Published in final edited form as:
Cogn Sci. 2011 August ; 35(6): 1105–1138. doi:10.1111/j.1551-6709.2011.01181.x.

Using variability to guide dimensional weighting: Associative mechanisms in early word learning

Keith S. Apfelbaum and
Dept. of Psychology, University of Iowa

Bob McMurray
Dept. of Psychology and Delta Center, University of Iowa

Abstract

At 14 months, children appear to struggle to apply their fairly well developed speech perception abilities to learning similar sounding words (e.g. *bih/dih*; Stager & Werker, 1997). However, variability in non-phonetic aspects of the training stimuli seems to aid word learning at this age. Extant theories of early word learning cannot account for this benefit of variability. We offer a simple explanation for this range of effects based on associative learning. Simulations suggest that if infants encode both non-contrastive information (e.g. cues to speaker voice) and meaningful linguistic cues (e.g. place of articulation or voicing), then associative learning mechanisms predict these variability effects in early word learning. Crucially, this means that despite the importance of task variables in predicting performance, this body of work shows that phonological categories are still developing in this age, and that the structure of non-informative cues has critical influences on word learning abilities.

Keywords

associative learning; word learning; language development; switch task; phonological development; variability

1.0 Introduction

It is now nearly dogma that infants learn the sound structure of their native language during the first year of life, just prior to the onset of word learning (Aslin, Werker & Morgan, 2002; Kuhl, Williams, Lacerda, Stevens & Lindblom, 1992; Werker & Curtin, 2005; Werker & Tees, 1984). However, for any given language, the structure of the lexicon determines which set of contrasts must be acquired: languages that lack word pairs like lake/rake have no need to distinguish /l/ from /r/. As a result, there is widespread interest in the idea that lexical contrast may shape phonological development (Charles-Luce & Luce 1990, 1995; Coady & Aslin, 2003; Metsala & Walley, 1998; Walley, Metsala & Garlock, 2003). Thus, the beginning of the second year presents a particularly interesting time for research on phonological development. This is the time when developing perceptual abilities first contact the emerging lexicon.

An important finding during this time is that while 14–month-old infants can discriminate the phoneme contrasts of their native language (Werker & Tees, 1984), at the same age infants are not sensitive to these same phonetic distinctions in word learning tasks (Stager &

Corresponding Author, Keith S. Apfelbaum, Dept. of Psychology, E11 SSH, University of Iowa, Iowa City, IA 52242, 319-335-0692, 319-335-0191 (fax), keith-apfelbaum@uiowa.edu.

Werker, 1997; Werker, Cohen, Lloyd, Casasola & Stager, 1998). For example, 14-month-olds can distinguish *b* from *d*, but struggle to learn that *bih* and *dih* refer to different objects (Stager & Werker, 1997).

Initial accounts of this effect emphasized difficulty. Either the demands of word learning (Fennell & Werker, 2003) or of the specific tasks used to assess it (Yoshida, Fennell, Swingley & Werker, 2009) strained infants' limited capacity preventing their use of such detail. However, recent work has implicated perceptual development by demonstrating that acoustic variability in the training stimuli can help infants overcome these difficulties with word learning (Rost & McMurray, 2009, 2010). This manipulation should have raised the difficulty of the task, yet infants learned the words.

The mechanism underlying this variability benefit is unclear. It is tempting to simply posit that variability is helpful: there is a large body of evidence showing that variability in the perceptual input improves infants' abilities to acquire categories (e.g. Oakes, Coppage & Dingel, 1997; Quinn, Eimas & Rosenkrantz, 1993). However, as we will describe, it may not be this simple: the benefits seen in early word learning appear dependent on variability in *irrelevant* dimensions (Rost & McMurray, 2010).

The goal of this report is to investigate a potential mechanism that could give rise to this effect. We examine simple associative learning and show that it can give rise to this unintuitive pattern of categorization performance. In the remainder of the introduction we will review the relevant empirical literature. We then present an associative account of a set of early word learning studies and demonstrate its power with a series of simulations. Our model suggests that associative learning may guide phonological development as word learning commences, and that the variability of different acoustic cues determines the outcome of these associative mechanisms.

