A comprehensive model of spoken word recognition must be multimodal: Evidence from studies of language mediated visual attention

Alastair C. Smith (alastair.smith@mpi.nl)
Max Planck Institute for Psycholinguistics, Wundtlaan 1,
Nijmegen, 6525 XD, The Netherlands

Padraic Monaghan (p.monaghan@lancaster.ac.uk)
Department of Psychology, Lancaster University,
Lancaster, LA1 4YF, U.K.

Falk Huettig (falk.huettig@mpi.nl)
Max Planck Institute for Psycholinguistics, Wundtlaan 1,
Nijmegen, 6525 XD, The Netherlands

Abstract

When processing language, the cognitive system has access to information from a range of modalities (e.g. auditory, visual) to support language processing. Language mediated visual attention studies have shown sensitivity of the listener to phonological, visual, and semantic similarity when processing a word. In a computational model of language mediated visual attention, that models spoken word processing as the parallel integration of information from phonological, semantic and visual processing streams, we simulate such effects of competition within modalities. Our simulations raised untested predictions about stronger and earlier effects of visual and semantic similarity compared to phonological similarity around the rhyme of the word. Two visual world studies confirmed these predictions. The model and behavioral studies suggest that, during spoken word comprehension, multimodal information can be recruited rapidly to constrain lexical selection to the extent that phonological rhyme information may exert little influence on this process.

Keywords: The Visual World Paradigm, Connectionist Modeling, Visual Attention, Spoken Word Comprehension.

Introduction

Words are complex entities that link representations across multiple modalities (e.g. phonological, orthographic, semantic, visual …). When processing spoken words in natural, 'real world' settings the cognitive system often has access to information from many modalities each of which may provide a reliable source of information that can be used to constrain lexical selection. Yet comparatively little is known about how these various sources of information interact and the temporal structure of this multimodal process.

Language mediated visual attention requires the ability to map between information activated by the visual environment and information activated by spoken language. Empirical studies of this process conducted using the visual world paradigm, in which participants view a visual display and simultaneously hear a spoken utterance while their eye gaze is recorded, have helped isolate the types of information involved in this mapping and the temporal structure of this process.

Huettig and McQueen (2007), presented participants with scenes that contained objects that overlapped with the spoken target word (e.g. beaker) in either a visual dimension (e.g. bobbin), a semantic dimension (e.g. fork) or shared their phonological onset (e.g. beaver). They observed that participants looked first towards items that shared their phonological onset, then once the speech signal disambiguated between phonological representations of the target and onset competitor, participants looked more towards items that overlapped in either a visual or semantic dimension.

Such eye gaze data has been used to index the timing of activation of specific forms of information by the speech signal and further has been offered as a measure of the relative influence of information types during spoken word comprehension. For example, the influence of phonological rhyme overlap on eye gaze in visual world studies is often only marginally significant, and less robust than onset overlap effects (Allopenna et al., 1998;McQueen & Huettig, 2012). This exposes features of the underlying mechanism that have been used to constrain models of speech processing. For example empirical evidence of rhyme effects has proved influential as they are predicted by continuous mapping models of spoken word recognition (e.g. TRACE: McClelland & Elman, 1986).

Evidence provided by these studies demonstrates that as a spoken word unfolds information associated with the word in phonological, visual and semantic dimensions is rapidly activated. Further this information can be utilized by the cognitive system to connect the unfolding speech signal to information that has been activated through other modalities. In the case of language mediated visual attention to map to information activated by the visual environment.

Although it has been established that phonological rhyme overlap can influence gaze behavior this has only been demonstrated under heavily constrained conditions in which a phonological mapping provides the only connection to spoken target words. Therefore its influence under more

natural conditions, when overlapping information in other modalities (e.g. visual or semantic) is available, is unknown. Within this study we examine the influence of phonological rhyme on language mediated visual attention under conditions in which competition is provided by items that share semantic and visual relationships with spoken target words. We first simulate the effects of this multimodal competition in a model of language mediated visual attention in which the processing of spoken words involves the parallel integration of information concurrently available in phonological, visual and semantic domains and then test whether such a model successfully predicts the behavioral consequences of the multimodal competition induced within two visual world studies. We discuss the results of these experiments both in respect to the mechanisms driving language mediated eye gaze and more broadly the possible implications for models of spoken word recognition.

