S1. TRACE simulations of anadrome effects

As we discussed in the main paper, several models of spoken word recognition predict activation for anadromes as a result of their overlap in the vowel position, as well as partial phonological similarity in the consonants. All of these models predict this activation on the basis of some slot-like match of positions in the input to positions in candidate words. Thus, the critical question is whether such models predict more activation for anadromes than overlapping words, since this activation could not derive from position-specific matching. Of the commonly used models of spoken word recognition, TRACE (McClelland & Elman, 1986) was the first to predict activation for other mismatched-onset competitors like rhymes (Allopenna, Magnuson, & Tanenhaus, 1998), and more importantly its architecture is particularly complex with respect to how it handles time. As a result, it was not immediately obvious that it would not show additional activation for anadromes beyond that of overlaps.

At the same time, it also seemed possible that TRACE would not be sensitive to anadromes. There are a number of reasons this could be. First, onset competitors (cohorts) receive significant activation in TRACE and inhibit other competitors (e.g., rhymes), preventing them from attaining much activation. So, even if the small amount of overlap present in an anadrome could affect activation, competition from other types of words could suppress it. Second, TRACE's time alignment system was designed (in part) to explicitly discriminate words from others with the same set of phonemes in different orders. To handle time, TRACE uses multiple copies of the network that "watch" the input at different time alignments; these interact with each other to produce the final output. For example, one copy may monitor for the word *tack* aligned at the first time-step[1] (/t/ at time 1, /æ/ at time 2, and /k/ at time 3), while another monitors for the same word at the second time-step (/t/ at time 2, /æ/ at time 3, and /k/ at time 4). There are no versions, however, that monitor for the word *tack* by looking for /t/ at time 3 and /k/ at time 1. As a result, anadromes should be difficult to activate, and it may be impossible to activate anadromes more than overlaps because these time-aligned networks essentially impose a slot-like system on the lexicon.

To determine which prediction was correct, we ran a series of simulations examining activation for phonemic anadromes and other closely related words.

Methods

We tested TRACE's handling of anadromes using jTRACE, a Java implementation of the TRACE model (Strauss, Harris, & Magnuson, 2007). Simulations were run using a 1375 word lexicon that included all words with a frequency rating greater than 60 in the WU Speech & Hearing Lab neighborhood database (Sommers, n.d.), along with 112 additional words that were considered for use in the experiment but were not already in the list. In order to represent all of the words, an

[1] Time in TRACE is not counted phoneme-by-phoneme, but rather uses a more fine-grained time-scale in which individual phonemes last for multiple time-slices (and overlap). We've simplified our discussion of time here by using just the index of the relevant phoneme (e.g., /t/ is the first phoneme in *tack*, /æ/ is the second, etc.*)* as the unit of time, even though this corresponds to multiple steps at TRACE's finer time-scale). When presenting results we will adopt this simplification as well, but indicate the actual time-slice (for purposes of replication) as T_{TRACE}.

expanded phoneme set was used (based on Mayor & Plunkett 2009, but changed to reflect the American English phoneme inventory).

Table S1. Words used in TRACE Simulations

Target	Cohort	Anadrome	Overlap	Unrelated
tack	tap	cat	cap	mill
lip	lid	pill	pin	cow
puck	putt	cup	cut	mail
mug	mud	gum	gut	fish
leap	leaf	peel	peak	moon
mad	map	dam	dad	shoes
sub	sun	bus	bun	well

We tested seven item-sets, which were based on the stimuli developed for the behavioral experiment (Table S1). Although 16 sets were used in the experiment, nine of the sets contained at least one word with an affricate or diphthong, which are represented as two phonemes in TRACE (e.g., "tS" and "aI"). This made it challenging to interpret the time-alignment for these phonemes. Thus we examined only the seven CVC sets.