1.1 Early Word Learning

Stager and Werker (1997) reported that 14-month-olds fail to learn similar sounding words despite functional phonological discrimination abilities. Their study and many subsequent investigations employed the Switch task (e.g. Fennell & Werker, 2003; Pater, Stager & Werker, 2004; Rost & McMurray, 2009, 2010; Werker et al, 1998; Werker, Fennell, Corcoran & Stager, 2002; though see Swingley & Aslin, 2002, 2007; Ballem & Plunkett, 2005). In this task, infants are habituated to pairings of auditory words and visual stimuli. After habituation, infants are tested in two conditions. In the match condition, a visual stimulus is displayed while the paired word is played; in the mismatch (or switch) condition, the same stimulus is paired with a different word. Infants' looking time indicates whether they have the correct word-referent linkages. If infants have learned both words and notice the difference, they should look longer to mismatching word-object pairs than to matching pairs. This is an elegant adaptation of habituation to word learning. Since the same materials (auditory and visual) are used for mismatching and matching trials, dishabituation cannot result from sensory differences; rather infants should only dishabituate on the basis of the word-object associations.

Werker et al (1998; see also Stager & Werker, 1997) have shown that after habituation to visual stimuli with the auditory labels *lif* and *neem*, 14-month olds show increased looking time when a visual stimulus is presented on mismatch trials. However, infants at the same age failed to notice mismatches when the trained words differed by only a single feature, place of articulation (*dih* vs. *bih*). Nonetheless, when infants at this age are tested in an auditory discrimination task, they easily discriminate these stimuli (Stager & Werker, 1997). Later work extended this to voicing (Pater et al, 2004, Rost & McMurray, 2009) and showed that even when two features differed (both voicing and place, but not manner), infants fail to

notice mismatches in a word learning situation (Pater et al, 2004). Thus, phonetic similarity (broadly construed) seems to hamper word learning at this age.

A number of accounts have been suggested for this failure. These include resource limitations (Fennell & Werker, 2003; Stager & Werker, 1997), phonological constraint reranking (Pater et al, 2004) and lexical competition (Swingley & Aslin, 2007). Generally, these suggest that some aspect of the way children are encoding words is insufficient to allow learning of close phonological neighbors in this task.

In contrast to the notion of a global capacity limit, the Switch task itself may be to blame. A number of studies have employed a two-alternative preferential looking procedure, in which the auditory stimulus is played while two visual items are presented, rather than the single alternative Switch task (Ballem & Plunkett, 2005; Swingley & Aslin, 2002). These tasks have demonstrated successful encoding of phonological detail of words at 14 months: infants show a reduced preference for the correct object when both known and recently learned words are mispronounced by a single feature. More recently, Yoshida et al (2009) showed that 14-month-olds can even learn two novel minimally different words with the same training given in the Switch task, when learning is assessed with this preferential looking task. Thus infants may not have a general capacity limit at this age. Instead, it appears that something about the Switch task may be causing infants' difficulties.

1.2 Variability and the development of speech perception

Capacity limits and task demands accounts offer a way to rectify the idea that perceptual development is relatively advanced by 14 months with infants' failures to take advantage of these skills during word learning. However, Rost and McMurray (2009) suggest that relatively incomplete perceptual processes may also be at play. They noted that previous studies all habituated children to a small number of auditory exemplars, typically from a single speaker. This training would be appropriate if children have well-delimited phonetic categories by this age. However, if phonetic categories were still developing, a training set with more variability might offer a better platform for learning new word-forms (e.g. Lively, Logan & Pisoni, 1995). Thus, they trained infants on two similar words (*buk/puk*) in a Switch design, but with stimuli recorded from 18 speakers. With this change, 14-month-olds learned both words.