Simulating multimodal competition in language mediated visual attention

Model

Simulations were conducted using the amodal shared resource model of language mediated visual attention which has previously been shown to replicate a broad range of word level effects reported in the visual world paradigm (see Smith, Monaghan & Huettig, 2013a; Smith, Monaghan & Huettig, 2013b). A diagram of the model's architecture is displayed in Figure 1.

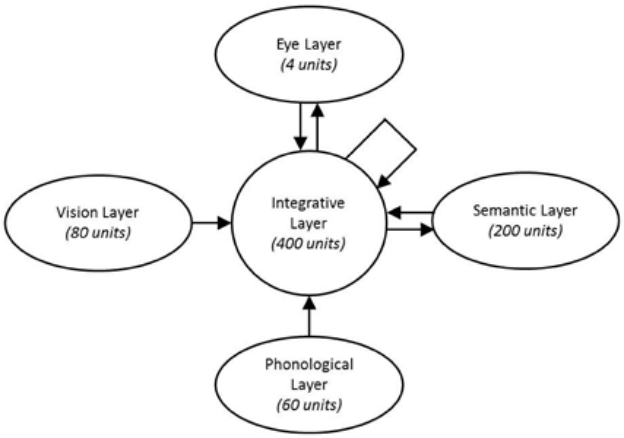

Figure 1: Network Architecture.

Architecture The network consists of four modality-specific layers connected via a central resource. Input is provided by visual, semantic and phonological layers. The visual layer simulates the extraction of visual information from four locations in the visual field, the phonological layer simulates the extraction of phonological information from the speech signal over time, and the semantic layer allows the network to associate semantic features with a given object or spoken word. Output behavior is represented by the eye layer which provides a measure of the models direction of attentional focus as a consequence of the integrated multimodal input. Each unit in the eye layer was associated with one of the four locations in the visual field. The models 'gaze' was interpreted as directed towards the location associated with the most activated eye layer unit.

Representations To train and test the network an artificial corpus was constructed consisting of 200 items, with each item assigned unique visual, phonological and semantic representations. Within the corpora were embedded 20 target items each of which was paired with a different semantic, visual and phonological rhyme competitor. The distance between targets and competitor representations was on average half the distance of that between a target and unrelated item in the modality that defined the competitor class (see Table 1).

Table 1: Relationships defining target, competitor (Com.) and distractor (Un.) items

Mod.	Item	Constraint (Features shared with target)	Cosine Distance
Pho.	Com.	Final 3 of 6 phonemes	.259
	Un.	Max. 2 consecutive phonemes	.496
Sem.	Com.	4 of 8 semantic features	.500
	Un.	Max. 1 semantic feature	.959
Vis.	Com.	Min. 5 of 10 visual features	.264
	Un.	Features shared with p = (0.5)	.506

Training The model was trained on four cross modal mapping tasks, these were: to activate an item's semantic representation when fixating on its visual representation; to activate an item's semantic representation when presented with its phonological representation; to activate the eye unit associated with the location of an item indicated by the presence of its phonological representation (phonological orienting) and to activate the eye layer unit associated with the location an item indicated by the presence of its semantic representation (semantic orienting). All four tasks were randomly interleaved although the task of mapping from speech to location was four times less likely to occur than all other tasks. Initial connection weights were randomized and adjusted over the course of 850 000 training trials using recurrent back propagation (learning rate 0.05). Once trained the model performed all training tasks accurately for over 99% of items in the training corpus.

