A time-invariant connectionist model of spoken word recognition

Thomas Hannagan (thom.hannagan@gmail.com)
CNRS & Aix-Marseille University
3, place Victor Hugo, 13331 Marseille, France

James S. Magnuson (james.magnuson@uconn.edu)
Department of Psychology, University of Connecticut, 406 Babbidge Road, Unit 1020, Storrs, CT 06269-1020 USA
and Haskins Laboratories, 300 George St., New Haven, CT 06511 USA

Jonathan Grainger (i.jonathan.grainger@gmail.com)
CNRS & Aix-Marseille University
3, place Victor Hugo, 13331 Marseille, France

Abstract

One of the largest remaining unsolved mysteries in cognitive science is how the rapid input of spoken language is mapped onto phonological and lexical representations over time. Attempts at psychologically-tractable computational models of spoken word recognition tend either to ignore time or to transform the temporal input into a spatial representation. This is the approach taken in TRACE (McClelland & Elman, 1986), the model of spoken word recognition that has the broadest and deepest coverage of phenomena in speech perception, spoken word recognition, and lexical parsing of multi-word sequences. TRACE reduplicates featural, phonemic, and lexical inputs at every time step in a potentially very large memory trace, and has rich interconnections (excitatory forward and backward connections between levels and inhibitory links within levels). This leads to a rather extreme proliferation of units and connections that grows dramatically as the lexicon or the memory trace grows. Our starting point is the observation that models of visual object recognition – including visual word recognition – have long grappled with the fundamental problem of how to model spatial invariance in human object recognition. We introduce a model that combines one aspect of TRACE – time-specific phoneme representations – and higher-level representations that have been used in visual word recognition – spatially- (here, temporally-) independent diphone and lexical units. This reduces the number of units and connections required by several orders of magnitude relative to TRACE. In this first report, we demonstrate that the model (dubbed TISK, for Time-Invariant String Kernel) achieves reasonable accuracy for the basic TRACE lexicon and successfully models the time course of phonological activation and competition. We close with a discussion of phenomena that the model does not yet successfully simulate (and why), and with novel predictions that follow from this architecture.

Keywords: Keywords: Spoken Word Recognition; Time invariance ; Computational models; TRACE.

Background

Could it be that despite very salient differences, the auditory and visual systems actually rely on the same mechanisms in order to recognize words? One signal has a temporal dimension and is carried by transient sound waves, the other is spatially extended and travels at the speed of light. One signal travels sequentially (over time) through the cochlear nerve, the other in parallel through the optic nerve. In their own dedicated primary cortical regions, however, both arrive at spatial representations – tonotopic for the auditory system, retinotopic for the visual system. What happens next, according to computational models of visual and spoken word recognition, further hints at some possible unification.

Modeling spoken and visual word recognition: TRACE and IA

From a psycholinguistic point of view, two early models of word recognition based on the same computational framework have been enormously successful. In the visual domain, the Interactive Activation (IA) model and its extensions (McClelland & Rumelhart, 1981; Grainger & Jacobs, 1996) can account for a large number of robust and sometimes counterintuitive behavioral findings, in a simple and elegant hierarchical structure where units at any level compete to represent the stimulus, and engage in "lobbying" up and down in the hierarchy. In the auditory domain, TRACE (an extension of the IA framework for speech; McClelland & Elman, 1986) continues to produce new insights into human behavior, including close fits to fine-grained estimates of the time course of spoken word recognition from the visual world paradigm (Allopenna et al., 1998; Dahan, Magnuson, Tanenhaus, & Hogan, 2001); Dahan, Magnuson, & Tanenhaus, 2001).[1]

One probably superficial difference between the two models is that between-level connections in IA models of reading typically include both inhibitory and excitatory connections, whereas between-level connections in TRACE