For each item-set, we measured activation for five types of words: the target, its anadrome, a cohort, an overlap word, and an unrelated object. We selected an overlap word that shared a vowel and one consonant (in the wrong position) with the target word. This was done to determine if activation for the anadrome was due either to vowel overlap or shared phonological features at onset.

Word activations from TRACE were transformed into predicted proportions of looks using the linking function described in Appendix C of McMurray, Samelson, Lee, and Tomblin (2010). This allows us to compare TRACE's output directly to the eye-tracking data we collected.

Results

As TRACE processes a string of phonemes, it uses multiple copies of the network corresponding to different time alignments with the input. Each word unit shows up once in each of these "time slices". Thus, if the model is given /tæk/ as an input, the word unit for *tack* will be most active at the time slice corresponding to the onset of the input. On the other hand, the unit for *at* will be most active for the alignment corresponding to the onset of the vowel.

As a result, it is complicated to determine the relevant activation level for a given word since there are actually multiple units corresponding to that word (at different time-alignments). One

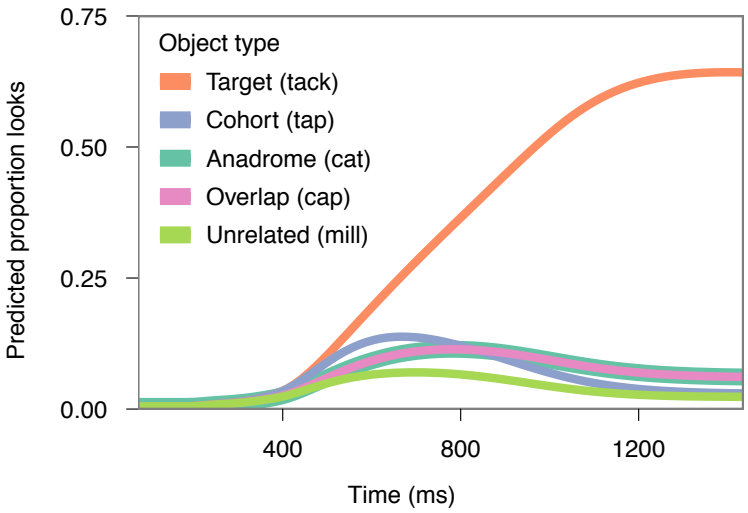

Figure S1. Predicted proportion of looks to the target, cohort, anadrome, overlap, and unrelated objects as a function of time when computed with the max-post-hoc rule. Activation is averaged over all seven item-sets (words in the legend are examples of one set). Note that the anadrome and overlap curves are nearly identical, but the line width of the anadrome curve is thicker for illustration.

way to handle this is to select the time slice that is most active for each word at the end of processing and use activations measured from that time slice throughout, the max post-hoc strategy. Figure S1 shows the average activation for each of the five words when computed in this way. The figure illustrates that the target and cohort are quite active early, and the cohort is gradually suppressed. As expected, there was little activation for the unrelated word. The anadrome showed some activation, which appears similar to our experimental findings. However, in contrast to our empirical results, the proportion of anadrome and overlap activation were almost identical (Table S2).

