DECIDING TO LOOK: REVISITING THE LINK BETWEEN LEXICAL ACTIVATIONS AND EYE MOVEMENTS IN THE VISUAL WORLD PARADIGM IN JAPANESE

by

HIDEKO TERUYA

A DISSERTATION

Presented to the Department of Linguistics
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

September 2018

DISSERTATION APPROVAL PAGE

Student: Hideko Teruya

Title: Deciding to Look: Revisiting the Link between Lexical Activations and Eye Movements in the Visual World Paradigm in Japanese

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Linguistics by:

Vsevolod Kapatsinski	Chairperson
Melissa Baese-Berk	Core Member
Eric Pederson	Core Member
Kaori Idemaru	Institutional Representative

and

Janet Woodruff-Borden	Vice Provost and Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded September 2018

DISSERTATION ABSTRACT

Hideko Teruya

Doctor of Philosophy

Department of Linguistics

September 2018

Title: Deciding to Look: Revisiting the Link between Lexical Activations and Eye Movements in the Visual World Paradigm in Japanese

All current theories of spoken word recognition (e.g., Allopenna et al., 1998; McClelland & Elman, 1986; Norris, 1994) suggest that any part of a target word triggers activation of candidate words. Visual world paradigm studies have relied on the linking hypothesis that the probability of looking at the referent of a word directly tracks the word's level of activation (e.g., Allopenna et al., 1998).

However, how much information is needed to trigger a saccade to a visual representation of the word's referent? To address this question, the present study manipulated the number and location of shared segments between the target and competitor words. Experimental evidence is provided by two visual world paradigm experiments on Japanese, using natural and synthesized speech. In both experiments, cohort competitor pictures were not fixated more than unrelated distractor pictures unless the cohort competitor shares the initial CVC with the target. Bayesian analyses provide strong support for the null hypothesis that shorter overlap does not affect eye movements. The results suggest that a listener needs to accumulate enough evidence for a word before a saccade is generated.

The human data were validated by an interactive computational model (TRACE: McClelland & Elman, 1986). The model was adapted to Japanese language to examine whether the TRACE model predicts competitor effects that fit human data. The model predicted that there should be effects when words share any amount with a target which confirms the current theory. However, the model did not fit the human data unless there is longer overlap between words. This indicates that eye movements are not as closely tied to fixation probabilities of lexical representations as previously believed.

The present study suggests that looking at a referent of a word is a decision, made when the word’s activation exceeds a context-specific threshold. Subthreshold activations do not drive saccades. The present study conclude that decision-making processes need to be incorporated in models linking word activation to eye movements.

This dissertation includes unpublished co-author material.

CURRICULUM VITAE

NAME OF AUTHOR: Hideko Teruya

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, Oregon
Southern Illinois University-Carbondale, Carbondale, Illinois
Kyushu Women's University, Fukuoka, Japan

DEGREES AWARDED:

Doctor of Philosophy, Linguistics, 2018, University of Oregon
Master of Arts, Applied Linguistics & TESOL, 2009, Southern Illinois University-Carbondale
Bachelor of Arts, Japanese Literature, 1997, Kyushu Women's University

AREAS OF SPECIAL INTEREST:

Psycholinguistics: Speech perception and production
First language and second language acquisition
Computational linguistics

PROFESSIONAL EXPERIENCE:

Teaching Assistant, University of Oregon, 2010 – 2015, 2018

Teaching Assistant, Southern Illinois University-Carbondale, 2007 - 2009

High school teacher, Okinawa, Japan, 2005 – 2006

Middle school teacher, Okinawa, Japan, 2003 - 2005

GRANTS, AWARDS, AND HONORS:

Graduate Teaching Fellowship, University of Oregon, 2010 – 2015, 2018

Teaching Assistant Fellowship, Southern Illinois University-Carbondale, 2007 - 2009

PUBLICATIONS:

Terurya, H., & Kapatsinski, V. (under review). Deciding to look: Revisiting the linking hypothesis for spoken word recognition in the visual world. *Language, Cognition and Neuroscience*.

ACKNOWLEDGMENTS

Completing the dissertation was not just the text you see in this paper. Along this journey in the entire Ph. D life in Eugene, I received exceptional support from the university as well as people outside of school.

I first wish to express sincere appreciation to my advisor, Professor Vsevolod Kapatsinski. Without his tremendous support and persistent guidance, this dissertation would not have been possible. He was always very respectful for any ideas I had even though they were novice opinions and very patient with me guiding me through every step of a way. At conferences, his keen and insightful comments on research shocked me and I learned how to be critical. He challenged me with the toughest questions and comments that nobody could ever give which made me a better researcher. I am deeply grateful to him for giving generously of his time and sharing his extraordinarily knowledge as a mentor. I could not ask for more and I truly feel lucky to have an advisor like him and seeing his exceptional work in person was invaluable experience as a researcher.

In addition, special thanks are due to my committee members who provided extensive and concrete suggestions. I am thankful to Professor Eric Pederson for opening my eyes to a bigger picture of the study; to Professor Melissa Baese-Berk for her meticulous and constructive comments to make arguments stronger and clearer; and to Professor Kaori Idemaru for wide-ranging research on discussion to make the study more interesting.

I would also like to thank my peers, Julia Trippe, Shahar Shirtz, Prakaiwan Vajrabhaya, Wook Kyun Choe, Ying Chen, Paul Olejarczuk, Charlie Farrington, Jason

McLarty, and many more. Their encouragements and support helped me to survive the program and inspired me with their enthusiasm about research.

My sincere appreciation also goes to the department of Linguistics. The Linguistics faculty members and staff showed a great care for students and guided us and supported us. Going through Ph. D program in a foreign country was very challenging.

During the rough time, my friends and family all over the world believed in me and supported me with love. Their warm encouragements and moral support encouraged me to get through this long journey.

Lastly, my deepest appreciation goes to my husband who was very patient for my long schooling and had sincere encouragements for everything I did. Without his constant encouragements and considerable support, completing the program would not have been possible.

TABLE OF CONTENTS

Chapter Page

LIST OF FIGURES

Figure Page

Figure Page

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

The visual world paradigm that is the focus of the present study involves looking at an array of visual stimuli while listening to a particular acoustic stimulus. The paradigm is thought to be useful as a window on spoken word recognition because eye movements are thought to directly track lexical activation (e.g., Allopenna, Magnuson, & Tanenhaus, 1998). All current theories of spoken word recognition (e.g., Allopenna et al., 1998; McClelland & Elman, 1986; Norris, 1994) suggest that any part of a target word triggers activation of candidate words. Visual world paradigm studies have relied on the linking hypothesis that the probability of looking at the referent of a word directly tracks the word's level of activation (e.g., Allopenna et al., 1998; McClelland & Elman, 1986; Norris, 1994). This assumption seems to be taken granted by everyone in the field.

The present study asks whether the lexical activation of spoken word recognition (e.g. TRACE: McClelland & Elman, 1986) is indeed *directly* reflected in fixation probabilities. The alternative linking hypothesis I consider is that a minimum amount of support for a word is necessary for the eye to be drawn to the word's referent. Rather than lexical activations being directly / faithfully mapped onto saccades, I argue that eyes may not move until the listener has accumulated enough evidence for a particular word being present in the speech signal. To address this question, the present study manipulated the number and location of shared segments between the target and competitor words to investigate how much information is needed to trigger a saccade to a visual representation of the word's referent. Experimental evidence for the alternative linking hypothesis is provided by two visual world paradigm experiments on Japanese, using natural and synthesized speech.

Previous work on the visual world paradigm has linked eye movement probabilities to lexical activations in an interactive activation model of spoken word recognition (TRACE: McClelland & Elman, 1986). The present study also simulates the link between lexical activations in TRACE and eye movement probabilities but revises the assumptions about this mapping. To this end, I developed a methodology for evaluating TRACE predictions for eye movements quantitatively, given a particular linking hypothesis, and used it to provide evidence for the alternative linking hypothesis that the mapping is mediated by a decision-making process. Because the present study examined word recognition in Japanese, the model was adapted to Japanese phonology and lexicon.

The present study suggests that looking at a referent of a word is a (unconscious) decision by a listener, made when the word's activation exceeds a context-specific threshold by accumulating evidence. That threshold may differ based on many different factors, including the participant's ability to see the alternative response choices without moving their eyes, the motoric effort that an eye movement will involve given the distance that needs to be traveled to look at a picture, and the probability that the eye will need to be moved again. Subthreshold activations do not drive saccades. That is, if one is not certain enough that a word is present in the signal, the eyes will not move to a picture of its referent. The present study suggests that linking eye movement data to word activations may require modeling the process of making a decision to make a saccade on the basis of accumulating evidence (e.g., Usher & McClelland, 2001).

The following sub-sections in this chapter describe the background of the present study topic as well as laying out motivations for the study based on issues and concerns regarding current theories and previous studies.

1.1. Spoken Word Recognition in the Eye Movement Studies

Word recognition is thought to be (largely) incremental (Allopenna et al., 1998; Arnold, Tomaschek, Sering, Lopez, & Baayen, 2017; Balling & Baayen, 2008; Cutler & Otake, 2002; Dahan, 2010; McClelland & Elman, 1986; Norris, 1994). This means that activation is cascading into a word representation and even the associated semantics as soon as there is *any* evidence for the word in the signal. In order to recognize a spoken word, a listener may use various acoustic cues present in the signal as well as contextual cues influencing which word one is more likely to be hearing. Incremental processing means that the listener utilizes individual cues to access words as soon as they become available, without waiting to integrate them into larger and potentially less ambiguous compounds. For example, as soon as a listener hears /bə/ in 'banana,' he or she starts accessing the meanings of words that begin with /bə/, without waiting until the end of the word. All current theories of spoken word recognition agree that incoming acoustic information activates a cohort of word candidates in the mental lexicon (e.g., Gaskell & Marslen-Wilson, 1997; Marslen-Wilson & Zwitserlood, 1989; McClelland & Elman, 1986; Norris, 1994; Norris & McQueen, 2008), though they vary in the extent to which detailed acoustic information is thought to be maintained for later re-interpretation (e.g., Bushong & Jaeger, 2017; Gwilliams, Linzen, Poeppel, & Marantz, 2018). While only one word eventually wins the competition for recognition, and is consciously identified as the target word, other words are activated along with it (e.g., Dahan & Gaskell, 2007;

Frauenfelder, Scholten, & Content, 2001; Marslen-Wilson, 1987) and may even continue to retain some residual activation after the target word is recognized (Kapatsinski, 2012; Kleinschmidt & Jaeger, 2015; Radeau, Morais, & Segui, 1995)

The paradigm is thought to be useful as a window on spoken word recognition because eye movements are thought to directly track lexical activation (e.g., Allopenna et al., 1998). While other behavioral data provide convergent evidence for incremental processing of spoken words and cascading activation of semantics from phonetics / phonology (e.g., Grosjean, 1980 for gating; Revill, Aslin, Tanenhaus, & Bavelier, 2008 for fMRI), the visual world paradigm has an important advantage in allowing the researcher to investigate lexical competition as it unfolds in real time. In a typical visual world experiment, a spoken target word is presented as the speaker is watching a display containing a depiction of the word's referent (picture) alongside the referents of other, similar-sounding words. Looks to depictions of the words' referents are thought to reflect activation of the words' semantic representations (Allopenna et al., 1998). In this way, we can use eye movements as a proxy for the activation levels of the words in real time and track changes in activation levels as the spoken word unfolds.

In the classic experiments by (Allopenna et al., 1998) as shown in Figure 1.1 below, a target word's semantic representation (e.g., the concept of a 'beaker') was activated from the beginning of the corresponding acoustic signal and its activation gradually increased throughout the time course of word recognition. Along with the target word activation, a cohort competitor (e.g., 'beetle') was also activated at the beginning of the word as strongly as the target, but its activation gradually decreased later on. Furthermore, the target word not only activated a cohort candidate, but also activated a word that shared the last part with the target word (a rhyme competitor). The rhyme effect (e.g., looks to a picture of a ‘speaker’) was observed somewhat later than the cohort effect and was much weaker than the cohort effect.

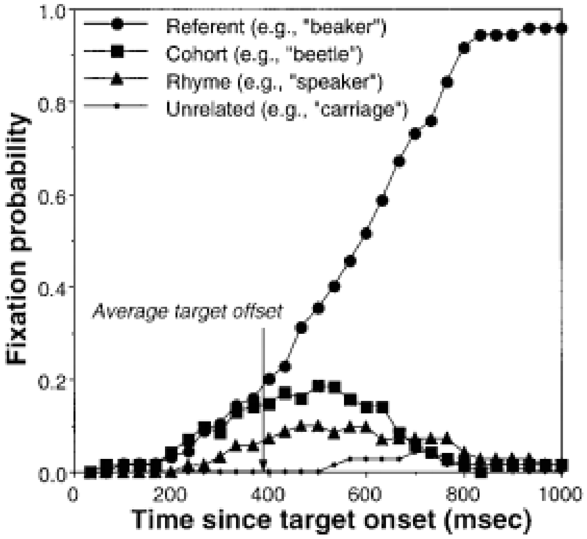


Figure 1.1. An example trial of the full competitor condition from Allopenna et al. (1998;428)

Activation of rhyme competitors following activation of cohort competitors does not necessarily contradict incremental processing. The listener may well use the onset information to activate a set of compatible words but continue to maintain uncertainty regarding the onset. Since perception is fallible, this behavior is rational (Kleinschmidt & Jaeger, 2015; Salasoo & Pisoni, 1985): one does not wish to erroneously rule out the possibility of having heard the word ‘beaker’ on the basis of erroneously misperceiving the initial [sp] as a [b] in a noisy environment. In fact, detailed acoustic information can be maintained and continues to be available for re-interpretation long after the word has ended (Bushong & Jaeger, 2017; Goldinger, 1996; Gwilliams et al., 2018; Palmeri, Goldinger, & Pisoni, 1993).

Maintenance of detailed acoustic information for indefinite periods of time does not contradict the use of this information to activate semantic representations as it becomes available. It does suggest that the listener may not decide what the word is immediately, although it is also compatible with the position that all such decisions are provisional and subject to revision (Gwilliams et al., 2018). In either case, the theoretical decision regarding the identity of the word is in principle separate from the decision I am interested in for the purposes of the current dissertation – the decision to move one's eyes to the referent of a word. The presence of looks to cohort competitors in the eye tracking record (e.g., Allopenna et al., 1998) indicates that eyes move to referents of words that the speaker can decide to move their eyes to the referent of a word that they then decide is not present in the speech signal. That is, the threshold I am interested in is generally lower than the threshold for consciously deciding that the word is present in the speech signal.

1.2. The TRACE Model and the Linking Hypothesis

The TRACE model is an interactive activation model. The model consists of three layers of units, including a feature level, a phoneme level, and a word level. Input, for example the /k/ in *kasa* 'umbrella,' is first represented on the feature level with feature values (e.g., strength levels for voiceless, sonorant, etc.), and the feature values activate phonemes that share them (e.g., /k/ and /g/ would be activated by a certain level of [sonorant]) and inhibit those that do not. The activated phonemes activate candidate words that contain them (e.g., *kasa* 'umbrella,' *kame* 'turtle,' *gomi* 'garbage.' etc.) and inhibit those that do not. The activated words feed activation back to the phoneme level, activating the phonemes they contain and inhibiting those they do not. Words also compete with each other via lateral inhibitory connections, as do phonemes. In other words, at the phoneme and word levels, there is inhibition within a level and bidirectional flow of activation and inhibition between levels. Since multiple features can be present simultaneously, features do not compete with each other for recognition, and top-down feedback does not affect feature activations, preventing hallucinations based on top-down input. The TRACE model proposes that any part of a target word activates other candidate words (McClelland & Elman, 1986).

Note that eye movements are not modeled by TRACE. The model simply exhibits the timecourse of lexical activations during recognition of a spoken word. A linking hypothesis is therefore required to lexical activations to eye movement probabilities. In this thesis, I am not arguing for or against the spoken word recognition model. There is extensive evidence for interactive activation at both neural and behavioral levels (e.g., Gow & Olson, 2015). Instead, I suggest that the standard linking hypothesis that transforms lexical activations directly into fixation probabilities using the Luce Choice Rule (Allopenna et al., 1998) is overly simple and needs to be reconsidered.

When coupled with the standard linking hypothesis, TRACE tends to predict that any amount of overlap with the target should increase the likelihood of fixating the referent of a word. Hypothetically if lexical activation and fixation probabilities are directly or faithfully linked, hearing the /ka/ in /kame/ may lead the listener to fixate a picture of a referent of any word starting with /ka/ more than a distractor picture. The direct link between TRACE activations and fixation probabilities is explicitly defended as the standard linking hypothesis for spoken word recognition in the visual world by

Allopenna et al. (1998) as well as by Tanenhaus, Magnuson, Dahan, & Chambers (2000). Figure 1.2 demonstrates an example trial of the full competitor condition predicted by TRACE with the standard linking hypothesis and those observed in human data by Allopenna et al. (1998). TRACE achieves an excellent fit to their human eye tracking data, providing evidence for the standard linking hypothesis.

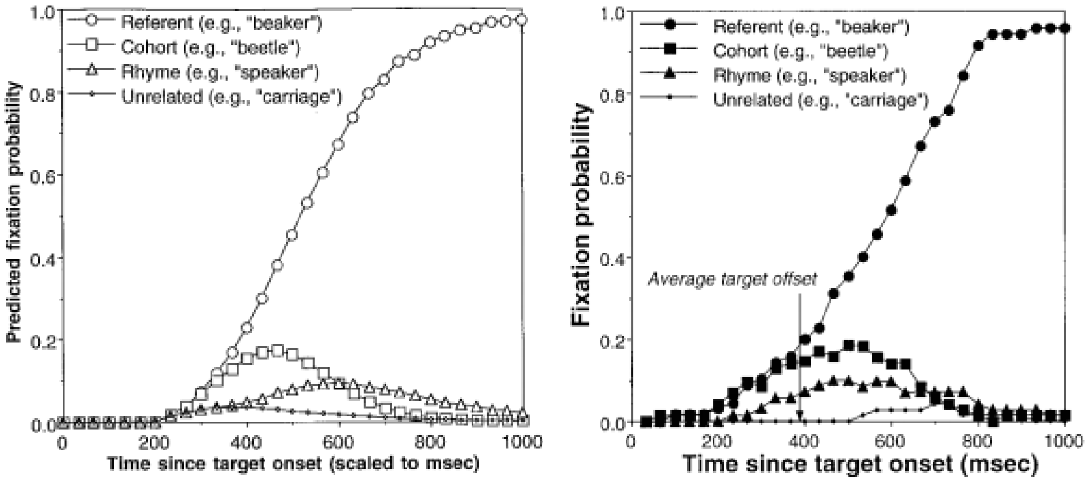


Figure 1.2. An example trial of the full competitor condition from Allopenna et al. (1998). TRACE model is on left (from p.425) and human data is on right (from p.428).

However, previous studies have tended to use a small set of target and competitor words and have not shown that *any* amount of segmental overlap is sufficient to observe lexical competition in the visual world paradigm. Table 1.1 is a summary of the characteristics of cohort stimuli and procedures of previous visual world studies that have provided evidence for cohort effects. The table indicates that studies have tended to examine monosyllabic or disyllabic words that shared several initial segments with a target. The cohort effect has been observed in both mono- and disyllabic words, although Simmons & Magnuson (2018) have recently reported that it was larger in monosyllables in their study. Since the studies did not directly manipulate the amount of overlap among words, it is unclear that any amount of overlap is sufficient to drive a saccade.

Table 1.1. Summary of stimulus characteristics and procedures in previous literature. Number of target-competitor sets having a certain number of syllables or a segmental overlap of a certain length shown in parentheses.

Study	# of segments overlap	# of syllables	# of trials for each condition	# of competitors in a trial	Pre-training using picture naming	Repetition of trials	Language
Allopenna et al. (1998)	2 (1) 3 (4) 4 (4)	2 (8)	6	1 or 2	Yes	Yes	English
Dahan et al. (2001a)	2 (8) 3 (8) 4 (1)	1 (2) 2 (14) 3 (1)	17	1	No	No	English
Dahan et al. (2001b)	2 (12) 3 (3)	1 (15)	15	1	Yes	No	English
Dahan & Gaskell (2007)	2 (16) 3 (10) 4 (2)	1 (23) 2 (3) 3 (2)	28	1	Unknown	No	Dutch
McMurray et al. (2010)	2 (18) 3 (17) 4 (5) 5 (1)	1 (21) 2 (20)	41	2	No	No	English
Mirman et al. (2011)	1 (1) 2 (5) 3 (4) 4 (1)	2 (11)	11	1	No	No	English

Note: A diphthong was counted as two segments

A plausible alternative hypothesis is therefore that some minimum amount of overlap with the target, as a proxy for amount of evidence from the acoustic signal, is required for activation of a competitor to be sufficient to draw an eye movement. The present study is intended to evaluate this alternative hypothesis by systematically varying the amount of segmental overlap between the target word and its competitors, whether these competitors share the beginning or the end with the target.

Because the standard linking hypothesis for the visual world paradigm directly connects eye movements to activation levels of words in the TRACE model (McClelland & Elman, 1986), the present study likewise utilizes the TRACE model to estimate the amount of support that alternative lexical candidates have at a given timepoint. While other models of spoken word recognition exist (e.g., Goldinger, 1998; Luce & Pisoni, 1998; Norris & McQueen, 2008), TRACE is the only one that is freely available. It also has the important advantage of generating real-time trajectories of activations that can in principle be mapped rather directly onto the trajectories of fixation probabilities (Allopenna et al., 1998). This enables the modeler to determine whether the differences in the extent and timecourse of competition between words differing in the location and extent of segmental of overlap mirror those predicted by the model. Furthermore, TRACE's interactive nature is consistent with neuroscientific findings on the presence of extensive top-down connections to early sensory processing areas (Bonte, Parviainen, Hytönen, & Salmelin, 2006; Eagleman, 2001; Gow & Olson, 2015; McClelland, Mirman, & Holt, 2006). The present study develops a rigorous model comparison approach that seeks to determine whether the between-condition differences in fixations predicted by lexical competition in TRACE are reflected in the eye movement record of human participants.

1.3. Pre-activation of Words in Visual World Studies

Importantly, lexical activation comes both from the signal and from top-down expectations (e.g., Benichov, Cox, Tun, & Wingfield, 2012; Broadbent, 1967; Goldiamond & Hawkins, 1958; Howes, 1957; Morton, 1964; Nittrouer & Boothroyd, 1990), which are captured by resting activation levels of lexical nodes in TRACE and other activation-based spoken word recognition models (Luce & Pisoni, 1998; McClelland & Elman, 1986; Morton, 1969). Previous studies have raised resting activation levels of both targets and competitors (not distractors) through pre-training on the small set of experimental materials before the experiment and repeating stimuli during the experiment. Typically, participants were asked to study the experimental words before the eye tracking experiment began to ensure the participants know the intended names for the pictures used in the experiment (e.g., Allopenna et al., 1998; Dahan, Magnuson, Tanenhaus, & Hogan, 2001b; Dahan & Tanenhaus, 2004). In addition, the words and picture sets were repeated throughout the experiment. By doing so, previous studies likely increased the likelihood that a limited amount of evidence in the signal would produce lexical activation observable in the eye movement proportions, increasing both cohort and rhyme effects (Huettig, Olivers, & Hartsuiker, 2011).

Pre-exposure and / or repetition are expected to increase top-down activation of previously encountered words (e.g., Goldiamond & Hawkins, 1958; Scarborough, Cortese, & Scarborough, 1977). This top-down activation in turn could increase cohort and rhyme effects – that is, magnify the differences between the competitors and distractors while minimizing the difference between the competitors and the target. Top-down expectations have been argued to help recognition the most for stimuli that have some bottom-up support but are not strongly supported by the signal (Broadbent, 1967;

Norris, Cutler, McQueen, & Butterfield, 2006; Norris & McQueen, 2008; Plaut & Booth, 2000). In particular, Plaut and Booth (2000) have argued that sigmoid node activation functions in connectionist models like TRACE predict their finding that bottom-up priming effects are significantly larger for words of an intermediate level of resting activation. Norris (2006) and Norris and McQueen (2008) have argued that top-down expectations should not override clear bottom-up evidence for or against a certain stimulus being present in the signal, in order to avoid hallucinations.

Given this reasoning, pre-exposure is expected to have relatively little influence on the activation level of the target word, which is strongly supported by the acoustic signal, and the distractors, which have essentially no bottom-up support, but could significantly boost cohort and rhyme competitors, which have limited bottom-up support, increasing cohort and rhyme effects. Pre-exposure and repetition are therefore best to avoid if we are to identify words whose activation is too low despite bottom-up support for a saccade to the referent to be triggered. The experiment was therefore designed to reduce exposure to experimental words and pictures as much as possible.

Although word recognition is incremental, some minimum level of lexical activation may be required to trigger a saccade to the word's referent. When the words are not pre-activated by top-down expectations, the activation necessary to drive a saccade must come from the signal, and therefore a word may require more support from the signal to trigger a saccade to its referent. If this expectation is upheld, and a substantial amount of signal support (e.g., several initial segments) is required to drive a saccade in the present experiments, then the linking hypothesis of a continuous mapping between activation levels and eye movement probabilities may need to be reconsidered.

1.4. Effect of Rhyme Competitors

Initial overlap between input spoken word and listeners' stored lexical representation strongly affects spoken word processing (e.g., Gaskell & Marslen-Wilson, 1997; Marslen-Wilson & Zwitserlood, 1989; McClelland & Elman, 1986; Norris, 1994; Norris & McQueen, 2008). However, although the activation of competitors with initial mismatch, the rhyme effect, was observed in natural data as well as in TRACE using a variety of methodological approaches to spoken word processing, the effect appears to be very sensitive to the type of task and the type of stimulus words examined. Some studies have found the rhyme effect while some have not.

For example, priming studies have observed priming between rhyme competitors but only with extensive overlap. Connine, Blasko, and Titone (1993) showed priming of a non-word whose initial segment was one or two phonological features away from the prime. This indicates that although the first incoming input may be weighted heavily (cohort effect) for word processing later information may still be helpful. However, the effect disappeared when words differed by more than a few features of the initial segment (Connine et al., 1993; Marslen-Wilson & Zwitserlood, 1989) and no priming was found when the words were shorter (i.e., monosyllable words: 'buns' and 'guns') (e.g., Gow, 2001).

In visual world eye tracking studies, Allopenna et al. (1998) found the rhyme effect between disyllabic words differing by more than a few features of a single segment: 'beaker-speaker,' 'carrot-parrot,' 'candle-handle,' pickle-nickel,' 'casket-basket,' 'paddle-saddle,' 'dollar-collar,' 'sandal-candle'. Simmons and Magnuson (2018) found a stronger rhyme effect in disyllables compared to monosyllables. Conversely, the effect of a cohort overlap in a certain number of segments was weaker in disyllables than monosyllables. Overall, these findings are consistent with the proposal that words are activated to the extent that they overlap with the target word (Kapatsinski, 2005; Simmons & Magnuson, 2018): a CVC monosyllabic rhyme neighbor shares 2/3 of its segments with the target, while a disyllable may share as much as 4/5. Conversely, a single-segment cohort competitor shares 1/3 with a CVC target but only 1/5 of a CVCVC one. These results are also predicted by the TRACE model, for a different reason: longer words face more competition than shorter words early on but yet receive more distinctive bottom-up input later in processing (Simmons & Magnuson, 2018). However, other inconsistencies remain. For example, Mirman, Yee, Blumstein, & Magnuson (2011) found a small rhyme effect for young college participants as a control group; however there seemed to be no effect for older participants (67 years old). Malins and Joanisse (2010) found no rhyme effect in Mandarin monosyllable words (e.g., chuang2 vs. huang2).

As observed by previous studies, the effect appears to be very sensitive to the type of task and the type of stimulus words examined.

1.5. No Mora Effect in Japanese Word Processing

Given the phonological structure of Japanese, a mora-timed language, one might expect the cohort effect to emerge in Japanese only when the competitor shares at least the initial CV (mora) with the target (e.g., Hayes, 1989; Labrune, 2012; Otake, Hatano, Cutler, & Mehler, 1993; Port, Dalby, & O'Dell, 1987; Vance, 1987). A single mora (light syllable) is constructed similarly to an English syllable that contains one nucleus (e.g., /e/

‘picture,’ /ke/ ‘hair,’/gjo/ ‘fish’). As in English, and other weight-sensitive languages, heavy syllables are two moras, while light syllables are one mora. For example, /ki.te/ ‘come here’ consists of two moras, but the geminate version, /kit.te/ ‘stamp,’ consists of three moras, because the first syllable is heavy and contains two moras. ‘Mosquito’ /ka/ is a single mora (a light syllable), but a nasal coda, /kaN/ ‘can’, adds a mora. Similarly, having a diphthong, /kai/ ‘sea shell’ is two moras. ‘Blood’ /tɕi/ is a single mora (light syllable), but having double vowels, /tɕii/ ‘status,’ is considered as two moras (e.g., Hayes, 1989; Port et al., 1987; Vance, 1987). Mora timing suggests that moras tend towards isochrony in production (Port et al., 1987), though the existence of a tendency towards moraic isochrony has been highly controversial (e.g., Beckman, 1992; Grabe & Low, 2002).

Whereas in English moras are thought to be constituents of the rhyme, so that onsets are non-moraic, belonging to no mora, the moraic isochrony hypothesis proposes that Japanese moras span onset-nucleus boundaries (e.g., Grabe & Low, 2002). Under this assumption, light syllables cannot be segmented any further, while heavy syllables are segmented between the nucleus and the coda, as in /ta.n/. Monitoring studies have provided support for this idea (e.g., Cutler & Otake, 1994; McQueen, Otake, & Cutler, 2001; Otake et al., 1993; Port et al., 1987). Otake et al. (1993) presented words such as /taniɕi/ and /tanɕi/, differing in whether /n/ is a separate mora as in /ta.n.ɕi/, or only part of a mora as in /ta.ni.ɕi/, and asked participants to detect either /ta/ or /tan/ in spoken Japanese words. Participants detected /ta/ equally easily in both word types but had difficulty detecting /tan/ when the /n/ formed the first part of a mora, as in /ta.ni.ɕi/. Note that these results cannot be due to syllable boundaries because the /ta/ in /tanɕi/ is not a syllable but easy to detect. Similarly, McQueen et al. (2001) suggested that words were easier to detect when their boundaries aligned with mora boundaries as opposed to falling inside a mora.

From these observations, one might think that the rhythmic unit, mora, can be a structural unit for spoken word recognition (i.e., word recognition in Japanese could proceed mora by mora). However, Cutler and Otake (2002) as well as Otake, Sakamoto, and Konomi (2004) argue that moras do not play a role in spoken word recognition (see also Content, Meunier, Kearns, & Frauenfelder, 2001 for the syllable in French). Cutler and Otake's (2002) argument is based on a study in which they presented a spoken non-word that was altered from a real word (e.g., *panorama*) by replacing a C, a V, or a CV (e.g., *panorema*, *panozama*, *panozema*) and asked a participant to change the non-word back into the real word (e.g., *panorama*). They found that words in which a single phoneme (C or V) was replaced, a part of the mora, were modified faster and more accurately than words in which the entire mora was replaced (CV). This suggests that moras are segmented into smaller units, and has been taken to imply that a word is not recognized mora by mora in Japanese. Nonetheless, this question remains somewhat open because monitoring and word modification studies are metalinguistic tasks whose relevance to spoken word recognition is uncertain.

The target and cohort competitors in the present study always have CVCV structure. Therefore, the initial CV constitutes a mora, as does the second CV. If Japanese spoken words are segmented into moras and recognized mora by mora, so that submoraic overlap is insufficient to drive lexical activation, then the cohort effect should be observed when the cohort competitor shares at least the initial CV with the target, but should not increase in magnitude when the two words overlap in an additional consonant (CVC overlap), which forms only part of the second mora, mirroring the monitoring results by Otake et al. (1993). In contrast, if the mora plays no special role, then the cohort effect should either continuously track the amount of evidence for a word in the speech signal, approximated here by segmental or acoustic overlap – according to the standard linking hypothesis – or should increase with segmental overlap once overlap is above a certain threshold. In either case, if any cohort effects at all are observed in the present study, non-moraic recognition would lead me to expect a difference between two-segment and three-segment overlap so that three-segment overlap produces a stronger cohort effect. Moraic recognition would predict no benefit from three segment cohort overlap compared to two segment overlap.

From the present theoretical perspective, one would not expect spoken word recognition in Japanese to proceed mora by mora – in TRACE, activation continuously cascades from acoustic features to lexical nodes and does not rely on recognition of sublexical units like segments or moras (e.g., McMurray, Tanenhaus, & Aslin, 2002). However, the issue remains empirically unsettled in Japanese, as previous studies arguing for or against moraic word recognition have not directly examined spoken word recognition using online measures.

1.6. Current Dissertation

The current dissertation directly investigates the segmental overlap required for word activation to be reflected in the eye movement record using natural speech stimuli. To accomplish this goal, the present study took advantage of the simple syllable structure of Japanese, which allows the experimenter to construct a relatively large number of competitor pairs varying in amount and location of overlap but sharing mora structure and other characteristics. In addition for enabling the construction of a relatively large number of comparable stimulus sets, Japanese has been underexamined in prior work on online spoken word recognition. As discussed above, previous studies of spoken word recognition in Japanese have employed metalinguistic tasks. In contrast, most previous studies using the visual world paradigm examined Indo-European languages. Therefore, a study of Japanese spoken word recognition extends the literature and helps ensure that the results are generalizable across languages and contribute to a general understanding of spoken word recognition as a whole.