[1] It is important to note that current, psychologically tractable models of spoken word recognition do not take real speech as their input. While Grossberg & Myers (2000) have modeled aspects of speech and word processing using real speech inputs, these efforts have not yet yielded a model that can handle speech input and a broad range of phenomena in spoken word recognition. In order to be able to address complex issues in word recognition without first solving all fundamental problems in speech perception, TRACE's inputs (for example) are "pseudo-spectral" acoustic-phonetic features that ramp on and off over time, with temporal overlap between adjacent phonemes providing a coarse analog of coarticulation.

are only excitatory. The evidence that this is superficial comes from demonstrations in visual letter identification that performance is at least as good without inhibitory connections between levels (Rey et al., 2009). A much more serious difference, however, is that the IA model can only recognize words at one location on the retina, whereas TRACE has to recognize words at any point in time.

But this impressive ability of TRACE is only achieved at the price of duplicating each unit for as many time slices as needed in the simulation. That is, the processing units in TRACE form a large memory, with units aligned with time ’slices’. Essentially, there is a copy of every feature, phoneme, and word unit at every time slice (the complete details are more complex – for example, words are only duplicated every 3 time slices; see McClelland & Elman, 1986, for details). When input begins, the first instant of the input aligns with and activates units in the first time slice in memory. As the input continues, it activates nodes aligned with specific time slices. Those units can become and remain active for a considerable time after the input has continued on. Conceptually, this is like marks on a page left by a seismograph – the memory banks contain a trace of the input that has come along. But these are not passive traces, since unit activations flux as a function of excitatory and inhibitory input from other units, and a decay parameter.

Having reduplicated units allows TRACE to solve the temporal alignment problem by brute force; given the input /dad/, it can tell that the phoneme /d/ should be activated twice and how far apart in time the two occurrences are – because the two instances of /d/ are encoded by completely independent /d/ detectors aligned with different points in time. The same applies at the word level; TRACE can tell that /dag/ (the TRACE representation of DOG) occurs twice in /dagitsdag/ (DOG EATS DOG) because the two instances are encoded by independent /dag/ detectors aligned with different points in time.

But this comes at a cost. Consider the number of units per slice: 63 x 3 features, 14 phonemes, and, in the basic TRACE lexicon, 212 words, for 415 units. If we ballpark the number of connections by assuming an average of 8 featural connections per phoneme, and 3 phonemes per word, and allowing for connections between units at adjacent time slices, we would have approximately 47,000 connections per time slice with a 200-word lexicon. If we make the trace approximately 2 seconds long (the duration of echoic memory), we need approximately 83 thousand units and 9.4 million connections. If we increase the lexicon to a more realistic size of 20,000 words and the phoneme inventory to 40, these figures reach approximately 4 million units and 80 billion connections.

One might argue that this may not be an unreasonable scale, given the number of neurons and connections in the brain. However, principles of parsimony (might there be a simpler solution?) and evolutionary pressures to minimize energy consumption would be reasonable motivations to seek a less costly solution to time-invariance. Exploring such an alternative is the purpose of this paper, and we report first results on a model that achieves decent performance using many fewer nodes and connections than TRACE. With a 2 second layer of time-invariant input nodes and TRACE's 14 phonemes and 212 words, TISK requires 9.7 thousand units and 62 thousand connections. This represents a 9-fold improvement over TRACE for units, and 2 orders of magnitude for connections. Critically, the orders of magnitude in improvement turn out to be proportional to lexicon size: with 20,000 words and 40 phonemes, TISK would require 48 thousand units (TRACE requires 84 times more) and 3.3 million connections (TRACE requires 24 *thousand* times more).

String kernels

In the machine learning literature, one computational technique that has been very successful at representing sequences of symbols independently of their position goes under the name of *string kernels* (Hofmann et al., 2007). Symbols can be amino-acids, nucleotides, or letters in a webpage: in every case the gist of string kernels is to represent strings (such as "TIME”) as points in a high-dimensional space of symbol combinations (for instance as a vector where each component stands for a combination of two symbols, and only the components for “TI”, “TM”, “TE”, “IM”, “IE”, “ME” would be non-zero). It is known that this space is propitious to linear pattern separations and yet can also capture the (domain-dependent) similarities between them. Although it has been argued in the visual modality that string kernels can account for masked priming effects and are thus likely involved in the early stages of processing, there has been very little investigation of String kernels in the auditory domain (Gales, 2009, being a yet unpublished exception).

