, A. (2003). Unfolding of phonetic information over time: A database of dutch diphone perception. *The Journal of the Acoustical Society of America*, *113*(1), 563–574.

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.

Sueur, J., Aubin, T., & Simonis, C. (2008). Seewave: A free modular tool for sound analysis and synthesis. *Bioacoustics*, *18*, 213–226. http://www.tandfonline.com/doi/abs/10.1080/09524622.2008.9753600

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., & Mulder, K. (2019). Analyzing Reaction Time and Error Sequences in Lexical Decision Experiments, In *Proc. Interspeech 2019*. https://doi.org/10.21437/Interspeech.2019-2611

ten Bosch, L., Giezenaar, G., Boves, L., & Ernestus, M. (2016). Modeling language-learners' errors in understanding casual speech, In *Errors by humans and machines in multimedia, multimodal, multilingual data processing*.

ten Bosch, L., Boves, L., & Ernestus, M. (2016). Combining data-oriented and process-oriented approaches to modeling reaction time data.

ten Bosch, L., Boves, L., Tucker, B. V., & 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*.

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*.

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*, Hyderabad, India: ISCA.

ten Bosch, L., Boves, L., & Ernestus, M. (2013). Towards an end-to-end computational model of speech comprehension: Simulating a lexical decision task, In *Interspeech 2013: 14th Annual Conference of the International Speech Communication Association.*

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).* University of Glasgow. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0480.pdfl

Tomaschek, F., Plag, I., Ernesus, M., & Baayen, R. H. (2019). Modeling the duration of word-final s in english with naive discriminative learning.

Traxler, M., & Gernsbacher, M. A. (2006). *Handbook of psycholinguistics*. Amsterdam: Elsevier/Academic Press.

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.

Vertanen, K. (2006). *Baseline WSJ acoustic models for HTK and Sphinx: Training recipes and recognition experiments* (tech. rep.). Cambridge, United Kingdom: Cavendish Laboratory.

Vitevitch, M. S., Siew, C. S. Q., & Castro, N. (2018). Spoken word recognition (S.-A. Rueschemeyer & G. Gaskell, Eds.). Oxford University Press. https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780198786825.001.0001/oxfordhb-9780198786825-e-2

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. http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Widrow, B., & Hoff, M. E. (1960). *Adaptive switching circuits* (tech. rep.). Stanford Univ Ca Stanford Electronics Labs.

Yap, M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language*, *60*(4), 502–529.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971–979.

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.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., & Woodland, P. (2006). The htk book (version 3.4). *Cambridge University Engineering Department*. http://htk.eng.cam.ac.uk/

Yuan, J., & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America*, *123*(5), 3878.

Ziegler, J. C., Besson, M., Jacobs, A. M., Nazir, T. A., & Carr, T. H. (1997). Word, pseudoword, and nonword processing: A multitask comparison using event-related brain potentials. *Journal of Cognitive Neuroscience*, *9*(6), 758–775.