Experiments reported in this dissertation examine competitor words that share the following amounts with a target word: the initial C, the initial CV, the initial CVC, the final CV and the final VCV. All current theories of spoken word recognition (e.g., Allopenna et al., 1998; McClelland & Elman, 1986; Norris, 1994) suggest that overlap with any part of a target word can increase activation levels of candidate words. The question I address is whether these activations are indeed directly / faithfully reflected in eye movements (fixation probabilities). If any amount of overlap results in detectable activation, then competitors will always be fixated more than unrelated distractors. Furthermore, competitors that overlap with the target in more segments will be fixated at a higher rate than those that overlap in fewer segments. Note, however, that – given the relative paucity of studies observing the rhyme effect – this expectation is weaker for the rhyme effect than for the cohort effect. The present study examines the role of initial segments in spoken word recognition by manipulating whether the cohort competitor overlaps with the target by the initial C, the initial CV, or the initial CVC. If the link between activations and eye movements is not as faithful as suggested by the standard linking hypothesis, then a candidate word can be activated by initial overlap without this overlap driving eye movements.

The alternative linking hypothesis I propose is that a minimum amount of support for a word is necessary for the eye to be drawn to the word's referent. Below that threshold (i.e., not enough support for a word), increases in activation do not affect fixation probabilities. Note that the cut-off point of the threshold for decision making whether a listener moves his / her eyes or not may change based on many different contextual factors (e.g., the physical or psychological state of the listener, types of tasks and stimuli, cost of response, etc.). In order to develop a complete theory of decision-making on spoken word recognition, one would need to examine every possible contextual factor that influences decision making. The present study is the first to show the evidence for a word needs to exceed a threshold to trigger an eye movement. However, the present study does not wish to argue that the amount of evidence observed to be sufficient in the present study would also be the same in any other language or context. Rather, I suggest that moving one's eyes to a picture is a decision that needs to be explicitly modeled in future work.

The proposal that eye movements are influenced, from the earliest point in processing, by a large number of contextual factors undoubtedly complicates the interpretation of visual world data. However, it brings the linking hypothesis for spoken word recognition in the visual world into conformity with what I take to be the take-home message of visual world research, i.e. that processing is task-specific, situated and interactive. For example, Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy (1995) have shown that the decisions of what the speaker means are immediately and continuously influenced by the visual scene the listener is looking at. Brown-Schmidt & Tanenhaus (2008) have shown that cohort competitors can be eliminated by top-down information coming from interlocutors and the shared visual world. Given this general conclusion of visual world studies, it is somewhat surprising that eye movement probabilities would ever directly reflect one source of information – the lexical activation of a word. The present dissertation can therefore be seen as an argument for feedback in research strategy: the conclusions of previous visual world studies suggest that processing is situated and interactive. It is time for the linking hypothesis connecting lexical activations to eye movements to reflect this.

1.7. Plan of the Dissertation

Chapter 2 presents the results of natural spoken word recognition experiment in Japanese. This experiment provides preliminary empirical support for the hypothesis that a competitor word will be fixated more than a distractor picture only after enough evidence of the input word has accumulated. Note that the unique prediction of this linking hypothesis, not shared with the standard linking hypothesis, is that there should be no difference in looks between a competitor picture and a distractor picture then the competitor-target overlap is below the required minimum. Supporting the alternative linking hypothesis therefore crucially requires obtaining support for a statistical null hypothesis: a difference in overlap does not affect eye movements (under certain conditions). It therefore requires a Bayesian approach to data analysis, which can distinguish between lack of evidence against the null and evidence in favor of the null (Kruschke, 2008; Wagenmakers, 2007). The approach, adopted from Wagenmakers (2007), is detailed in Chapter 2 and used throughout the dissertation.[1]

Chapter 3 examines the influence of coarticulation on spoken word recognition. The spoken words from Experiment 1 in Chapter 2 were synthesized using diphone synthesis to eliminate long distance coarticulation (V-to-V coarticulation in CVCV words). With respect to the linking hypotheses referenced above, if any amount of overlap with the acoustic signal leads to detectable word activation, we should expect increased looks to cohort competitors in this experiment compared to Experiment 1 now that the beginning of the target word contains no coarticulatory cues to its end, which distinguishes it from the cohort competitors. If instead listeners wait to move their eyes until they have accumulated enough evidence for a word being present in the acoustic signal, then we should expect fewer looks at the target and cohort competitor early on than that in Experiment 1, because the absence of the coarticulatory cues and the relative degradedness of synthesized speech provides these words with less support.

Chapter 4 examines whether the TRACE model of spoken word recognition can explain the human data (both synthesis and natural). The linking hypothesis for the visual world paradigm proposed by Allopenna et al. (1998) suggests that fixation probabilities directly track TRACE activation levels. However, no studies have yet systematically tested the assumption that activations of words are *always* reflected in saccades. A few studies have successfully demonstrated that TRACE activation trajectories provide a

[1] Note that, as argued by Kruschke, the decision to adopt a Bayesian approach to data analysis does not imply endorsing a Bayesian approach to cognition. Bayesian data analysis allows for rational inference from the observed data. Whether humans decision-making is rational in this way is an open question that is beyond the scope of this dissertation.

good match to fixation probabilities in natural data when the competitor words shared several segments with the target word (e.g., Allopenna et al., 1998; Dahan & Gaskell, 2007; Dahan, Magnuson, Tanenhaus, et al., 2001a). However, it is unclear whether this is also true when the overlap is more limited, and whether TRACE would make the same predictions in Japanese.

To this aim, in Chapter 4, I develop a methodology to assess whether between-condition differences in trajectories predicted by TRACE are supported by the human data by comparing them to a baseline model that retains the average temporal dynamics of the TRACE activation curves while eliminating the between-condition differences. The TRACE trajectories are approximated using Generalized Additive Models (Wood, 2003) that include condition as a predictor and do an excellent job at reproducing the trajectories. The baseline model is then derived by simply eliminating the condition predictor, generating one curve to fit TRACE predictions from both conditions under comparison. The baseline model is then compared to the full TRACE model using the BIC approximation to the Bayes Factor (Wagenmakers, 2007), which allows me to distinguish between evidence for the full TRACE model, evidence in favor of the baseline model and lack of definitive evidence for either model.

In addition, Chapter 4 argues that in order to evaluate the predictions of TRACE for another language (i.e., Japanese); it is necessary to modify the segment specifications of TRACE to fit the phonetics of that language. The process of developing a language-appropriate acoustic feature specification in TRACE is illustrated for Japanese in Chapter 4 where I develop descriptions for the complete set of Japanese phonemes. My hope is that this procedure can be replicated in other languages, opening them up to computational modeling of real-time dynamics of spoken word recognition.

Chapter 5 explores parameter manipulation in the TRACE model to examine what plausible parameter changes could achieve a better fit to the human data and describe the difference between synthesized and natural speech, where synthesized speech is less clear than natural speech and has no long-distance co-articulation.

CHAPTER II

EXPERIMENT 1: NATURAL SPEECH STIMULI

The work presented in this chapter is also reported in a co-authored article invited for resubmission to the journal *Language, Cognition and Neuroscience*

2.1. Introduction

Experiment 1 examines how much segmental overlap is needed to trigger a saccade to a visual representation of the word's referent in response to natural speech. Experimental stimuli systematically vary the length and location of segmental overlap between the target word and its competitors.

All current theories of spoken word recognition (e.g., Allopenna et al., 1998; McClelland & Elman, 1986; Norris, 1994) suggest that overlap with *any* part of a target word triggers activation of candidate words. The question I address is about linking hypothesis whether these activations are directly reflected in eye movements (fixation probabilities). If any amount of overlap results in detectable activation, then competitors will always be fixated more than unrelated distractors. Furthermore, competitors that overlap with the target in more segments will be fixated at a higher rate than those that overlap in fewer segments.

Previous eye tracking studies have shown that a target word triggers activation of cohort candidate words when those words share a few phonemes with the target word (e.g., Allopenna et al., 1998; Dahan & Gaskell, 2007; Dahan et al., 2001a; 2001b). However, it is unclear how much overlap is needed for a cohort effect to be observed.

Given that previous research has not manipulated the extent of overlap, it is possible that two or even three segments are needed, as well as that even a single initial consonant is sufficient. Similarly, it is unclear how much final overlap is enough to observe the rhyme effect. Previous priming and visual world eye tracking studies have shown that words differing from the target word by the initial onset compete with it (e.g., Allopenna et al., 1998; Connine et al., 1993; Simmons & Magnuson, 2018). However, it is unclear whether competition might occur between words that differ by more than the initial onset.

To address this question, the present study manipulated the number and location of shared segments between the target and competitor words and conducted Bayesian analyses that allowed us to investigate whether a particular amount of evidence for the presence of a form in the acoustic signal increases the probability of fixating the referent of the form. The advantage of these analyses for the present purposes is their ability to provide evidence for the null hypothesis – in this case, the hypothesis that consistency with the acoustic signal does not affect saccades when that consistency is below a certain threshold. In other words, the evidence for a word needs to exceed a threshold to drive a saccade to the word's referent.

Because the alternative linking hypothesis I intend to evaluate proposes that some minimum amount of activation is necessary for the listener to implicitly decide to fixate the referent of a word, I attempted to minimize signal-external sources of stimulus activation in this experiment. In previous studies, participants were often asked to study the stimulus pictures before the experiment and the same picture trial set was repeatedly used during the experiment (Allopenna et al., 1998; Dahan et al., 2001b). While this procedure is effective in ensuring that the participants know the words that the pictures are intended to correspond to, it raises significant concerns with linking fixation proportions to TRACE activation levels, as repetition effects are not incorporated into the TRACE models of the task.

Another design choice motivated by the alternative linking hypothesis is that participants in the present study were required to look at the fixation cross at the beginning of each trial. This is not a unique feature of this experiment (see also Allopenna et al., 1998; Dahan et al., 2001b; McMurray, Samelson, Lee, & Bruce Tomblin, 2010; Mirman et al., 2011) as it is commonly used to ensure that looking at any one referent requires a saccade – and therefore a decision to move one's eyes that can be time-aligned with some event in the acoustic signal. Without having a fixation cross, one can look at a picture at 25% chance among four pictures and simply there is 50% of chance that a participant has already fixated at a target picture or a competitor picture at onset of a target word. Studies that do not require participants to look at the fixation cross at the beginning of a trial commonly discard all trials in which the participant happens to already be looking at the target and the competitor, resulting in a significant data loss of 25% of trials (e.g., Huettig & Altmann, 2011 for two pictures on the screen). Dahan et al. (2001b) only discarded trials in which the target picture was fixated at the target onset and then continued to be fixated (4.7%). It could be problematic to include about 36% of trials on which participants already fixated the target or a competitor at target onset. For the present purposes, requiring looks to the fixation cross has the additional advantage of making looks to referents relatively costly: the participant will have to move their eyes back to the fixation point after clicking on a referent. This design therefore makes continuing staring at the fixation cross the behavior with the lowest motor cost, which is sometimes observed in our participants (leading to exclusion of such participants from the experiment). Indeed, the high incidence of this behavior in a pilot experiment led me to move the pictures further apart, so they are more difficult to perceive with peripheral vision. The relatively high separation between the pictures makes the saccades even more costly. When saccades are costly, the lexical activation level necessary to drive a saccade to the referent may be higher, resulting in a greater likelihood of detecting that activation of a word can be insufficient to drive a saccade to its referent. While the present experiment focuses on the influence of degree of bottom-up support / phonological overlap with the target on eye movements, the alternative linking hypothesis proposed in the present study claims that saccades are decisions[2], and that the costs[3] and benefits of a saccade can therefore have a strong influence on eye movement behavior, which opens up a new area for research on eye movements in the visual world (see also Meier & Blair, 2013, whose participants sample the visual features in a search task in a way that minimizes the number of saccades).

2.2. Methods

2.2.1. Participants

Thirty one native speakers of Japanese, all students at the University of Oregon, participated in this experiment. They were either paid or earned course credit for their participation. All of them reported normal hearing and eyesight. Most of the participants

[2] The present study does not define the ‘decision making’ as a conscious decision. Rather eye movement decisions are unconsciously made by a listener when s/he accumulates enough evidence that a word is present in the signal.

[3] Moving eyes is less costly than moving the hand to a picture. However, it is still more costly than doing nothing.

were Japanese college students (*M* = 21 years old) who came to the States for a study abroad program for a few terms to study English (24 out of 30 subjects). Since the study abroad program is set to a few terms, most of the subjects had lived in the States less than a year at the time of the experiment (1-6 months = 18 subjects, 7-12 months = 8 subjects, 12-24 months = 1 subject, longer than 24 months = 3 subjects).

2.2.2. Stimuli

Experiment 1 contained a total of 137 trials consisting of 59 critical trials, 59 control trials, 16 filler trials, and three practice trials. Each trial featured a set of four colored pictures depicting the referents of a target word, a competitor word, and two unrelated words (see Figure 2.1 below) on critical trials, or four unrelated words on filler trials.

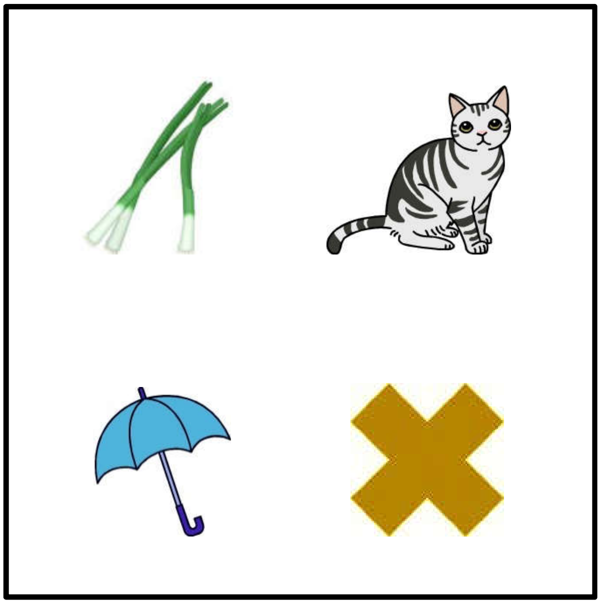


Figure 2.1. Example of a critical trial in Experiment 1 and 2. The target word refers to *negi* “a green onion”, the competitor word refers to *neko* “a cat”, the two unrelated words refer to *batsu* “a cross” and *kasa* “an umbrella”.

Pictures were selected from Google Japan image website to ensure Japanese participants' familiarity with the objects. For the visual stimuli, colored pictures were used. Colors and shapes of the four pictures in a trial set differed from one another to avoid visual similarities that may drive saccades to pictures related to the target visually (Dahan & Tanenhaus, 2005; Huettig & Altmann, 2011; Huettig & McQueen, 2007). The experiment was pilot-tested by native Japanese speakers and ambiguous pictures were replaced with less ambiguous ones.

Table 2.1 displays five phonological conditions differing in the number and location of shared segments between the target word and the competitor word. Target words in critical trials mostly consisted of two mora CVCV words but the filler trials contained 1, 2, or 3 mora words. All critical target words were consonant-initial but filler and control target words occasionally began with a vowel to avoid restricting the participants' expectations to consonant-initial words. Cohort 1 condition comprised 11 trials. Each trial contained a target word and a cohort competitor word that shared the initial consonant with the target word (C _ _ _) and also contained two unrelated words that were phonologically and semantically unrelated to either the target or the competitor. Cohort 2 condition had 12 trials. In this condition, the target word and the competitor word shared the initial mora (CV_ _). Cohort 3 condition was comprised of 9 trials. In this condition, the target word shared the initial three segments (CVC _) with the competitor, which means that the two words mismatched only in the final segment. Rhyme 2 condition contained 12 trials, and a target word shared the final mora with the competitor (_ _ CV). In other words, they differed by the initial mora. Rhyme 3 condition had 11 trials, and each target word shared the final three segments with its competitor (_ VCV), which means that they differed only by the initial segment.

Table 2.1. Conditions varying locations of phonological sharing (See Appendix A for the complete sets).

Conditions	Target	Competitor	Unrelated	Unrelated	Average audio duration
Cohort 1 (11 sets)	*kumo* 'spider'	*kata* 'shoulder'	*batsu* 'x-mark'	*hari* 'needle'	489 ms
Cohort 2 (12 sets)	*nasu* 'eggplant'	*nabe* 'pot'	*kumo* 'cloud'	*tsuru* 'crane'	460 ms
Cohort 3 (9 sets)	*kamo* 'duck'	*kame* 'turtle'	*roba* 'donkey'	*fugu* 'puffer fish'	458 ms
Rhyme 2 (12 sets)	*negi* 'green onion'	*yagi* 'goat'	*hato* 'pigeon'	*kasa* 'umbrella'	449 ms
Rhyme 3 (11 sets)	*futa* 'lid'	*buta* 'pig'	*hana* 'flower'	*maru* 'circle'	453 ms

As Table 2.2 below shows, there was one more critical condition that differed from the ones just mentioned. In this Cohort & Rhyme Mixed condition, two competitor pictures were present on the same trial. A trial included the target, the unrelated distractor and two competitor words; a Cohort 1 competitor (e.g., *hebi* and *hone*) and a Rhyme 3 competitor (e.g., *hebi* and *ebi*). Previous studies including off-line and on-line studies show the rhyme effect for words differ by an initial phoneme (e.g., Allopenna et al., 1998; Connine et al., 1993; McMurray et al., 2010). A competitor word shared the final three segments with a target word, but the initial segment was deleted (e.g., *hebi* and *ebi*). Even in the case of deletion, the rhyme effect may be observed since the portion of overlap still remains as words in the Rhyme 3 condition that differ by the initial phoneme. On the contrary, if the initial segment plays an important role on recognition, we may not observe the rhyme effect for the mis-matching rhyme words, but instead observe a cohort effect for a word matches the initial segment with a target word.

Table 2.2. Conditions varying locations of phonological sharing and having two competitors.

	Target	Competitor 1	Competitor 2	Unrelated	Average audio
		Cohort 1	Rhyme 3		duration
Cohort & Rhyme (4 sets)	*hebi* ‘snake’	*hone* ‘bone’	*ebi* ‘shrimp’	*saru* ‘monkey’	451 ms

The experiment also contained 8 filler trials for which all of the four words were unrelated to one another phonologically and semantically (e.g., *ka* ‘mosquito’, *niji* ‘rainbow’, *hata* ‘flag’, & *tamago* ‘egg’). These trials involved a shorter target word in the critical trials or a longer target word in the control.

The six critical conditions and one unrelated filler (base) condition added up to a total of 67 trials. These same picture sets were used on control trials. On the control trials, the target word was chosen from one of the two unrelated words on the critical trial. This made the target word unrelated to any of the other three words. For instance, the critical trial for the example Cohort 1 condition illustrated in Table 2.1 used *kumo* ‘spider’ as the target word. For the corresponding control trial, the same picture set was used but one of the unrelated pictures, *hari* ‘needle’, became the target word. The participants therefore were exposed to the same picture set twice, first while hearing a critical target word (e.g., *kumo* ‘spider’) and then while hearing an unrelated target word (e.g., *hari* ‘needle’). Critical trials were administered first to avoid pre-exposing participants to the stimulus sets.

In total, participants were exposed to the same trial set twice in the entire experiment. Participants did encounter the same picture more than twice throughout the experiment. For example, a picture of a bear *kuma* was used for two different trials in Cohort 1 and Rhyme 2. The experiment contained a total of 141 pictures. 122 pictures appeared twice in the critical trials and also twice in the control trials, since the picture sets on the critical trials were identical to the control trials. 16 pictures only appeared once in the critical trials and the control trials. Three pictures appeared three times in both trial types. The repeated usage of pictures was necessary to increase the number of word / picture pairs in order to create 67 critical trials within one experiment. Though participants saw each picture multiple times, they heard the name of each picture at most once (if that picture depicted the referent of a target word).

Four words in each trial set typically had the same pitch accent (52 / 67 trials). The Japanese stimulus words were recorded by a female Japanese native speaker who was born and raised in the Tokyo area. The words were presented at the 70dB level. Each word was presented in isolation rather than in a carrier phrase since the carrier phrase creates anticipatory coarticulation providing cues to identify a word before a target word would become available.

The initial phonemes of the experimental words included variety of consonants (e.g., 2 nasals, 5 stops, 2 fricatives, 2 liquids in Cohort 1 condition). They were distributed similarly within a condition to avoid having too many of the same initial consonant within a condition. Note that there are more stops than other initial consonants in Japanese, which yielded more stop-initial words within a condition (see Appendix A for the complete stimulus sets). In addition, overlap in vowels was minimized across words in a trial. For example, one of the Cohort 1 trials consisted of *mame* ‘bean,’ *mikan* ‘tangerine,’*hooki* ‘bloom,’ and *tsuri* ‘fishing,’ the vowel that is shared at the same location across words is only /i/ in *hooki* ‘bloom’ and *tsuri* ‘fishing’ which are both unrelated words. Because the experiment contained many conditions, and the aim was to maximize lexical diversity in the stimulus set rather than restricting the stimuli to a handful of perfectly matched word sets, it was impossible to strictly control word frequency, phonological similarity, semantic similarity (e.g., knife & cutting board), semantic categorical similarity (e.g., animals, tools), pitch accent similarity, shape and color similarity, initial consonant similarity and vowel similarity in each trial. Instead, statistical control for individual characteristics of words was attempted using the random effect of word, accompanied by visual examination of the by-word random intercepts to search for outliers.

2.2.3. Procedure

Participants were tested individually in a quiet room. A participant was seated in front of a computer screen; an eye tracking device and a computer mouse were located in front of the screen. First, participants were guided to put their chin and forehead onto the headrest. The experimenter then calibrated the eye tracker (EyeLink 1000) using a 9-point calibration procedure focusing on the participant’s right eye. After appropriate adjustment and calibration, participants were instructed to listen to isolated words through the headphones they wore and then click on a picture on the screen which matched the word they heard. To ensure that participants were in a Japanese language processing mode, all instructions were presented in Japanese, and the experiment was conducted by the first author, a native Japanese speaker (cf. Canseco-Gonzalez et al., 2010). There were three blocks in the experiment: practice trials, critical trials, and control trials in sequential order. The trials and locations of pictures in each block were randomized. On each trial, the set of the four pictures appeared for 1000 ms; then a red fixation cross was presented at the center of the screen along with the four pictures. Once the participant had fixated the red cross, a target word was played through the headphones. The participants did not study any of the pictures prior to experiment. There was no time limit for the participants to select the picture. The eye position was recorded, and the identity of the interest area it fell into identified, by EyeLink every millisecond.

Whereas several studies have ensured that the participants name pictures in the intended way by pre-training the participants on the intended names of the pictures, it was considered important, given the aims of the present study, to avoid pre-training in order to reduce the likelihood of exaggerating lexical competition effects. For this reason, the present study ensured that participants named the pictures in the intended way by asking them to name the pictures (one by one) after the experiment was completed. Previous studies without pre-training have not asked participants how they named the pictures; therefore experimenters never knew if pictures in their experiment were named as experimenters intended (Dahan & Gaskell, 2007; Dahan et al., 2001a; Dahan et al., 2001b; McMurray et al., 2010; Mirman et al., 2011). Though some of these studies normed the experimental pictures using a different group of subjects (Dahan & Gaskell, 2007; Dahan et al., 2001a; Mirman et al., 2011), and saw 90% between-subject agreement in naming, it is impossible to know for sure whether the actual experimental participants would name the pictures in the intended way and therefore whether the competitor words were indeed overlapping with the target as much or as little as intended. Given the focus on amount of overlap in this study, it was felt that name agreement is crucial.

2.2.4. Data Processing

Trials were excluded from data analysis when participants named pictures differently than expected, and the difference affected phonological similarity relations in a trial. For example, naming a picture of a boot *buutsu* as a shoe *kutsu* changed the phonological similarity relations of a trial in which the word *buutsu* was intended to be a cohort competitor of the target word *buta* 'pig' ineligible to be included in the analysis. In such a case, the trial was excluded. However, not all of the different namings affected the phonological similarity relations of a trial. For example, naming a picture of a killer whale *shachi* as a dolphin *iruka* did not affect the phonological similarity relations of a trial on which the target word was *hato* 'pigeon' and the competitor was *hebi* 'snake' because both *shachi* and *iruka* were unrelated to both the target and the competitor and were therefore appropriate as distractors. On average, two trials per participant had to be excluded on this basis. One participant was excluded from data analysis due to a high number of unexpected picture names (43 / 141 words). Table 2.3 is a summary of trial exclusion in the experiment. As mentioned above, one subject was excluded from the analysis which yielded 134 trials (3.23%). 69 trials (1.66%) were excluded due to unintended picture namings. 269 trials (6.48%) were excluded due to absence of looks to target pictures during the window of audio onset to the mouse click. After these exclusions, 88.64% of the trials (3682 / 4154) remained. Note that there were no recognition errors (clicks on an incorrect picture).

Table 2.3. Summary of trial exclusions.

	Excluded subjects	Excluded trials due to wrong naming	Excluded trials due to absence of looks to target pictures in the analysis window	Total # of included trials in the experiment
Number of trials	134 (1 subject)	69	269	3682
%	3.23 %	1.66 %	6.48 %	88.64 %

2.2.5. Analysis

The data for each trial was aggregated in 20 ms intervals. The dependent variable was then the proportion of time within each interval spent fixating a particular picture. Data were analyzed using growth curve analysis (Mirman, 2014; Mirman, Dixon, & Magnuson, 2008). The model consisted of a fixed effect of Picture Type (Competitor vs. Distractor) interacting with time represented as a weighted sum of fourth-order orthogonal polynomials (centered time, time^2, time^3 and time^4), random intercepts for Items (trials) and Subjects and random slopes for Picture Type by Subject. Time window for the analysis is from 200 ms to 1000 ms in 20 ms time interval. 200 ms is usually considered to be the minimal time required for planning a saccade (Matin, Shao, & Boff, 1993, though cf. Altmann (2011). Fixation proportion of target pictures reached maximum at 1000 ms after target onset on average. Average time to click a target picture was 1242 ms for Experiment 1 (and 1630 ms for Experiment 2).

The competitor item is related to the target whereas the distractor is not. The crucial effects of interest are therefore the difference between looks to the competitor and the distractor and its interactions with time on critical trials. If looks to a competitor significantly outnumber looks to the distractor in a particular time interval, evidence that the competitor indeed competes with the target for recognition during that time period is obtained. The present study refers to significant evidence for competition between cohort competitors as the 'cohort effect' and significant evidence for competition between rhyme competitors as the 'rhyme effect'.

The alternative linking hypothesis crucially predicts null cohort effects for small amounts of overlap between the target and the competitor. To quantify evidence for the null hypothesis (no effect), a Bayesian hypothesis test was performed. The test estimates the Bayes Factor for the alternative hypothesis H_1 (a non-zero effect of a predictor) and the null hypothesis H_0 (no effect) by comparing the BIC (Bayesian Information Criterion) values of a model that includes the predictor and one that excludes it (Wagenmakers, 2007). For the analysis, looks to Competitor pictures and Distractor pictures of a certain type (e.g., Cohort 1) were examined. The model embodying the alternative hypothesis included the predictor Picture Type (Competitor or Distractor) while that embodying the null hypothesis did not. Supporting the null hypothesis means that it is more probable than not, given the data, that looks to a picture are unaffected by whether the name of the picture is phonologically similar to the target word presented on a trial; i.e., whether there is evidence for the name in the acoustic signal. As shown by Wagenmakers (2007), there is a direct relationship between the BIC difference between two models and the Bayes Factor, under the assumption that the two models are equally probable a priori, which allows us to use the probability of the data given a model instead of the probability of the model given the data. In the current case, the two models are nested such that there is a null model (H0) and an alternative model (H1), which differs from it by the addition of a parameter. The BIC difference between the two models ($\Delta\mathrm{BIC}_{10}$) is then the BIC of the null model subtracted from the BIC of the alternative model. According to Wagenmakers (2007), the relationship between the BIC difference and the Bayes Factor is as follows, where *D* is the observed human data, and *BF* is the Bayes Factor,

$$BF_{01} \approx \frac{\mathrm{Pr}_{\mathrm{BIC}}\,(D|H_0)}{\mathrm{Pr}_{\mathrm{BIC}}(D|H_1)} = \exp(\Delta\mathrm{BIC}_{10}\,/2)$$

The estimated BF value was then converted into the posterior probability of H_0 (see below) given the experimental data (D):

$$Pr_{BIC}\,(H_0|D) = \frac{BF_{01}}{BF_{01}+1}$$

The posterior probability of H_0 was interpreted using the heuristic degree of evidence cut-offs provided by Raftery (1995) (Table 2.4).

Table 2.4. Explanation of posterior probability by Raftery.

Bayes Factor BF_{01}	Pr (H_0 | D)	Evidence
1 – 3	.50 - .75	weak
3 – 20	.75 - .95	positive
20 -150	.95 - .99	strong
> 150	> .99	very strong

2.3. Results

Figure 2.2 demonstrates raw data, mean fixation proportions for each picture type (Target, Competitor, & Distractor) averaged across subjects and trials in a specific condition.

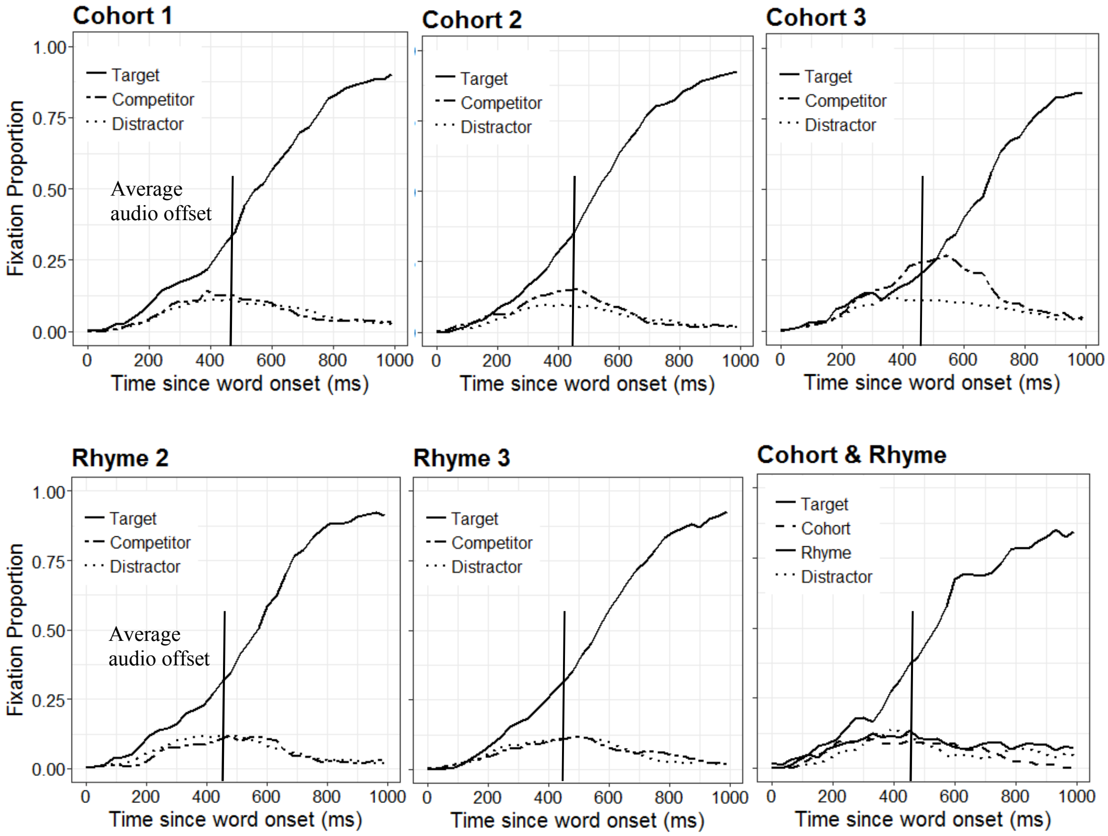


Figure 2.2. Fixation proportion of targets, competitors, and distractors for each condition.

Visual inspection of the plots suggests that there may not be a competitor effect except in the Cohort 3 condition and perhaps the Cohort 2 condition. Competitor pictures were fixated more than distractor pictures for those two conditions while both competitor pictures and distractor pictures were fixated equally for the other conditions. In addition, looks to referents of targets and non-targets diverged early, as soon as 200 ms after word onset. This divergence appeared to occur earlier in the Cohort 1 (100 ms) condition than in the Cohort 2 (200 ms) condition, and even earlier in the Rhyme 2 (50 ms) condition, where the competitor did not share the beginning with the target word, indicating that the cohort-target overlap may influence looks to the target. However, the target quickly diverged from *both* the unrelated distractors and the related competitors, except in the Cohort 3 condition.

The growth curve analysis indicated that there was no significant difference in looks between the competitor picture and the distractor picture in the Cohort 1 condition (sharing the initial single segment, e.g., *kumo* ‘spider’ and *kata* ‘shoulder’) (*Estimate* = -.035, *SE* = .163, *p* = .829) and the Cohort 2 condition (sharing initial two segments, e.g., *nasu* ‘eggplant’ and *nabe* ‘pot’) (*Estimate* = -.230, *SE* = .216, *p* = .285). The posterior probability of the null hypothesis for both the Cohort 1 and Cohort 2 conditions was close to .99, which constitutes very strong evidence for the null hypothesis (no effect). This indicates that sharing the initial one or two segments with the target did not increase the likelihood of fixating the referent of a word in the present experiment.

This is also true for the Rhyme conditions and the Cohort and Rhyme Mixed condition. There was no significant difference in looks between competitor pictures and distractor pictures for the Rhyme 2 condition (e.g., *negi* ‘green onion’ and *yagi* ‘goat’) (*Estimate* = .228, *SE* = .141, *p* = .107) or the Rhyme 3 conditions (e.g., *futa* ‘lid’ and *buta* ‘pig’) (*Estimate* = -.035, *SE* = .183, *p* = .847). Subjects fixated the referent of an unrelated word as frequently as the referent of a word that shared a few segments with the target. This means that sharing the final 2 or even 3 segments with the target did not generate enough activation to draw an eye movement to the competitor.