Simulations

Simulation 1 The visual input presented to the model at trial onset consisted of the visual representations of a rhyme competitor and three unrelated distractors. 5 time steps followed display onset to enable pre-processing of visual information before the first phoneme of the target word was presented to the phonological layer. An additional phoneme was provided at each subsequent time step until the entire phonological representation of the target word has unfolded. The most active unit in the eye layer was recorded at each time step. In total 480 trials were simulated, with all 20

target – competitor pairs tested with distractors in all possible locations in the visual field (n = 24).

Figure 2 presents the change in the proportion of fixations directed towards each category of item (rhyme competitor or average distractor) from word onset.

To examine whether the model fixated rhyme competitors more than distractors we first calculated the proportion of time steps in the display preview period (ts 0 - 5) in which the rhyme competitor was fixated. We then calculated the same measure for the unrelated distractors. As there were three distractors and only one competitor, distractor fixation proportions were divided by 3. We then compared using paired t-tests the mean log-transformed ratio of the proportion of fixations toward rhyme competitors / proportion of fixations toward unrelated distractors in the preview period to the same ratio calculated for each time step post word onset (ts 6-29). Results of the analysis showed that the model first displayed a bias toward fixating rhyme competitors over distractors from time step 13 [M = -0.247, SD = 0.207, $t_{simulation}(7) = -3.374$, p = 0.012; M = -0.258, SD = 0.352, $t_{item}(19) = -3.267$, p = 0.004]. This bias continued for all subsequent time steps.

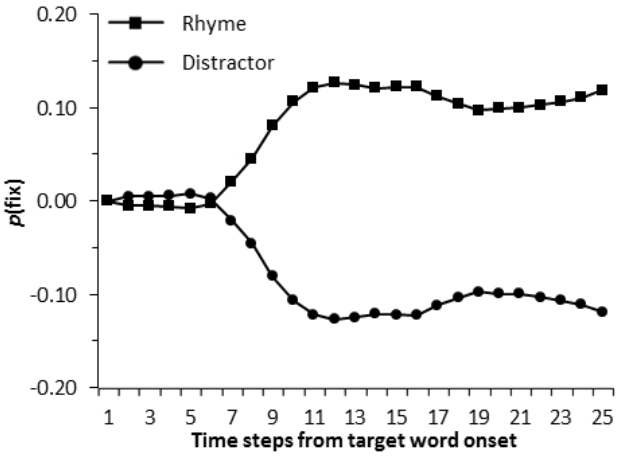

Figure 2: Change in fixation proportions from trial onset for simulations of Experiment 1.

Simulation 2 An identical procedure was followed as described for simulations 1 yet scenes now contained a rhyme competitor, a visual competitor, a semantic competitor and an unrelated distractor. Figure 3 displays the change in the proportion of fixations to each category of item from word onset.

For each competitor type separate analysis was conducted following the same procedure used in simulation 1. Comparing visual competitor – distractor ratios showed that the model first fixated visual competitors more than distractors in time step 12 [M = -0.262, SD = 0.276, $t_{simulation}(7) = -2.682$, p = 0.031; M = -0.275, SD = 0.281, $t_{item}(19) = -4.375$, p < 0.001], the model continued to fixate visual competitors more than distractors for all subsequent time steps.

Comparing semantic competitor – distractor ratios showed increased fixation of semantic competitors from time step 13 [M = -0.324, SD = 0.299, $t_{simulation}(7) = -3.068$, p = 0.018 ; M = -0.291, SD = 0.241, $t_{item}(19) = -5.393$, p < 0.001], semantic competitors remained fixated above distractor levels for all remaining time steps.

Finally, comparing rhyme competitor – distractor ratios showed that rhyme competitors were first fixated more than distractors in time step 17 [M = -0.376, SD = 0.366, $t_{simulation}(7) = -2.904$, p = 0.029; M = -0.349, SD = 0.411, $t_{item}(19) = -3.796$, p = 0.001], rhyme competitors continued to be fixated more than distractors in all subsequent time steps.