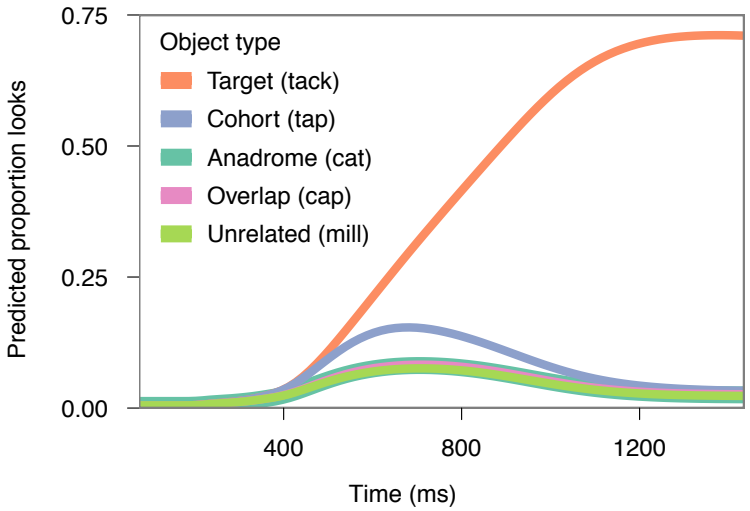

Figure S2. Predicted proportion of looks as a function of time when measured with words aligned at the time-slice corresponding to the first phoneme (T_{TRACE}= 2). (Anadrome curve drawn thicker for illustration.)

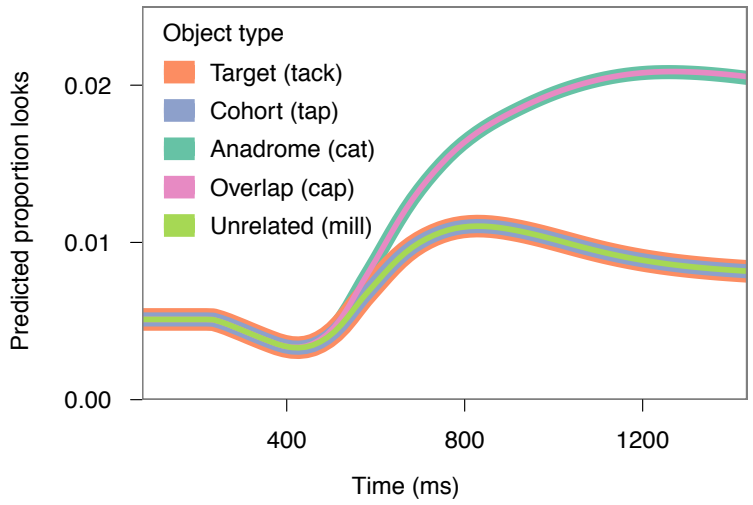

Figure S3. Predicted proportion of looks as a function of time for words aligned at the final phoneme (T_{TRACE}=7). Target, cohort and unrelated activation were nearly identical, and overlap and anadrome activation were nearly identical. (Anadrome, target, and cohort curves drawn thicker for illustration.)

A consequence of the max post-hoc strategy is that different words (e.g., targets and anadromes) can have a different maximum time alignment. The target and cohort both showed maximum activation when aligned with the first consonant in the input, indicating that the model considered the target (e.g., *tack*) and its cohort (*tap*) at the alignment beginning at the onset of the /t/. However, the anadrome and overlap were most active when aligned with the onset of the final consonant for every item-set. Activation at this later time slice suggests that when the model heard *tack*, *cat* was active as if the model ignored the /t/ and /æ/ and heard something like ##k.

Indeed, when we examined activation for the anadrome at the time-slice aligned with the initial consonant (T_{TRACE}=2), the activation patterns for both anadrome and overlap are quite low and closely follow that of the unrelated word

(Figure S2). [2] While the activation for these words is slightly higher than for the unrelated word, it is important to note that raw activation for both words was well below zero meaning they were inhibited and, if anything, the overlap word is slightly more active than the anadrome (Table S2) while empirical results showed the opposite pattern.

Table S2. Mean transformed activation in each simulation.

Object	Max post-hoc	T_{TRACE}=2	T_{TRACE}=7
Target	0.37148765	0.41125660	0.006896546
Cohort	0.05660876	0.06239214	0.006897673
Anadrome	0.06260065	0.03889392	0.012135984
Overlap	0.06266062	0.03976531	0.012161048
Unrelated	0.03507031	0.03635223	0.006896879

We also examined activation at the time slice aligned with the final phoneme (T_{TRACE}=7). Figure S3 shows that the timecourse of anadrome activation is the same as that of the overlap, and the overall mean proportion of activation for the two words is almost identical (Table S2). In addition, activation for these two words appears to be driven by the fact that they both begin with the final phoneme of the target word.