There was also no significant difference in looks between competitor pictures and distractor pictures in the Cohort and Rhyme Mixed condition, where the cohort competitor (Cohort 1) and the rhyme competitor (Rhyme 3) were present on the same trial (e.g., *hebi*, ‘snake,’ *hone* ‘bone,’ and *ebi* ‘shrimp’) (Cohort: *Estimate* = .254, *SE* = .343, *p* = .459, Rhyme: *Estimate* = .189, *SE* = .340, *p* = .578). The Bayes Factor analysis very strongly supports the null hypothesis (no competitor effects) for the mixed condition (posterior probability of the null hypothesis >.99). As in the Cohort 1 and the Rhyme 3 conditions (with only one competitor), the data in the Cohort 1 & Rhyme 3 Mixed condition strongly suggest that the name of a picture sharing the initial one segment or the final three segments with the presented word did not influence eye movements to the picture.

However, three initial segments appeared to provide sufficient lexical activation: there was a cohort effect in the Cohort 3 condition. When the competitor shared the initial three segments with the target (e.g., *kamo* ‘duck’ and *kame* ‘turtle’), the competitor picture was fixated significantly more than the distractor picture (*Estimate* = -.901, *SE* = .209, *p* < .001), suggesting that words sharing three initial segments competed each other for recognition. In addition, a significant difference in the quadratic time term indicated a greater curvature in the trajectory of looks to the cohort picture that was not there in looks to the distractor picture (*Estimate* = .791, *SE* = .079, *p* < .001).

Figure 2.3 illustrates the very early divergence between target and other pictures. Figure 2.3 shows the mean time (in milliseconds) spent fixating each picture for each 20 ms time interval in the 200-400 ms time window for all the conditions except Cohort 3, where the looks indicate equal consideration of the target and cohort competitors early on. Looks to target pictures quickly diverged from looks to the competitor pictures and the distractor pictures (around 200 ms). There was a significant difference in looking proportions between the target picture and the competitor picture (*Estimate* = -1.331, *SE* = .249, *p* < .001) as well as between the target picture and the distractor picture (*Estimate* = -1.285, *SE* = .223, *p* < .001). This indicates that the target quickly diverged from *both* the unrelated distractors and the related competitors, except in the Cohort 3 condition.

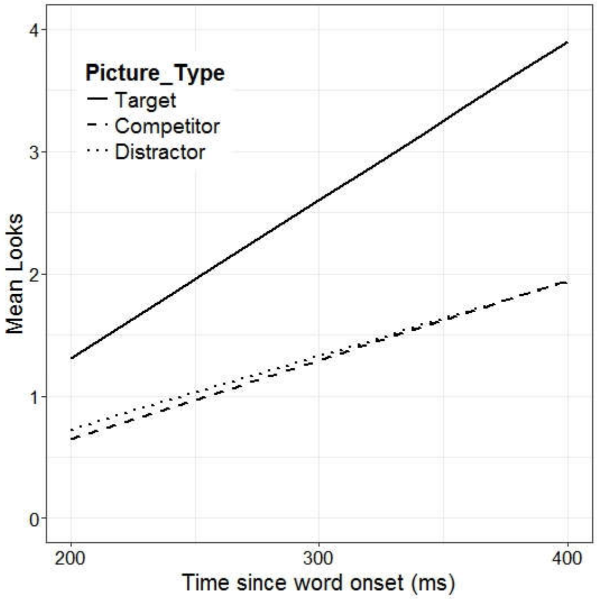


Figure 2.3. Mean time spent fixating targets (solid line), competitors (dashed line), and distractors (dotted line) in each 20 ms time interval for the 200-400 ms time window.

2.4. Discussion

Previous research has suggested that spoken words are processed incrementally, with lexical representations and even the associated semantics activated as the spoken form of the word is being perceived, with only a constant 200 ms delay for programming eye movements (e.g., Matin et al., 1993). In theory, overlap with any part of a target word can lead to activation of a candidate word during processing (e.g., Allopenna et al., 1998; McClelland & Elman, 1986; Norris, 1994), even when the shared parts are not initial. Furthermore, these activations are thought to be directly reflected in eye movements to pictures of referents 200 ms later (Allopenna et al., 1998; Tanenhaus et al., 2000). Thus, some studies have reported effects for ‘sub’ vs. ‘bus’ (Toscano, Anderson, & McMurray, 2013) or rhyme sharing (Allopenna et al., 1998; Dahan et al., 2001a; 2001b; Simmons & Magnuson, 2018). Experiment 1 asked whether any amount of bottom-up information is sufficient to trigger a saccade to a visual representation of the word’s referent when top-down activation of competitors is minimized. In addition, it asked how immediately a saccade occurs. Previous studies did not control the amount of overlap and have not systematically explored how much segmental overlap is necessary to observe lexical competition in the visual world paradigm. Experiment 1 was intended to fill this gap by systematically varying the amount of segmental overlap between the target word and its competitors, as well as whether these competitors share the beginning or the end with the target. Experiment 1 found that looks to a picture boosted when the name of the picture shared the three initial segments (the initial CVC) with the presented word. However, final overlap had no effect, and neither did shorter initial overlap. The results suggest either that: 1) lexical semantic representations are activated by Japanese listeners only after three segments of a spoken disyllabic word are perceived, which appears unlikely, or 2) relatively strong activation is required to drive an eye movement in the present task. The results provide no evidence for an influence of the mora on spoken word recognition: participants’ eye movements were affected by overlap in the initial three segments of a CVC but not in the initial two segments, CV, with a significant difference in the looks to the competitor in the two conditions. Mora-by-mora recognition would instead predict that there should be a significant effect of CV (one-mora) overlap with no additional effect of additional submoraic segmental overlap in the Cohort 3 condition (Cutler & Otake, 2002). The present results are therefore inconsistent with mora-by-mora recognition, and no additional phonological units (e.g., mora) therefore appear to be necessary to include in the TRACE model for Japanese.

The results of the present experiment are in fact not inconsistent with the classic study by Allopenna et al. (1998), which first documented lexical competition in spoken word recognition using the visual world paradigm. Seven of the 8 stimuli presented by Allopenna et al. (1998) to their participants, except for the famous *beaker-beetle* example, involved overlap in three or more initial segments. Likewise, most stimuli in Dahan, Tanenhaus, & Chambers (2002) involve substantial initial overlap involving more than two segments. Cohort effects in Dahan et al. (2001a), Creel, Aslin, and Tanenhaus (2008), and Canseco-Gonzalez et al. (2010) are somewhat more problematic because most of their stimuli showed two-segment overlap. However, as noted above, the procedures adopted in previous work have involved pre-exposing participants to the words included in the experiment, which may have increased the extent of lexical competition compared to the present experiment.

Amount of shared information can be counted as length or proportion of overlap. Thus, we could consider one-segment overlap in the present study as an instance of 25% cohort overlap. One might argue that proportion of overlap may matter more than the number of shared segments (Kapatsinski, 2005; Simmons & Magnuson, 2018). For example, the proportion of overlap differs between monosyllabic words and disyllabic words when two phonemes overlap. The proportion of overlap can be 66% for CVC words like ‘cat’ and ‘cap’ and only 40-50% for words for CVC(V)C words like ‘beaker’ and ‘beetle’ though for both cases, two phonemes are shared. Greater overlap could therefore show a stronger cohort effect because it results in a greater relatedness proportion. However, both the present study and the study by Allopenna et al. (1998) examined disyllabic words with 50% overlap (Cohort 2 condition in the present study). Yet, the present study did not exhibit the cohort effect that was observed in Allopenna et al.’s study. The stimulus words in their study consisted of five to six segments and shared two to three segments whereas the words in the present study consisted of four segments and shared two segments, thus exhibiting greater proportion of overlap. Thus ‘how much’ bottom-up information is needed to activate candidate words enough to evince an eye movement is likely affected by other factors such as the amount of top-down activation in the experiment. As the present experiment minimized top-down activation of lexical candidates, it is unsurprising from the interactive activation perspective of models like TRACE that the cohort effects are weaker than in Allopenna et al.’s study, which used pre-training and repetition of words throughout the experiment.

At first glance, lack of overlap effects in conditions other than Cohort 3 appears inconsistent with theories claiming that any part of a target word can activate multiple candidate words during processing (e.g., TRACE; McClelland & Elman, 1986). A small or non-initial part of the target word (Cohort 1 & 2, Rhyme 2 & 3, and Mixed competitor) did not seem to activate candidate words in Experiment 1. However, several eye tracking studies using the visual world paradigm demonstrated a cohort effect as well as a rhyme effect, suggesting that both word-initial and non-initial acoustics activate the words that contain them (Allopenna et al., 1998). Furthermore, there are good theoretical reasons for spoken word recognition to work this way. Given that any part of a spoken word can be obscured by noise, misperceived or mispronounced, word recognition needs to be robust enough to recognize the word using any set of partial cues (Salasoo & Pisoni, 1985). Furthermore, given that words vary in length and duration, and there are few clear acoustic cues to word boundaries in continuous speech, the listener cannot in general wait until they hear a certain number of segments from such a boundary before they start entertaining lexical hypotheses (McClelland & Elman, 1986; Norris & McQueen, 2008). The results of the present study can be reconciled with these theoretical considerations, as embodied by the TRACE model, and with the results of previous studies as long as a certain level of activation is necessary to drive a saccade to the referent of a word. Pre-exposure to and repeated presentation of the words in previous studies may have served to increase their activation levels enough to drive saccades to their referents with a lower degree of bottom-up support from the signal (i.e., lower overlap with the presented target word). The present study does not conclude that listeners will always need three initial segments of a word to decide to look at a picture of its referent. Rather, it proposes that the listener needs to accumulate evidence for a word before a saccade is generated, i.e., there is a threshold below which the word's activation is not high enough to drive a saccade and will not be reflected in the eye tracking record.

As in previous studies, looks to referents of targets and non-targets diverged early, as soon as 200 ms after word onset. This divergence appeared to occur earlier in the Cohort 1 (100 ms) condition than in the Cohort 2 (200 ms) condition, and even earlier in the Rhyme 2 (50 ms) condition, where the competitor does not share the beginning with the target word, indicating that the cohort-target overlap does influence looks to the target. However, the target quickly diverged from *both* the unrelated distractors and the related competitors, except in the Cohort 3 condition (Figure 2.2). That is, participants' willingness to fixate any signal-consistent referent was consistently affected by the presence of more than one such referent on screen, but words corresponding to some referents were not activated enough to draw eye movements to themselves. They merely decreased the likelihood of a saccade from the fixation cross to the target referent.

Most researchers in the visual world paradigm have assumed that 200 ms is the minimum time necessary to plan a saccade and that therefore effects observed 200 ms after stimulus onset reflect processing of the very beginning of the auditory stimulus (e.g. Allopenna et al., 1998; Tanenhaus et al., 2000); but cf. Altmann (2011). The divergence between the looks to the target and looks to other stimuli therefore appears to be driven by the very beginning of the word. Yet, signal support for three initial segments appears to be necessary to drive a saccade to a referent in the present experiment. One may therefore wonder how looks can be driven by the identity of the initial CVC, when that CVC has only started to be articulated.

The likely explanation for the early divergence between target and competitors in the Cohort 1 and 2 conditions is coarticulation. Because of coarticulation inherent to natural speech, the beginning of a CVCV stimulus will provide information about its end and, in particular, the second consonant. Therefore, the listener may perceive the identity of the initial three segments of the target from the very beginning of the target stimulus, resulting in immediate suppression of looks to cohort competitors that do not share the second consonant with the target, and are therefore inconsistent with the speech signal. It is known from previous studies that listeners do utilize coarticulatory information in word recognition, and that this information can drive saccades in the visual world paradigm (Beddor, McGowan, Boland, Coetzee, & Brasher, 2013; Dahan et al., 2001b; Salverda, Kleinschmidt, & Tanenhaus, 2014; Tobin, Cho, Jennett, & Magnuson, 2010). Given these observations, listeners may not need to hear the initial 3 segments (e.g., /kam/) to trigger an eye movement to the visual representation of an auditory word (e.g., *kame*). Rather, listeners may be actually making an eye movement upon hearing the initial 1 or 2 segments (e.g., /k/ or /ka/), as long as the initial 1 or 2 segments provide coarticulatory information identifying the following segment of the word (e.g., /kam/). This would still suggest that the signal needs to provide evidence for multiple segments to trigger a saccade to the word's referent. The early divergence (200 ms) between looks to the target and the others (Figure 2.3, page 42) would, however, be explained by coarticulation. One could also argue that looks to the target were suppressed by uncertainty regarding its identity caused by the presence of competitors. Even though the activation of competitors was insufficient to drive a saccade to their referents, it could have inhibited, and therefore suppressed saccades to, the target.

To reduce coarticulation effects, Experiment 2 replicated the present experiment using auditory stimuli produced with a diphone synthesizer (MBROLA, Dutoit, Pagel, Pierret, Bataille, & Vrecken, 1996). The diphone synthesizer limits coarticulation to adjacent segments. This means that listeners in Experiment 2 will NOT receive information about the second consonant from the very beginning of the auditory stimulus. If a CVC match is necessary for a Japanese listener to look at the referent of a word in the present experiment, looks to both targets and Cohort 3 competitors should be delayed (i.e., later than 200 ms) in synthesized speech.

Furthermore, if the acoustic signal needs to provide information consistent with the initial CVC of a word to drive saccades to its referent, participants in Experiment 2, just like participants in Experiment 1, should continue to fixate Cohort 1 and Cohort 2 competitors with the same probability as unrelated distractors, despite the beginning of the synthesized stimuli providing no disambiguating information that would allow the listener to rule these lexical candidates out.

On the other hand, because synthesized speech does not match the listeners' stored representations of spoken words as well as natural speech does, there may be increased uncertainty about the speech signal. If eye movements are truly "promiscuous" (Tanenhaus et al., 2000), and the listener looks at the referents of all words that are consistent with the signal, one might expect this uncertainty to lead to increased looks to non-target referents. In particular, looks to cohort competitors may have been *suppressed* in the Cohort 1 and 2 conditions of Experiment 1 by the presence of early coarticulatory information about the end of the target word (e.g., the second consonant in CV$\underline{C}$V). That is, listeners may have looked at all referents consistent with the signal, but the signal was not consistent with cohort competitors early on due to the presence of coarticulatory information. If participants are able to recognize the target word early on, the target word may not be confusable with a competitor word that shares the initial segments with the target. If looks to the Cohort 1 and Cohort 2 competitors fail to occur in Experiment 1 because they are suppressed by coarticulatory information about the end of the word, participants may look at Cohort 1 and Cohort 2 competitors when the acoustic signal in the initial CV transition is as consistent with these competitors as with the target. The absence of coarticulatory information about the end of the target word in Experiment 2 should then increase looks to Cohort 1 and Cohort 2 competitors, allowing cohort effects to emerge with shorter overlap. If, on the other hand, the signal needs to provide information identifying the initial CVC of a Japanese word to drive a saccade to the referent of that word, then we should not observe Cohort 1 and Cohort 2 effects in Experiment 2, replicating the results of Experiment 1.

CHAPTER III

EXPERIMENT 2: SYNTHESIZED SPEECH STIMULI

The work presented in this chapter is also reported in a co-authored article invited for resubmission to the journal *Language, Cognition and Neuroscience*

3.1. Introduction

As discussed above, Experiment 2 examines word recognition in speech synthesized using diphone concatenation, to evaluate whether elimination of long-distance coarticulation would affect looks to competitors and, if so, whether it would increase or decrease looks to cohort competitors compared to natural speech. Synthesized speech increases ambiguity in the acoustic signal, which could have two consequences. If listeners look at everything that is consistent with the signal – i.e., eye movements are maximally promiscuous – then there should be more looks to cohort competitors in synthesized speech compared to natural speech. In synthesized speech, the beginning of the stimuli provides no information that favors the target over the cohort competitors, which therefore predicts that the listener should be equally likely to look at both. However, the alternative linking hypothesis suggests that that equal likelihood should be near zero: participants should continue looking at the fixation cross until there is *enough* evidence for a particular word being present in the acoustic signal. Given the results of Experiment 1, enough evidence in the present population and task – with its costs and benefits and lack of repetition – means enough to identify the second consonant in a CVCV. Since the information about the second consonant is delayed by diphone synthesis, we also expect that looks to the target will diverge from looks to unrelated distractors (and insufficiently related competitors) later than in Experiment 1. Previous work using synthesized speech has provided evidence consistent with this prediction (Farris-Trimble, McMurray, Cigrand, & Tomblin, 2014; McMurray, Farris-Trimble, & Rigler, 2017). However, this work examined speech produced using a cochlear implant simulator, which produces highly degraded spectra, drastically reducing segment discriminability, but does not affect coarticulation. The present study instead reduces coarticulation while leaving non-coarticulatory spectral cues to segments largely intact.

3.2. Methods

3.2.1. Participants

Thirty seven Japanese students at the University of Oregon, native speakers of Japanese, participated in this experiment. They were either paid or earned course credit for their participation. All of them reported normal hearing and eyesight. The participants did not take part in Experiment 1. As in Experiment 1, most of the participants were Japanese college students (*M* = 21 years old) who came to the States for a study abroad program for a few terms to study English (20 out of 35 subjects). Most of the subjects had lived in the States less than six months at the time of the experiment (1-6 months = 26 subjects, 7-12 months = 6 subjects, longer than 24 months = 4 subjects).

3.2.2. Stimuli

The words, pictures, and phonological conditions were identical to Experiment 1 except the audio was resynthesized from the natural speech audio used in Experiment 1. Synthesis was performed to reduce a long-distance coarticulation, vowel-to-vowel (V-to-V) coarticulation, in CVCV because that coarticulation may help listener to anticipate the end of the target word early. Synthesis was performed using a diphone synthesizer, MBROLA (Dutoit et al., 1996). Because diphone synthesis blends diphones (i.e., C-V-C-V), it removes long distance coarticulation (i.e., there is no coarticulation between the first V and the second V or the first C and the second C; the cues to the second consonant first emerge during the second half of the first vowel). The input to MBROLA included the duration of each segment in each word from Experiment 1 and the word's pitch contour, with points measured in increments comprising 5% of each segment's duration. The 'jp 2' voice (a female speaker of Japanese that the program listed) was used. To fine-tune the audio for naturalness, pitch was then manually adjusted by the author. Ten ms of silence was added before and after a word. The naturalness of resynthesized words was rated by 11 native listeners of Japanese. Each listener rated words on a percentage scale, 100% representing the audio being as similar as possible to natural speech. The naturalness of resynthesized speech was 74% on average, indicating that synthesized speech is somewhat degraded relative to natural speech. However, word recognition accuracy in the eye tracking study was 100%, indicating that enough cues are eventually provided by synthesized speech for all words to be accurately identified and the competitors to be successfully ruled out.

3.2.3. Procedure

Procedure was identical to Experiment 1.

3.3.4. Data Processing and Analysis

The criteria for discarding trials were the same as Experiment 1. Table 3.1 is a summary of trial exclusion in the experiment. Ninety five trials (1.92%) were excluded due to unintended picture names and two participants (268 trials, 1.92%) were excluded due to a technical problem with the eye tracker, which failed to record fixations for half of these participants' trials. 241 trials (4.86%) were excluded due to absence of looks to target during the window of audio onset to the mouse clicking. After these exclusions, 87.84% of the trials (4355 / 4958) remained. The statistical analysis was the same as in Experiment 1.

Table 3.1. Summary of trial exclusions.

	Excluded subjects	Excluded trials due to unintended picture naming	Excluded trials due to absence of looks to target pictures in the analysis window	Total # of included trials in the experiment
Number of trials	268 (2 subjects)	95	241	4355
%	5.41	1.92	4.86	87.84

3.3. Results

Figure 3.1 demonstrates raw fixation proportion data for each picture type (Target, Competitor, & Distractor) averaged across subjects and trials in a specific condition. The patterns of fixation proportions appear largely similar to those in natural speech data except for the Cohort & Rhyme Mixed condition where the rhyme pictures were fixated more than in natural speech.

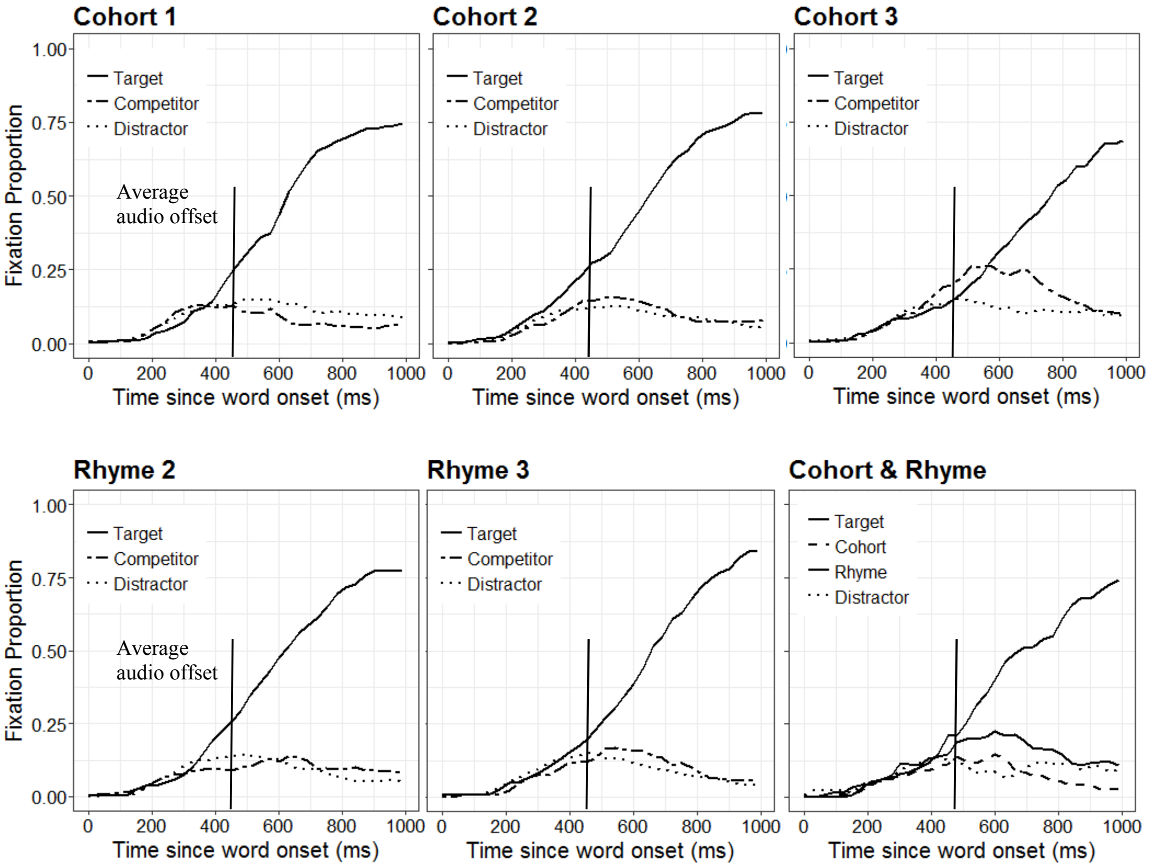


Figure 3.1. Fixation proportion of targets, competitors, and distractors for each condition.

Growth curve analysis demonstrated that, as in Experiment 1, there was no cohort effect in the Cohort 1 condition (*Estimate* = .481, *SE* = .277, *p* = .082) and the Cohort 2 condition (*Estimate* = -.21, *SE* = .333, *p* = .529). As in Experiment 1, Bayesian analysis indicates very strong evidence for the absence of a cohort effect in the Cohort 1 and 2 conditions (posterior probability of the null hypothesis >.99).

As in Experiment 1, there was a cohort effect in the Cohort 3 condition (initial three segments shared). Competitor pictures were fixated significantly more than distractor pictures (*Estimate* = -.893, *SE* = .232, *p* < .001), indicating that words sharing the initial three segments with the target were activated enough to drive saccades to their referents. In addition, a significant difference in the quadratic time term indicates there was a difference in the timecourse of looks between distractors and Cohort 3 competitors (*Estimate* = .965, *SE* = .074, *p* < .001).

As for the Rhyme 2 & 3 conditions, there was no significant difference in fixations between competitor pictures and distractor pictures for Rhyme 2 (e.g., *negi* and *yagi*) (*Estimate* = .038, *SE* = .147, *p* = .796) and Rhyme 3 (e.g., *futa* and *buta*) (*Estimate* = -.007, *SE* = .156, *p* = .963) conditions. The Bayes factor analysis also provide very strong evidence for the absence of an effect of final overlap (posterior probability of the null hypothesis >.99). Subjects fixated the referent of an unrelated word as frequently as the referent of a word that shared a few segments with the target.

There are a few differences between Experiments 1 and 2. First, as expected, the probability of looking to the signal-consistent referents diverged from the probability of looking at signal-inconsistent referents later when the listener is presented with synthesized speech. Second, there was a rhyme effect in Experiment 2, but only in the Cohort & Rhyme Mixed condition.

As noted above, later divergence was expected between looks to the target and looks to other referents in synthesized speech. In Experiment 1, looks to target pictures diverged from looks to other pictures around 200 ms after target onset (Figure 2.3, page 42) whereas in Experiment 2 the divergence occurred around 400 ms (Figure 3.2 below). Figure 3.2 on the left (Experiment 2) illustrates the mean time (in milliseconds) spent fixating each picture for each 20 ms time interval in the 200-400 ms time window for all conditions except Cohort 3. In response to synthesized speech, looks to target pictures did not diverge from looks to competitor and distractor pictures until ~400 ms after stimulus onset (left panel), although they quickly diverged from looks to other pictures in natural speech (right panel). Unlike in Experiment 1, there was no significant difference in looks between the target picture and the competitor picture (*Estimate* = -.204, *SE* = .249, *p* = .41) as well as between the target picture and the distractor picture (*Estimate* = -.026, *SE* = .192, *p* = .89) in the 200-400 ms time interval in the current experiment (left panel). The Bayesian hypothesis tests provided very strong evidence for the absence of an effect of Picture Type (posterior probability of the null hypothesis >.99). All the pictures were fixated equally often in the 200-400ms time window.

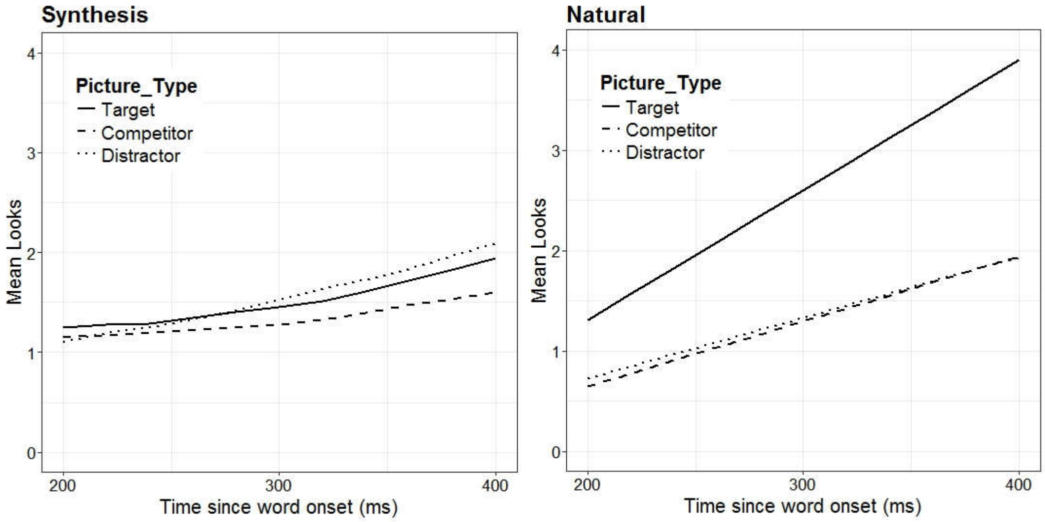


Figure 3.2. Mean time spent fixating targets (solid line), competitors (dashed line), and distractors (dotted line) in each 20 ms time interval for the 200-400 ms time window.

Furthermore, there was a significant difference in looks to target pictures for the 200-400ms time window between natural speech and synthesized speech (*Estimate* = -1.153, *SE* = 0.571, *p* < 0.05). Figure 3.3 shows that target pictures were looked at significantly more in natural speech than in synthesized speech, implying that there was an early divergence in natural speech and late divergence in synthesized speech. Diphone synthesis eliminated coarticulatory cues to the second consonant (CVCV) from the first consonant (CVCV) and the first half of the first vowel (CVCV). As a result, saccades to referents whose names contained that second consonant were delayed by about the same time. This result is consistent with the speech signal driving participants' looks to a picture in both experiments only if the speech signal provided evidence for the initial three segments of the picture's name.

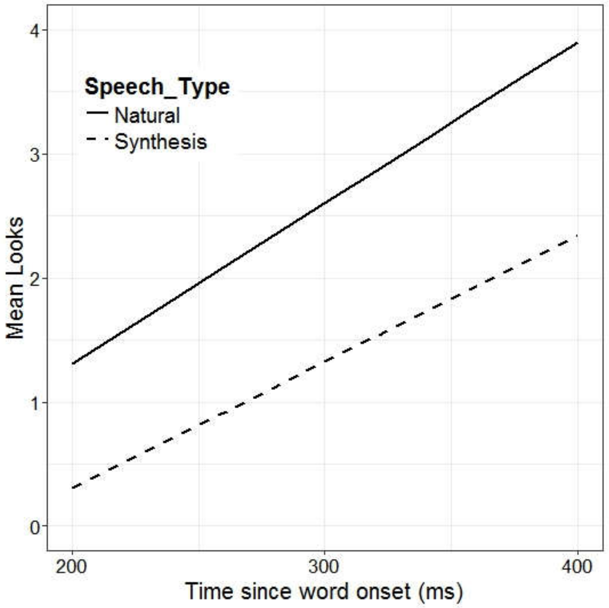


Figure 3.3. Mean time spent fixating targets in natural speech (solid line) and in synthesized speech (dashed line) in each 20 ms time interval for the 200-400 ms time window.

In addition, there was also a delay in the reduction of looks to the competitor in the Cohort 3 condition for synthesized speech (See Figure 3.1, page 55). The difference concerns the timepoint at which looks to cohort competitors and targets begin to diverge. For natural speech, both the target word and the competitor word were fixated equally until ~500 ms after target onset, at which point looks to cohort competitors started to decline (Figure 2.2, page 39). However, for synthesized speech, both words were fixated until ~600 ms (Figure 3.1, page 55). The realization point (600 ms) is about 140 ms after the target word offset. The divergence between the target and the Cohort 3 competitor is driven by cues to the final vowel of the target (CVC<u>V</u>). When these cues are delayed in synthesized speech, the divergence is delayed as well.

Figure 3.4 below shows a comparison between natural speech and synthesized speech for looks to competitor pictures in the Cohort 3 condition from 200 ms to 800 ms. As just discussed, the peak of the looks to the competitor for natural speech is at about 500 ms after target onset while the peak for synthesized speech is at about 600 ms.

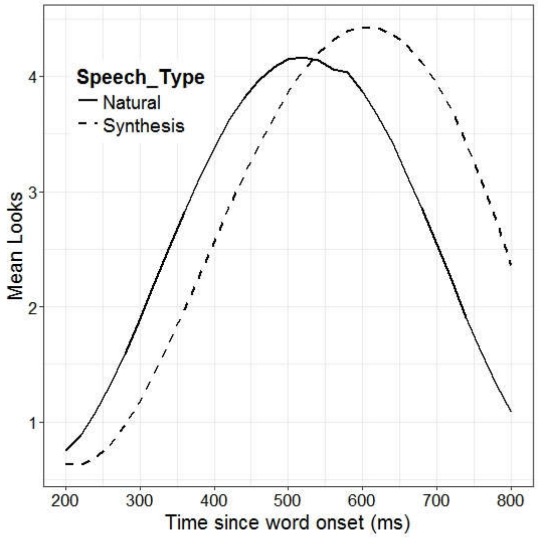


Figure 3.4. Mean time spent fixating competitors for natural speech (solid line) and synthesized speech (dashed line) in each 20 ms time interval.

Significant effects on the quadratic term between the natural speech and the synthesized speech indicates that there was a steeper peak in response to natural speech (*Estimate* = -.43, *SE* = .12, *p* < .001). This result can be attributed to the greater clarity of natural speech, which means that both the cues that favor the cohort competitor and the target over the unrelated distractor and those that favor the target over the cohort competitor are clearer in natural speech. The clear cues that distinguish the cohort competitor from the distractor may result in greater listener confidence that the cohort competitor is present until the cues that distinguish it from the target are perceived. The greater clarity of the lat(t)er cues then produces greater listener confidence that the cohort competitor is absent from the signal, resulting in a steeper rise and fall of looks to the cohort referent.

Another difference between the experiments was that there was a rhyme effect when two competitors were present in the same trial (Cohort & Rhyme Mixed) in Experiment 2 (synthesized speech). The difference in response patterns between Experiments 1 and 2 can be observed if one compares the rhyme curves in Figures 2.2 and 3.1. For the Cohort & Rhyme Mixed condition (e.g., *hebi*, *hone*, and *ebi*) in Experiment 2, cohort pictures were fixated equally as distractor pictures (*Estimate* = -.166, *SE* = .238, *p* = .486) as was seen in Experiment 1. However, rhyme pictures were fixated significantly more than distractor pictures (*Estimate* = .729, *SE* = .327, *p* = .026), suggesting that words sharing three final segments competed each other for recognition. The quadratic time term indicating a greater curvature in the trajectory of looks to the rhyme picture that was not there in looks to the distractor picture (*Estimate* = -.790, *SE* = .118, *p* < .001). Unlike in Experiment 1, there was a rhyme effect in Experiment 2 for the mixed condition.