Given the demonstrated versatility of the technique, there is every reason to suspect that string kernels could also work in spoken word recognition, where symbols would then be discrete and time-specific phonemes, which would be turned into vectors in the space of time-invariant phoneme combinations. This would entail that the same type of representations are in fact at work in spoken and visual word recognition. However, while one can find some appeal in this unification (this would for instance pave the way to establishing connections between sublexical orthography and sublexical phonology), there remains the nagging problem of how to turn sequences of time-specific phonemes into time-invariant phoneme combinations – that is, how to compute the string kernel for spoken words. Thinking in the unified framework of string kernels suggests that similar problems across modalities can receive similar solutions, and we now introduce our time-invariant alternative to the TRACE model, which handles the transition between time-specific and time-invariant units in much the same way as location-specific and location-invariant units are activated in the visual modality, through the use of symmetry networks (Shawe-Taylor, 1989).

Model

Architecture and dynamics

The model is illustrated in Figure 1. It uses the same lexicon and basic activation dynamics as the TRACE model, but a radically different architecture. It is comprised of four levels: inputs, phonemes, nphones (currently, nphones are single phones or diphones) and words. Inputs consist of a bank of time-specific feature units as in TRACE, through which a wave of transient activation pattern travels. However, this input layer is deliberately very simplified compared to its TRACE analog, given that at any time there is always at most one input unit active – inputs do not overlap in time, and do not code for phonetic similarity (that is, the /d/ unit is equally similar to /a/ and /t/, as each unit can either be on or off; we will address phonetic grain in future work). This input level sends activation forward to the phoneme level. The time-specific phoneme level consists of 10 banks of 14 phonemes that serve as input to the network (the limitation to 10 is completely arbitrary, but sufficient for single-word recognition; there are only 14 phonemes because we are using the 14 phonemes implemented in TRACE). Input phonemes are introduced one at a time and activate the time-invariant nphone level via feedforward connections. Phoneme-to-nphone connection weights are set according to a gradient weighting scheme that we will shortly describe. The nphone level consists of 196 + 14 units, one for each phoneme and for each of the 142 possible diphones that can be composed with the set of phonemes. Units at this level compete with one another via lateral inhibition, and send activation forward to the time invariant word level through excitatory connections, whose weights were normalized by the number of nphones of the destination word. The word level consists of 212 units, one for each of the original words in the TRACE lexicon, with lateral inhibition between words, and feedback excitatory connections from words to nphones.

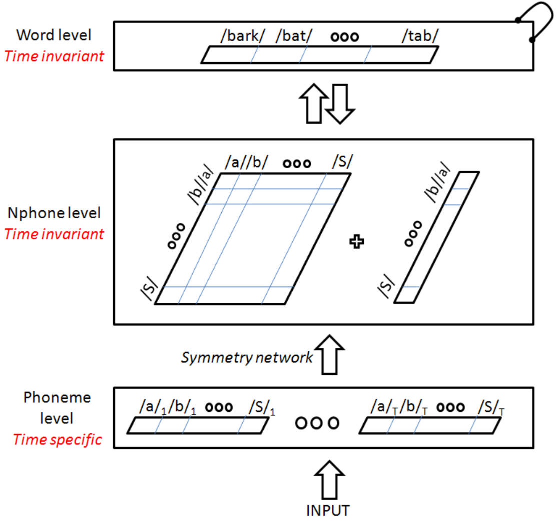

Figure 1: The TISK model - a time-invariant architecture for spoken word recognition.