Table S3. Peak max-post-hoc activation for individual item-sets.

Item-set	Anadrome	Overlap
leap/leaf/peel/peak/moon	0.118182	0.118182
lip/lid/pill/pin/cow	0.111898	0.111898
mad/map/dam/dad/shoes	0.115054	0.115054
mug/mud/gum/gut/fish	0.113768	0.113768
puck/putt/cup/cut/mail	0.117590	0.117590
sub/sun/bus/bun/well	0.10436	0.10436
tack/tap/cat/cap/mill	0.117965	0.117965

It was also possible that the different item-sets (which show different degrees of featureal overlap among the consonants) would show different effects. To examine this, we looked at each item-set individually. The results are shown in Table S3. Anadromes and overlaps receive identical activation regardless of the degree of phonological similarity with the target at onset, in contrast to what we observed with listeners. That is, in TRACE, *cat* and *cap* compete with each other equally when *tack* is the target, as do *bus* and *bun* when *sub* is the target.

Thus, in no case does it appear that the anadrome and overlap are differentially activated. The fact that all phonemes are shared between the target and the anadrome does not give the anadrome an advantage in activation.

Discussion

Our TRACE simulations show more activation for anadromes and overlaps than for unrelated

[2] Because the simulations for T_{TRACE}=2 and T_{TRACE}=7 include only a single time-slice, the transformed activations will not necessarily map directly onto the observed proportions in the eye-tracking data. However, the same transformation is applied here for comparison with the max post-hoc simulation.

words, but the pattern of activation was the same for both types of competitors, in contrast to the results observed with human listeners. Moreover, anadrome and overlap activation was highest for lexical units aligned with the last phoneme of each input word. This suggests that the anadrome activation we observed in TRACE is due to the fact that, when the model heard *tack,* it also slightly considered that the final /k/ might be the onset of a new word beginning with /k/.

Interestingly, our simulations also showed activation for overlap words, something that has not been predicted or observed empirically but has also not been tested in TRACE. However, the most important finding from this series of simulations is that TRACE, due to its time alignment process, does not show activation for anadrome competitors beyond the activation shown for other words that share the target's final consonant in initial position.

S2. Statistical model comparisons

For all of the statistical models used in this study, we used linear mixed-effects models, implemented in the LME4 package (Bates & Sarkar, 2011) in R. Our dependent variable was the empirical-logit-transformed proportion of looks within a particular time window, and the fixed effect was contrast-coded object-type with the object-type that was expected to receive fewer fixations coded as -0.5, and the object-type expected to receive more as +0.5 (e.g., anadrome/unrelated, +0.5/-0.5). There are two random effects in our experimental design: the participant and the item-set that appeared on each trial. This permitted a range of potential statistical approaches in the mixed-effects framework, including models with either by-subject effects (corresponding to F_1 analyses in ANOVA), by-item effects (F_2 analyses), or both.

The design of our experiment, however, was not ideal for detecting effects in models that look at differences by-item. First, we were limited in the number of picturable CVC anadrome pairs in English. As a result, we have only 16 item-sets, which may not provide sufficient power to detect small competitor effects. In some ways, item-set does not really serve as a random effect at all: we have included almost all the existing items in English and do not have enough power for this analysis. In addition, there are a number of differences between the item-sets that we were not able to control (e.g., differences in word frequency, the amount of phonological feature overlap in the consonants, the quality of cohort competitors, visual salience of the pictures). As discussed in the main paper, some of these differences are likely to affect the degree of activation for anadromes and overlaps and the differences in activation between them. Thus, some item-sets may show effects, while others do not.