3.4. Discussion

The results of Experiment 2 largely paralleled the findings with natural speech. In particular, both speech types produced a cohort effect only for the Cohort 3 condition (3 segment sharing). In Experiment 2, the beginning of the speech signal contains no coarticulatory cues that would allow the listener to distinguish the target from Cohort 1 and Cohort 2 competitors. The listener nonetheless does not look at these competitors any more than at unrelated distractors. One-segment and two-segment overlap is apparently not enough to drive a saccade in the present experiments. The lack of looks to Cohort 1 and 2 competitors in Experiment 2 indicates that this absence of looks is not caused by coarticulatory cues to the end of the word allowing the listener to quickly rule out these competitors as inconsistent with the speech signal. Participants don't look at these cohort competitors even when the speech signal contains no cues to rule them out. Instead of looking at the referents of all words consistent with the signal, participants look only at referents of words for which the signal provide substantial evidence.

The target words in the present experiment contain points of disambiguation that separate them from Cohort 1, Cohort 2, and Cohort 3 competitors. These points of disambiguation are delayed by the absence of coarticulation in Experiment 2. Because of this, it was expected the trajectory of looks to the Cohort 3 competitor and target to diverge from looks to distractors later in Experiment 2 compared to Experiment 1. This expectation was confirmed by the timecourse analyses: the likelihood of looks to signal-consistent pictures increased beyond that of looks to signal-inconsistent pictures diverged about 200 ms later in response to synthesized speech compared to natural speech. Furthermore, looks to the target also diverged from looks to the Cohort 3 competitors later in synthesized speech compared to natural speech.

A somewhat unexpected difference between the two experiments was the presence of a rhyme effect in the Cohort & Rhyme Mixed condition in Experiment 2 (synthesized speech). In this condition, the rhyme competitor is missing the initial segment of the target (e.g., *hebi* vs. *ebi).* Notably, there was no rhyme effect in the Rhyme 3 condition where the target and the competitor differed in the identity of the initial segment (e.g., *futa* vs. *buta*).[4] The likely explanation for these differential effects is that the presence of the initial consonants ([h], [k], or [s]) in the Cohort & Rhyme Mixed condition relies on detecting a period of high-frequency aperiodic noise corresponding to a fricative or the release of a voiceless stop, a sound that is acoustically hard to integrate with the following speech (Remez, Rubin, Berns, Pardo, & Lang, 1994). In the visual world paradigm, Galle (2014) has shown that the impact of an initial fricative in the acoustic signal is significantly delayed compared to the impact of other cues, even ones that follow the fricative in the speech stream. In relatively unnatural synthesized speech, which makes speech sound less like speech, this integration may be especially likely to be delayed, making words like *hebi* temporarily homophonous with words like *ebi*.

McQueen & Huettig (2012) and Brouwer & Bradlow (2011) have previously shown that signal degradation increases the strength of rhyme effects in the visual world paradigm, comparing high-overlap cohort stimuli and rhyme stimuli differing from the target by the initial consonant. Their rhyme stimuli were of the *futa* / *buta* type and not of the *hebi* / *ebi* type and therefore would not be expected to elicit the rhyme effect based on the data here. However, the type of degradation used in these experiments (replacing or obscuring speech with noise) was different from the type used here. Whereas the present noise may have led participants to have integration difficulties, replacing or obscuring sounds with noise may lead listeners to instead increase the estimated likelihood of misperceiving or missing a consonant, including an onset, increasing rhyme effects

[4] Note that the first vowels in the stimuli are all fully voiced, meaning that the vowels are not reduced vowels or voiceless vowels even though Japanese phonology allows for reduction and devoicing in this environment. This ensures that the target and the competitor differ only in the initial consonant.

across the board. Exposure to reduced forms may similarly decrease the degree of overlap the listener required for a saccade (Brouwer, Mitterer, & Huettig, 2012). The human data suggest that a listener needs to accumulate enough evidence for a saccade rather than saccades mapping directly to lexical activation levels. The following chapter presents a computational investigation of whether the TRACE model of spoken word recognition can account for the human data (for both synthesized and natural speech) when equipped with the standard linking hypothesis.

CHAPTER IV

EXPERIMENT 3: COMPUTATIONAL SIMULATION

(TRACE MODEL)

4.1. Introduction

All current models of spoken word recognition suggest that any part of a target word activates candidate words (e.g., Allopenna et al., 1998; Luce & Pisoni, 1998; McClelland & Elman, 1986; Norris, 1994; Salasoo & Pisoni, 1985). The standard linking hypothesis for the visual world paradigm appears to suggest that these activation differences – standardly modeled using TRACE – should be directly reflected in fixation probabilities (Allopenna et al., 1998; Tanenhaus et al., 2000). Sharing some parts of target words in the present study seemed to be insufficient to either an eye movement or an increase in lexical activation. However, I have not yet shown that TRACE activations do indeed reflect segmental overlap differences manipulated in Experiments 1-2. It is therefore possible that the phonetics or lexical statistics of Japanese lead TRACE not to predict that competitor activation should exceed distractor activation in the Cohort 1, 2, and Rhyme conditions of the present experiments, even though TRACE predicts such effects for comparable English stimuli examined by Allopenna et al. (1998).

No studies have parametrically varied overlap between candidate words, or systematically examined TRACE's sensitivity to overlap. The spoken word recognition data based on natural speech stimuli (Experiment 1) and synthesized speech stimuli (Experiment 2) were therefore compared to the predictions of the TRACE model

(McClelland & Elman, 1986), to evaluate the model’s ability to predict the patterns in the data.

Experiment 3 asks the following questions: 1) Does the TRACE model predict competitor effects in all conditions? 2) Do the predicted competitor effects in the TRACE model match those observed in the human data? I intend to show that the results of the behavioral studies reported above can be reconciled with TRACE and with the results of previous studies if we assume that a certain level of activation is necessary to drive a saccade to the referent of a word.

The model consists of three layers of units, including a feature level, a phoneme level, and a word level. Input, for example the /k/ in *kasa* ‘umbrella’ is first represented on the feature level with feature values (e.g., strength levels for voiceless, sonorant, etc.), and the feature values activate phonemes that share them (e.g., /k/ and /g/ would be activated by a certain level of [sonorant]) and inhibit those that do not. The activated phonemes activate candidate words that contain them (e.g., *kasa* ‘umbrella,’ *kame* ‘turtle,’ *gomi* ‘garbage’. etc.) and inhibit those that do not. The activated words feed activation back to the phoneme level, activating the phonemes they contain and inhibiting those they do not. Words also compete with each other via lateral inhibitory connections, as do phonemes. In other words, at the phoneme and word levels, there is inhibition within a level and bidirectional flow of activation and inhibition between levels. Since multiple features can be present simultaneously, features do not compete with each other for recognition, and top-down feedback does not affect feature activations, preventing hallucinations based on top-down input. The architecture of the model is described further in the method section.

4.2. Methods

4.2.1. Stimuli

Stimulus words and trials used for the TRACE simulation are identical to those presented to human participants in Experiment 1 (natural speech) and Experiment 2 (synthesized speech).

4.2.2. Procedure

The simulation was conducted using the software jTRACE (Strauss, Harris, & Magnuson, 2007), which is available online from the Computational Cognitive Neuroscience of Language Lab at the University of Connecticut (https://magnuson.psy.uconn.edu/jtrace/). Like the original TRACE, jTRACE contains 14 phonemes as follows: /b, p, d, t, g, k, s, ʃ, ɹ, l, ɒ, u, i, ʌ/ and /-/ a silence, which has no featural overlap with any of the 14 phonemes. In order to test Japanese words, the phonemes in jTRACE were revised to include 24 consonants and 5 vowels as shown in Table 4.1. Previous studies limited lexica to the words that can be represented with phonemes similar to the existing phonemes in jTRACE (e.g., Dahan et al., 2001a; Dahan et al., 2001b; Marslen-Wilson & Warren, 1994) or used only words that can be found in the existing jTRACE lexica (e.g., McMurray et al., 2010; Mirman et al., 2011), except for a recent Mandarin simulation that modified the phonemes for the Mandarin phonemic inventory (Shuai & Malins, 2017). TRACE has not been previously applied to Japanese. The present study is therefore the first extension of the TRACE model to Japanese spoken word recognition, which required adding new phonemes and modifying featural descriptions of some existing phonemes.

Table 4.1. The Japanese phonemic inventory incorporated into TRACE in the present study.

Consonants

	Bilabial	Dental[5]	Alveolar	Alveolo-palatal	Palatal	Velar	Uvular	Glottal
Plosive	p b	t d				k g		
Nasal	m	n				ŋ	N	
Tap			ɾ					
Fricative	ɸ		s z	ɕ ʑ	ç			h
Approximant					j	w		
Affricate			ts dz	tɕ dʑ				

Vowels

	Front	Central	Back
Close	i		ɯ
Close-mid	e		o
Open-mid			
Open			ɑ

Note: These phonemes were used to transcribe the experimental words in the present study. A few phonemes that appear in Japanese words (e.g., /ɲ ɣ β /) were not used because the stimulus words did not contain these phonemes.

Phoneme feature specifications were also revised to better match the acoustics of Japanese. TRACE uses acoustic features definitionally similar to those proposed by Jakobson, Fant, & Halle (1951), including sonority, anteriority, height, diffusion, acuteness, voicing and burst amplitude[6]. The Mandarin TRACE-T model (Shuai & Malins, 2017) contained roundness, place of articulation, manner of articulation, tongue

[5] /t/, /d/, and /n/ in Japanese are considered dental (IPA handbook) or more front than the alveolar English coronals (Vance, 1987)

[6] The definition of the Burst feature provided by McClelland and Elman (1986) was as follows: “The amplitude of the burst of noise that occurs at the beginning of word initial stops, was included to provide an additional basis for distinguishing the stop consonants, which otherwise differed from each other on only one or two dimensions.” (McClelland and Elman, 1986: p.14)

height, tongue position, voicing and tone. Because the features are supposed to be variable acoustic representations, TRACE represents each feature as a scale of multiple values, which allows in-between ambiguous segments to be represented. Thus the strength of each feature ranges from level 1 to level 8, level 1 being the weakest level and level 8 being the strongest level of the feature. Note that level 9 was assigned only to a silence /-/. For example, sonority ranges between low vowels (8) and stops (1). At each level of strength (1-8), feature values in particular segments vary between from 0 to 1. This representation allows TRACE to represent each segment's feature value as a distribution over several possible feature values that is intended to reflect the way in which acoustic realizations of the segment vary (see Appendix B for complete charts for TRACE: McClelland & Elman, 1986; see Appendix C for jTRACE: Strauss et al., 2007; and the present study in Appendix D). For instance, Table 4.2 below displays the features of /t/. On the sonority dimension (Son.), the highest value (1.0) at the strength level of 1 indicates weakest sonority. This is appropriate for Japanese but would not be appropriate for English where /t/ is commonly reduced to a much more sonorous approximant (as in some realizations of 'butter'), which would mean that an English /t/ would be represented by values spread over a wide range of sonority levels. This example shows that simply reusing English feature specifications in applying TRACE to Japanese would be inappropriate. The major modification for Japanese was to include /ɾ/, which can be transcribed as [ɾ ɹ ɽ ʀ l d] depending on its location in a word and the type of an adjacent vowel, and pronunciations can vary across individuals (e.g., Labrune, 2014). Therefore, the values of features for /ɾ/ were spread widely to account for the variation. The values were determined by referring to TRACE (McClelland & Elman, 1986) and jTRACE

(Strauss et al., 2007) feature and segment definitions, which are based on phonological features by Jakobson et al. (1951), as well as to Japanese phonological features (Matsui, 2017; Vance, 1987).

Table 4.2. Phoneme feature dimensions for /t/.

Level	Sonority	Anterior	Height	Diffuseness	Acuteness	Voiced	Burst
1	1.0			1.0			
2						1.0	
3							
4							
5							0.2
6							1.0
7		1.0			1.0		
8							

In TRACE, an input word excites features first. For example, /t/ in /toɾa/ has values for each feature as shown in Table 4.3. When these values are detected in the signal, corresponding phoneme(s) will be activated. In this case, both /t/ and /d/ will be activated to some extent at the phoneme level, then activating words that contain /t/ and /d/ such as /tana/, /toɾa/, and /take/.

Table 4.3. Phoneme feature dimensions and values for /t/ and /d/.

	Sonority	Anterior	Height	Diffusion	Acuteness	Voiced	Burst
t	1 (1.0)	7 (1.0)	0	1 (1.0)	7 (1.0)	2 (1.0)	5 (0.2) 6 (1.0)
d	1 (1.0)	7 (1.0)	0	1 (1.0)	7 (1.0)	7 (1.0)	5 (1.0) 6 (0.2)

Several parameters (e.g., word layer inhibition, feature decay, etc.) can be manipulated in the simulation to account for differences between participants or tasks. In the present study, one experiment used natural speech while the other used synthesized speech, which is relatively degraded. To account for this degradation, we can turn up the input noise parameter (which slows lexical activation) in hopes of accounting for 1) the slower recognition of the target and 2) the presence of the rhyme effect in the Mixed competitor condition in synthesized speech. However, default parameter settings were used for the analysis in this chapter for two reasons. First, parameter changes compared to previous simulations need to be theoretically motivated by the way in which the present experiments objectively differ from previous spoken word recognition experiments in the visual world paradigm. Otherwise, any data pattern could be captured by TRACE with parameter tweaks (see also Norris & McQueen, 2008). For example, McMurray et al. (2010) show that some parameter value combinations in TRACE can actually *reverse* cohort effects so that related stimuli are activated less than unrelated ones, a result that makes little theoretical sense. For most parameters, there is no convincing reason to expect the settings to differ between this experiment and its predecessors, making it questionable to tweak them to fit the data. Second, as shown later in Chapter 5, manipulation of the parameters plausibly affected by the current experimental design do not explain the task differences (natural and synthesized speech). Further investigations and explanations of parameter manipulation are addressed in the next chapter.

4.2.3. Data Processing, Analysis and Results

This section illustrates the procedure of TRACE simulation on a Japanese spoken word and demonstrates the issues that arise in the statistical analysis of this data and their solutions.

An input word (e.g., /taki/ 'waterfall') is fed into TRACE sequentially, segment by segment from beginning to end ('left to right'). Each segment is converted into the corresponding acoustic features, whose activations rise and fall over the duration of the segment as shown below. Figure 4.1 below is an example input feature distribution of /taki/ over its time course.

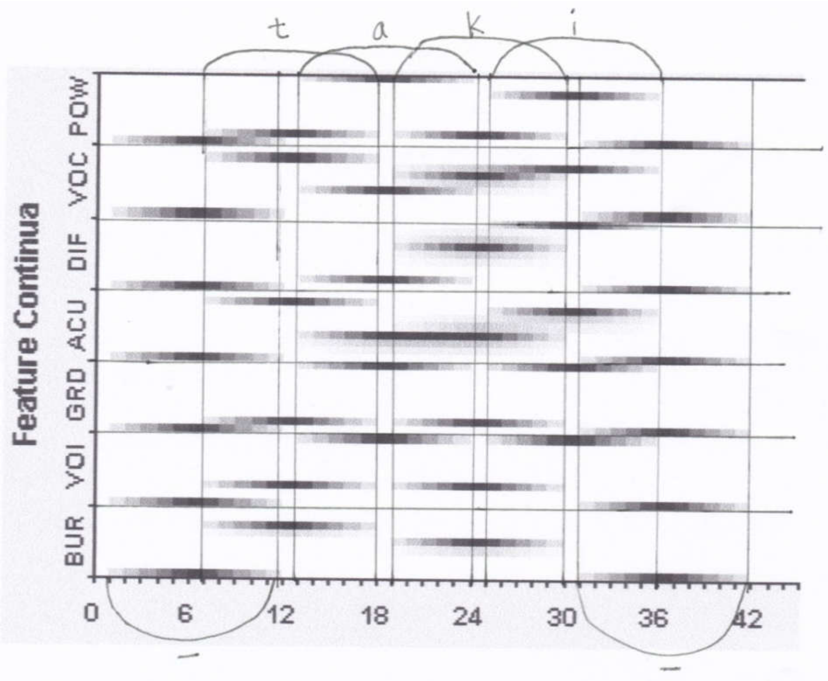


Figure 4.1. Example input feature distribution of /taki/ in time course.

X-axis represents time since input word onset (0 to 42 cycle time). Y-axis is blocked by seven features where each feature block represents strength levels (9 to 1 from bottom to up). Each phoneme corresponds to 11 cycles of time and a silence, /-/, is added at the beginning and at the end of a word. For example, the first and the last 11-cycle represent a silence which has a strength level of 9 for every feature. The dashed line, activation of feature, sits on the bottom (level 9) for each feature block. Adjacent phonemes overlap by five cycles. TRACE therefore exhibits local coarticulation analogous to that produced by the diphone synthesizer in Experiment 2, but no long-distance coarticulation. The duration of overlap (coarticulation) is identical for any combination of phonemes unlike the duration of phonemes and coarticulation differ in natural speech. Figure 4.1 demonstrates coarticulation in the word, /taki/. The initial five time cycles of /t/ overlap with the last five time cycles of silence and the last five time cycles of /t/ overlap with the initial five time cycles of the following phoneme, /a/.

In order to compare the TRACE model and the human data in real time, cycle time was rescaled to match real time in milliseconds. Average duration of stimuli words, 460 ms, was divided by the average number of cycles in a word, 42.4 cycles, which yielded one time cycle being 10.8 ms.

The simulations provided trajectories of activation levels of each word over time. Activations, a, for each item *i* were then transformed into response strengths, *S*, following (Allopenna et al., 1998) and (Dahan et al., 2001a). The free parameter *k* determines the extent to which this transformation increases the differences between activations. It was set to 7 following Allopenna et al. (1998) and Dahan et al. (2001a).

$$S_i = e^{ka_i} \text{ (Allopenna et al. 1998, p.424)}$$

The response strengths (*S*) were converted into response probabilities (*L*) by using the Luce choice rule (Luce, 1959), shown in the equation below. The Luce choice rule ensures that the entire probability mass is divided between the j possible response options. That is, the response probabilities range from 0 to 1. For example, at the beginning of a target word, the probability of looking at a picture is ¼ for each of the four pictures on the screen.

$$L_i = \frac{s_i}{\Sigma s_j} \text{ (Allopenna et al. 1998, p.424)}$$

However, the current experimental task (as well as that in Allopenna et al., 1998) required a participant to look at the fixation point (at center of the display) before s/he can look at any pictures. This yielded probabilities of looking at each picture close to zero at the beginning of the trial in human data before the listener made a saccade to a referent, since they were looking at something else rather than a picture of one of the four objects. To deal with this issue, Allopenna et al. rescaled looking probabilities by Δ_t, which is a free parameter fit to the human data. To calculate Δ_t, the maximum activation at each time slice was divided by the maximum activation across all time slices of the same trial. Then, Δ_t was multiplied by the looking probability derived from TRACE activations via the Luce Choice Rule above, L_i to generate expected Response (R) (here, fixation) Proportions.

$$\Delta_t = \frac{\max(a_{i,t})}{\max(a_i)} \text{ (Allopenna et al., 1998, p.424)}$$

$$p(R_i) = \Delta_t L_i \text{ (Allopenna et al., 1998, p.425)}$$

This rescaling means that the speaker has some probability of moving eyes away from the fixation point at any given time. This probability is based on how strongly activated the *most likely* lexical candidate is at that point in time. Once the decision to move the eyes is made, the speaker then decides *where* to move them, which is based on how strongly *each* of the candidates is activated. That is, the probability mass allocated to the decision to move is divided between the possible locations one can decide to move to. This is somewhat unintuitive: the speaker decides to move based on how activated one candidate is even when the decision is to move to a depiction of another candidate. It would be more consistent to replace the maximum function above with the average or the sum of activations so that the decision to move is based on all the candidates pulling one away from the fixation cross. However, we retain the max function here for comparability with Allopenna et al. (1998) and following work in this paradigm.

Figure 4.2 (next page) depicts curves for the TRACE-predicted fixation probabilities for each picture type in each condition. For Cohort conditions, Response (fixation) probabilities for cohort competitor pictures become higher as the numbers of shared phonemes increases. This is also observed for Rhyme conditions although the difference between the conditions seems to be slight. For the Mixed competitor condition, the probability of the cohort fixation is higher than for rhyme or distractor pictures and the rhyme effect is not predicted even though the rhyme competitor shares most with the target word. The trajectory of the probability of looking at a target picture diverges from those for distractors around 200 ms for TRACE data, which is similar to the figure for natural speech (Figure 2.3, page 42). The divergence starts later in synthesized speech (400 ms) than in TRACE (200 ms). This is somewhat surprising given that the degree of coarticulation in TRACE mirrors the degree of coarticulation in synthesized speech in the present study and suggests that synthesized speech may be causing participants to delay commitment to an eye movement decision.

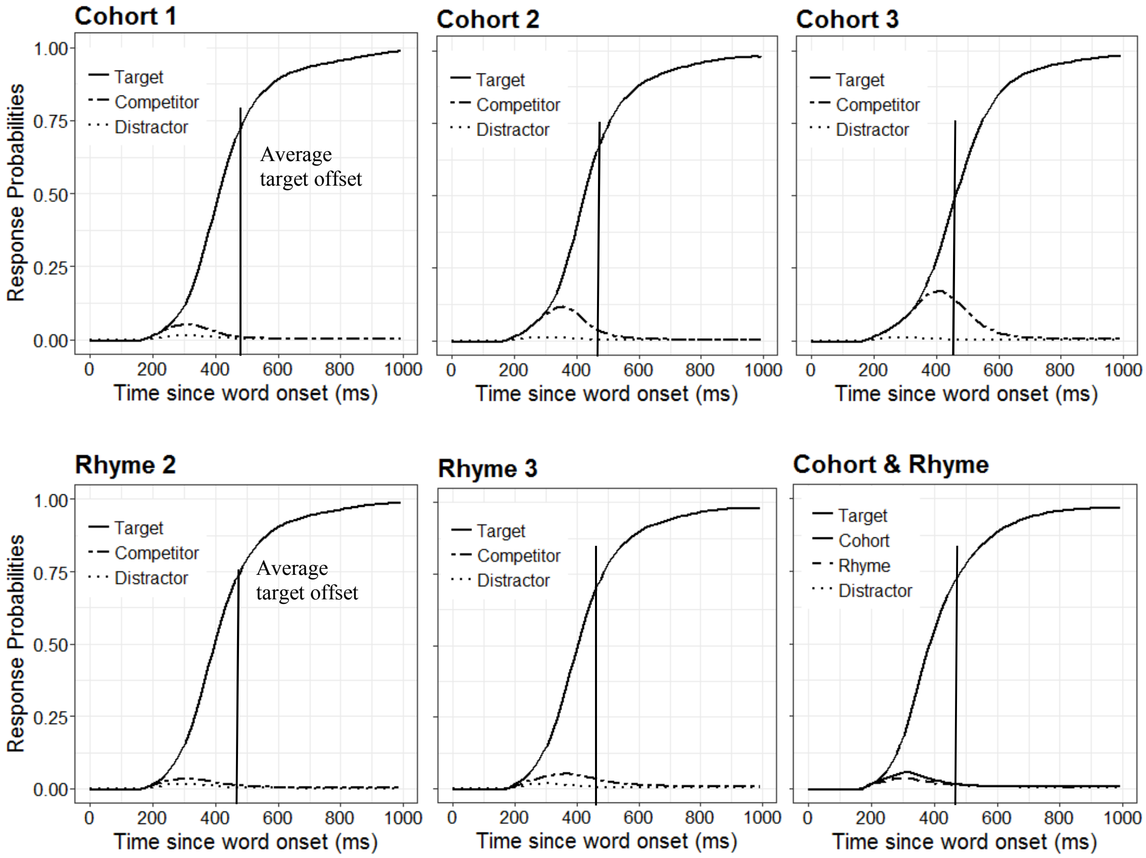


Figure 4.2. Growth curve results of TRACE data for each condition.

While previous studies have largely limited themselves to the type of qualitative visual analysis described so far (Dahan et al., 2001b; Shuai & Malins, 2017), this is insufficient to determine whether the predictions of TRACE match human data. In particular, Figure 4.2 shows small but reliable Cohort effects within TRACE. However, it is not clear whether these predicted effects are large enough to be detectable in the human data. In other words, to evaluate the predictions of TRACE it is important to know whether the predicted differences between conditions are distinguishable from lack of condition differences (null hypothesis).

Two types of quantitative comparison between the TRACE model and the human data were for this purpose. The first analysis examined the fit of the TRACE model to the human speech data (Natural & Synthesized) within a single condition (e.g., Cohort 1, Cohort 2, etc.) while the second analysis examined the fit between conditions (e.g., Cohort 1 vs. Cohort 2).

For these analyses, generalized additive models (GAMs; fit using the gam function in the mgcv package; Wood, 2007; in R) were used to summarize the curves generated by a TRACE model. GAM is a recent alternative to polynomial models for curve-fitting (here, fitting the trajectory of fixation proportions as a non-linear function of time). Although polynomial growth curve models have the advantage of providing the modeler with interpretable parameters, which motivated the choice to apply them to the human data in Chapters 2 and 3, GAM models tend to produce better fits and better extrapolation performance than polynomial models (e.g., Baayen, van Rij, de Cat, & Wood, 2018; Wieling, Nerbonne, & Baayen, 2011). This was also the case here: the fits to the TRACE curves generated by GAM were markedly better than polynomial growth curve fits. In the analyses reported in this chapter, a statistical model was developed to be intended to stand in for TRACE in model comparisons. This model must fit TRACE curves very well if it is to represent it.

Bayes Factor comparisons were then used to compare a GAM model with the condition effect predicted by TRACE and a near-identical GAM model missing the effect of condition. This approach retains the overall timecourse predictions from the TRACE model in both GAM models, focusing solely on whether the *condition differences* (such as the one between a competitor and a distractor, or a Cohort 1 competitor vs. Cohort 2 competitor) are as predicted by TRACE or not. I now describe the analytic approach step by step, while identifying the issues that arise in the analysis and the proposed solutions for these issues.

The first step is to build a GAM model of the TRACE-predicted fixation trajectories. This involves predicting the dependent variable Fixation Proportion based on the independent variable (Picture Type, competitor(s) vs. distractor, for within-condition analysis and Competitor Type for between-condition analyses of competitor fixations). Because fixation proportions change over time in complex non-linear ways, and the trajectory of change is different across the levels of the independent variable, a smooth for time within each level of the predictor variable was also included. This (alternative) model (H1) [7] allows for a difference in the overall level or in change over time of Fixation Proportion across Picture types or Competitor types. A comparable null model (H0) is fit to the TRACE predictions by omitting the independent variable (Picture Type of Competitor Type). This model fits a single smooth for time, which means that it retains the overall trajectory of fixation proportions predicted by TRACE across conditions, by pooling the data across conditions and fitting a single non-linear function of time to the resulting fixation proportions. In other words, the null model therefore retains the overall timecourse predictions of TRACE while removing the predicted effect of condition.

[7] H1 <- gam (Fixation Proportion ~ s (Time, by = Picture Type) + Picture Type, data = TRACE data)
H0 <- gam (Fixation Proportion ~ s (Time), data = TRACE data)

Example figures of H1 and H0 model fits to the TRACE predictions in the Cohort 1 condition are shown in Figure 4.3 below. The thicker lines (both solid & dashed) represent the GAM model fit while the thinner lines represent the raw data of TRACE predictions, averaged across trials. As seen in the figure, GAM provides an excellent fit to the Fixation Proportion curves, which means that its predictions can stand in for TRACE predictions in fitting the human data for the H1 model.[8]

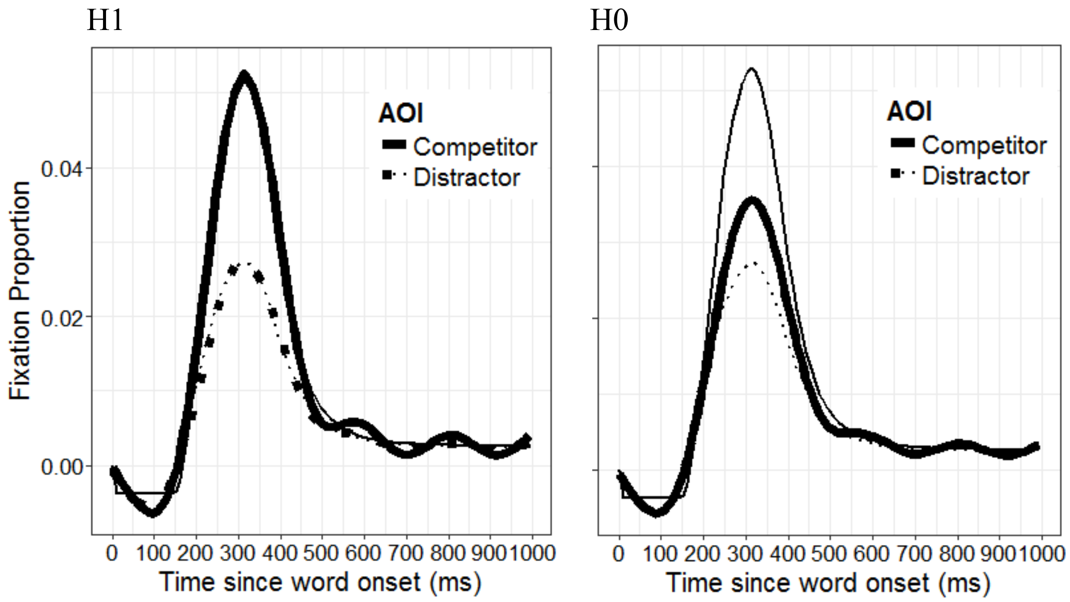


Figure 4.3. Plot of the GAM fits to TRACE predictions for the Cohort 1 condition (left = H1, which includes the effect of AOI, and right = H0, which excludes it)

The second step is to generate predictions from the H1 and H0 models for human data, using the predict() function in R[9]. Since the TRACE model was trained on the same

[8] Spurious wiggliness is still present in the GAM fit where the TRACE curves are smooth at the edges of the graph, reflecting the bump-like thin-plate regression smooth basis function used in this version of GAM (Wood, 2003). However, this wiggliness accounts for very little variance and is greatly reduced compared to polynomial growth curve fits.

[9] H1.prediction <- predict (H1, newdata = human data)
H0.prediction <- predict (H0, newdata = human data)

items, these predictions are intended to capture some item variability, and the H1 model is intended to capture the condition effect.

Thirdly, new GAM models were fit to human data by combining the predicted values from TRACE models (H1 & H0) with the random effects of Subject and Item.[10] The random effects are necessary to produce a reasonable fit to the human data, as suggested by Farris-Trimble & McMurray's (2013) finding that there are sizeable and stable individual differences in eye movement behavior in the visual world paradigm. In this initial analysis, I also followed recent work in the Baayen and Milin labs (e.g., Milin, Feldman, Ramscar, Hendrix, & Baayen, 2017) in allowing the GAM model to fit non-linear interactions between TRACE activations and time, because the predicted fixation trajectories in TRACE do not have quite the shape observed in human data. For example, the figure below shows that the fixation probability curves in human data are quite skewed, and so not quite go down to zero in the way that TRACE curves do. These shape differences could be due to extraneous reasons like the assumption that all segments are equally long and equally coarticulated, and hence were thought to be worth abstracting away from. However, the better fit to human data associated with a GAM link between TRACE predictions and fixation proportions comes at a significant cost, as we will see later. Figure 4.4 (next page) exhibits the outputs of the new models (thick smooth lines) and the human raw data averaged across subjects and items for the Cohort 1 condition (thin lines). In the H1 model, the difference in looks between competitors and distractors

[10]
```
newdata.H1 <- gam (Fixation Proportion ~ te (H1.prediction, Time) + s (SUBJECT, bs="re")
                + s (ITEM, bs = ”re”), data = human data)
newdata.H0 <- gam (Fixation Proportion ~ te (H0.prediction, Time) + s (SUBJECT, bs="re"),
                + s (ITEM, bs=”re”), data = human data)
```

is very slight while there was no difference in looks between the two pictures in the H0 model, which represents the null hypothesis.

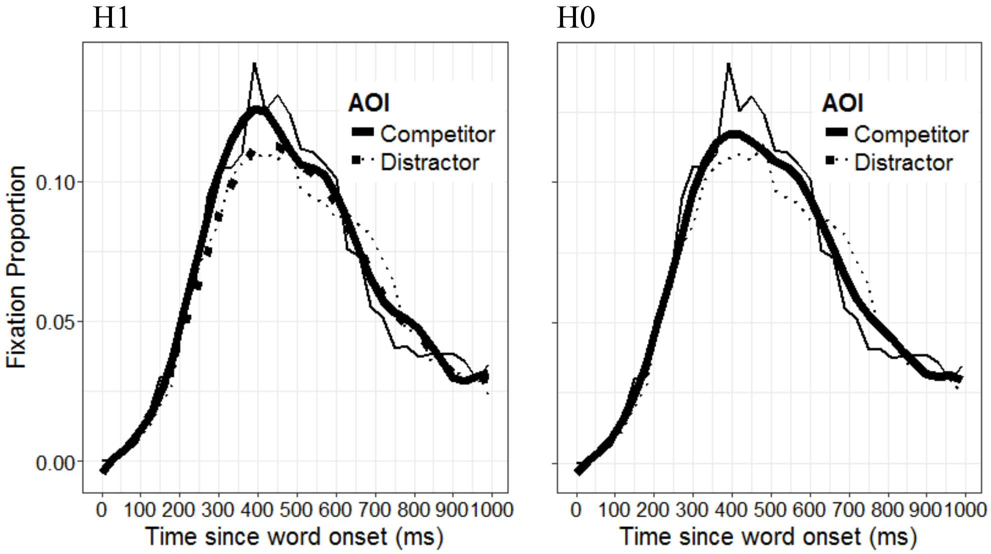


Figure 4.4. Plots of the GAM models of human data as a function of TRACE predictions for the Cohort 1 condition (left = H1 and right = H0).