We also calculated for each competitor the competitor – distractor ratio over the entire post word onset period (ts 6 – 29) and compared this between competitor types using paired t-tests. We observed a greater visual competitor effect than either semantic or rhyme effects [visual/semantic: M = 0.121 SD = 0.164, $t_{simulation}(7) = -2.098$, p = 0.074; M = 0.151, SD = 0.204, $t_{item}(19) = 3.305$, p = 0.004; visual/rhyme: M = -0.500, SD = 0.231, $t_{simulation}(7) = 6.134$, p < 0.001; M = 0.536, SD = 0.305, $t_{item}(19) = 7.860$, p < 0.001], while the semantic competitor effect was greater than the rhyme effect [semantic/rhyme: M = 0.379, SD = 0.272, $t_{simulation}(7) = 3.938$, p = 0.006; M = 0.386, SD = 0.253, $t_{item}(19) = 6.822$, p < 0.001].

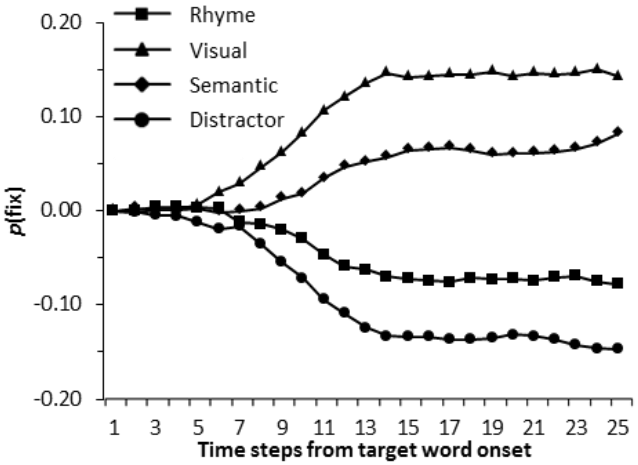

Figure 3: Change in fixation proportions from trial onset for simulations of Experiment 2.

Discussion

Simulations predict that the effect of visual and semantic overlap will be greater than that of rhyme overlap and that visual effects will precede semantic effects, with both preceding rhyme effects. Finally, the model predicts that added competition from semantic and visual competitors will lead to a delay in the emergence of rhyme effects.

Experiment 1: Effects of rhyme overlap in target absent scenes

An initial visual world experiment was conducted to establish that the materials constructed captured the effect of phonological rhyme overlap on language mediated visual attention.

Participants 40 participants aged between 18 and 30 (mean = 21.6) participated in this experiment. All were native speakers of Dutch.

Materials 15 experimental trials were constructed each consisting of a visual display and spoken Dutch sentence. Each sentence consisted of a target word embedded in a neutral carrier sentence in which the target word was not predictable (e.g. Dutch: "Zij begrepen niet waarom de roos verwelkt was." English translation: "They could not understand why the rose was withered.").

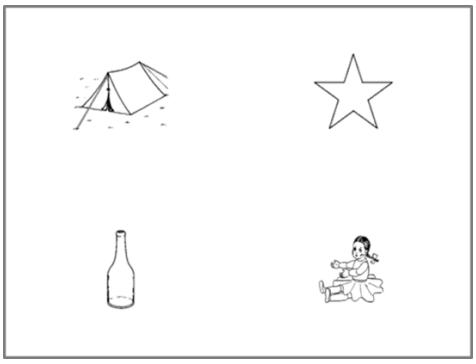

Figure 4: Example of a visual display from an experimental trial in Experiment. In this trial the target word was *cent*, the rhyme competitor *tent* and the three unrelated distractors *pop* (doll), *ster* (star) and *fles* (bottle).

Experimental displays contained a phonological rhyme competitor and three unrelated distractors. Phonological rhyme competitors differed only in their initial phoneme from the target words phonological representation. The following relationships were controlled between the target word, and competitors and distractors: word frequency, number of syllables, number of letters, number of shared phonemes, visual similarity, semantic similarity. Separate semantic similarity and visual similarity norming studies were conducted to gain measures of the similarity between each of the images presented in the visual display and the paired target word. All images used within the study were black and white line drawings. To ensure that the names attributed to displayed images were well motivated a picture name correspondence pre-test was conducted.