With this in mind, we examined various random effects structures to determine (1) which model was the most complex one justified by the data, and (2) whether we see effects for by-subject and/or by-item analyses. We constructed four nested models and compared model fit using chi-square goodness-of-fit tests. In particular, we compared models with by-subject random intercepts, by-subject and by-item intercepts, by-subject slopes and by-item intercepts, and both by-subject and by-item slopes. Crucially, as described in detail below, the results differed depending on whether by-item slopes were added or not. To presage the results, we found robust evidence that all of the

effects reported in the main paper generalize across participants, but in the analyses more sensitive to item-level variability there was mixed evidence for some of the effects. This suggests that it is safe to conclude that most participants will show the anadrome effects we report, but given the variability among item-sets, not all items will.

To summarize the set of models that were run, the *within-trial-type* analyses examined the relative proportion of looks to two different types of objects on the same trial-type (e.g., between the anadrome and unrelated objects on the Cohort/Anadrome trials). The *within-visual-stimulus* analyses examined looks to the same object (e.g., *bus*) when it was serving different roles *across trial-types* (e.g., it is an anadrome when *sub* is the target, but an overlap when *sun* is the target). Finally, the *maximally-different-consonant* analyses looked within-trial-type, but only included the four item-sets with a three-feature difference between consonants. Overall, we wanted to establish (1) whether there is evidence for more fixations to the anadromes than the unrelated words, (2) whether overlaps receive more fixations than unrelated words (to assess whether this could be a viable source of the anadrome effect), and (3) whether anadromes receive more fixations than overlaps (which is necessary to show that the phonemes in the wrong position are contributing to the activation).

Table S3 shows the results of the model comparisons for each analysis presented in the main paper. In each case, the model with both by-subject and by-item random slopes provided the best fit. Models with by-subject slopes and by-item intercepts did not provide a better fit than models with random intercepts for both factors, though for the initial anadrome-overlap comparison (including all item-sets), the improvement in fit was marginally better ($\chi^2(2)=5.3$, $p=0.071$). For all the other analyses, including by-subject random slopes did not change the pattern of results.

We next evaluated the fixed effect of object-type in the different models. While typical practice in mixed effects modeling is to only evaluate the maximal model justified by the data, there are cases in which this is not as informative as one might like. When running an F_1/F_2-style analysis, one can make independent judgments about whether fixed effects generalize across subjects and items. Here, however, in a model that contains both by-subject and by-item effects, it is not possible to precisely characterize which random effects can be generalized. In this case, since we knew that the design was under-powered by items and that there was substantial variability by item, we wanted to look at the significance of the fixed effects for models with and without by-item slopes. We used model comparisons to evaluate the effect of object-type in each of the models. Specifically, we compared models with no fixed-effect (but the same random effects structure) to models with object-type as the sole fixed effect.

Table S4 shows the results of the analyses. When we examined the effect of object-type in the models with only random intercepts for item (and either random-slopes or random-intercepts for participants), we saw a consistent pattern of results. Anadromes were fixated significantly more than unrelated objects in both the within-trial-type analysis, and in the same-visual-stimulus analysis; they were also fixated more than overlap objects in the same-visual-stimulus and maximally-different-consonant analyses.

However, in the models with by-item random slopes, we only saw significant effects of object-type in some cases. For the within-trial-type comparisons, none of the objects showed any differences from each other (i.e., looks to anadromes were not different from looks to overlaps, which were not different from looks to unrelated objects). However, for the same-visual-stimulus comparisons, we found significant anadrome-unrelated and anadrome-overlap differences, but not overlap-unrelated differences. For the maximally-different-consonant comparisons, we did not find any differences between the anadromes and overlaps for the within-trial-type model, and the same-visual-stimulus model did not successfully converge.

Table S3. Results of model comparisons for each of the analyses presented in the main paper. Each column represents the improvement in fit for increasingly complex models (from left to right), starting with a model with only by-subject intercepts.