These models are then compared using the BIC approximation to the Bayes Factor (Wagenmakers, 2007) to determine whether the predictions of H1 or H0 provide a better fit to the human data. Support for H1 means that the human data evidence the condition difference predicted by TRACE, while support for H0 means that human data provide evidence against the predicted condition difference. It is also possible for the model comparison to be inconclusive, in which case the human data are insufficient to distinguish between the two hypotheses. For example, in the graph above, the predictions of H1 and H0 are nearly identical because 1) both retain the overall timecourse predictions of TRACE and the random subject and item effects, and 2) the difference between looks to distractor and competitor in TRACE was expected to be small. The data should therefore be inconclusive regarding whether H1 and H0 are supported.

Table 4.4 describes the results of the model comparisons within each condition (Cohort 1, Cohort 2, Rhyme etc.) for natural speech stimuli.

Table 4.4. BIC differences between H1 and H0 models and evidence strength of the model in natural speech for each condition type.

Condition	Cohort 1	Cohort 2	Cohort 3	Rhyme 2	Rhyme 3	Mixed[11]
BIC differences	2	48	286	5	0	38
Evidence for H1 or H0	Neither	H1	H1	H0	Neither	H0
posterior probability of H1 or H0	0.25 (H0)	1 (H1)	1 (H1)	0.93 (H0)	0.44 (H0)	1 (H0)

For example, the difference of BIC values between H1 and H0 models is five in the Rhyme 2 condition and the H0 model is supported, meaning that the BIC value of the H0 model is smaller than that of the H1 model. The H0 model was supported for the Rhyme 2 and the Mixed competitor conditions, which means that human data were better described by a model that says there is no competitor / distractor difference (in proportion of looks or proportion and time-course or time-course depending on H0) than by the model that says the difference between distractor and competitor is the same as in TRACE. In other words, there is a difference in fixation proportion between rhyme pictures and distractor pictures in the TRACE rhyme 2 condition (H1), which generates predictions that are distinct from those of the corresponding H0; however human data are more consistent with H0. Likewise, the difference in activations between the rhyme

[11] The analysis includes Fixation Proportion of Cohort, Rhyme and Distraction.

competitors and unrelated distractors in the Mixed condition in TRACE had no effect in natural speech data.

H1 model was supported for the Cohort 2 and Cohort 3 conditions, meaning that human data were better described by a model that says the difference between competitor and distractor is the same as in TRACE than by a model that says there is no difference. In other words, the difference in fixation proportion trajectories between distractor and competitor pictures in the Cohort 3 condition is more consistent with the difference predicted by TRACE than with no difference: both TRACE and human data have a big difference in fixation proportion trajectories between the pictures (cohort effect). TRACE therefore successfully explained the human data. The subtle differences in BIC values between H1 and H0 models for the Cohort 1 and Rhyme 3 conditions are inconclusive. This inconclusiveness is largely due to absence of observable effects of relatedness in TRACE predictions for these conditions, making H0 and H1's predictions for the human data near identical.

Table 4.5 suggests support for either H1 or H0 in each condition of the study utilizing synthesized speech stimuli.

Table 4.5. BIC differences between H1 and H0 models and evidence strength of the model in synthesized speech.

Condition	Cohort 1	Cohort 2	Cohort 3	Rhyme 2	Rhyme 3	Mixed
BIC differences	73	16	219	32	9	2
Evidence for H1 or H0	H0	H0	H1	H0	H1	Neither
posterior probability of H1 or H0	1 (H0)	0.99 (H0)	1 (H1)	0.99 (H0)	0.99 (H1)	0.80 (H0)

The H0 model was supported for Cohort 1, Cohort 2, and Rhyme 2, which indicates failures of TRACE predictions, while the H1 model was supported for Cohort 3 and Rhyme 3, indicating support for TRACE predictions. The model comparison was inconclusive for the Mixed competitor condition.

Table 4.6 shows the results of model comparisons for the differences between conditions for both speech types. The results were the same across speech types in that H0 was supported for the comparisons between the Rhyme 2 and the Rhyme 3 conditions, which means TRACE failed to predict the absence of differences between these conditions in human data, while H1 was supported for the comparisons between the Cohort 2 and the Cohort 3 conditions, indicating that TRACE succeeded in predicting the differences between these conditions. The model comparisons for the Cohort 1 vs. Cohort 2 conditions were inconclusive.

Table 4.6. Results of between conditions for both speech types.

Speech Types	Natural speech			Synthesized speech		
Condition	1C vs. 2C	2C vs. 3C	2R vs. 3R	1C vs. 2C	2C vs. 3C	2R vs. 3R
BIC differences	1	32	65	1	112	50
Evidence for H1 or H0	Neither	H1	H0	Neither	H1	H0
posterior probability of H1 or H0	0.6 (H0)	1 (H1)	1 (H0)	0.64 (H0)	1 (H1)	1 (H0)

The analyses reported above provided some evidence that TRACE predicts certain effects of overlap that are absent from human data. When the Bayesian model comparisons are conclusively favoring H0, the human data suggest that a zero effect of overlap is more probable than the effect predicted by TRACE. Above, this holds for several comparisons between unrelated distractors and phonological competitors that overlap with the target in synthesized speech. In particular, both final (rhyme) overlap and initial (cohort) overlap in 1 or 2 segments is more likely to have no effect on eye movements than to have the effect predicted by TRACE. In contrast, initial or final overlap in 3 segments does have the effect predicted by TRACE. Together, these results suggest that overlap in fewer than 3 segments is insufficient to drive a saccade in the synthesized speech data. In contrast, overlap in 2 initial segments does appear to be sufficient to increase the probability of a saccade in response to natural speech, while the results for final overlap are inconclusive.

The condition comparisons show that evidence for 3 initial segments of a word is more likely to produce a saccade to the word's referent than evidence for 2 initial segments with both types of stimuli (natural and synthesized), although in synthesized speech there is evidence that 2 segments are insufficient to increase saccade probability over 0 segments whereas in natural speech 2-segment overlap does appear to be sufficient to increase fixation probabilities.

Caution is warranted, however, in interpreting these results. As noted above, the GAM models of human data allowed for a non-linear tensor product interaction between TRACE predictions and time, which results in an excellent fit to human fixation trajectories. This has become a standard approach in recent psycholinguistic work. For example, Milin et al. (2017) argue for a discriminative model of word learning by using lexical activations from this model as input to a GAM model of word recognition data. The GAM using model activations achieves a better fit than a GAM using measures like word frequency. However, this kind of argumentation is problematic because the GAM model allows the predicted values to have any kind of relationship to the observed values; even a non-monotonic one (see also Kapatsinski, 2017, for a similar critique of random-forest analyses). Allowing the GAM model to treat observed values as any kind of smooth function of predicted values means that a model may also be favored simply to the extent that its predictions are variable across datapoints because the predicted differences between datapoints can then be rescaled arbitrarily by the GAM model. In this way, the greater variability of the predictions of the H1 model, which has an additional predictor, may make it unduly favored in model comparisons (see Figure 4.5 below). Although the curvature patters of fixation proportion for each picture type are almost identical between the models, the more complex model, H1, was disfavored because it contained an additional predictor.

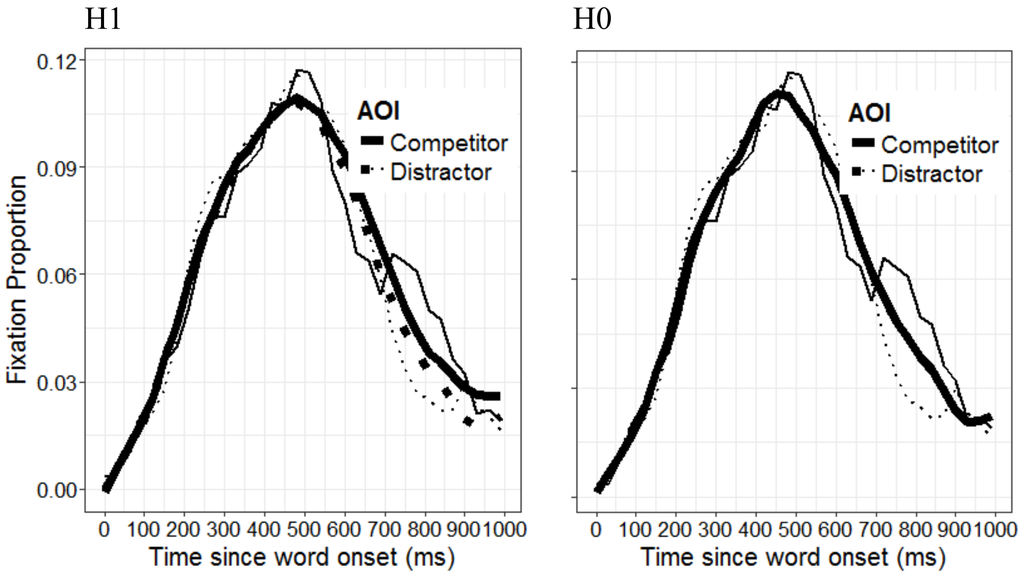


Figure 4.5. Example outputs for the models that demonstrated similar curves, but H0 was favored for the Rhyme 3 condition (left = H1 and right = H0).

Because H1 and H0 predictions are rescaled by GAM for each model comparison, the H1 and H0 models' predictions can also have a different relationship to the data in different condition comparisons. This can introduce inconsistencies into the results. For example, the model comparisons in synthesized speech provide strong support for H1 over H0 in the comparison of the Rhyme 3 competitor to the corresponding distractor, and strong H0 support for the comparison of the Rhyme 2 competitor over distractor, and yet H0 is supported over H1 for the difference between the Rhyme 2 and Rhyme 3 conditions. For these reasons, I would like to advocate against fitting smooths to model predictions in comparing those predictions to observed data. Thus, in the models below I use H1 and H0 predictions as simple linear predictors in GAM.

Second, the above model comparisons treated the H1 and H0 models as equally complex because both have the same number of predictors once applied to human data. However, the H1 model of the TRACE trajectories is in fact more complex, which gives it greater flexibility in accounting for human data. As shown in the figure above, it produces two distinct values for each time point, whereas H0 produces only 1. We would like to capture the fact that the H1 model's predictions for human data result from a model with one extra parameter, by punishing it for the extra complexity in BIC calculations. To accomplish this, H1 model's predictions were residualized[12] on H0 model's predictions.

[12] Residuals <- lm (H1.predict ~ H0.predict, data=Natural data) $ residuals

The residuals were then entered as an additional predictor of human data unique to the H1 model[13]. The shared variance between the H0 and H1 models is then attributed to H0. Therefore the H1 model must earn its keep by showing that the variance that is unique to it is predictive of human data. The H0 model[14] lacks the additional predictor. The differences in fit between the two approaches are illustrated below.

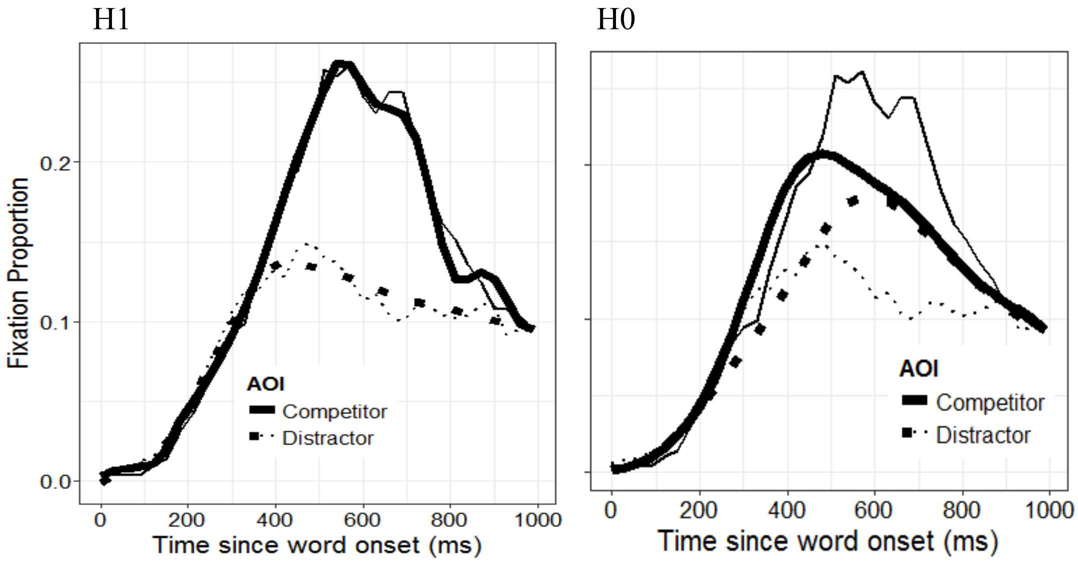


Figure 4.6. Example fits of the model in which observed values are a smooth non-linear function of TRACE predictions (on the left) and the revised model, in which observed values are a linear function of the predictions (on the right).

[13] newdata.H1 <- gam (Fixation Proportion ~ H0.predict +
Residuals.H1 +
s (Time) +
s (H0.predict, SUBJECT, bs = "re") +
s (H0.predict, ITEM, bs = "re") +
s (SUBJECT, bs = "re") +
s (ITEM, bs = "re"), data= human data)

[14] newdata.H0 <- gam (Fixation Proportion ~H0.predict +
s (Time) +
s (H0.predict, SUBJECT, bs="re") +
s (H0.predict, ITEM, bs = "re") +
s (SUBJECT, bs="re") +
s (ITEM, bs = "re"), data = human data)

The left panel in Figure 4.6 above is the fit of the H1 model in the original analysis, which I argue to be too excellent due to the reshaping of model predictions with GAM, while the right panel is the H1 model in the alternative analysis, which does not allow GAM to reshape model predictions. Although the output on the right panel is not fitting the human data as well as the model on the left, it provides a fairer evaluation of the underlying TRACE model.

The revised analysis yielded the ultimate model comparison results for within-condition comparisons, as shown in Table 4.7 for natural speech stimuli. Between-condition comparisons still suffer from a separate issue addressed later. The H0 model was supported for Cohort 1, Rhyme 3 and Mixed conditions, in which the human data were more consistent with the absence of a phonological overlap effect on fixations than with the effect predicted by TRACE. The H1 model was supported for Cohort 2 and Cohort 3 conditions, where the effect of overlap was more consistent with the predictions of TRACE than with zero difference. The Cohort 3 condition in human data has a significant cohort effect (fixation proportion difference between competitor and distractor) while the cohort effect was not significant in the Cohort 2 condition. However, the present results show that the Cohort 2 data are in fact more consistent with a small non-zero effect similar to the one predicted by TRACE than with zero effect predicted by the null hypothesis. This difference in analysis outcomes is further discussed at the end of this chapter. Finally, there was little difference in BIC values between the models for the Rhyme 2 condition; therefore the results of model comparison are inconclusive. The observed effect is in between the (small) rhyme effect predicted by TRACE in this condition and zero effect.

Table 4.7. BIC differences between H1 and H0 models and evidence strength of the model in natural speech.

Condition	Cohort 1	Cohort 2	Cohort 3	Rhyme 2	Rhyme 3	Mixed
BIC differences	5	37	179	2	9	10
Evidence for H1 or H0	H0	H1	H1	Neither	H0	H0
posterior probability of H0	0.93 (H0)	1 (H1)	1 (H1)	0.25 (H0)	0.98 (H0)	0.99 (H0)

In synthesized speech (Table 4.8 below), H0 was supported for Cohort 1, Cohort 2, and Rhyme 3 conditions, which is consistent with the analyses of human data that also showed no effect of overlap. H1 was supported for Cohort 3 and Rhyme 2, and the model comparison was inconclusive for the Mixed condition.

Table 4.8. BIC differences between H1 and H0 models and evidence strength of the model in synthesized speech.

Condition	Cohort 1	Cohort 2	Cohort 3	Rhyme 2	Rhyme 3	Mixed
BIC differences	9	9	36	6	7	0
Evidence for H1 or H0	H0	H0	H1	H1	H0	Neither
posterior probability of H1 or H0	0.99 (H0)	0.99 (H0)	1 (H1)	0.97 (H1)	0.97 (H0)	0.49 (H0)

The results of model comparisons for the effect of conditions are shown below (Table 4.9). The H1 model was supported for the comparison between the Cohort 2 and 3 conditions for natural speech. As predicted by TRACE, participants looked at the competitor more when the competitor overlapped with the target in three segments than when it overlapped in only two. The H0 model was supported for the comparison between the Cohort 1 and 2 conditions and the comparison between the Rhyme 2 and 3 conditions for natural speech, indicating no effect of cohort short-overlap and rhyme overlap. However, the comparisons for any condition types are inconclusive for synthesized speech. These results are subject to a pernicious analytical issue that I discuss next.

Table 4.9. Results of between conditions for both speech types.

Speech types	Natural speech			Synthesized speech		
Condition	1C vs. 2C	2C vs. 3C	2R vs. 3R	1C vs. 2C	2C vs. 3C	2R vs. 3R
BIC differences	14	36	9	1	2	1
Evidence for H1 or H0	H0	H1	H0	Neither	Neither	Neither
posterior probability of H1 or H0	0.99 (H0)	1 (H1)	0.98 (H0)	0.43 (H0)	0.75 (H0)	0.66 (H0)

Barth and Kapatsinski (2018) reported Monte Carlo simulations showing that fit (e.g., log-likelihood) comparisons between mixed-effects models can be misleading. In particular, they show that randomly reordering the values of a real predictor can result in a model with the same fit as the original model, as long as different levels of the original predictor are associated with different levels of a random-effects predictor and the random-effects predictor does have some effect. In that case, the random-effects predictor can 'step up' to capture the variance that was really generated by the fixed-effects predictor. In the present data, different conditions have different items, and Item is a random-effects predictor in the GAM model. As a result, the H0 model can attribute whatever between-condition variance it can't capture to the random effect of Item.

A clear example of this phenomenon is shown in Figure 4.7. The left panel shows the fit of the H1 model for the comparison between Cohort 2 and Cohort 3 conditions, whereas the right panel is the H0 model. Although the H0 model is supposed to be a null hypothesis that says there is no difference in looks between the conditions, H0 model does in fact predict a difference in looks between conditions. In other words, even though the H0 model excluded the fixed effect of Condition, the random effect of Item captured the difference between the conditions, producing separate predicted value curves for the Cohort 3 and Cohort 2 conditions. As argued by Barth and Kapatsinski (2018), these results indicate that mixed-effects models should be compared on their generalization performance on withheld levels of the random effect(s), e.g., using a cross-validation analysis.

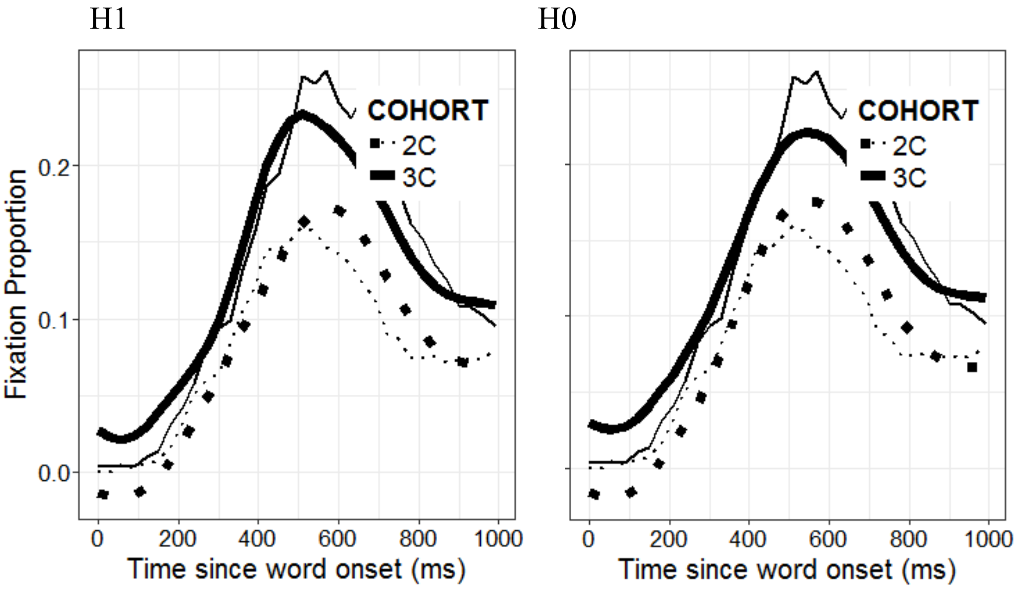


Figure 4.7. Example plots of models that captured the Cohort Type effect in H0 model (left = H1 and right = H0).

Table 4.10 illustrates the types of t-tests (BIC comparison) that were performed for between conditions analysis (1C vs. 2C, 2C vs. 3C, & 2R vs. 3R) in both speech types (natural & synthesized).

Table 4.10. Summary of BIC comparisons.

Analysis	Analysis type	Model Type	Testing data type	Random effect of Item in TRACE model	Random effect of Item in the final model
Analysis A	1. H1FYY	H1	Familiar	Yes	Yes
	2. H0FYY	H0	Familiar	Yes	Yes
Analysis B	3. H1NYN	H1	New	Yes	No
	4. H0NYN	H0	New	Yes	No
Analysis C	5. H1FNN	H1	Familiar	No	No
	6. H0FNN	H0	Familiar	No	No
Analysis D	7. H1NNN	H1	New	No	No
	8. H0NNN	H0	New	No	No

For example, H1FYY means that H1 (model) with F(amiliar item), Y(es for the random effect of Item in TRACE model), Y(es for the random effect of Item in the final model).

The new analysis was performed using a cross-validation analysis using the following steps; 1) split the data into training and testing, 2) build predictive models (H0 and H1) based on the training data, 3) apply the predictive model to the testing data, and 4) compare BIC values of models to determine the more likely model given the test data. Whereas previous analysis used the same items for fitting and testing the models, the new analysis randomly extracted 60% of the TRACE data from both conditions as training data for GAM models with the random effect of Item (H1 and H0). Similarly to the previous analysis, the TRACE prediction models were then applied to the test human data. In contrast to previous analyses, the human data came either from the same items as the training data (Analysis A: familiar items) or from the 40% of items that were not selected for training the model (Analysis B: new items). Analysis A is similar to the previous analysis that used the same items are used throughout the analysis with random Item effects, which creates a problem due to random Item effects capturing the fixed effect of Condition. In contrast, Analysis B is a cross-validation analysis because it contains different items between training and testing data so that random Item effects cannot capture the effect of Condition. Each analysis was repeated 100 times and BIC values of H1 and H0 were collected each time. Analyses C and D were performed using the same procedure as Analysis A and B except that the models lacked the random effect of Item. Even if a model is to be tested on new items, a random effect of item can be useful in training the model because inclusion of a real random effect helps estimate the coefficients for correlated fixed effects (e.g., Barth & Kapatsinski, 2018). However, as shown by Baayen, Vasishth, Kliegl, & Bates (2017), inclusion of non-significant random effects in a GAMM model can lead to mis-estimation of fixed effects. The analyses seek to answer the following questions: 1) was the GAMM model with a random effect of Item (Analysis A: Analysis types 1 & 2) performing better than the model without random effects (Analysis B: Analysis types 3 & 4)? And if so, 2) was it only overfitting and so only doing better on familiar items (Analysis types 1, 2, 5,& 6) or was it also doing better on new items (Analysis types 3, 4, 7, & 8), therefore helping to accurately estimate fixed effects?

For each analysis below, residuals of the two models were compared using a t-test to determine whether the models with a random effect of item resulted in models with significantly smaller residuals, i.e., a better fit to the data. Every test turned to be non-significant. Having the random effect of Item did not improve any model. There is therefore no reason to include the random effect of Item (Baayen et al., 2017). Thus, the random effect of Item was excluded from the final analysis.

Figure 4.8 shows that the difference between conditions (Cohort 2 and 3) in the final Analysis (D) is captured only by the H1 model and not by the H0 model, meaning that there is a difference in looks between 2C and 3C in the H1 model while there is no difference between 2C and 3C in the H0 model.

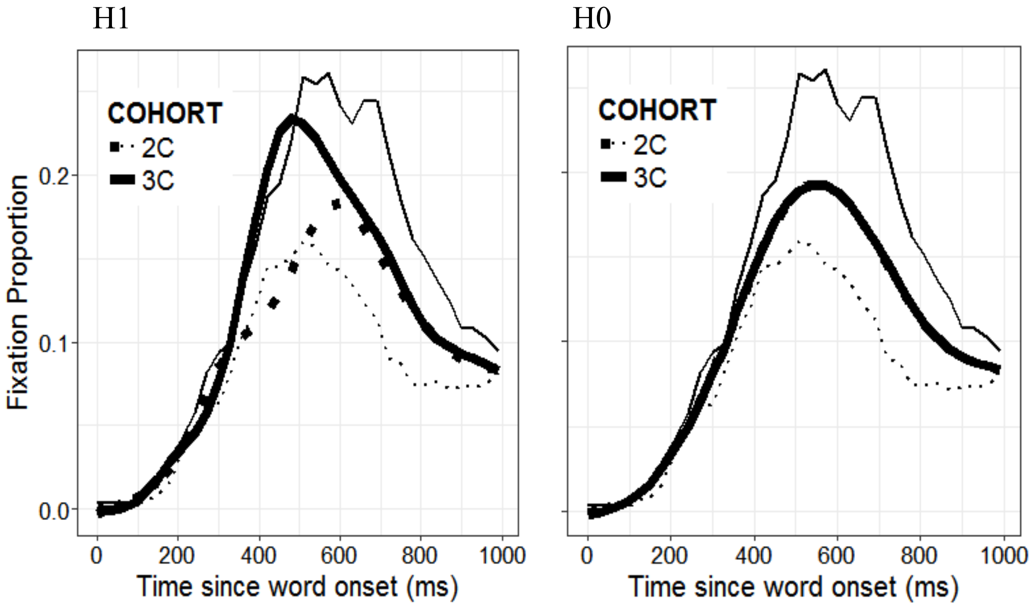


Figure 4.8. Example plots of models that eliminated the Cohort Type effect in H0 model (left = H1 and right = H0).

Due to the random effect of Item in the previous analyses capturing condition differences in the null model, condition comparisons were often inconclusive. As Table 4.11 shows, the results without Item effects provided evidence for the H0 model in the comparison between Cohort 1 and 2 for both speech types. Human data were not explained by the difference between Cohort 1 and 2 in TRACE predictions, i.e., the bigger cohort effect in the Cohort 2 condition in TRACE.

Table 4.11 below is results of the final Bayesian analysis between condition for both speech types (natural and synthesized speech). The H1 model was supported for the comparison between the Cohort 2 and 3 conditions for both speech types. As predicted by TRACE, participants looked at the competitor more when the competitor overlapped with the target in three segments than when it overlapped in only two. There was a difference across speech types for the comparison between the Rhyme 2 and 3 conditions. The H0 model was supported for natural speech, indicating no effect of rhyme overlap, but the comparison was inconclusive for synthesized speech.

Table 4.11. Results of the final Bayesian analysis between conditions for both speech types.

Speech types	Natural speech			Synthesized speech		
Condition	1C vs. 2C	2C vs. 3C	2R vs. 3R	1C vs. 2C	2C vs. 3C	2R vs. 3R
BIC differences	8	132	9	9	61	2
Evidence for H1 or H0	H0	H1	H0	H0	H1	Neither
posterior probability of H1 or H0	0.98 (H0)	1 (H1)	0.99 (H0)	0.99 (H0)	1 (H1)	0.27 (H0)

Figure 4.9 shows raw data between condition for natural speech, synthesized speech and TRACE as reference for the Cohort between-conditions. Whereas TRACE predicts a bigger cohort effect when the cohort competitor overlaps with the target in 2

segments, this is not observed in the human data, which are more consistent with no effect of segmental overlap than with the difference predicted by TRACE.[15]

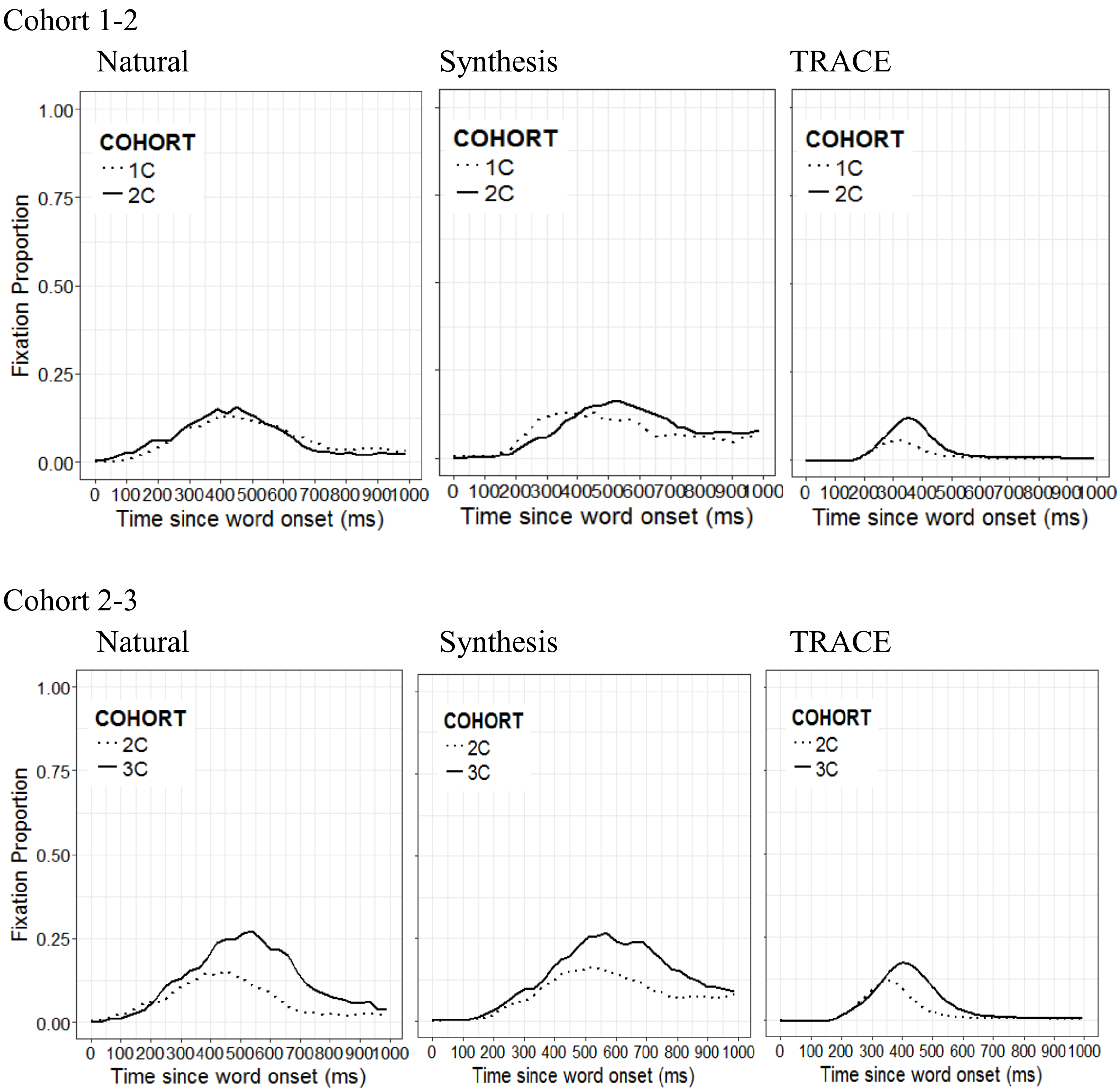


Figure 4.9. Plots of raw data in fixation proportion of competitor pictures across all items. The top row is plots of a comparison between Cohort 1 & 2, the middle row is plots for between Cohort 2 & 3 conditions.

[15] Note that the within-condition analysis in the Cohort 2 condition in natural speech suggested that the cohort effect (difference between competitor and distractor) was explained the model. However, the result of between conditions analysis suggest that looks to the competitor pictures between the conditions are the same. Preference for the H1 model in the Cohort 2 condition may therefore be due to the paucity of looks to the distractor in that condition rather than to increased looks to the competitor.

Figure 4.10 shows raw data between condition for natural speech, synthesized speech and TRACE for rhyme conditions. It appears to be slight or no difference between conditions for the human data and TRACE.

Rhyme 2-3

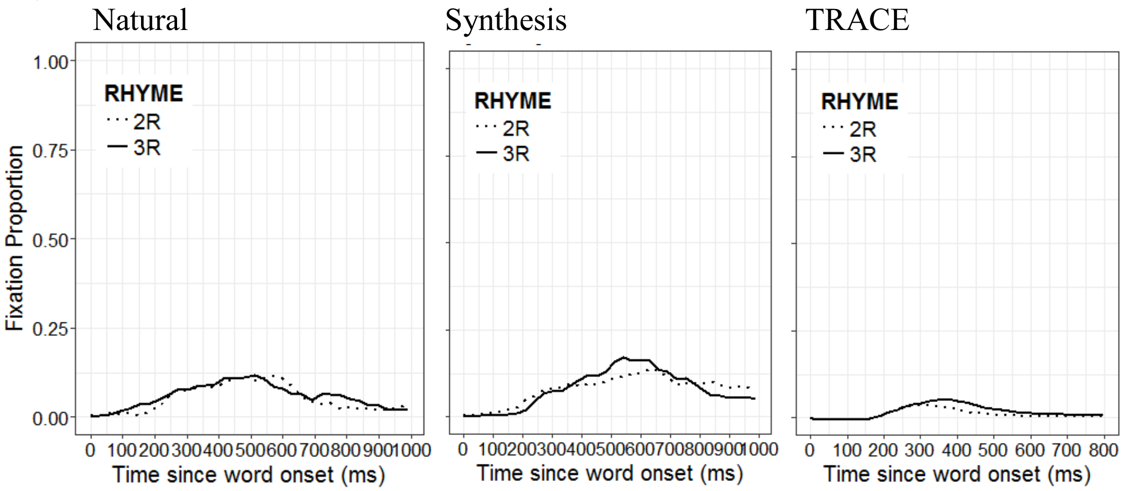


Figure 4.10. Plots of raw data in fixation proportion of competitor pictures across all items for rhyme conditions.