In addition to the 15 experimental trials, 15 target absent and 15 target present filler trials were constructed. These also consisted of a spoken target word embedded in a neutral sentence paired with a display containing four black and white line drawings. In target present displays an image whose name corresponded to the spoken target word was presented along with the images of three unrelated distractor objects. In target absent displays images of four unrelated distractors were presented.

Procedure A tower mounted eye tracker was used to record participant's eye movements as they viewed displays on a computer monitor and listened to sentences through headphones. Participants were asked to look at the display while listening to the spoken sentences. Within the experiment the 15 experimental trials, 15 target absent trials and 15 target present trials were randomly interleaved. Each trial proceeded as follows, replicating the procedure in Huettig & McQueen, 2007). First a fixation cross appeared in the center of the screen for 500 ms, this was followed by a blank screen for 600 ms. Then a scene consisting of four images was presented, at which point a spoken sentence began to unfold. The location of each item was randomized across items and participants.

Results Figure 5 presents a time course graph recording the difference from target word onset, in the proportion of fixations directed towards phonological rhyme competitors and unrelated distractors for the first 1600ms post target word onset. Average distractor fixation proportions were calculated by dividing the total proportion of fixations towards unrelated distractors by three.

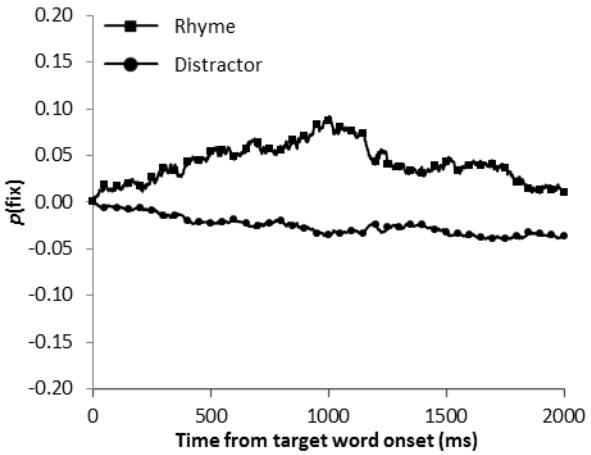

Figure 5: Change in the proportion of fixations directed towards rhyme competitors and average unrelated distractor from word onset displayed by participants in Experiment 1.

For analysis we divided the first 1600 ms period post word onset into four 400 ms bins and compared behavior in each of these periods to behavior in the 400 ms prior to target word onset. For each bin, in each test trial we calculated the empirical log odds of fixating each category of item (competitor or distractor). To form our dependent variable we calculated the difference between the log-odds of fixating the rhyme competitor and the log-odds of fixating a distractor, this variable reflects the difference in fixation behavior as a consequence of phonological rhyme

overlap. This measure was then analyzed using linear mixed effects models to examine whether for each 400ms window post target word onset the difference between fixation of distractors and rhyme competitors differed from the difference observed in the baseline window prior to target word onset. The model used a fixed effect of window and random effects of subject and item including random intercepts and slopes for time window both by subject and item. To derive p-values we assume t-values were drawn from a normal distribution.

We observed a significant effect of rhyme overlap 801–1200 ms post word onset [$\beta = 0.68$; $t = 2.22$; $p = 0.03$] with participants fixating rhyme competitors more than distractors in this window compared to the baseline period. There was also a marginal increase in fixation of phonological rhyme competitors in the window 1201–1600 ms post word onset in comparison to the baseline period [$\beta = 0.05$; $t = 1.85$; $p = 0.06$].

Experiment 1 establishes that the level of phonological rhyme overlap embedded within the materials is sufficient to induce increased fixation of rhyme competitors over distractors replicating previous rhyme effects reported in the literature.