Note that feedback serves several functions, as does lexical-phonemic feedback in TRACE: it provides a basis for lexical effects on phoneme decisions; it makes the model more efficient and robust against noise (Magnuson et al., 2005); and it provides an implicit sensitivity to phonotactics – the more often a phoneme or nphone occurs in lexical items, the more feedback it potentially receives. Feedback in models of spoken word recognition is a controversial topic (see McClelland et al., 2006; McQueen et al., 2006; Mirman et al., 2006), which we do not address here; our aim is to see whether a model with a radically simpler computational architecture compared to TRACE can (begin to) account for a similar range of phenomena in spoken word recognition.

Units in the model are leaky integrators: at each cycle, all units are activated according to the net input they receive and to their previous activation, minus a decay term, as described in equation 1:

$$A_i(t) = \begin{cases} A_i(t-1)*(1-Decay)+Net_i(t)*(1-A_i(t-1)), \\ if Net_i > 0 \\ A_i(t-1)*(1-Decay)+Net_i(t)*A_i(t-1), \\ if Net_i < 0 \end{cases}$$

and where the net input of unit i at time t is given by:

$$Net_i(t) \quad = \quad SUM_{j=1}^{k} w_{ij} a_j(t)$$

The python code for the model as well as the list of parameters are available upon request to the first author. We now describe how the connections between phonemes and nphones are set in the model.

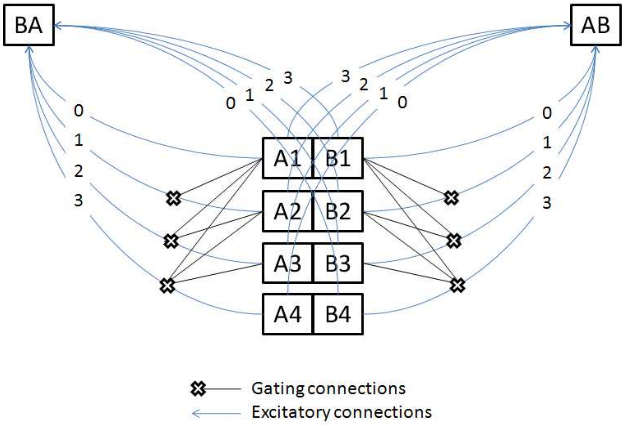

Figure 2: A symmetry network for time-invariant nphone recognition that can distinguish between anaphones.

A symmetry network for phonological string kernels

The problem we are confronting here is to achieve time-invariant recognition while still distinguishing between transposed phoneme combinations. Since we must recognize a succession of phonemes like [/a/$_t$, /b/$_{t+1}$] whatever time "t" is, we need to be able to recognize each phoneme /a/ and /b/ at any "t". But since each unit must activate at any time, how then can we activate unit /ab/ rather than /ba/ at the nphone level?

This issue of selectivity (here, between "anaphones": words with the same phonemes in different order) versus invariance (here, to position-in-time) has long been identified in the fields of visual recognition and computer vision, and has recently received attention in a series of articles investigating invariant visual word recognition (Dandurand, Grainger, & Dufau, 2010; Dandurand, Hannagan, & Grainger, 2010; Hannagan, Dandurand & Grainger, 2011). The solution adopted in the present model is illustrated in Figure 2, and was inspired by what has been learned through this recent work on the way various backpropagation networks deal with the selectivity versus invariance dilemma (to our knowledge this solution has not yet been proposed in spoken word recognition models). Briefly stated, this consists of correlating phoneme-to-nphone connection strengths with phoneme position-in-time, as illustrated in Figure 2 (blue excitatory connections). If the weights from phoneme units /a/$_1$, /a/$_2$,..., /a/$_T$ to diphone unit /ab/ decrease linearly from T-1 to zero, and if on the contrary the weights from phoneme units /b/$_1$, /b/$_2$,..., /b/$_T$ to diphone unit /ab/ increase at the same pace from zero to T-1, then presenting the input sequence [/a/$_t$, /b/$_{t+1}$] will always result in a constant net input for /ab/ whatever the time "t" is, and it will result in a smaller constant net input to /ba/. By setting the weights from these phoneme units to the transposed diphone /ba/ in exactly the opposite pattern, and by setting once and for all a common activation threshold for every diphone unit anywhere between these two net inputs, one can ensure that the network can always neatly distinguish between /ab/ and /ba/. To prevent sequences with repeated phonemes like [/b/$_1$, /a/$_2$, /b/$_3$] from activating large sets of unwanted nphones like /bi/, /b^/), it is however necessary to introduce gating connections (black connections in Figure 2), whereby for instance /b/$_1$ disables the connection between all future /b/$_{t>1}$ and diphones /*b/ (where "*" stands for any phoneme but b).