While this is difficult to account for in a task demands framework, it accords well with a range of studies showing that manipulating the variability of the items during training plays a crucial role in studies in which infants must acquire categories during laboratory training procedures. For example, in visual categorization, infants form exclusive categories for objects when the stimuli have close perceptual similarity (e.g. cats), while being less discriminatory for categories with more perceptual variability (e.g. dogs; Oakes, Coppage & Dingel, 1997; Quinn, Eimas & Rosenkrantz, 1993). Crucially, however, variability here has the opposite effect: more variability leads to less specific categories. Further, it is unclear how or if this applies to speech as these studies taught infants a single category (not two). Thus, while these studies implicate variability of some kind in perceptual learning (reinforcing Rost & McMuray's argument that speech categories are still developing), they do not offer an explanation for these results.

More relevant to our discussion of acquisition of speech categories, Maye, Werker and Gerken (2002) showed that 9-month-old infants are sensitive to the distributional structure of phonetic categories, such that bimodal distributions of items are classified as two distinct categories, while distributions without distinct modes are treated as a single category. Interestingly, in a subsequent analysis of the tokens used during Rost and McMurray's (2009) training suggested that the training set contained a clear bimodal distribution of

Voice Onset Time (VOT, the continuous cue that distinguishes /b/ from /p/)—exactly what Maye et al (2002) predict would enhance speech categories via distributional learning (Rost & McMurray, 2010).

This suggests that only the variability in the contrastive cue (VOT) is required to support learning in the multiple speaker condition; variability in the non-contrastive cues (e.g. talker identity) may not have been important. Rost and McMurray (2010) explicitly tested this by varying only the contrastive dimension of VOT (along a similar bimodal distribution) while holding speaker constant. Across two studies, no evidence for learning was seen. However, when the VOT was held constant (within each category) and the non-contrastive cues (talker) varied, learning was observed. This somewhat non-obvious finding conflicts with the aforementioned findings on variability by suggesting that *greater variability* on *irrelevant* dimensions is what is important.

Indeed there is some evidence for this effect in other learning paradigms. Gomez (2002) found that variability aids statistical sequence learning, such that variability in an intervening element cues infants to learn non-adjacent dependencies. Similar results abound in early work on learning theory, wherein the variability of irrelevant cues appears to help attune focus onto useful cues (Bourne & Restle, 1959; Bush & Mosteller, 1951; Restle, 1955). Finally, there is some related evidence in infant speech perception tasks: variability in lexical stress appears to aid segmentation abilities, even though lexical stress is not a reliable cue to lexical identity (Bortfeld & Morgan, 2010). Infants trained with words presented in multiple stress patterns were better able to recognize novel exemplars of these words in running speech thereafter. It appears that learning is most effective when there is high variability in non-contrastive dimensions, and when contrastive dimensions display consistent, lawful variation.

This fundamentally argues that discrimination does not fully capture phonological development in infancy. This is underscored by Dietrich, Swingley and Werker (2007) who showed that while both English and Dutch infants can discriminate vowel length (which is only phonemic in Dutch), only Dutch infants map two different vowel lengths onto two different objects. However, Rost and McMurray (2010) build on this to offer a suggestion as to what other developmental processes are necessary. In this case, infants still need to identify which cues are useful for word learning (c.f., Werker & Curtin, 2005), and the relative variability among cues may play a crucial role in this.

1.3 Mechanisms for Harnessing Variability

One possible account of these findings is that *perceptual learning* mechanisms that are tuned to the statistics of a dimension could act to up-weight or down-weight entire perceptual dimensions. After hearing a series of highly variable pitches, for example, listeners simply pay less attention to pitch. Approaches to perceptual learning based on this weighting-by-reliability approach (e.g. Atkins, Fiser & Jacobs, 2001; Ernst & Banks, 2002) predict exactly this finding. Under this account, highly variable dimensions receive less weight than dimensions that are more consistent. Toscano and McMurray (2010) recently showed how this approach can be applied to the statistics of speech categories. Critically, this account predicts that variability within perceptual dimensions is sufficient to down- or up-weight an entire perceptual dimension: dimensions with clustered values receive greater weights, while dimensions with less structured variability are effectively ignored.