	Analysis	Subject intercepts vs. Subject + Item intercepts	Subject slopes + Item intercepts vs. Previous	Subject + Item slopes vs. Previous
Within Trial-Type	Anadrome vs. Unrelated (cohort/anadrome trials)	$\chi^2(1)=151.5$, $p<0.001$	$\chi^2(2)=0.005$, $p=0.997$	$\chi^2(2)=68.1$, $p<0.001$
	Overlap vs. Unrelated (cohort/overlap trials)	$\chi^2(1)=97.3$, $p<0.001$	$\chi^2(2)=0.280$, $p=0.869$	$\chi^2(2)=58.2$, $p<0.001$
	Anadrome vs. Overlap (anadrome/overlap trials)	$\chi^2(1)=217.1$, $p<0.001$	$\chi^2(2)=5.3$, $p=0.071$	$\chi^2(2)=60.4$, $p<0.001$
Same Visual Stimulus	Anadrome vs. Unrelated (same-visual-stimulus)	$\chi^2(1)=130.5$, $p<0.001$	$\chi^2(2)=2.3$, $p=0.310$	$\chi^2(2)=103$, $p<0.001$
	Overlap vs. Unrelated (same-visual-stimulus)	$\chi^2(1)=77.3$, $p<0.001$	$\chi^2(2)=0.13$, $p=0.938$	$\chi^2(2)=65.3$, $p<0.001$
	Anadrome vs. Overlap (same-visual-stimulus)	$\chi^2(1)=196.2$, $p<0.001$	$\chi^2(2)=1.4$, $p=0.499$	$\chi^2(2)=36.3$, $p<0.001$
Max. Diff. Cons.	Anadrome vs. Overlap (within-trial-type; max-diff-cons)	$\chi^2(1)=74.1$, $p<0.001$	$\chi^2(2)=0.032$, $p=0.984$	$\chi^2(2)=11.3$, $p=0.004$
	Anadrome vs. Overlap (same-visual-stimulus; max-diff-cons)	$\chi^2(1)=12.6$, $p<0.001$	$\chi^2(2)=1.2$, $p=0.538$	Model did not converge successfully

Thus, the results of these analyses are mixed: when object-type is examined within individual trial-types, we do not see any differences between the competitors; when differences are examined across trial-types (controlling for visual stimulus differences), we do see effects. In the traditional framing of subject- and item-analyses, the lack of robust effects in the more complex model by-item model suggests that the effect of object-type may not robustly generalize across items. The next goal is to figure out why – which items show the effect, and which do not, and what factors layered within the item-set are driving the effects. For our hypotheses, the most critical anadromes are those that have minimal feature overlap in the bracketing consonants, so it is crucial to determine whether they show the appropriate effects of object-type.

First, we investigated the visual properties of the objects. Some pictures may have been more salient, easier to recognize, or less likely to be mapped onto two names. Indeed, when this is controlled (by examining fixations to the same object across positions in the *same-visual-stimulus* analyses), we see that the effects of object-type emerge and are robust in the random slopes models: anadromes are fixated significantly more than both unrelated and overlap objects.

The second factor we looked at was phonological overlap among the consonants. Here, item-sets like *cat/tack/tap* have consonants (/t/ and /k/) that share two features (place and voicing) and as a result are fairly confusable. This should increase looks to both anadrome and overlap objects. Confusability in the final consonant (across overlaps and targets) could also play a role, further creating variability between item-sets. In contrast, for *bus/sub/sun*, there is less phonological overlap in the initial consonant. This variation across item-sets could be contributing to the lack of significant effects when by-item slopes are added. Fortunately, we can narrow the range of possible analyses by focusing on the subset of item-sets that have no phonological feature overlap with the target at onset – these are the most critical for evaluating our hypothesis. These analyses showed more looks to anadromes than to overlaps in the model with random intercepts, but not in the model with random slopes (though there are only four item-sets in this analysis, so the model is far too underpowered to detect such an effect by-item, and since we are not asking about four specific items, in some ways an item analysis is not necessary). Thus, one reason the more complex, by-item random slope model did not show an effect of object-type is that some item-sets showed a difference, while others did not. Critically, the most theoretically important ones show the effect.