Other architectures exist that can operate the transition between time-specific phonemes and time-invariant nphones, but the symmetry network we introduce within this model builds on a solution found by the backpropagation algorithm, and has thus arguably a headstart in learnability. It also seems to provide a faithful implementation of the "string kernel" code recently described by Hannagan & Grainger (2011).

Results

Recognition rate

We presented the model with every word in its 212-word lexicon. A word was counted as correctly recognized if it had been the most active lexical unit for ten cycles in a row before the deadline, which was set to 100 cycles. Recognition performance was similar across different operational measures of recognition. With these settings, the model correctly recognizes 98% of the 216 words. We consider this satisfactory for a first test of a new computational approach, although the TRACE model reaches 100% recognition. A consideration of the few unrecognized words, like /triti/ and /st^did/, is instructive in that they were often confused with their cohort candidates (e.g. /trit/ and /st^di/), which activate exactly the same nphones but one (resp. /ti/ and /id/). This confusion can only happen in the current model when two phonemes are closely repeated at the end of a relatively long word, since the importance of any one nphone for recognition of a word is currently inversely proportional to how many nphones it activates. We note that a model whose nphone-to-word weights would be set following other criteria (for instance, the conditional probability of the word given the nphone) would give more importance to diagnostic nphones and reach perfect accuracy.

Competitor effects: Cohort, rhyme and embedding

Figure 3 shows the average cohort and rhyme effects in the model (left panel) and in TRACE (right panel). The curves were calculated by averaging across trials the activation levels of all targets ("target" curve in black), of all words that started with the same phonemes as the target ("Cohorts" curve in red), of all words that ended with the same phonemes as the target ("Rhymes" curve in blue), of all words contained in the target ("Embeddings" curve in purple) and of all other words ("Mean of all words" curve in grey).

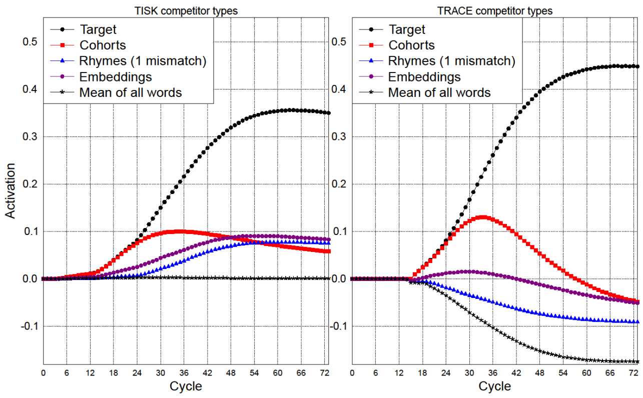

Figure 3: Average activations in the lexicon, when partitioned for each trial as Target, Cohort words, Rhyming words, embedded words and All words.
Left panel: TISK model.
Right panel: TRACE model.

Apart from superficial differences in zero-valued versus negative resting levels, Figure 3 shows that the agreement between models is good on competitor effects. Indeed the magnitude and ordering of the Cohorts, Rhymes and Embeddings effects is similar in the two models, relative to the baseline Mean of all words.

The behavior exhibited by both models also mirrors the cohort and rhyme effects that have been reported in humans performing for instance the so-called "visual world" task. In a nutshell, overall candidate words that begin like the target are more active early on during processing while those that end like the target are more active later one during processing, without ever rising to the activation level of the target, or going below the activation level of unrelated words.