In summary, the TRACE model coupled with the standard linking hypothesis for the visual world paradigm (Allopenna et al., 1998) predicted that there should be sometimes subtle but detectable effects of both initial and final phonological overlap so that participants look at competitors more when they share two segments with the target than when they share only one, and more when they share three than when they share two. However, whereas the predicted difference between 2 and 3 segment initial overlap is observed in human data, the data suggest that the activation difference between 1 and 2 segment initial overlap does not affect fixations. There are also no detectable effects of initial overlap in a single segment. In addition, TRACE predicted a single-segment cohort overlap effect in the Mixed competitor condition but human eye movements showed a three-segment rhyme overlap effect instead, though only when presented with synthesized speech. Overall, this pattern of results is consistent with the proposal that a minimal amount of lexical activation is necessary to drive a saccade to the referent of a word. In the present population of participants exposed to the present task, that minimal amount corresponds to bottom-up evidence for three segments of the word one might consider fixating.

4.3. Discussion

Previous work on spoken word recognition in the visual world has provided convincing demonstrations that listeners exhibit both cohort and rhyme effects, leading to the proposal that any part of a target word can activate multiple candidate words during processing and that the activations will be reflected directly in probabilities of looking at pictures of the words' referents (e.g., Allopenna et al., 1998). The present work has tested the limits of this proposal, investigating whether even the initial phoneme or two would trigger activations of possible competitor words. Behavioral data described in the preceding chapters suggested that this is not the case. In this chapter, I confirmed that TRACE, coupled with the standard linking hypothesis, would indeed predict that two initial phonemes would be sufficient to result in lexical activation that would be noticeable in the eye movement record with the present Japanese stimuli. TRACE predictions that overlap will influence eye movements were then compared with the null hypothesis that there is no effect of overlap on eye movements using Bayesian analyses, which allow the modeler to distinguish between lack of evidence against the null and evidence for the null (Wagenmakers, 2007).

Generally speaking, TRACE predictions were accurate for cases when target-competitor overlap had a significant effect in the behavioral human data, because there was a numerical competitor effect for all conditions in TRACE. Even so, the predicted rhyme effect was very slight in the Rhyme 2 condition and the Mixed competitor condition in TRACE. Previous literature on spoken word recognition with offline tasks provided evidence of rhyme effect if the words only differed by a few features of the initial phoneme (Connine et al., 1993). Visual world eye tracking studies found the rhyme effect even though the initial phonemes were more than one phoneme feature away (Allopenna et al., 1998; Mirman et al., 2011). However, no studies observed the rhyme effect with words that differ by more than a word onset. In that respect, the TRACE prediction of near-zero rhyme effect with two-phoneme overlap and two-phoneme difference appears accurate, and is supported by the present study. However, the absence of a significant rhyme effect in the Rhyme 3 condition and the presence of a significant rhyme effect in the mixed condition in synthesized speech were not predicted by TRACE. TRACE predicted a rhyme effect for the Rhyme 3 condition (e.g., *nasu* & *basu*) but not in the Mixed condition (e.g., *kame* & *ame*). TRACE in the present study favored competitor words that differ by initial phoneme than competitor words that deleted the initial phoneme from the target word. This appears to be the opposite of human data. As discussed on Chapter 3, the rhyme effect in the Mixed condition (with deletion neighbors) may be due to late integration of initial fricatives into the word percept (Galle, 2014), which is not incorporated into the incremental input processing at the feature level in TRACE.

In summary, this chapter investigated the behavior of the TRACE model and confirmed that it predicts robust cohort effects and somewhat less robust rhyme effects. However, the behavior of the model appeared to diverge in systematic ways from that of human participants. In particular, cohort effects are less robust in human participants in the present study when the cohort competitor shares only one or two segments with the target. This lack of sensitivity to low amounts of overlap is the crucial evidence that looking at a visual representation of a word’s referent is a decision, made only when the word’s activation exceeds a context-specific threshold. Subthreshold activations do not drive saccades. The following chapter explores the parameter manipulation of the TRACE model to examine what plausible parameter changes could achieve a better fit to the human data and describe the difference between speech types.

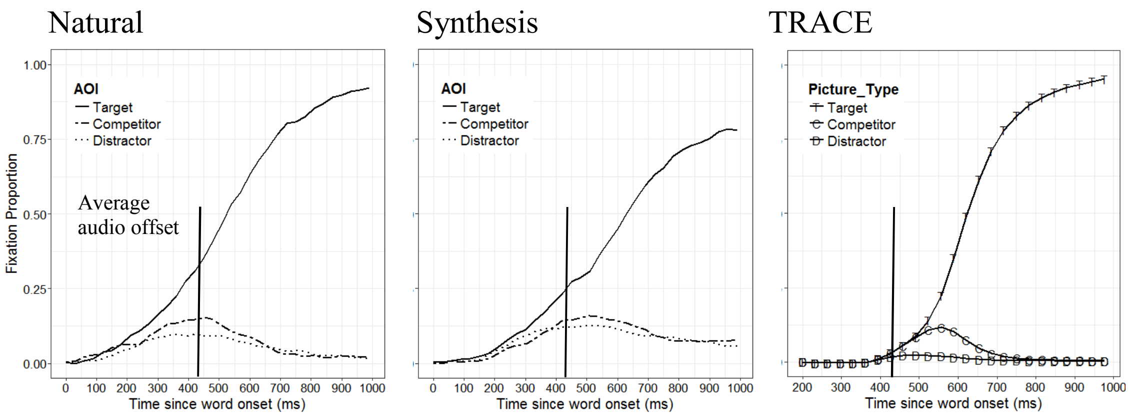
CHAPTER V

EXPERIMENT 4: PARAMETER MANIPULATION OF THE TRACE MODEL

5.1. Introduction

As noted earlier, the present study did not manipulate most of the parameters one might consider to be affecting word recognition, with the exception of the contrast between synthesized and natural speech, where synthesized speech is less clear than natural speech and has no long-distance co-articulation. There are other manipulations one could do in order to examine the effects of task manipulations on TRACE predictions and human behavior. For example, one could ask participants to perform a secondary task during word recognition to reduce word activation. The TRACE simulations reported above are based on default parameter settings because unmotivated parameter manipulation provides TRACE with virtually unlimited flexibility to fit any data pattern (McMurray et al., 2010; Norris & McQueen, 2008). However, it is worthwhile to examine what plausible parameter changes could achieve a better fit to the human data and describe the difference between speech types in the Mixed condition, where a rhyme effect is observed only in the synthesized speech condition.

There are apparent visual differences in plots between TRACE predictions and human data, as follows: 1) there are fewer Cohort fixations in the Cohort 1 & 2 conditions in human data than in TRACE data (Figure 5.1 for the Cohort 2 condition);

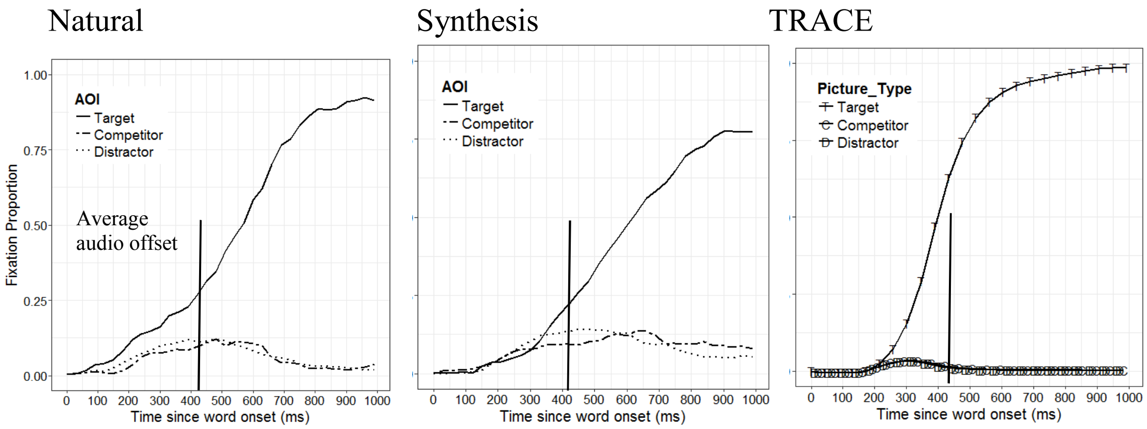

Natural

Synthesis

TRACE

Figure 5.1. An illustration of the cohort effects across speech types (Cohort 2).

2) there is a rhyme effect in the Mixed condition in synthesized speech data that is absent from both natural speech data and TRACE (Figure 5.2);

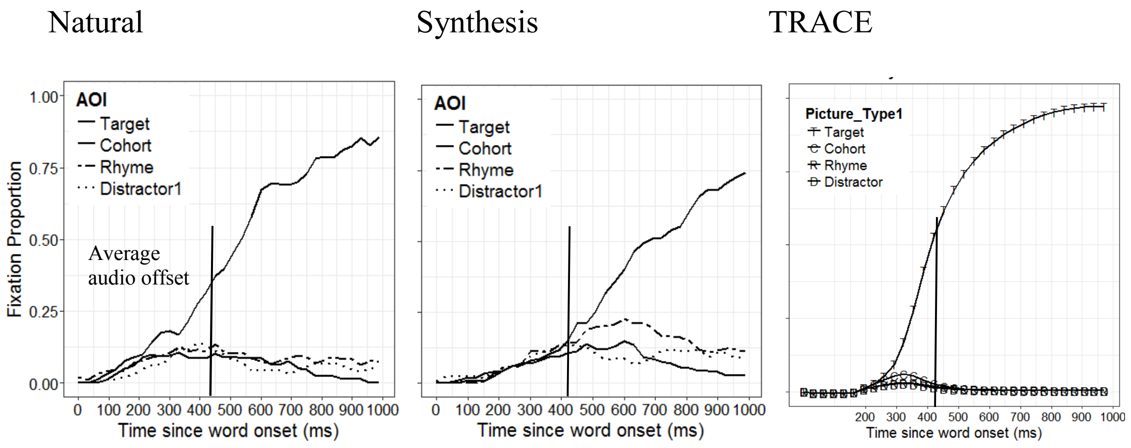

Natural

Synthesis

TRACE

Figure 5.2. An illustration of the divergence difference across speech types (Rhyme 2).

3) there is a slower divergence of looks to target pictures from other pictures in the synthesized speech data (about 400 ms) than in natural speech data and in TRACE about 200 ms (Figure 5.3).



Natural Synthesis TRACE

Figure 5.3. Rhyme effect in the Mixed condition by speech types.

5.2. Methods

5.2.1. Parameters

There are 40 parameters in jTRACE that can be manipulated. However for both practical and theoretical reasons and based on investigation of previous studies (McMurray et al., 2010; Mirman et al., 2011), six parameters were chosen for investigation. These parameters were selected to possibly account for the failure of the model (no effect for the Cohort 1 & 2 conditions), stimulus difference (natural & synthesis), the rhyme effect for synthesized speech, and divergence time difference between speech types.

The input noise parameter adds noise to the acoustic (featural) input. Synthesized speech used in the present study did not contain actual noise. However, synthesized speech could be perceived less clearly than natural speech, which could lead to delayed recognition and increased rhyme effects (Farris-Trimble et al., 2014; McMurray et al., 2017). As discussed in Chapter 3 in reporting on the synthesized speech experiment, less clearly spoken word may increase rhyme effects in the Mixed condition in synthesized speech. In fact, Mirman et al. (2011) found increasing rhyme effects in TRACE as values of input noise increased and decreasing cohort effect as input noise increased. In addition, inaccurate perception in input may cause a delay of activation which could be a source of later divergence of target fixation in synthesized speech (Farris-Trimble et al., 2014; McMurray et al., 2017).

The attention parameter controls quickness of response to input. In TRACE, when attention to the auditory signal decreases, the rise in activation based on bottom-up perceptual input also slows (Mirman et al., 2011). Slower activation of the input may be a cause of later divergence of looks to the target and cohort competitors from looks to unrelated distractors in synthesized speech data.

The rest.w parameter is responsible for the degree to which resting activation (i.e., top-down expectations) influences the activation of a candidate word. Resting activations that are greater than 0 reduce competition (interaction), which reduces cohort/rhyme effect (Mirman et al., 2011). A greater reliance on top-down expectations in humans may account for the reduced competitor effects in human data compared to TRACE.

The gamma.w parameter specifies the speed of deactivation of word competitors which is controlled by the strength of inhibition between words. This does not appear to affect the rhyme effect (e.g., Mirman et al., 2011), possibly because the rhyme is not deactivated until late during word recognition. When inhibition of word candidates decreases, the cohort effect increases. On the other hand, when inhibition of word candidates increases, the cohort effect decreases (e.g., McMurray et al., 2010; Mirman et al., 2011). The value of this parameter may be higher when speech is degraded because the listener should take perceptual evidence against a word with a grain of salt when it is unreliable (e.g., Gwilliams et al., 2018).

The aLPHA[fp] (Alpha_fp in the present paper) parameter handles the phoneme activation rate from feature input. When activation of phonological representations from features decreases, fixation of target and cohort pictures is reduced and activation of the word become slower (McMurray et al., 2010). The lower clarity of synthesized speech may cause participants to rely more on bottom-up processing than top-down processing, reducing top-down activation flow.

The aLPHA[pw] (Alpha_pw) parameter is similar to aLPHA[fp] in relation to activation of words from phonemes. When activation of word representations from phonemes decreases, fixations of target and cohort pictures are reduced and activation of the word becomes slower. (McMurray et al., 2010). Table 5.1 below summarizes parameter manipulation.

Table 5.1. Summary of parameter manipulation.

Parameter (jTRACE)	Parameter explanation	Predicted outcome	Default value	Tested values
Input Noise	Added noise over input (lower perceptual fidelity)	- Cohort effects decrease and Rhyme effects increase as values increase - Slower divergence of looks to target picture from looks to other pictures - The parameter may be higher for synthesized speech	0	0 0.3 0.6 0.9

(Table 5.1. continued.)

Parameter (jTRACE)	Parameter explanation	Predicted outcome	Default value	Tested values
Attention	Responsiveness to input	- Slower activation of input with decreasing values - Slower divergence of looks to target picture from other pictures as values decrease - The parameter may be lower for synthesized speech	1.0	0.4 0.8 1.0 1.2
rest.w	Resting activation of word candidates	- Decreased Cohort & Rhyme effects as values increase - The parameter may be higher in synthesized speech	-0.01	-0.025 -0.0175 -0.01 -0.0025 0.005
gamma.w	Deactivation of word competitors	-Decreasing cohort effects as values increase - The parameter may be lower for synthesized speech	0.03	0.01 0.02 0.03 0.04 0.05
aLPHA [fp]	phoneme activation rate from feature	- Reduced fixations to target and cohort pictures as values decrease - Slower activation of words as values decrease - The parameter may be lower for synthesized speech	0.02	0.0025 0.0075 0.01 0.02 0.035
aLPHA [pw]	word activation rate from phoneme	- Reduced fixations to target and cohort pictures as values decrease - Slower activation of words as values decrease - The parameter may be lower for synthesized speech	0.05	0.01 0.03 0.05 0.07

5.2.2. Data Processing and Analysis

Predictions of the TRACE model with a particular set of parameter settings were compared to the human data from each condition. Unfortunately, because the space of possible parameter settings is so large, and jTRACE requires rerunning the model manually for each item for every combination of settings, it was only feasible to manipulate the parameters one by one. That is, all parameter values not mentioned below remain at their default values. While this means that it should be possible to find a better combination of parameter settings by changing multiple parameters from their default values, the aim here is simply to show how the changes in the parameters influence the behavior of the model rather than to find an optimal combination of parameter settings.[16]

Since part of what the present study aims to explain is the time point at which looks to the target diverge from looks to unrelated words, Target fixation were included in this analysis. The prediction model[17] was applied to the human data to obtain predicted values of Fixation Proportion in order to evaluate the model fit.[18] Then the root mean square errors (RMSE) were calculated to evaluate the model fit for each parameter setting. The fit to all data obtained from participants presented within a particular condition within a particular speech type is evaluated.[19] While it appears impossible for participants to set parameters of their mental models to different values between conditions, as trials from different conditions are all part of the same randomly ordered block, results are reported separately for each condition to show what would be required for TRACE to capture the results observed in that condition. Parameter settings that hold across

[16] Prior work on parameter settings in TRACE (Mirman et al., 2011) has limited itself to examining TRACE predictions for three items. The present study selected one trial that was closest to the average within a condition based on random effects in the human data analyses in Chapters 2-3, which resulted in manually running a single simulation 126 times (1 trial x 6 conditions x 21 manipulations); with 3 items, 378 simulations would be required.

[17] Prediction model <- gam (Fixation Proportion ~ s (Time, by=Picture Type) + Picture Type + s(Time), data=TRACE)

[18] Prediction values <- predict (Prediction model, newdata=Natural data)

[19] One could instead evaluate the model fit separately for each picture type. For example, examining fixation proportions of Target pictures to see which parameter settings result in the best model fit. However, this would allow different parameter settings to explain looks to different picture types. For example, Input_Noise may be heavily involved in fitting fixations to Target pictures, whereas Alpha_pw may explain looks to competitor pictures. This kind of result would be difficult to interpret psychologically.

conditions within a speech type can then be interpreted as the settings that speech type may effect in the participants exposed to it.

5.2.3. Results

Tables 5.2 below shows the parameters whose settings are most important to change to improve the fit to the natural speech data and Table 5.3 shows the results for the synthesized speech data (see Appendix E & F for a complete ranking table with RMSE values within each condition for each speech type). For both speech types, the same parameters are involved in improving fit to human data, which are Alpha_pw, Alpha_fp, Attention and Input_Noise in order. Decreasing values of these parameters produced a better fit of the models, except for the Input noise parameter whose value needs to be increased from the default to add noise.

Table 5.2. Parameter settings that result in the five best fits in natural speech data. For each setting, all other parameters are set to default values.

	Cohort 1		Cohort 2		Cohort 3	
Parameters and their values	1.Input_Noise	0.6	1.Alpha_fp	0.0075	1.Alpha_pw	0.01
	1.Alpha_pw	0.01	1.Attention	0.4	2.Attention	0.4
	2.Alpha_fp	0.0075	1.Alpha_pw	0.01	3.Alpha_fp	0.0075
	2.Input_Noise	0.3	1.Alpha_fp	0.01	4.Alpha_fp	0.0025
	2.Input_Noise	0.9	1.Input_Noise	0.9	5.Alpha_fp	0.01

	Rhyme 2		Rhyme 3		Mixed	
Parameters and their values	1.Alpha_fp	0.0075	1.Alpha_fp	0.0075	1.Alpha_pw	0.01
	2.Alpha_fp	0.01	2.Input_Noise	0.6	2.Alpha_fp	0.0025
	2.Attention	0.4	2.Alpha_pw	0.01	3.Alpha_fp	0.0075
	3.Alpha_pw	0.01	3.Attention	0.4	4.Alpha_fp	0.01
	3.Alpha_pw	0.03	4.Alpha_fp	0.01	4.Attention	0.4

Note: the number specified to the left of the parameter name represents the rank of the best fit in each condition. Some parameters ranked equally.

Table 5.3. The best five parameter settings for synthesized speech data.

	Cohort 1		Cohort 2		Cohort 3	
Parameters and their values	1.Alpha_pw	0.01	1.Alpha_pw	0.01	1.Alpha_pw	0.01
	1.Input_Noise	0.9	2.Alpha_fp	0.0025	2.Alpha_fp	0.0025
	2.Input_Noise	0.6	3.Alpha_fp	0.0075	3.Attention	0.4
	2.Alpha_fp	0.0025	3.Attention	0.4	4.Alpha_fp	0.0075
	3.Alpha_fp	0.0075	4.Input_Noise	0.9	5.Alpha_fp	0.01

	Rhyme 2		Rhyme 3		Mixed	
Parameters and their values	1.Alpha_pw	0.01	1.Alpha_pw	0.01	1.Attention	0.4
	2.Alpha_fp	0.0075	2.Alpha_fp	0.0025	2.Input_Noise	0.6
	3.Alpha_fp	0.0025	3.Input_Noise	0.6	2.Alpha_fp	0.01
	3.Attention	0.4	3.Attention	0.4	2.Alpha_fp	0.0075
	3.Alpha_fp	0.01	4.Alpha_fp	0.0075	2.Alpha_pw	0.03

Note: the number specified to the left of the parameter name represents the rank of the best fit in each condition. Some parameters ranked equally.

Decreasing bottom-up activation flow to word / phoneme representation from lower levels can explain slower increases in fixation proportion observed in the human data. Divergence of looks to target from other pictures becomes later in time course with parameter manipulation. Alpha_pw and Alpha_fp were expected to be higher values in natural speech to account for earlier divergence of looks to the target compared to synthesized speech. However, this was not the case. Setting the parameters to lower values captured the late divergence for synthesized speech but also improved fit to the natural speech data, which actually had an earlier divergence.

Reducing Alpha_pw to 0.01 (Default is 0.05) produced the best fit to human data with both speech types. In addition, reducing Alpha_fp to 0.0075 (Default is 0.02) produced the best fit to natural speech data. However, there is some evidence that noise appears to be greater in synthesized speech: increasing noise shows up in the top five ranking of parameter changes slightly more often in synthesized speech (for 4 conditions) than in natural speech (for 3 conditions). It appears that a high setting of the Input_Noise parameter describes synthesized speech better than other parameters, which suggests that the Input_Noise parameter may be responsible for the difference between the speech types.

However, it is not clear that this is sufficient: different types of speech degradation are likely to change spoken word recognition in different ways. In particular, I have argued that diphone synthesized speech may make fricatives and stop bursts harder to integrate with the rest of the speech signal. This is quite different from spectral degradation (noise vocoding) examined by Farris-Trimble and colleagues (Farris-Trimble et al., 2014), which appears to be a better fit to the noise parameter. In fact, visual inspection of plots (Figure 5.4 below) of predicted fixation proportions suggests that the optimized settings of Input_Noise (0.9, right panel) or Alpha_pw (0.01, left panel) did not capture the human data well, i.e., the thick lines, the model predictions, did not track the thin lines. In particular, setting the input noise parameter to a high level produces a poor fit to the trajectory of looks to the target and generally underestimates competition between the target and other lexical candidates.

It cannot therefore be concluded that the Input_Noise parameter was responsible for the slower divergence of looks to signal-consistent and inconsistent referents and the stronger rhyme effect in the Mixed condition in synthesized speech. Reduced attention to the bottom-up input is also somewhat consistent with the effects of synthesized speech. However, neither parameter manipulation captures the difference between the rhyme effects in the Mixed condition in natural vs. synthesized speech, providing some additional support for the proposal that this difference could be due to the difficulty of integrating a diphone synthesized fricative or stop into the speech stream. As stream segregation is not modeled by TRACE, it has little hope of capturing this difference.

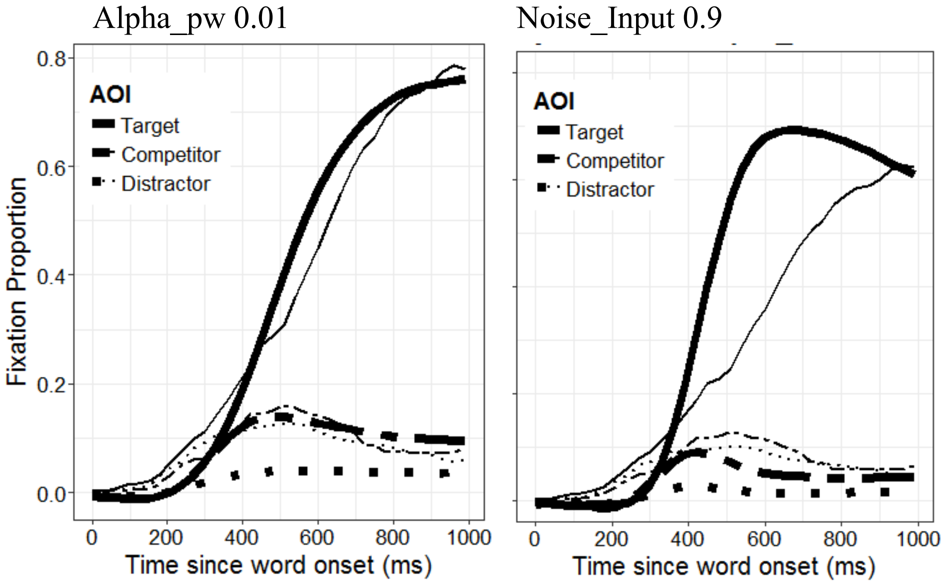


Figure 5.4. The best parameter setting, Alpha_pw 0.01, in the left panel and the second best setting, Input_Noise 0.9, in the right panel for the Cohort 2 condition in synthesized speech.

In addition, the competitor effects in Cohort 2 and 3 conditions are weaker in human behavioral data than in TRACE. The fit of TRACE was expected to be improved in this respect by the parameter manipulations, but this was not the case. As argued earlier, the weakness of the cohort effects may be due to participants imposing a threshold on activations so that when activation is too low, the eyes do not move. For some participants, even three segments may not be enough, resulting in eye movements that go directly to the target referent despite lexical competition

In summary, reducing the Alpha_pw parameter and the Alpha_fp parameter generated the best model fit for both speech types. However, it did not explain the difference between synthesized and natural speech because the same setting of this parameter was optimal for both speech types. Likewise, the Input_Noise parameter did not explain the difference between the speech types. Moreover, no parameter setting captured the rhyme effect in the Cohort and Rhyme Mixed condition for synthesized speech, which was missing in natural speech and TRACE.

5.3. Discussion

TRACE predictions could potentially be improved by changing parameter settings from their default values, which could also help explain task / stimulus set differences. Mirman et al. (2011) and McMurray et al. (2010) investigated whether parameter setting differences in TRACE can explain the effects of language deficits on spoken word recognition. In the study of McMurray et al. (2010), items presented to participants in the experiment and those in the TRACE simulation differed. Nonetheless, TRACE captured the qualitative patterns in human data, which has led the cuthors to conclude that it provides a plausible account of individual differences.

Parameter manipulations of lexical decay, lexical activation rate (Alpha_pw), and generalized slowing (Alpha_fp & Alpha_pw) had the greatest role in explaining the individual differences in human data in McMurray et al. (2010) for specific language / cognitive impaired participants. Interestingly, the same parameters, except lexical decay, also improved the fit between TRACE and human data in the present study the most. These results therefore support the possibility that these parameters may vary across studies. While one needs to be careful with making strong conclusions in favor of a model this complex based on its ability to fit the human data under some combination of parameter settings, the fact that the same few parameters appear to be controlling differences in human behavior across studies is encouraging.

Input_Noise and Attention parameters in the present study were expected to explain the slower divergence of looks to target and an increase of the rhyme effect in the Mixed condition in synthesized speech. Instead decreasing bottom-up activation flow to words and phoneme representations (Alpha_pw & Alpha_fp) provided the best fit to human data. The parameter manipulations slowed the activations of words. Because looks to each picture rise more slowly in human data than in TRACE, parameters that can slow down the rise in activation improve the fit more than other parameters. However, fixations to each picture may rise more slowly in human data than in TRACE because of averaging over individuals: some people may rapidly look at pictures while other people take longer to respond, and some may fixate a picture for longer than others. Therefore, after averaging all the subjects' trials, the curves of fixation rise and fall more slowly than in TRACE (see also Gallistel, Fairhurst, & Balsam, 2004, for averaging artifacts in modeling learning curves). If the slow rise and fall of fixations in averaged human data is indeed an averaging artifact, then alpha parameters can capture but perhaps not explain the curve shapes.

While it is possible that the slower activation and persistence of fixation in the human data may be due to word activation rate from phoneme (Alpha_pw), no parameters seemed to explain the experimental manipulation in the present study (speech type). One condition had normal (natural) speech while the other had no long-distance coarticulation and slight degradation of auditory cues due to synthesis. The effects of diphone synthesis appeared to be similar to those previously observed with cochlear implant simulations (Farris-Trimble et al., 2014; McMurray et al., 2017), despite the much lower level of spectral degradation in the present study. The effect of speech type was not explained by the parameter manipulations: the same parameter settings demonstrated the best fit for both speech types. No parameter specifically improved the ability of the model to fit the larger rhyme effect in the Mixed competitor condition in synthesized speech and the later divergence of looks to signal-consistent and inconsistent lexical candidates in synthesized speech. I now discuss the findings of the dissertation in the context of previous work on spoken word recognition in the visual world.

CHAPTER VI

GENERAL DISCUSSION

Numerous studies suggest that the initial sound(s) of a word activate multiple candidate lexical representations (e.g., Allopenna et al., 1998; Gaskell & Marslen-Wilson, 1997; McClelland & Elman, 1986). In theory, any part of a target word can activate multiple candidate words during processing (e.g., Allopenna et al., 1998; McClelland & Elman, 1986; Norris, 1994). Several eye tracking studies using the visual world paradigm demonstrated a cohort effect as well as a rhyme effect, suggesting that both word-initial and non-initial acoustics activate the words that contain them (Allopenna et al., 1998). Furthermore, these have been used to be consistent with the TRACE model of spoken word recognition, which proposes continuous, bidirectional activation flow between words and sublexical units (McClelland & Elman, 1986). TRACE is usually described as predicting both cohort and rhyme effects, though the presence of the rhyme effect is crucially dependent on the rhyme competitor not being strongly inhibited by the target (e.g., McMurray et al., 2010). However, the fit of TRACE to human data has not often been analyzed in quantitative detail in prior work, leaving open the question of whether the cohort and rhyme effects are predicted by TRACE precisely when they are exhibited by humans. The present study has developed a methodology for evaluating TRACE predictions quantitatively.

Allopenna et al. (1998) and Tanenhaus et al. (2000) link fixation probabilities at a point in time directly to activation levels of all lexical representations given the signal experienced until that point. In this formulation, saccades can be triggered by any amount of overlap and distractors (without overlap) should always be fixated less than competitors (with overlap). Thus, for example, experiencing [k…] should drive the listener to divide his / her visual attention among the referents of all and only [k]-initial words present on the screen. They should not fixate the referents of other words (e.g., distractors), unless they are activated by top-down contextual information or have a high a priori probability / resting activation level. Because the competitors on critical trials are distractors on control trials in the present experiment and there was no pre-experiment exposure, the top-down influences and priors are controlled between critical and control trials. This means that, if the linking hypothesis proposed by Allopenna et al. (1998) holds, distractors should therefore always be fixated less than competitors on critical trials.

An alternative hypothesis proposed here is that consistency with the acoustic signal does not affect saccades when that consistency is below a certain threshold. In other words, the evidence for a word needs to exceed a threshold to drive a saccade to the word's referent. Since supporting this hypothesis means supporting the null, we conducted Bayesian analyses (Wagenmakers, 2007) that allowed us to investigate whether a particular amount of evidence for the presence of a form in the acoustic signal is more consistent with the effect predicted by TRACE or with the absence of an effect.

The present experiments provided evidence that the extent of overlap between the presented word and a lexical representation matters in the way predicted by TRACE when the overlapping parts are long but not when they are short (i.e., when the competitor word and the target shared the initial segment or the initial two segments). These results indicate that eye movements are not as closely tied to fixation probabilities of lexical representations as previously believed. Specifically, the present results are not consistent with the Allopenna et al. (1998) proposal except when the target overlaps with the competitor in three initial segments.

Note that the present study cannot be generalized as claiming that three initial segments will always constitute the minimum amount of overlap necessary to drive a saccade to the word's referent because the threshold will change based on many other contextual factors (see also Brown-Schmidt & Tanenhaus, 2008). We return to this issue later in this discussion.

While TRACE failed to explain the human data in the study, it successfully explained human data in several previous visual world studies (e.g., Allopenna et al., 1998; Dahan et al., 2001a; McMurray et al., 2010; McMurray, Tanenhaus, & Aslin, 2009; Mirman et al., 2011). The major reason for this difference in conclusions appears to be that previous studies investigated TRACE predictions for human data that exhibited a difference between looks to related competitors and unrelated distractors. However, most of the competitor conditions in the present study did not exhibit this competitor-distractor difference. TRACE was unable to predict this lack of differences because the model is based on the assumption that any part of a target word can activate multiple candidate words. Predictions of TRACE appeared to be robust to plausible manipulations of parameter settings. While this appears to be a failure of TRACE, the conclusion to take from this work is not, in my view, that spoken word recognition is not characterized by continuous activation of competing words. Rather, the failure should be traced back to the overly simple linking hypothesis that transforms lexical activations directly into fixation probabilities using the Luce Choice Rule (Allopenna et al., 1998). Rather, moving one’s eyes to a picture is a decision that needs to be explicitly modeled in future work.