However, this may not be sufficient to account for the pattern of development during early word learning. In the Toscano and McMurray (2010) model, for example, any cues which show consistent clustering will be treated as meaningful. However, the acoustic environment of the infant includes many cues that vary lawfully, yet are not phonologically meaningful.

For example, the fundamental frequency (F0) of voices will cluster into categories of lower F0s for male speakers, and higher F0s for female speakers. However, the infant must gradually learn to ignore this cue with respect to acquiring representations of words. Thus, the presence of statistical clusters may not predict the relatively low weighting that this cue must eventually get for word learning.

Moreover, it is not clear that one would want to weight a perceptual dimension uniformly high or low for all purposes. A cue like first formant frequency is a weak cue for voicing (Summerfield & Haggard, 1977), while it fairly robust cue to vowel height (Peterson & Barney, 1952). Thus, it may need different weights for different phonetic features or for different purposes. However, even this may be insufficient: Ranbom and Connine (2007), for example, suggest that certain phonetic variants of /t/ are linked directly to specific words in adults. Indeed Dietrich et al's (2007) results support this by suggesting that vowel length is not ignored – it is just not used in some cases for word learning.

Perhaps most pertinently, 14 month-olds can discriminate familiar minimal pair words (e.g. *ball/doll*) in the Switch task. If irrelevant cues like speaker voice are down-weighted enough to allow this discrimination, why does this ability not extend to novel words? It would be difficult to account for this dissociation if speaker cues were down-weighted uniformly across all words. Thus, there may need to be a mechanism for determining the importance of perceptual cues for specific phonetic features or even for specific words.

Here, we investigate an alternative: associative learning. Under this view, early words are associated with individual acoustic cue values, and these associations allow the system not just to show sensitivity to raw variability, but rather to seek lawful variation *with respect to some other representation* (the word or object in this case). In learning words, infants can track which cues vary consistently within and across words, and which seem to have no connection to the words they are learning. For example, every time an infant sees a *ball* and hears a parent name it, some phonetic cues are always present (things like the rising formants at onset, low VOT, and high formant frequencies in the vowel). Meanwhile, the F0 will vary depending on who is saying it, and thus will not be consistent across multiple utterances of *ball*. A model in which associations are formed between cue-values and visual categories can thereby bootstrap phonological development, by highlighting which cues are relevant for word identity, and which are unrelated. Indeed, associative training has recently been shown to be more effective than purely perceptual training in speech categorization development for 9-month olds (Yeung & Werker, 2009).

We suggest that the differences in performance across studies of infant word learning (e.g. Rost & McMurray, 2009, 2010; Stager & Werker, 1997; Yoshida et al, 2009) are not indicative of infants failing to encode sufficient detail in difficult learning environments; instead, we suggest that the structure of variability in the different training paradigms alters the way acoustic cues are mapped to words. Rather than underspecified phonological representations, infants encode all available information, whether this information is phonological or not. Their uncertainty about which cues are contrastive leads to *overspecification* of word representations (c.f. Werker & Curtin, 2005). Variable training exemplars serve to ensure that no non-contrastive cues become strongly linked to word identity.

Our goal is to instantiate this account computationally to explore its ramifications. We start by outlining the theoretical account in more detail, and then describe its instantiation in a model. We then simulate many of the existing findings on early word learning to show how this simple account offers a more unified explanation of many of the patterns of data. Finally, we discuss some of the implications and limitations of this approach.

2.0 Theoretical Account

In Rost and McMurray's (2009, 2010) studies, variability along non-phonemically contrastive dimensions (e.g. talker voice or pitch) appears to help infants use contrastive cues. Why would non-contrastive information affect learning? Rost and McMurray suggest that infants may not know which dimensions are meaningful in their language by 14 months. As a result, infants associate both non-contrastive information and phonetically-relevant information with newly learned words. For example, when associating a new word like *buk* with its acoustic instantiation, infants may associate it with its low VOT, but also with its high pitch and 200ms duration.