Discussion

The previous results tentatively suggest that a time- specific model of spoken word recognition like TRACE could in principle be replaced by a time-invariant alternative (TISK). This raises the questions of whether there is indeed any kind of evidence for time-invariant phonological representations in the brain, above and beyond considerations of parsimony, and whether one could find predictions that would allow us to uniquely distinguish between the time-invariant and time-specific candidate models. We now address these two questions.

Neural evidence for time invariant spoken word recognition?

Researchers interested in the neural representations for visual words are blessed with the Visual Word Form Area, a well-defined region in the brain that sits at the top of what is still known as the ventral visual stream, and is demonstratively the locus of our ability to read words (Gaillard et al., 2006), but critically not to hear them. Until recently, the common wisdom was that by the mere virtue of its situation in the brain – if not by its purported hierarchical architecture with increasingly large receptive fields – the VWFA was bound to achieve complete location invariance for word stimuli. However recent fMRI studies show that, and computational modeling explains why, a significant degree of sensitivity to location is present in the VWFA Rauschecker et al. (2011). A trained, functional model of location invariance for visual words explains why this can be so: in this model the conflicting requirements for location invariant and selectivity conspire with limited resources, and force the model to develop in a symmetry network with broken location symmetry on its weights. This in turn produces "semi-location invariant" distributed activity patterns, which are more sensitive to location for more confusable words (Hannagan, Dandurand & Grainger, 2011). Thus brain studies have already been highly informative and have helped constrain our thinking on how the brain recognizes visual words.

But attempts to proceed in the same way for the auditory modality quickly run into at least two brick walls. The first is that there is no clear homologue of the VWFA for spoken words. This might be because the speech signal varies in more dimensions than the visual signal corresponding to a visual object; a VWFA homologue for speech might need to provide invariance not just in temporal alignment, but also across variation in rate, speaker characteristics, etc. While there have been reports of hints of such invariance and/or multidimensional sensitivity in the superior (Salvata et al., in press) and medial (Chandrasekaran et al., 2011) temporal gyri, a VWFA homologue for speech has not yet been detected.

The second is that paradigms for testing time-invariance are less easily designed than those which test location-invariance in the visual case. Varying on Rauschecker et al. (2011) however, we can propose a task that tests for the presence of time-specific word representations, in which subjects would be presented with a sequence of meaningless sounds where one spoken word would be embedded. By manipulating the position of this word in the sequence, one could then test whether a "blind" classifier could be trained on the evoked fMRI activation patterns to discriminate which activation patterns correspond to which positions-in-time. A clear demonstration that a classifier is unable to "blindly" map phonological patterns to position-in-time would be good evidence for the model we have introduced. In the alternative scenario, a successful blind classifier would be a smoking gun for this model. Following on our work in the visual modality, we would then need to consider a revised version with limited and shared units that could possibly achieve semi-time invariant representations.

Specific predictions

A specific prediction of this model concerns the treatment of repeated phonemes in a word. As we have seen, the TRACE model deals with both cases by assigning activation to different time-specific units, whereas the model we have introduced must activate for instance the same "na" unit in "banana"at two different times. Finding evidence against this central feature would plainly falsify the model. However it is still unclear at this point how this would really manifest in the model (for instance would words with repeated diphones such as "banana" get more activation from the diphone level than in the TRACE model?). In fact one critical test for the current model will reside in its ability to handle such inputs in a way that is consistent with humans. If the expected differences with TRACE are indeed obtained, experimental evidence could then be gathered with the "visual world paradigm" by presenting targets and distractors with or without repeated diphones. Similarly, one would expect the same phenomena to be within reach of empirical investigations for repeated words in a sentence.