Participants in the present study appear reluctant to fixate a picture unless the acoustic signal provides evidence for the initial CVC of the picture’s name. As a result, participants look at pictures of unrelated distractors as much as they look at pictures of cohort competitors when the cohort competitor shares only the initial C or CV with the target word. In the case of natural speech stimuli, information about the initial CVC is likely present from the very beginning of the stimulus, allowing for early saccades to target pictures and cohort competitors sharing the initial CVC with the target. In the case of speech produced by diphone synthesis, the first consonant and initial half of the first vowel do not provide any information about the second consonant of the CVC. Consequently, looks to the target and cohort competitor do not start increasing above the level of looks to the distractor until 200 ms after the middle of the first vowel, ~400 ms after stimulus onset. This is a significant delay relative to previous studies, where looks to cohort competitors and targets begin to diverge from looks to distractors approximately 200 ms after stimulus onset (Allopenna et al., 1998; Dahan et al., 2001a; Dahan et al., 2001b; Tanenhaus et al., 2000) or even sooner (Altmann, 2011). The divergence between target and competitors occurs earlier in the natural speech stimuli than in the synthesized speech stimuli. This is likely explained by coarticulation. Synthesized speech eliminated long-distance coarticulation and perhaps also reduced coarticulation between adjacent segments comparing to natural speech. The stronger coarticulation in natural speech provides information about the end of a target word, in particular, the second consonant. This helps the listener perceive the identity of the initial three segments of the target from the very beginning of the target word. The following results can be explained by coarticulation: 1) very early divergence of looks to the target from looks to unrelated distractors in natural speech than in synthesized speech – well before the three initial segments apparently necessary to fixate a word's referent are perceived; and 2) earlier divergence between the target and distractors in the Cohort 1 condition compared to the Cohort 2 condition in natural speech stimuli. The later divergence in synthesized speech supports the coarticulation explanation for these natural speech effects.

When stimuli and procedures used in previous studies are examined in sufficient detail, most of the results are consistent with three initial segments being necessary for a word to be activated enough to drive a saccade to its referent. Most target-cohort competitor pairs used in previous studies exhibiting cohort effects in spoken word recognition have involved at least that much overlap (including all but one of the stimuli in Allopenna et al., 1998, which were often reused in follow-up studies). As previous studies did not examine how the magnitude of cohort effects varied across stimuli, it is not clear whether stimuli featuring extensive overlap were responsible for these effects. Furthermore, in a typical visual world study, participants study pictures and their intended names shortly before an experiment. Participants who are trained in this way may be ready to activate candidate words based on very little information coming from the signal, resulting in activating cohort competitors sharing only the initial CV with the target (e.g., Dahan et al., 2001b). Making the experience more realistic by introducing noise and / or variability in word form realization may also make participants more lenient with respect to the level of support a word must receive from the signal to be plausibly present in the signal (Brouwer & Bradlow, 2011; Brouwer et al., 2012;

McQueen & Huettig, 2012). The present study did not pre-expose participants to the stimuli, minimized repetition of trials as well as the words and pictures that comprise them, and provided participants with a relatively clear signal. In some ways, then, we may have led the participants to rely on the signal for driving saccade decisions more than they would in many other situations. Evidence for such expectation-driven effects in visual world studies is provided by the finding that pre-activation from a predictive context can cause the listener to fixate a word's referent earlier than they otherwise would (e.g., Altmann & Kamide, 1999; Arai, van Gompel, & Scheepers, 2007) as well as by the existence of word frequency effects in the paradigm (Magnuson, Dixon, Tanenhaus, & Aslin, 2007; Magnuson, Tanenhaus, Aslin, & Dahan, 2003). While no studies have directly explored the effects of pre-exposure, some have raised the possibility that pre-exposure to the pictures may increase the activation of competitors (Huettig et al., 2011). If this is indeed the case for pictures, pre-exposure to the words is also potentially problematic in the same way. Uncontroversially, activation of a word is a function of the resting activation level, which is boosted by a recent experience with the word, and the support the word is receiving from the acoustic signal (e.g., Allopenna et al., 1998; McClelland & Elman, 1986; Norris & McQueen, 2008). In the absence of pre-exposure to the words, the acoustic signal may need to provide a substantial degree of support to the word for the listener to generate a saccade to a depiction of the word's referent.

Note that the present study does not claim that there is something special about the initial CVC. Specifically, the study does not claim that the initial CVC acts as a discrete 'unit of lexical access' in Japanese. Phonological analyses of Japanese posit no role for the initial CVC: it comprises the initial mora plus the onset of the following one, and is therefore larger than a mora (CV) but smaller than two morae (CVCV). It is thus an a priori implausible unit of recognition (see also Cutler & Otake, 2002). Furthermore, current models of spoken word recognition show that segmentation into discrete sublexical units is unnecessary (e.g., Arnold et al., 2017; Baayen, Shaoul, Willits, & Ramscar, 2016; Cutler & Otake, 2002; Goldinger & Azuma, 2003; McMurray et al., 2002). The amount of overlap necessary to activate a word to a level sufficient to drive a saccade to a depiction of its referent will likely vary across experiments as a function of many factors, including how reliable the acoustic signal is perceived to be (Huettig & McQueen, 2009), variability in the acoustic realizations of a word (Brouwer et al., 2012), and contextual information regarding the word's identity (Altmann & Kamide, 1999).

The present study does not conclude that listeners will always need three initial segments of a word to decide to look at a picture of its referent. Rather, it proposes that the listener needs to accumulate evidence for a word before a saccade is generated, i.e., there is a threshold below which the word's activation is not high enough to drive a saccade and will not be reflected in the eye tracking record. The existence of such a threshold is strongly supported by Bayesian analyses: an initial C or CV does not influence eye movements of the participants in the present study. The threshold itself may vary with the demands and payoffs of the task, individual differences between speakers, and characteristics of the auditory and visual stimuli used in the experiment. The impact of all these factors on the threshold deserves careful consideration and modeling. At the end of the day, making a saccade to a word's referent requires making a decision. Linking hypotheses connecting spoken word recognition to eye movements in the visual world paradigm cannot assume that eye movements will always faithfully reflect continuous differences in activation levels and ought to incorporate models of making decisions based on accumulating evidence (e.g., Mazurek, Roitman, Ditterich, & Shadlen, 2003; Ratcliff & McKoon, 2008; Usher & McClelland, 2001) as well as the costs and benefits associated with moving vs. staying put (Meier & Blair, 2013).

APPENDICES

APPENDIX A

CRITICAL TRIAL STIMULUS SETS

The word frequencies were obtained from the *Balanced Corpus of Contemporary Written Japanese* (BCCWJ) (Maekawa et al., 2014). Freq. represents word frequency per million words. Note that the average word frequency was combined between the Unrelated 1 and Unrelated 2 words.

Cohort 1

	Target	Freq.	Competitor	Freq.	Unrelated 1	Freq.	Unrelated 2	Freq.
1	mame ‘bean’	9.8	mikaN ‘tangerine’	8.2	hooki ‘bloom’	3.2	tsɯɾi ‘fishing’	12.5
2	nabe ‘pot’	29.1	netto ‘net’	30.1	tsɯɾɯ ‘crane’	5.7	kaki ‘persimmon’	4.5
3	baɾa ‘rose’	20.2	bɯɯtsɯ ‘boot’	6.9	tokee ‘clock’	25.2	ɯsagi ‘rabbit’	14.6
4	bɯdoo ‘grape’	7.1	batta ‘grasshopper’	1.7	hooki ‘bloom’	3.2	kani ‘crab’	8.2
5	kɯbi ‘neck’	115.0	kago ‘basket’	10.5	zoo ‘elephant’	11.1	saiɸɯ ‘wallet’	13.1
6	kɯmo ‘spider’	7.1	kata ‘shoulder’	82.8	batsɯ ‘x-mark’	4.1	haɾi ‘needle’	17.0
7	kame ‘turtle’	10.4	kiŋgjo ‘goldfish’	5.6	netto ‘net’	30.1	hoN ‘book’	164.0
8	hato ‘pigeon’	5.9	hebi ‘snake’	15.0	ɕatɕi ‘killer whale’	0.9	kiŋgjo ‘goldfish’	5.6
9	ɕita ‘tongue’	28.5	ɕoojɯ ‘soy sauce’	22.8	kani ‘crab’	8.2	ɯde ‘arm’	79.2
10	ɾokkaa ‘locker’	2.6	ɾiboN ‘ribon’	12.6	mado ‘window’	77.6	jagi ‘goat’	4.3
11	ɾoba ‘donkey’	3.4	ɾemoN ‘lemon’	7.4	hanabi ‘firework’	9.6	batto ‘bat’	6.1
Average Freq.		21.74		18.51			23.09	

Cohort 2

	Target	Freq.	Competitor	Freq.	Unrelated 1	Freq.	Unrelated 2	Freq.
12	nasɯ ‘eggplant’	9.0	nabe ‘pot’	29.1	kɯmo ‘cloud’	38.2	tsɯɾɯ ‘crane’	5.7
13	negi ‘green onion’	10.6	neko ‘cat’	64.8	kasa ‘umbrella’	13.8	batsɯ ‘x-mark’	4.1
14	neko ‘cat’	64.8	neʑi ‘screw’	5.1	gamɯ ‘gum’	2.7	kata ‘shoulder’	82.8
15	taki ‘waterfall’	14.5	tana ‘shelf’	15.1	ɯsagi ‘rabbit’	14.6	ɾiŋgo ‘apple’	19.5
16	bɯta ‘pig’	13.4	bɯɯtsɯ ‘boot’	6.9	sake ‘sake’	72.6	kagi ‘key’	41.4
17	kaba ‘hippopotamus’	0.7	kaki ‘oyster’	4.5	ito ‘thread’	24.9	tɕizɯ ‘map’	30.6
18	kɯtsɯ ‘shoe’	37.9	kɯtɕi ‘mouth’	220.1	tamago ‘egg’	46.0	aɾi ‘ant’	5.0
19	tsɯki ‘moon’	82.3	tsɯme ‘nail’	21.3	saiɸɯ ‘wallet’	13.1	ɾakko ‘sea otter’	0.9
20	haɕi ‘bridge’	29.9	hane ‘feather’	14.4	booɾɯ ‘ball’	46.7	ika ‘squid’	11.1
21	ɸɯne ‘ship’	72.0	ɸɯgɯ ‘puffer fish’	3.4	ame ‘rain’	95.1	tako ‘octopus’	8.0
22	ɾibon ‘ribon’	12.6	ɾisɯ ‘squirrel’	2.7	tako ‘octopus’	8.0	neʑi ‘screw’	5.1
23	ɾemoN ‘lemon’	7.4	ɾetasɯ ‘lettuce’	4.8	naiɸɯ ‘knife’	15.4	hanabi ‘firework’	9.6
Average Freq.		29.59		32.68			25.62	

Cohort 3

	Target	Freq.	Competitor	Freq.	Unrelated 1	Freq.	Unrelated 2	Freq.
24	toɾi ‘bird’	36.8	toɾa ‘tiger’	9.5	same ‘shark’	4.8	ɕika ‘deer’	7.2
25	hako ‘box’	31.0	haka ‘grave’	20.8	otɕa ‘tea’	33.9	noɾi ‘glue’	6.3
26	hane ‘feather’	14.4	hana ‘nose’	49.5	kiɾiN ‘giraffe’	3.2	sɯika ‘watermelon’	7.5
27	kamo ‘duck’	6.0	kame ‘turtle’	10.4	ɾoba ‘donkey’	3.4	ɸɯgɯ ‘puffer fish’	3.4
28	kɯma ‘bear’	16.4	kɯmo ‘spider’	7.1	wani ‘alligator’	2.5	saɾɯ ‘monkey’	14.5

29	soɾa ‘sky’	88.7	soɾi ‘sled’	3.2	kamo ‘duck’	6.0	wani ‘alligator’	2.5
30	take ‘bamboo’	11.7	taki ‘waterfall’	14.5	çiza ‘knee’	42.6	ɕoojɯ ‘soy sauce’	22.8
31	kago ‘basket’	10.5	kagi ‘key’	41.4	hatɕi ‘bee’	6.3	tsɯme ‘nail’	21.3
32	kɯmo ‘cloud’	38.2	kɯma ‘bear’	16.4	batto ‘bat’	6.1	paN ‘bread’	33.0
Average Freq.		28.19		19.20			12.63	

Rhyme 2

	Target	Freq.	Competitor	Freq.	Unrelated 1	Freq.	Unrelated 2	Freq.
33	semi ‘cicada’	5.9	kami ‘paper’	38.6	booɾɯ ‘ball’	46.7	hana ‘flower’	167.8
34	toɾa ‘tiger’	9.5	saɾa ‘plate’	22.8	aɾi ‘ant’	5.0	kami ‘hair’	61.6
35	kaki ‘persimmon’	5.5	tsɯki ‘moon’	82.3	çige ‘mustache’	13.8	inɯ ‘dog’	86.7
36	kɯtɕi ‘mouth’	220.1	hatɕi ‘pot’	8.5	inɯ ‘dog’	86.7	tokee ‘clock’	25.2
37	gomi ‘garbage’	34.5	kami ‘hair’	61.6	ebi ‘shrimp’	12.3	çiza ‘knee’	42.6
38	hata ‘flag’	10.9	ɕita ‘tongue’	28.5	bɯdoo ‘grape’	7.1	kiɾiN ‘giraffe’	3.2
39	negi ‘green onion’	10.6	jagi ‘goat’	4.3	hato ‘pigeon’	5.9	kasa ‘umbrella’	13.8
40	hone ‘bone’	35.5	jane ‘roof’	25.5	tɕizɯ ‘map’	30.6	ɾetasɯ ‘lettuce’	4.8
41	jane ‘roof’	25.5	ɸɯne ‘ship’	72.0	kaba ‘hippopotamus’	0.7	naiɸɯ ‘knife’	15.4
42	ɾisɯ ‘squirrel’	2.7	basɯ ‘bus’	42.5	haɾi ‘needle’	17.0	gamɯ ‘gum’	2.7
43	mimi ‘ear’	104.6	semi ‘cicade’	5.9	ɸɯta ‘lid’	25.7	kawa ‘river’	8.7
44	jɯki ‘snow’	70.2	waki ‘armpit’	28.9	taɾɯ ‘barrel’	2.0	momo ‘peach’	15.3
Average Freq.		44.63		35.12			29.82	

Rhyme 3

	Target	Freq.	Competitor	Freq.	Unrelated 1	Freq.	Unrelated 2	Freq.
45	naɕi ‘pear’	5.3	haɕi ‘bridge’	29.9	kɯtsɯ ‘shoe’	37.9	zoo ‘elephant’	11.1
46	basɯ ‘bus’	42.5	nasɯ ‘eggplant’	9.0	ɾokkaa ‘locker’	2.6	ɕatɕi ‘killer whale’	0.9
47	niʑi ‘rainbow’	6.7	çiʑi ‘elbow’	10.3	ɾakko ‘sea otter’	0.9	otɕa ‘tea’	33.9
48	taɾɯ ‘barrel’	2.0	maɾɯ ‘circle’	19.0	ɯɕi ‘cow’	22.9	çige ‘mustache’	13.8
49	tana ‘shelf’	15.1	hana ‘nose’	49.5	itɕigo ‘strawberry’	9.3	booɕi ‘hat’	20.6
50	kɯɾi ‘chestnut’	7.5	tsɯɾi ‘fishing’	12.5	ɕika ‘deer’	7.2	isɯ ‘chair’	47.9
51	ɸɯta ‘lid’	25.7	bɯta ‘pig’	13.4	hana ‘flower’	167.8	maɾɯ ‘circle’	19.0
52	same ‘shark’	4.8	mame ‘bean’	9.8	tokee ‘clock’	25.2	hatɕi ‘bee’	6.3
53	saɾa ‘plate’	22.8	baɾa ‘rose’	20.2	ika ‘squid’	11.1	mikaN ‘tangerine’	8.2
54	jɯbi ‘finger’	69.6	kɯbi ‘neck’	115.0	momo ‘peach’	15.3	sake ‘sake’	72.6
55	waki ‘armpit’	28.9	kaki ‘persimmon’	5.5	isɯ ‘chair’	47.9	ɯma ‘horse’	66.7
Average Freq.		20.99		26.74			30.61	

Unrelated

	Target	Freq.	Competitor	Freq.	Unrelated 1	Freq.	Unrelated 2	Freq.
56	ka ‘mosquito’	8.4	niʑi ‘rainbow’	6.7	hata ‘flag’	10.9	tamago ‘egg’	46.0
57	tsɯkɯe ‘desk’	34.7	kaki ‘persimmon’	5.5	ɸɯde ‘brush’	13.6	çiʑi ‘elbow’	10.3
58	ki ‘tree’	147.2	hoN ‘book’	164.0	ɯde ‘arm’	79.2	itɕigo ‘strawberry’	9.3
59	ɸɯe ‘whistle’	6.2	hako ‘box’	31.0	tombo ‘dragonfly’	4.4	aɕi ‘leg’	164.9
60	kome ‘rice’	14.7	ha ‘tooth’	45.1	aɕi ‘leg’	164.9	ɸɯde ‘brush’	13.6
61	ito ‘thread’	24.9	kami ‘paper’	38.6	booɕi ‘hat’	20.6	haka ‘grave’	20.8

62	mado	77.6	soɾa	88.7	jɯbi	69.6	kaeɾɯ	6.6
	‘window’		‘sky’		‘finger’		‘flog’	
63	sɯzɯ	5.3	hatɕi	8.5	kɯɾi	7.5	tombo	4.4
	‘whistle’		‘pot’		‘chestnut’		‘dragonfly’	
Average Freq.		39.88		22.61			40.41	

Cohort & Rhyme

	Target	Freq.	Cohort	Freq.	Rhyme	Freq.	Unrelated	Freq.
64	hebi	15.0	hone	35.5	ebi	12.3	saɾɯ	14.5
	‘snake’		‘bone’		‘shrimp’		‘monkey’	
65	kɯɕi	3.4	kawa	8.7	ɯɕi	22.9	ɾiŋgo	19.5
	‘chestnut’		‘river’		‘cow’		‘apple’	
66	soɾi	3.2	sɯika	7.5	oɾi	4.3	take	11.7
	‘slid’		‘watermelon’		‘cage’		‘bamboo’	
67	kame	10.4	kɯɕi	3.4	ame	95.1	sɯzɯ	5.3
	‘turtle’		‘comb’		‘rain’		‘bell’	
Average Freq.		8.00		13.78		33.65		12.75

APPENDIX B

PHONEME FEATURE SPECIFICATIONS

(TRACE)

Phoneme feature specification and values in TRACE (McClelland & Elman, 1986) (1 = very low, 8 = very high)

	Consonantal	Vocalic	Diffuseness	Acuteness	Voiced	Power	Burst
1	ɒ, i, u, ʌ	p/b t/d k/g	ɹ	ɒ, ʌ	p, t, k s, ʃ		
2			k/g l ɒ	p/b ɹ u			
3	l, ɹ			k/g			g
4		s, ʃ		ʃ l		p/b t/d k/g	k
5	s, ʃ		ʌ				d
6			ʃ u			s, ʃ	t
7		l, ɹ	p/b t/d s	t/d	b, d, g	l, ɹ ʌ	b
8	p/b t/d k/g	ɒ, i, u, ʌ	i	s i	l, ɹ ɒ, i, u, ʌ	ɒ, i, u	p

APPENDIX C

PHONEME FEATURE SPECIFICATIONS

(jTRACE)

Phoneme feature specification and values in jTRACE (Strauss et al., 2007) (1 = very low, 8 = very high). Note: the phonemes that are not specified a value in a parenthesis represent the value of 1. Cons. = Consonantal, Voc = Vocalic, Diff = Diffuseness, Voi = Voicing, Pow = Power, & Bur. = Burst. The strength levels 1 to 8 may be reversed in jTRACE (1 = very high, 8 = very low).

	Cons.	Voc.	Diff.	Acuteness				Voi.	Pow.	Bur.
1	i	l, r ɒ, i, u, ʌ		s		i		ɒ, i, u, ʌ	p b(0.2)	ɒ, i, u
2	p/b t/d s	b, d, g	ɒ, i, u, ʌ	s(0.3) t/d		i(0.3)		l, ɹ	p(0.2) b	l, ɹ
3	ʃ u			s(0.1)	ʃ(0.1)	i(0.1)			t d(0.2)	s, ʃ
4	ʌ		s, ʃ		ʃ(0.3)		k/g(0.1)		t(0.2) d	
5			l, ɹ		ʃ	u(0.1)	k/g(0.3)	s, ʃ	k g(0.2)	p/b t/d k/g

6					ʃ (0.3)	u(0.3)
						ɒ,
						ʌ(0.1)
							k/g		k(0.2)
									g	

7	k/g l ɹ(0.5)			p/b	ʃ (0.1)	u ɒ, ʌ(0.3)	k/g(0.3)			
8	ɒ l(0.5) ɹ	p, t, k s, ʃ	p/b t/d k/g			u(0.3) ɒ, ʌ	k/g(0.1)	p/b t/d k/g		

APPENDIX D

PHONEME FEATURE SPECIFICATIONS

(PRESENT STUDY)

Phoneme feature specification and values in the TRACE simulation in the present study (1 = very low, 8 = very high). Note: the phonemes that are not specified a value in a parenthesis represent the value of 1.

	Sonority	Anterior	Height	Diffuseness	Voiced	Power
1 L O W	p/b t/d k/g	h	a, N	p/b t/d k/g	s, ts, ɕ, tɕ, ç, h, ɸ	tɕ(0.2) dɕ

2	s, h, ç, ɸ, ɕ/ʑ	N, o	ŋ	ɾ	p/t/k	ts(0.2)
						dz
						tɕ
						dɕ(0.2)

3	m, n, ŋ, N	ŋ, w, a k/g(0.1) ɾ(0.1)	k/g(0.1)	ɸ, h		k(0.2), g ts dz(0.2)
4	ɾ	ɯ k/g(0.3) ɾ(0.2)	k/g(0.3) e, o	ç		k, g(0.2)
5	j, w ɾ(0.5)	ç, e k/g ɾ(0.3)	k/g	s, ts/dz, ɕ/ʑ, tɕ/dʑ		t(0.2) d
6	i, ɯ	i, j k/g(0.3) ɾ(0.5)	k/g(0.3)	m/n/ŋ/N	dz, dʑ, ʑ	t, d(0.2)
7	e, o	t, d, n, ts, dz ɾ	k/g(0.1)	j, w	b/d/g ɾ	p(0.2) b
8 H I G H	a	p, b, m, ɸ	i, j ɯ, w ɕ/ʑ, tɕ/dʑ ç, ɾ	a/i/ɯ/e/o	a/i/ɯ/e/o m/n/ŋ/N j, w	p, b(0.2)

	Acuteness						
1 L O W			h	k/g(0.1)		a/o/ɯ/w/ɸ(0.1)	N(0.5)
2		ɕ/ʑ/tɕ/dʑ(0.1)	p, b, m	k/g(0.3)	ŋ(0.5)	a/o/ɯ/w/ɸ(0.3)	N
3		ɕ/ʑ/tɕ/dʑ(0.3)	ɾ(0.5)	k/g	ŋ	a/o/ɯ/w/ɸ	
4		ɕ/ʑ/tɕ/dʑ	ɾ	k/g(0.3)		e(0.5)	i/j/ç(0.1)
5		ɕ/ʑ/tɕ/dʑ(0.3)		k/g(0.1)		e	i/j/ç(0.3)
6	s/ts/dz(0.1)	ɕ/ʑ/tɕ/dʑ(0.1)					i/j/ç
7	s/ts/dz(0.3)		t, d, n				
8 H I G H	s/ts/dz						

APPENDIX E

RMSE FOR EACH CONDITION

(NATURAL SPEECH)

Cohort 1: Natural Speech

	Speech_Type	Condition	Parameter	Value		RMSE
1	Natural	1C	Input_Noise	0.6		0.278275
2	Natural	1C	Alpha_pw	0.01		0.279562
3	Natural	1C	Alpha_fp	0.0075		0.281567
4	Natural	1C	Input_Noise	0.3		0.283662
5	Natural	1C	Input_Noise	0.9		0.283828
6	Natural	1C	Attention	0.4		0.286839
7	Natural	1C	Alpha_fp	0.01		0.287192
8	Natural	1C	Alpha_pw	0.03		0.292937
9	Natural	1C	Alpha_fp	0.0025		0.299916
10	Natural	1C	Gamma.w	0.04		0.303377
11	Natural	1C	Gamma.w	0.05		0.303771
12	Natural	1C	Attention	0.8		0.304899
13	Natural	1C	Input_Noise	0	Default	0.306326
14	Natural	1C	Attention	1	Default	0.306326
15	Natural	1C	Gamma.w	0.03	Default	0.306326
16	Natural	1C	Alpha_fp	0.02	Default	0.306326
17	Natural	1C	Alpha_pw	0.05	Default	0.306326
18	Natural	1C	Rest.w	-0.01	Default	0.306327
19	Natural	1C	Rest.w	0.005		0.308121
20	Natural	1C	Rest.w	-0.0025		0.308285
21	Natural	1C	Attention	1.2		0.30859
22	Natural	1C	Rest.w	-0.0175		0.308879
23	Natural	1C	Rest.w	-0.025		0.309176
24	Natural	1C	Gamma.w	0.02		0.312967
25	Natural	1C	Alpha_pw	0.07		0.315228
26	Natural	1C	Gamma.w	0.01		0.317234
27	Natural	1C	Alpha_fp	0.035		0.325507

Cohort 2: Natural Speech

	Speech_Type	Condition	Parameter	Value		RMSE
1	Natural	2C	Alpha_fp	0.0075		0.272534
2	Natural	2C	Attention	0.4		0.273224
3	Natural	2C	Alpha_pw	0.01		0.275124
4	Natural	2C	Alpha_fp	0.01		0.276343
5	Natural	2C	Input_Noise	0.9		0.277489
6	Natural	2C	Alpha_pw	0.03		0.280186
7	Natural	2C	Input_Noise	0.6		0.28071
8	Natural	2C	Input_Noise	0.3		0.290166
9	Natural	2C	Attention	0.8		0.290286
10	Natural	2C	Rest.w	-0.01	Default	0.292535
11	Natural	2C	Input_Noise	0	Default	0.292535
12	Natural	2C	Attention	1	Default	0.292535
13	Natural	2C	Gamma.w	0.03	Default	0.292535
14	Natural	2C	Alpha_fp	0.02	Default	0.292535
15	Natural	2C	Alpha_pw	0.05	Default	0.292535
16	Natural	2C	Rest.w	-0.0025		0.295573
17	Natural	2C	Rest.w	0.005		0.295655
18	Natural	2C	Attention	1.2		0.295801
19	Natural	2C	Rest.w	-0.0175		0.295992
20	Natural	2C	Gamma.w	0.01		0.296005
21	Natural	2C	Rest.w	-0.025		0.296175
22	Natural	2C	Gamma.w	0.04		0.296988
23	Natural	2C	Gamma.w	0.05		0.298771
24	Natural	2C	Gamma.w	0.02		0.300412
25	Natural	2C	Alpha_fp	0.0025		0.301936
26	Natural	2C	Alpha_pw	0.07		0.302755
27	Natural	2C	Alpha_fp	0.035		0.30488

Cohort 3: Natural Speech

	Speech_Type	Condition	Parameter	Value	RMSE
1	Natural	3C	Alpha_pw	0.01	0.303672
2	Natural	3C	Attention	0.4	0.30871
3	Natural	3C	Alpha_fp	0.0075	0.311656
4	Natural	3C	Alpha_fp	0.0025	0.31571
5	Natural	3C	Alpha_fp	0.01	0.3191
6	Natural	3C	Alpha_pw	0.03	0.321868
7	Natural	3C	Input_Noise	0.6	0.32724
8	Natural	3C	Gamma.w	0.01	0.32759
9	Natural	3C	Attention	0.8	0.334012

10	Natural	3C	Input_Noise	0.3		0.339003
11	Natural	3C	Rest.w	-0.01	Default	0.339048
12	Natural	3C	Input_Noise	0	Default	0.339048
13	Natural	3C	Attention	1	Default	0.339048
14	Natural	3C	Gamma.w	0.03	Default	0.339048
15	Natural	3C	Alpha_fp	0.02	Default	0.339048
16	Natural	3C	Alpha_pw	0.05	Default	0.339048
17	Natural	3C	Gamma.w	0.04		0.341107
18	Natural	3C	Gamma.w	0.05		0.342174
19	Natural	3C	Rest.w	-0.0025		0.342182
20	Natural	3C	Attention	1.2		0.342223
21	Natural	3C	Rest.w	-0.0175		0.342302
22	Natural	3C	Rest.w	-0.025		0.342366
23	Natural	3C	Rest.w	0.005		0.345318
24	Natural	3C	Gamma.w	0.02		0.345383
25	Natural	3C	Alpha_fp	0.035		0.351836
26	Natural	3C	Input_Noise	0.9		0.417075

Rhyme 2: Natural Speech

	Speech_Type	Condition	Parameter	Value		RMSE
1	Natural	2R	Alpha_fp	0.0075		0.267594
2	Natural	2R	Alpha_fp	0.01		0.269494
3	Natural	2R	Attention	0.4		0.269945
4	Natural	2R	Alpha_pw	0.01		0.271117
5	Natural	2R	Alpha_pw	0.03		0.272007
6	Natural	2R	Input_Noise	0.6		0.274793
7	Natural	2R	Gamma.w	0.05		0.277714
8	Natural	2R	Gamma.w	0.04		0.277834
9	Natural	2R	Input_Noise	0.3		0.278603
10	Natural	2R	Attention	0.8		0.278943
11	Natural	2R	Rest.w	-0.01	Default	0.279723
12	Natural	2R	Input_Noise	0	Default	0.279728
13	Natural	2R	Attention	1	Default	0.279728
14	Natural	2R	Gamma.w	0.03	Default	0.279728
15	Natural	2R	Alpha_fp	0.02	Default	0.279728
16	Natural	2R	Alpha_pw	0.05	Default	0.279728
17	Natural	2R	Attention	1.2		0.280588
18	Natural	2R	Rest.w	0.005		0.280704
19	Natural	2R	Rest.w	-0.025		0.281429
20	Natural	2R	Rest.w	-0.0025		0.281429
21	Natural	2R	Input_Noise	0.9		0.284031

22	Natural	2R	Alpha_pw	0.07		0.287924
23	Natural	2R	Gamma.w	0.02		0.290337
24	Natural	2R	Alpha_fp	0.035		0.29782
25	Natural	2R	Gamma.w	0.01		0.298532
26	Natural	2R	Alpha_fp	0.0025		0.300733

Rhyme 3: Natural Speech

	Speech_Type	Condition	Parameter	Value		RMSE
1	Natural	3R	Alpha_fp	0.0075		0.276516
2	Natural	3R	Input_Noise	0.6		0.278039
3	Natural	3R	Alpha_pw	0.01		0.278126
4	Natural	3R	Attention	0.4		0.281415
5	Natural	3R	Alpha_fp	0.01		0.282047
6	Natural	3R	Input_Noise	0.3		0.283726
7	Natural	3R	Alpha_pw	0.03		0.284359
8	Natural	3R	Attention	0.8		0.287788
9	Natural	3R	Input_Noise	0	Default	0.293606
10	Natural	3R	Attention	1	Default	0.293606
11	Natural	3R	Gamma.w	0.03	Default	0.293606
12	Natural	3R	Alpha_fp	0.02	Default	0.293606
13	Natural	3R	Alpha_pw	0.05	Default	0.293606
14	Natural	3R	Rest.w	-0.01	Default	0.293606
15	Natural	3R	Rest.w	-0.0025		0.298672
16	Natural	3R	Attention	1.2		0.29871
17	Natural	3R	Rest.w	-0.0175		0.298816
18	Natural	3R	Rest.w	-0.025		0.298948
19	Natural	3R	Alpha_fp	0.0025		0.299363
20	Natural	3R	Rest.w	0.005		0.300214
21	Natural	3R	Gamma.w	0.04		0.300487
22	Natural	3R	Gamma.w	0.01		0.301327
23	Natural	3R	Gamma.w	0.05		0.302249
24	Natural	3R	Gamma.w	0.02		0.305183
25	Natural	3R	Alpha_pw	0.07		0.308412
26	Natural	3R	Alpha_fp	0.035		0.312789
27	Natural	3R	Input_Noise	0.9		0.374872

Cohort & Rhyme Mixed: Natural Speech

	Speech_Type	Condition	Parameter	Value		RMSE
1	Natural	Mix	Alpha_pw	0.01		0.308799
2	Natural	Mix	Alpha_fp	0.0025		0.313205
3	Natural	Mix	Alpha_fp	0.0075		0.320116
4	Natural	Mix	Alpha_fp	0.01		0.32524
5	Natural	Mix	Attention	0.4		0.326825
6	Natural	Mix	Alpha_pw	0.03		0.330265
7	Natural	Mix	Input_Noise	0.6		0.332067
8	Natural	Mix	Gamma.w	0.05		0.332902
9	Natural	Mix	Input_Noise	0.3		0.333388
10	Natural	Mix	Gamma.w	0.04		0.336025
11	Natural	Mix	Rest.w	0.005		0.336629
12	Natural	Mix	Attention	0.8		0.336765
13	Natural	Mix	Input_Noise	0	Default	0.3373974
14	Natural	Mix	Attention	1	Default	0.3373974
15	Natural	Mix	Gamma.w	0.03	Default	0.3373974
16	Natural	Mix	Alpha_fp	0.02	Default	0.3373974
17	Natural	Mix	Alpha_pw	0.05	Default	0.3373974
18	Natural	Mix	Rest.w	-0.01	Default	0.3373974
19	Natural	Mix	Rest.w	-0.003		0.337864
20	Natural	Mix	Attention	1.2		0.338269
21	Natural	Mix	Rest.w	-0.018		0.338701
22	Natural	Mix	Rest.w	-0.025		0.339066
23	Natural	Mix	Alpha_pw	0.07		0.343892
24	Natural	Mix	Gamma.w	0.02		0.35091
25	Natural	Mix	Input_Noise	0.9		0.352198
26	Natural	Mix	Alpha_fp	0.035		0.353774
27	Natural	Mix	Gamma.w	0.01		0.358957