Although the results of the analyses using by-item random slopes are mixed, as noted above, once we controlled for the known sources of variability between item-sets, we saw the predicted effects. Given that the pool of possible anadromes was restricted (both in the number of item-sets and in their phonetic and semantic forms), we did not expect to have sufficient power to detect small effects in an item analysis.

To summarize, we found strong anadrome effects for by-subject analyses (models with both by-subject intercepts and slopes show effects) but mixed results for by-item analyses (models with by-item intercepts show effects, as do models with by-item slopes for same-visual-stimulus comparisons; models with by-item slopes for within-trial-type comparisons did not show any effects). Since the by-item analyses may not have enough power given our design and the known variability among our items, both models with random intercepts and models with random slopes

are reported in the main paper. The results here suggest that, while generalizations across items may be less robust than across subjects in this dataset, we still find some evidence for these effects in models with by-item random slopes.

Table S4. Results for effect of object-type for each of the models examined. P-values calculated via chi-square goodness-of-fit between models with and without a fixed effect of object-type. Significant effects are highlighted in **bold**.

	Analysis	By-subject intercept By-item intercept	By-subject slope By-item intercept	By-subject slope By-item slope
Within Trial-Type	Anadrome vs. Unrelated (cohort/anadrome trials)	**b=0.025, SE=0.006, p<0.001**	**b=0.025, SE=0.006, p<0.001**	b=0.025, SE=0.017, p=0.131
	Overlap vs. Unrelated (cohort/overlap trials)	b=0.001, SE=0.006, p=0.847	b=0.001, SE=0.006, p=0.847	b=0.001, SE=0.014, p=1
	Anadrome vs. Overlap (anadrome/overlap trials)	b=0.011, SE=0.006, p=0.092	b=0.011, SE=0.006, p=0.124	b=0.011, SE=0.016, p=0.523
Same Visual Stimulus	Anadrome vs. Unrelated (same visual stimulus)	**b=0.044, SE=0.006, p<0.001**	**b=0.044, SE=0.006, p<0.001**	**b=0.044, SE=0.018, p=0.024**
	Overlap vs. Unrelated (same visual stimulus)	**b=0.016, SE=0.005, p=0.003**	**b=0.016, SE=0.005, p=0.004**	b=0.016, SE=0.014, p=0.271
	Anadrome vs. Overlap (same visual stimulus)	**b=0.028, SE=0.006, p<0.001**	**b=0.028, SE=0.006, p<0.001**	**b=0.028, SE=0.013, p=0.045**
Max Diff. Cons.	Anadrome vs. Overlap (within-trial-type; max-diff-cons)	**b=0.055, SE=0.015, p<0.001**	**b=0.055, SE=0.015, p<0.001**	b=0.055, SE=0.040, p=0.183
	Anadrome vs. Overlap (same-visual -stimulus; max-diff-cons)	**b=0.046, SE=0.012, p<0.001**	**b=0.046, SE=0.013, p<0.001**	Model did not converge successfully

References

Allopenna, P., 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*, 419-439.

Bates, D., & Sarkar, D. (2011). lme4: Linear mixed-effects models using S4 classes.

Mayor, J. & Plunkett, K. (2009). Using TRACE to Model Infant Sensitivity to Vowel and Consonant Mispronunciations. In: N. Taatgen et al. (Eds.) *Proceedings of the 31st Annual Cognitive Science Society* (pp 1816-1821). Austin, TX: Cognitive Science Society.

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

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition, 86*(2), B33-B42.

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*, 1-39.

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*, 19-30.

Sommers, M. (nd). WU Speech & Hearing Lab neighborhood database. [Online] Available from: http://neighborhoodsearch.wustl.edu/Neighborhood/Home.asp [accessed 26 August 2011]