Experiment 2: Effects of Rhyme overlap in scenes containing Semantic and Visual competitors

A second experiment examined the influence of competition from items that overlapped in either a visual or semantic dimension on the previously observed rhyme effect.

Participants 39 participants aged between 18 and 30 (mean = 25.3) participated in experiment 2. All participants were native speakers of Dutch.

Materials Experiment 2 used the same materials as used in experiment 1 other than in each experimental display one unrelated distractor was replaced by a visual competitor, while a second was replaced by a semantic competitor.

To ensure that visual and semantic competitors shared greater visual and semantic similarity respectively with their paired target word, visual and semantic similarity norming studied were conducted on all experimental scenes. Additional controls ensured that competitor sets only differed from distractors in their relationship to the target word in the dimension intended.

Procedure The procedure applied in experiment 1 was replicated in experiment 2.

Results Figure 6 displays the difference in the proportion of fixations from display onset directed toward each category of item displayed.

We applied the same method of analysis to the result of experiment 2 as used in experiment 1. The difference between the log-odds of fixating a visual competitor and those of fixating an unrelated distractor did not differ between the baseline window and the 400ms directly following word onset. Fixation of visual competitors in relation to distractors did differ from baseline levels in windows 401 – 800ms [$\beta = 0.67$; $t = 2.24$; $p = 0.03$], 801 – 1200ms [$\beta = 0.80$; $t = 2.68$; $p = 0.01$] and 1201 – 1600ms [$\beta = 0.58$; $t = 2.01$; $p = 0.05$]. Throughout this period there was an effect of visual similarity with visual competitors fixated more than distractors in comparison to behavior in the baseline window. Differences between semantic competitor fixations and unrelated distractor fixations did not differ from baseline levels in first 800 ms post word onset. There was however a marginal difference in the 801-1200ms window [$\beta = 0.45$; $t = 1.79$; $p = 0.07$], with semantic overlap leading to increased fixation in this period compared to the baseline window. This increase over baseline levels was significant in the final 400ms window [1201 – 1600 ms: $\beta = 0.66$; $t = 2.51$; $p = 0.01$]. There was however no evidence for an influence of phonological rhyme overlap at any point post word onset.

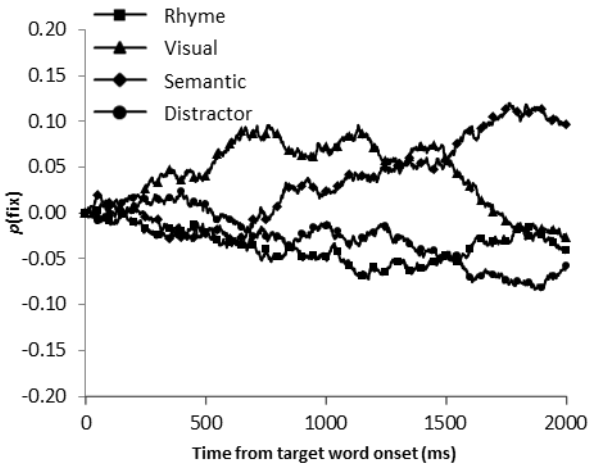

Figure 6: Change in the proportion of fixations directed towards rhyme competitors, semantic competitors, visual competitors and unrelated distractors from word onset displayed by participants in Experiment 2.

When presented with scenes containing a visual competitor, a semantic competitor, a rhyme competitor, and an unrelated distractor participants displayed a bias towards fixating visual competitors from 401 – 800ms post word onset. Participants also displayed at later time points a bias toward fixating semantic competitors from 801 – 1200 ms. However, in contrast to the results of experiment 1, no evidence was found for an influence of phonological rhyme overlap on fixation behavior.