APPENDIX F

RMSE FOR EACH CONDITION

(SYNTHESIZED SPEECH)

Cohort 1: Synthesized Speech

	Speech_Type	Condition	Parameter	Value		RMSE
1	Synthesis	1C	Alpha_pw	0.01		0.285844
2	Synthesis	1C	Input_Noise	0.9		0.28685
3	Synthesis	1C	Input_Noise	0.6		0.292877
4	Synthesis	1C	Alpha_fp	0.0025		0.296276
5	Synthesis	1C	Alpha_fp	0.0075		0.305261
6	Synthesis	1C	Input_Noise	0.3		0.306565
7	Synthesis	1C	Attention	0.4		0.31184
8	Synthesis	1C	Alpha_fp	0.01		0.314386
9	Synthesis	1C	Alpha_pw	0.03		0.322198
10	Synthesis	1C	Gamma.w	0.05		0.33531
11	Synthesis	1C	Gamma.w	0.04		0.33531
12	Synthesis	1C	Attention	0.8		0.336578
13	Synthesis	1C	Rest.w	-0.01	Default	0.338462
14	Synthesis	1C	Input_Noise	0	Default	0.338462
15	Synthesis	1C	Attention	1	Default	0.338462
16	Synthesis	1C	Gamma.w	0.03	Default	0.338462
17	Synthesis	1C	Alpha_fp	0.02	Default	0.338462
18	Synthesis	1C	Alpha_pw	0.05	Default	0.338462
19	Synthesis	1C	Rest.w	0.005		0.341086
20	Synthesis	1C	Rest.w	-0.0025		0.341152
21	Synthesis	1C	Attention	1.2		0.341474
22	Synthesis	1C	Rest.w	-0.0175		0.341775
23	Synthesis	1C	Rest.w	-0.025		0.34208
24	Synthesis	1C	Gamma.w	0.02		0.345806
25	Synthesis	1C	Alpha_pw	0.07		0.348086
26	Synthesis	1C	Gamma.w	0.01		0.350756
27	Synthesis	1C	Alpha_fp	0.035		0.359871

Cohort 2: Synthesized Speech

	Speech_Type	Condition	Parameter	Value		RMSE
1	Synthesis	2C	Alpha_pw	0.01		0.290921
2	Synthesis	2C	Alpha_fp	0.0025		0.300695
3	Synthesis	2C	Alpha_fp	0.0075		0.309671
4	Synthesis	2C	Attention	0.4		0.311033
5	Synthesis	2C	Input_Noise	0.9		0.312765
6	Synthesis	2C	Alpha_fp	0.01		0.319024
7	Synthesis	2C	Input_Noise	0.6		0.32454
8	Synthesis	2C	Alpha_pw	0.03		0.325754
9	Synthesis	2C	Input_Noise	0.3		0.339063
10	Synthesis	2C	Attention	0.8		0.339895
11	Synthesis	2C	Rest.w	-0.01	Default	0.342478
12	Synthesis	2C	Input_Noise	0	Default	0.342478
13	Synthesis	2C	Attention	1	Default	0.342478
14	Synthesis	2C	Gamma.w	0.03	Default	0.342478
15	Synthesis	2C	Alpha_fp	0.02	Default	0.342478
16	Synthesis	2C	Alpha_pw	0.05	Default	0.342478
17	Synthesis	2C	Rest.w	-0.0025		0.346042
18	Synthesis	2C	Rest.w	0.005		0.346122
19	Synthesis	2C	Attention	1.2		0.346256
20	Synthesis	2C	Rest.w	-0.0175		0.346419
21	Synthesis	2C	Gamma.w	0.01		0.34654
22	Synthesis	2C	Rest.w	-0.025		0.346571
23	Synthesis	2C	Gamma.w	0.04		0.347311
24	Synthesis	2C	Gamma.w	0.05		0.349212
25	Synthesis	2C	Gamma.w	0.02		0.351388
26	Synthesis	2C	Alpha_pw	0.07		0.353918
27	Synthesis	2C	Alpha_fp	0.035		0.356643

Cohort 3: Synthesized Speech

	Speech_Type	Condition	Parameter	Value	RMSE
1	Synthesis	3C	Alpha_pw	0.01	0.309632
2	Synthesis	3C	Alpha_fp	0.0025	0.312697
3	Synthesis	3C	Attention	0.4	0.327746
4	Synthesis	3C	Alpha_fp	0.0075	0.337965
5	Synthesis	3C	Alpha_fp	0.01	0.348917
6	Synthesis	3C	Alpha_pw	0.03	0.35319
7	Synthesis	3C	Gamma.w	0.01	0.35377
8	Synthesis	3C	Input_Noise	0.6	0.359299

9	Synthesis	3C	Attention	0.8		0.368288
10	Synthesis	3C	Rest.w	-0.01	Default	0.373372
11	Synthesis	3C	Input_Noise	0	Default	0.373372
12	Synthesis	3C	Attention	1	Default	0.373372
13	Synthesis	3C	Gamma.w	0.03	Default	0.373372
14	Synthesis	3C	Alpha_fp	0.02	Default	0.373372
15	Synthesis	3C	Alpha_pw	0.05	Default	0.373372
16	Synthesis	3C	Input_Noise	0.3		0.373908
17	Synthesis	3C	Gamma.w	0.04		0.375539
18	Synthesis	3C	Gamma.w	0.05		0.376481
19	Synthesis	3C	Rest.w	-0.0025		0.376598
20	Synthesis	3C	Attention	1.2		0.376652
21	Synthesis	3C	Rest.w	-0.0175		0.376747
22	Synthesis	3C	Rest.w	-0.025		0.376818
23	Synthesis	3C	Rest.w	0.005		0.379985
24	Synthesis	3C	Gamma.w	0.02		0.380503
25	Synthesis	3C	Alpha_fp	0.035		0.387489
26	Synthesis	3C	Input_Noise	0.9		0.401669

Rhyme 2: Synthesized Speech

	Speech_Type	Condition	Parameter	Value		RMSE
1	Synthesis	2R	Alpha_pw	0.01		0.287074
2	Synthesis	2R	Alpha_fp	0.0075		0.299364
3	Synthesis	2R	Alpha_fp	0.0025		0.302308
4	Synthesis	2R	Attention	0.4		0.304078
5	Synthesis	2R	Alpha_fp	0.01		0.305034
6	Synthesis	2R	Alpha_pw	0.03		0.310508
7	Synthesis	2R	Input_Noise	0.6		0.314768
8	Synthesis	2R	Input_Noise	0.9		0.315906
9	Synthesis	2R	Input_Noise	0.3		0.318776
10	Synthesis	2R	Gamma.w	0.04		0.319256
11	Synthesis	2R	Gamma.w	0.05		0.319404
12	Synthesis	2R	Attention	0.8		0.320165
13	Synthesis	2R	Input_Noise	0	Default	0.320435
14	Synthesis	2R	Attention	1	Default	0.320435
15	Synthesis	2R	Gamma.w	0.03	Default	0.320435
16	Synthesis	2R	Alpha_fp	0.02	Default	0.320435
17	Synthesis	2R	Alpha_pw	0.05	Default	0.320435
18	Synthesis	2R	Rest.w	-0.01	Default	0.320436
19	Synthesis	2R	Attention	1.2		0.321627
20	Synthesis	2R	Rest.w	0.005		0.321744

21	Synthesis	2R	Rest.w	-0.025	0.32263
22	Synthesis	2R	Rest.w	-0.0025	0.32263
23	Synthesis	2R	Alpha_pw	0.07	0.329801
24	Synthesis	2R	Gamma.w	0.02	0.334255
25	Synthesis	2R	Alpha_fp	0.035	0.342261
26	Synthesis	2R	Gamma.w	0.01	0.343175

Rhyme 3: Synthesized Speech

	Speech_Type	Condition	Parameter	Value		RMSE
1	Synthesis	3R	Alpha_pw	0.01		0.282339
2	Synthesis	3R	Alpha_fp	0.0025		0.290771
3	Synthesis	3R	Input_Noise	0.6		0.295724
4	Synthesis	3R	Attention	0.4		0.296773
5	Synthesis	3R	Alpha_fp	0.0075		0.308164
6	Synthesis	3R	Alpha_fp	0.01		0.318537
7	Synthesis	3R	Input_Noise	0.3		0.321288
8	Synthesis	3R	Alpha_pw	0.03		0.322388
9	Synthesis	3R	Attention	0.8		0.326763
10	Synthesis	3R	Input_Noise	0	Default	0.334838
11	Synthesis	3R	Attention	1	Default	0.334838
12	Synthesis	3R	Gamma.w	0.03	Default	0.334838
13	Synthesis	3R	Alpha_fp	0.02	Default	0.334838
14	Synthesis	3R	Alpha_pw	0.05	Default	0.334838
15	Synthesis	3R	Rest.w	-0.01	Default	0.334838
16	Synthesis	3R	Gamma.w	0.01		0.337569
17	Synthesis	3R	Rest.w	-0.0025		0.341363
18	Synthesis	3R	Attention	1.2		0.341367
19	Synthesis	3R	Rest.w	-0.0175		0.341451
20	Synthesis	3R	Rest.w	-0.025		0.341561
21	Synthesis	3R	Rest.w	0.005		0.343347
22	Synthesis	3R	Gamma.w	0.04		0.343389
23	Synthesis	3R	Gamma.w	0.05		0.345273
24	Synthesis	3R	Gamma.w	0.02		0.348696
25	Synthesis	3R	Alpha_pw	0.07		0.352387
26	Synthesis	3R	Alpha_fp	0.035		0.357547
27	Synthesis	3R	Input_Noise	0.9		0.363778

Cohort & Rhyme Mixed: Synthesized Speech

	Speech_Type	Condition	Parameter	Value		RMSE
1	Synthesis	Mix	Attention	0.4		0.288699
2	Synthesis	Mix	Input_Noise	0.6		0.291236
3	Synthesis	Mix	Alpha_fp	0.01		0.291263
4	Synthesis	Mix	Alpha_fp	0.0075		0.291687
5	Synthesis	Mix	Alpha_pw	0.03		0.29195
6	Synthesis	Mix	Input_Noise	0.3		0.292184
7	Synthesis	Mix	Attention	0.8		0.293387
8	Synthesis	Mix	Rest.w	0.005		0.293676
9	Synthesis	Mix	Input_Noise	0	Default	0.2938751
10	Synthesis	Mix	Attention	1	Default	0.2938751
11	Synthesis	Mix	Gamma.w	0.03	Default	0.2938751
12	Synthesis	Mix	Alpha_fp	0.02	Default	0.2938751
13	Synthesis	Mix	Alpha_pw	0.05	Default	0.2938751
14	Synthesis	Mix	Rest.w	-0.01	Default	0.2938751
15	Synthesis	Mix	Gamma.w	0.05		0.294217
16	Synthesis	Mix	Rest.w	-0.0025		0.294218
17	Synthesis	Mix	Gamma.w	0.04		0.29434
18	Synthesis	Mix	Attention	1.2		0.294436
19	Synthesis	Mix	Rest.w	-0.0175		0.294672
20	Synthesis	Mix	Rest.w	-0.025		0.294889
21	Synthesis	Mix	Alpha_pw	0.01		0.294903
22	Synthesis	Mix	Alpha_pw	0.07		0.29696
23	Synthesis	Mix	Gamma.w	0.02		0.300729
24	Synthesis	Mix	Alpha_fp	0.035		0.302528
25	Synthesis	Mix	Gamma.w	0.01		0.305587
26	Synthesis	Mix	Alpha_fp	0.0025		0.308483
27	Synthesis	Mix	Input_Noise	0.9		0.341121

REFERENCES CITED

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. https://doi.org/10.1006/jmla.1997.2558

Altmann, G. T. M. (2011). Language can mediate eye movement control within 100milliseconds, regardless of whether there is anything to move the eyes to. *Acta Psychologica*, *137*(2), 190–200. https://doi.org/10.1016/j.actpsy.2010.09.009

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. https://doi.org/10.1016/S0010-0277(99)00059-1

Arai, M., van Gompel, R. P. G., & Scheepers, C. (2007). Priming ditransitive structures in comprehension. *Cognitive Psychology*, *54*(3), 218–250. https://doi.org/10.1016/j.cogpsych.2006.07.001

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. https://doi.org/10.1371/journal.pone.0174623

Baayen, 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. https://doi.org/10.1016/j.jml.2016.11.006

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. https://doi.org/10.1080/23273798.2015.1065336

Baayen, R. H., van Rij, J., de Cat, C., & Wood, S. (2018). Autocorrelated Errors in Experimental Data in the Language Sciences: Some Solutions Offered by Generalized Additive Mixed Models. In *Mixed-Effects Regression Models in Linguistics* (pp. 49–69). Springer, Cham. https://doi.org/10.1007/978-3-319-69830-4_4

Balling, L. W., & Baayen, R. H. (2008). Morphological effects in auditory word recognition: Evidence from Danish. *Language and Cognitive Processes*, *23*(7–8), 1159–1190. https://doi.org/10.1080/01690960802201010

Barth, D., & Kapatsinski, V. (2018). Evaluating Logistic Mixed-Effects Models of Corpus-Linguistic Data in Light of Lexical Diffusion. In *Mixed-Effects Regression Models in Linguistics* (pp. 99–116). Springer, Cham. https://doi.org/10.1007/978-3-319-69830-4_6

Beckman, M. E. (1992). Evidence for speech rhythms across languages. In *Speech perception, production and linguistic structure* (pp. 457–463). Tokyo: Omsha and Amsterdam: IOS Press.

Beddor, P. S., McGowan, K. B., Boland, J. E., Coetzee, A. W., & Brasher, A. (2013). The time course of perception of coarticulation. *The Journal of the Acoustical Society of America*, *133*(4), 2350–2366. https://doi.org/10.1121/1.4794366

Benichov, J., Cox, L. C., Tun, P. A., & Wingfield, A. (2012). Word recognition within a linguistic context: Effects of age, hearing acuity, verbal ability and cognitive function. *Ear and Hearing*, *32*(2), 250–256. https://doi.org/10.1097/AUD.0b013e31822f680f

Bonte, M., Parviainen, T., Hytönen, K., & Salmelin, R. (2006). Time Course of Top-down and Bottom-up Influences on Syllable Processing in the Auditory Cortex. *Cerebral Cortex*, *16*(1), 115–123. https://doi.org/10.1093/cercor/bhi091

Broadbent, D. E. (1967). Word-frequency effect and response Bias. *Psychological Review*, *74*(1), 1–15.

Brouwer, S., & Bradlow, A. R. (2011). The influence of noise on phonological competition during spoken word recognition. In *Proceedings of the 17th International Congress of Phonetic Sciences. International Congress of Phonetic Sciences* (Vol. 2011, pp. 364–367).

Brouwer, S., Mitterer, H., & Huettig, F. (2012). Speech reductions change the dynamics of competition during spoken word recognition. *Language and Cognitive Processes*, *27*(4), 539–571. https://doi.org/10.1080/01690965.2011.555268

Brown-Schmidt, S., & Tanenhaus, M. K. (2008). Real-Time Investigation of Referential Domains in Unscripted Conversation: A Targeted Language Game Approach. *Cognitive Science*, *32*(4), 643–684. https://doi.org/10.1080/03640210802066816

Bushong, W., & Jaeger, T. F. (2017). Maintenance of perceptualinformation in speech perception. *Proceedings of the Annual Conference of the Cognitive Science Sociery*, *39*, 186–191.

Canseco-Gonzalez, E., Brehm, L., Brick, C. A., Brown-Schmidt, S., Fischer, K., & Wagner, K. (2010). Carpet or Cárcel: The effect of age of acquisition and language mode on bilingual lexical access. *Language and Cognitive Processes*, *25*(5), 669–705. https://doi.org/10.1080/01690960903474912

Connine, C. M., Blasko, D. G., & Titone, D. (1993). Do the Beginnings of Spoken Words Have a Special Status in Auditory Word Recognition? *Journal of Memory and Language*, *32*(2), 193–210. https://doi.org/10.1006/jmla.1993.1011

Content, A., Meunier, C., Kearns, R. K., & Frauenfelder, U. H. (2001). Sequence detection in pseudowords in French: Where is the syllable effect? *Language and Cognitive Processes*, *16*(5–6), 609–636. https://doi.org/10.1080/01690960143000083

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heeding the voice of experience: The role of talker variation in lexical access. *Cognition*, *106*(2), 633–664. https://doi.org/10.1016/j.cognition.2007.03.013

Cutler, A., & Otake, T. (1994). Mora or Phoneme? Further Evidence for Language-Specific Listening. *Journal of Memory and Language*, *33*(6), 824–844. https://doi.org/10.1006/jmla.1994.1039

Cutler, A., & Otake, T. (2002). Rhythmic categories in spoken-word recognition. *Journal of Memory and Language*, *46*(2), 296–322. https://doi.org/10.1006/jmla.2001.2814

Dahan, D. (2010). The time course of interpretation in speech comprehension. *Current Directions in Psychological Science*, *19*(2), 121–126. https://doi.org/10.1177/0963721410364726

Dahan, D., & Gaskell, M. G. (2007). The temporal dynamics of ambiguity resolution: Evidence from spoken-word recognition. *Journal of Memory and Language*, *57*(4), 483–501. https://doi.org/10.1016/j.jml.2007.01.001

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001a). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology*, *42*(4), 317–367. https://doi.org/10.1006/cogp.2001.0750

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001b). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes*, *16*(5–6), 507–534. https://doi.org/10.1080/01690960143000074

Dahan, D., & Tanenhaus, M. K. (2004). Continuous Mapping From Sound to Meaning in Spoken-Language Comprehension: Immediate Effects of Verb-Based Thematic Constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 498–513. https://doi.org/10.1037/0278-7393.30.2.498

Dahan, D., & Tanenhaus, M. K. (2005). Looking at the rope when looking for the snake: Conceptually mediated eye movements during spoken-word recognition. *Psychonomic Bulletin & Review*, *12*(3), 453–459. https://doi.org/10.3758/BF03193787

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & Van der Vrecken, O. (1996). The MBROLA project: towards a set of high quality speech synthesizers free of use for non commercial purposes. In *, Fourth International Conference on Spoken Language, 1996. ICSLP 96. Proceedings* (Vol. 3, pp. 1393–1396 vol.3). https://doi.org/10.1109/ICSLP.1996.607874

Eagleman, D. M. (2001). Visual illusions and neurobiology. *Nature Reviews Neuroscience*, *2*(12), 920–926. https://doi.org/10.1038/35104092

Farris-Trimble, A., & McMurray, B. (2013). Test–Retest Reliability of Eye Tracking in the Visual World Paradigm for the Study of Real-Time Spoken Word Recognition. *Journal of Speech, Language, and Hearing Research*, *56*(4), 1328–1345. https://doi.org/10.1044/1092-4388(2012/12-0145)

Farris-Trimble, A., McMurray, B., Cigrand, N., & Tomblin, J. B. (2014). The process of spoken word recognition in the face of signal degradation. *Journal of Experimental Psychology. Human Perception and Performance*, *40*(1), 308–327. https://doi.org/10.1037/a0034353

Frauenfelder, U. H., Scholten, M., & Content, A. (2001). Bottom-up inhibition in lexical selection: Phonological mismatch effects in spoken word recognition. *Language and Cognitive Processes*, *16*(5–6), 583–607. https://doi.org/10.1080/01690960143000146

Galle, M. (2014). Integration of multiple and asynchronous acoustic cues to word initial fricatives and context compensation in 7-year-olds, 12-year-olds and adults. *Theses and Dissertations*. Retrieved from https://ir.uiowa.edu/etd/1320

Gallistel, C. R., Fairhurst, S., & Balsam, P. (2004). The learning curve: Implications of a quantitative analysis. *Proceedings of the National Academy of Sciences*, *101*(36), 13124–13131. https://doi.org/10.1073/pnas.0404965101

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. https://doi.org/10.1080/016909697386646

Gaskell, M. G., & Marslen-Wilson, W. D. (2001). Lexical Ambiguity Resolution and Spoken Word Recognition: Bridging the Gap. *Journal of Memory and Language*, *44*(3), 325–349. https://doi.org/10.1006/jmla.2000.2741

Goldiamond, I., & Hawkins, W. F. (1958). Vexierversuch: The Log Relationship Between Word-Frequency and Recognition Obtained in the Absence of Stimulus Words. *Journal of Experimental Psychology*, *56*(6), 457.

Goldinger, S. D. (1996). Words and voices: episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *22*(5), 1166–1183.

Goldinger, S. D. (1998). Echoes of Echoes? An Episodic Theory of Lexical Access. *Psychological Review*, *105*(2), 251–279. https://doi.org/10.1037/0033-295X.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), 305–320. https://doi.org/10.1016/S0095-4470(03)00030-5

Gow, D. W. (2001). Assimilation and Anticipation in Continuous Spoken Word Recognition. *Journal of Memory and Language*, *45*(1), 133–159. https://doi.org/10.1006/jmla.2000.2764

Gow, D. W., & Olson, B. B. (2015). Lexical mediation of phonotactic frequency effects on spoken word recognition: A Granger causality analysis of MRI-constrained MEG/EEG data. *Journal of Memory and Language*, *82*, 41–55. https://doi.org/10.1016/j.jml.2015.03.004

Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. *Papers in Laboratory Phonology*, *7*, 515–546.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*, *28*(4), 267–283. https://doi.org/10.3758/BF03204386

Gwilliams, L., Linzen, T., Poeppel, D., & Marantz, A. (2018). In spoken word recognition the future predicts the past. *Journal of Neuroscience*, 0065–18. https://doi.org/10.1523/JNEUROSCI.0065-18.2018

Hayes, B. (1989). Compensatory Lengthening in Moraic Phonology. *Linguistic Inquiry*, *20*(2), 253–306.

Howes, D. (1957). On the Relation between the Intelligibility and Frequency of Occurrence of English Words. *The Journal of the Acoustical Society of America*, *29*(2), 296–305. https://doi.org/10.1121/1.1908862

Huettig, F., & Altmann, G. T. M. (2011). Looking at anything that is green when hearing “frog”: How object surface colour and stored object colour knowledge influence language-mediated overt attention. *The Quarterly Journal of Experimental Psychology*, *64*(1), 122–145. https://doi.org/10.1080/17470218.2010.481474

Huettig, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of Memory and Language*, *57*(4), 460–482. https://doi.org/10.1016/j.jml.2007.02.001

Huettig, F., & McQueen, J. M. (2009). AM radio noise changes the dynamics of spoken word recognition. In *15th Annual Conference on Architectures and Mechanisms for Language Processing*.

Huettig, F., Olivers, C. N. L., & Hartsuiker, R. J. (2011). Looking, language, and memory: Bridging research from the visual world and visual search paradigms. *Acta Psychologica*, *137*(2), 138–150. https://doi.org/10.1016/j.actpsy.2010.07.013

Jakobson, R., Fant, G., & Halle, M. (1951). *Preliminaries to Speech Analysis. The distinctive features and their correlates*. MIT Press, Cambridge MA.

Kapatsinski, V. (2005). Constituents can exhibit partial overlap: Experimental evidence for an exemplar approach to the mental lexicon. In *Proceedings from the Annual Meeting of the Chicago Linguistic Society* (Vol. 41(2), pp. 227–242).

Kapatsinski, V. (2012). What Statistics Do Learners Track? Rules, Constraints and Schemas in (Artificial) Grammar Learning. In *Frequency effects in language: Learning and processing* (Vol. 244.1, pp. 53–82). de Gruyter Mouton.

Kapatsinski, V. M. (2017). Heike Behrens and Stefan Pfänder: Experience Counts: Frequency Effects in Language. *Cognitive Linguistics*, *28*(2), 349–359. https://doi.org/10.1515/cog-2016-0097

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148–203. https://doi.org/10.1037/a0038695

Kruschke, J. K. (2008). Bayesian approaches to associative learning: From passive to active learning. *Learning & Behavior*, *36*(3), 210–226. https://doi.org/10.3758/LB.36.3.210

Labrune, L. (2012). Questioning the universality of the syllable: evidence from Japanese *. *Phonology*, *29*(1), 113–152. https://doi.org/10.1017/S095267571200005X

Labrune, L. (2014). The phonology of Japanese /r/: A panchronic account. *Journal of East Asian Linguistics*, *23*(1), 1–25.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: the neighborhood activation model. *Ear and Hearing*, *19*(1), 1–36.

Luce, R. D. (1959). *Individual choice behavior; a theoretical analysis.* New York: Wiley.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., … Den, Y. (2014). Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, *48*(2), 345–371. https://doi.org/10.1007/s10579-013-9261-0

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The Dynamics of Lexical Competition During Spoken Word Recognition. *Cognitive Science*, *31*(1), 133–156. https://doi.org/10.1080/03640210709336987

Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., & Dahan, D. (2003). The time course of spoken word learning and recognition: Studies with artificial lexicons. *Journal of Experimental Psychology: General*, *132*(2), 202–227. https://doi.org/10.1037/0096-3445.132.2.202

Malins, J. G., & Joanisse, M. F. (2012). Setting the tone: An ERP investigation of the influences of phonological similarity on spoken word recognition in Mandarin Chinese. *Neuropsychologia*, *50*(8), 2032–2043. https://doi.org/10.1016/j.neuropsychologia.2012.05.002

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1–2), 71–102. https://doi.org/10.1016/0010-0277(87)90005-9

Marslen-Wilson, W., & 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., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(3), 576–585. https://doi.org/10.1037/0096-1523.15.3.576

Matin, E., Shao, K. C., & Boff, K. R. (1993). Saccadic overhead: Information-processing time with and without saccades. *Perception & Psychophysics*, *53*(4), 372–380. https://doi.org/10.3758/BF03206780

Matsui M. (2017). 日本語における音韻要素の内部構造. *Theoretical and applied linguistics at Kobe Shoin*, (20), 89–126.

Mazurek, M. E., Roitman, J. D., Ditterich, J., & Shadlen, M. N. (2003). A Role for Neural Integrators in Perceptual Decision Making. *Cerebral Cortex*, *13*(11), 1257–1269. https://doi.org/10.1093/cercor/bhg097

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0

McClelland, J. L., Mirman, D., & Holt, L. L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Sciences*, *10*(8), 363–369. https://doi.org/10.1016/j.tics.2006.06.007

McMurray, B., Farris-Trimble, A., & Rigler, H. (2017). Waiting for lexical access: Cochlear implants or severely degraded input lead listeners to process speech less incrementally. *Cognition*, *169*(Supplement C), 147–164. https://doi.org/10.1016/j.cognition.2017.08.013

McMurray, B., Samelson, V. M., Lee, S. H., & Tomblin, J. B. (2010). Individual differences in online spoken word recognition: Implications for SLI. *Cognitive Psychology*, *60*(1), 1–39. https://doi.org/10.1016/j.cogpsych.2009.06.003

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*(2), B33–B42. https://doi.org/10.1016/S0010-0277(02)00157-9

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. https://doi.org/10.1016/j.jml.2008.07.002

McQueen, J. M., & Huettig, F. (2012). Changing only the probability that spoken words will be distorted changes how they are recognized. *The Journal of the Acoustical Society of America*, *131*(1), 509–517. https://doi.org/10.1121/1.3664087

McQueen, J. M., Otake, T., & Cutler, A. (2001). Rhythmic Cues and Possible-Word Constraints in Japanese Speech Segmentation. *Journal of Memory and Language*, *45*(1), 103–132. https://doi.org/10.1006/jmla.2000.2763

Meier, K. M., & Blair, M. R. (2013). Waiting and weighting: Information sampling is a balance between efficiency and error-reduction. *Cognition*, *126*(2), 319–325. https://doi.org/10.1016/j.cognition.2012.09.014

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., & Baayen, R. H. (2017). Discrimination in lexical decision. *PLOS ONE*, *12*(2), e0171935. https://doi.org/10.1371/journal.pone.0171935

Mirman, D. (2014). *Growth curve analysis and visualization using R*. Boca Raton, Florida: CRC Press.

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, *59*(4), 475–494. https://doi.org/10.1016/j.jml.2007.11.006

Mirman, D., Yee, E., Blumstein, S. E., & Magnuson, J. S. (2011). Theories of spoken word recognition deficits in Aphasia: Evidence from eye-tracking and computational modeling. *Brain and Language*, *117*(2), 53–68. https://doi.org/10.1016/j.bandl.2011.01.004

Morton, J. (1964). The effects of context on the visual duration threshold for words. *British Journal of Psychology*, *55*, 165–180.

Morton, J. (1969). Interaction of Information in Word Recognition. *Psychological Review*, *76*(2), 165–178.

Nittrouer, S., & Boothroyd, A. (1990). Context effects in phoneme and word recognition by young children and older adults. *The Journal of the Acoustical Society of America*, *87*(6), 2705–2715.

Norris, D. (1994). Shortlist: a connectionist model of continuous speech recognition. *Cognition*, *52*(3), 189–234. https://doi.org/10.1016/0010-0277(94)90043-4

Norris, D. (2006). The Bayesian Reader: Explaining Word Recognition as an Optimal Bayesian Decision Process. *Psychological Review*, *113*(2), 327–357. https://doi.org/10.1037/0033-295X.113.2.327

Norris, D., Cutler, A., McQueen, J. M., & Butterfield, S. (2006). Phonological and conceptual activation in speech comprehension. *Cognitive Psychology*, *53*(2), 146–193. https://doi.org/10.1016/j.cogpsych.2006.03.001

Norris, D., & McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395. https://doi.org/10.1037/0033-295X.115.2.357

Otake, T., Hatano, G., Cutler, A., & Mehler, J. (1993). Mora or Syllable? Speech Segmentation in Japanese. *Journal of Memory and Language*, *32*(2), 258–278. https://doi.org/10.1006/jmla.1993.1014

Otake, T., Sakamoto, Y., & Konomi, Y. (2004). Phoneme-based word activation in spoken-word recognition: evidence from Japanese school children (pp. 337–340). Presented at the INTERSPEECH-2004.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic Encoding of Voice Attributes and Recognition Memory for Spoken Words. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *19*(2), 309.

Plaut, D. C., & Booth, J. R. (2000). Individual and developmental differences in semantic priming: empirical and computational support for a single-mechanism account of lexical processing. *Psychological Review*, *107*(4), 786–823.

Port, R. F., Dalby, J., & O'Dell, M. (1987). Evidence for mora timing in Japanese. *The Journal of the Acoustical Society of America*, *81*(5), 1574–1585. https://doi.org/10.1121/1.394510

Radeau, M., Morais, J., & Segui, J. (1995). Phonological Priming Between Monosyllabic Spoken Words. *Journal of Experimental Psychology. Human Perception and Performance*, *21*(6), 1297–1311. https://doi.org/10.1037/0096-1523.21.6.1297

Raftery, A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology*, *25*, 111–163. https://doi.org/10.2307/271063

Ratcliff, R., & McKoon, G. (2008). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., & Lang, J. M. (1994). On the perceptual organization of speech. *Psychological Review*, *101*(1), 129–156.

Revill, K. P., Aslin, R. N., Tanenhaus, M. K., & Bavelier, D. (2008). Neural correlates of partial lexical activation. *Proceedings of the National Academy of Sciences*, *105*(35), 13111–13115. https://doi.org/10.1073/pnas.0807054105

Salasoo, A., & Pisoni, D. B. (1985). Interaction of knowledge sources in spoken word identification. *Journal of Memory and Language*, *24*(2), 210–231. https://doi.org/10.1016/0749-596X(85)90025-7

Salverda, A. P., Kleinschmidt, D., & Tanenhaus, M. K. (2014). Immediate effects of anticipatory coarticulation in spoken-word recognition. *Journal of Memory and Language*, *71*(1), 145–163. https://doi.org/10.1016/j.jml.2013.11.002

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(1), 1–17. https://doi.org/10.1037/0096-1523.3.1.1

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. https://doi.org/10.3758/s13428-015-0690-0

Simmons, E., & Magnuson, J. (2018). Word length, proportion of overlap, and phonological competition in spoken word recognition. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1064–1069). Madison, WI.

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. https://doi.org/10.3758/BF03192840

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science (New York, N.Y.)*, *268*(5217), 1632–1634.

Tanenhaus, M. K., Magnuson, J. S., Dahan, D., & Chambers, C. (2000). Eye Movements and Lexical Access in Spoken-Language Comprehension: Evaluating a Linking Hypothesis between Fixations and Linguistic Processing. *Journal of Psycholinguistic Research*, *29*(6), 557–580. https://doi.org/10.1023/A:1026464108329

Tobin, S. J., Cho, P. W., Jennett, P. M., & Magnuson, J. S. (2010). Effects of anticipatory coarticulation on lexical access. In *Proceedings of the Annual Conference of the Cognitive Science Society* (Vol. 32, pp. 2200–2205).

Toscano, J. C., Anderson, N. D., & McMurray, B. (2013). Reconsidering the role of temporal order in spoken word recognition. *Psychonomic Bulletin & Review*, *20*(5), 981–987. https://doi.org/10.3758/s13423-013-0417-0

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological Review*, *108*(3), 550–592.

Vance, T. (1987). *An introduction to Japanese phonology*. Albany, NY: SUNY Press.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/BF03194105

Wieling, M., Nerbonne, J., & Baayen, R. H. (2011). Quantitative Social Dialectology: Explaining Linguistic Variation Geographically and Socially. *PLOS ONE*, *6*(9), e23613. https://doi.org/10.1371/journal.pone.0023613

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*(1), 95–114. https://doi.org/10.1111/1467-9868.00374