Conclusions

We have presented a computational model of spoken word recognition (TISK) that achieves a close-to-perfect word recognition rate (98%), while also exhibiting the ability to account for basic aspects of phonological competition – the time course of cohort and rhyme effects. This time-invariant alternative uses vastly (orders of magnitude) less computational resources than its time-specific counterpart, TRACE, the economy in number of units being inversely proportional to the number of time steps allowed as input and (in TRACE) memory at all levels or (in our model) at the phoneme level. A notable property of the model is that the computational mechanisms involved – string kernels and symmetry networks – are exactly the same as have been proposed in the visual word recognition literature, paving the way to a possible unified account of word recognition across modalities. Finally we have pointed to where we think specific predictions of the model should arise, and we have put forward a new task that makes the model more easily falsifiable.

Acknowledgments

This work was conducted under the European Research Council grant #230313 awarded to Jonathan Grainger.

References

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

Chandrasekaran, B., Chan, A.H.D., & Wong, P.C.M. (2011). Neural processing of what and who information during spoken language processing. *Journal of Cognitive Neuroscience, 23(10),* 2690-2700.

Cohen, L., Dehaene, S., Naccache, L., Lehericy, S., Dehaene-Lambert, G., Henaff, M., et al. (2000). The visual word-form area: spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Cognition and Brain Theory, 123*, 291 – 307.

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

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

Dandurand, F., Grainger, J., & Dufau, S. (2010). Learning location invariant orthographic representations for printed words. *Connection Science, 22(1),* 25 – 42.

Dandurand, F., Hannagan, T., & Grainger, J. (2010). Neural networks for word recognition: Is a hidden layer necessary? In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, 688-693.

Gaillard, R., Naccache, L., Pinel, P., Clémenceau, S., Volle, E., Hasboun, D., et al. (2006). Direct intracranial, fMRI, and lesion evidence for the causal role of left inferotemporal cortex in reading. *Cognition and Brain Theory, 50(2),* 191 – 204.

Gales, M. J. F. (2009). Sequence kernels for speaker and speech recognition. In Proc. Technology Workshop at Johns Hopkins University, Baltimore.

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple readout model. *Psychological Review, 103,* 518 – 565.

Grossberg, S., & Myers, C. W. (2000). The resonant dynamics of speech perception: Interword integration and duration-dependent backward effects. *Psychological Review, 107,* 735 – 767.

Hannagan, T., Dandurand, F., & Grainger, J. (2011). Broken symmetries in a location invariant word recognition network. *Neural Computation, 23 (1),* 251–283.

Hofmann, T., Scho □lkopf, B., & Smola, A. J. (2007). Kernel methods in machine learning. *Annals of Statistics, 36,* 1171 – 1220.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review , 114* , 1 – 37.

Magnuson, J. S., Strauss, T. J., & Harris, H. D. (2005). Interaction in spoken word recognition models: Feedback helps. In *Proc. Ann. Cognitive Science Society*.

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

McClelland, J. L., Mirman, D., & Holt, L. L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Science, 10,* 363 – 369.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. an account of basic findings. *Psychological Review, 88,* 375 – 407.

McQueen, J., Norris, D. & Cutler A. (2006). Are there really interactive processes in speech perception? *Trends in Cognitive Sciences, 10(12),* 533.

Mirman, D., McClelland, J. L., & Holt L.L. (2006). Theoretical and empirical arguments support interactive processing. *Trends in Cognitive Sciences, 10(12),* 534.

Rauschecker, A. M., Bowen, R. M., Parvizi, J., & Wandell, B. A. (2011). Position-sensitivity in the VWFA measured using fMRI pattern-classification and intracranial recordings in humans. In *Society for Neuroscience Proc*.

Rey, A., Dufau, S., Massol, S., & Grainger, J. (2009). Testing computational models of letter perception with item-level ERPs. *Cognitive Neurospsychology, 26,* 7 – 22.

Salvata, C., Blumstein, S. E., & Myers, E. B. (in press). Speaker invariance for phonetic information: An fMRI investigation. *Language & Cognitive Processes*.

Shawe-Taylor, J. (1989). Building symmetries into feedforward networks. In *Proceedings of the First IEE Conference on Artificial Neural Networks*, London.