General Discussion

Our two visual world studies demonstrate that visual and semantic relationships exert a greater influence on language mediated visual attention than phonological rhyme. Further they show that the presence of visual and semantic overlap can eliminate the influence of phonological rhyme on such

behavior. For such effects to be observed visual and semantic information relating to the unfolding spoken target word must have been activated and available to map onto pre-activated information, before phonological rhyme overlap could begin to exert an influence on eye gaze, which from experiment 1 we know that it can. Our results therefore provide further evidence that visual and semantic information can be activated rapidly when processing spoken words and that this information can be used to constrain selection of items in the visual environment.

Our simulations predicted a greater influence of visual and semantic overlap on visual attention in comparison to phonological rhyme overlap. In the model the rhyme competitor effect is delayed when visual and semantic competitors are also present in the scene. This is driven by the model's gaze being directed towards only the most salient item. In scenes in which rhyme provides the only connection to the target, rhyme competitors are likely to be the most salient item and therefore attract attention. However, overlapping visual and semantic features of the corresponding competitors are more salient than features of the overlapping phonological rhyme. Therefore, when such items are present in the scene rhyme competitors are less likely to be the most salient item and therefore less likely to be fixated.

What drives the visual and semantic competitor advantage within the model is the contrast in the temporal structure of representations across modalities. Unlike visual and semantic representations, phonological representations possess a temporal component. In the model initial phonemes present a good predictor of the intended target and are therefore more influential than phonemes in the rhyme, that become available after the system has sufficient information to identify the target item. The current study indicates that such a mechanism leads to a greater saliency of overlapping visual or semantic information than phonological rhyme, to the extent that phonological rhyme information did not have an observable behavioral influence on fixation behavior when in competition with visually or semantically overlapping items.

We know from earlier modelling (see Smith, Monaghan and Huettig, 2013a) that the addition of noise to the phonological input in the training stage modulates the level to which such initial phonemes predict the target item, and therefore the weight the model places on the influence of none overlapping phonemes on fixation of rhyme competitors. Previous empirical work has demonstrated that the amount of noise in the speech signal can modulate the influence of phonological rhyme (McQueen & Huettig., 2012). It would therefore be interesting to examine whether under conditions in which information across multiple modalities can be recruited to map between items, noise is able to dynamically adapt the influence of information types, does visual noise lead to an increase in visual overlap effects, and does auditory noise lead to an emergence of phonological rhyme effects.

Our results provide further evidence for a rapid recruitment of multimodal information to constrain lexical processing. Based on our findings we propose that spoken word comprehension in natural settings, in which semantic and visual information is likely to be pre-activated, semantic and visual information is recruited rapidly by the cognitive system to the extent that later phonological information often exerts little or no influence on this process.

Although models of spoken word recognition that focus purely on phonological information processing (e.g. TRACE: McClelland & Elman, 1986; Shortlist B: Norris & McQueen, 2008) provide an important contribution to our understanding of the manner in which information carried in the acoustic signal is used to constrain lexical processing, our results indicate that as models of day to day spoken language processing, such models provide a narrow view of the cognitive processes involved, describing the contribution of only one component of a complex multimodal system. Our data suggests that under conditions in which information from other modalities is available to constrain lexical access the role of phonological processing post word uniqueness point is often marginal or inconsequential. Spoken word recognition models must therefore be multimodal to offer a comprehensive description of day to day spoken language processing.

References

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

Huettig, F., and McQueen, J. M. (2007). The tug of war between phonological, semantic, and shape information in language-mediated visual search. *Journal of Memory and Language. 54*, 460–482.

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

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,* 509-517.

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

Smith, A. C., Monaghan, P., & Huettig, F. (2013a). An amodal shared resource model of language-mediated visual attention. *Frontiers in Psychology, 4*, 00528.

Smith, A.C., Monaghan, P., & Huettig, F. (2013b). Modelling language-vision interactions in the hub-and-spoke framework. In J. Mayor, & P. Gomez (Eds.), *Computational Models of Cognitive Processes: Proceedings of the 13th Neural Computation and Psychology Workshop (NCPW13). Singapore: World Scientific Publishing.*