The studies in which infants fail in the Switch task have all used stimuli with similar non-contrastive information. This has important consequences when we consider how these associations interact with the Switch task. Because the auditory stimuli are from a single speaker, factors like pitch and indexical cues will be consistent across the both of the two words. Repeated presentation of the words with the same pitch and indexical cues, thus leads the child to associate these specific cue-values with *both words*. Figure 1A shows the result of learning in such a case. While VOT values become selectively linked to the correct visual items, the trained pitch and indexical values become strongly associated with both visual items.

Critically, however, the Switch task requires infants to reject an incorrect pairing. That is, on a mismatch trial, they hear *buk* but see the object corresponding to *puk.* Successful learning would be shown by increased looking (surprise). When this mismatching word has the same pitch and indexical cue-values as the correct word, it will be partially activated, leading infants to have difficulty rejecting this mis-pairing. Thus infants' difficulty occurs because they receive several cues that are consistently associated with both words (e.g. the same pitch is associated with both words) and only a single cue (e.g. VOT of the onset phoneme) that disambiguates them.

This builds on task demands accounts. For example, Yoshida et al (2009) also argue that infants' difficulty derives from the fact that both competitors are active due to their phonological overlap (the fact that both *buk* and *puk* share the same coda). As a result, both visual targets are partial matches to the input. However, this cannot explain how variability in speaker voice plays a role. Our account agrees in the importance of co-activation, but we base this co-activation additionally on a failure to ignore the unnecessary speaker cues during learning. Thus, a critical component of infants' difficulty in learning the words is their overspecification (inclusion of too much detail) in lexical representations.

How does this account predict a benefit for variability? With variable exemplars of a word, no single non-contrastive cue-value ever becomes highly associated with either word. As each successive token presents different values of these features, associations are spread out across a range of values. Only the phonologically relevant VOT cue consistently covaries with individual words (see Figure 1B). As a result, during testing infants will recognize whether the auditory target matches or mismatches their expectation of the correct target word, as only the contrastive cue-values will have formed strong connections to word identity.

Thus, two factors are critical: the relative correlations between a particular non-contrastive cue-value (e.g. a single pitch or indexical value) and each of the two words (the phonetic cues will only be correlated with one); and the way that variability in these cues causes learning to spread out across the space. Together, these influence how acoustic cues (both contrastive and non-contrastive) are mapped to words. This in turn shapes the activation for the two lexical candidates that influences Switch task performance.

We illustrate this account of switch task performance here using a series of simulations employing a simple connectionist architecture. There are a number of models of word learning in the literature that employ similar principles (Gliozzi, Mayor, Hu & Plunkett, 2009; MacWhinney, 1989; McMurray, Horst, Toscano & Samuelson, 2009; Regier, 2005; Schafer & Mareschal, 2001). The present simulations are not meant to replace them, nor do we purport to offer a complete model of word learning. Instead, the computational demonstrations are used to explore a theoretical perspective and explain why certain behaviors during word learning appear anomalous. Rather than building a word learning model, we are trying to distill a set of basic learning principles that may be common to many models and determine how their interaction with the input and the task could drive this pattern of performance. In this regard, the existing associative models strengthen this account—their use of similar principles of learning means that our findings here are likely to be found embedded within the rich set of findings one would observe with a more complete model.

3.0 Computational Demonstrations

3.1 Simulation 1: Variability and Invariance

3.1.1 Architecture of the model—The initial system used three sets of auditory units representing contrastive cues (VOT[1]), and two forms of non-contrastive information (F0 and indexical information[2]), along with two visual units, representing the visual stimuli presented during training (*buk* and *puk*). Figure 2 presents a schematic of this model. Each auditory set consisted of 10 units topographically representing values along its dimension. Learning occurs as the model forms associations between auditory information and visual targets. We visualize this as a network, but it is meant to distill a larger process, not model word learning. Such an approach is similar to approaches taken in early explorations of learning theory; rather than trying to develop a full scale model of behavior, we aim to investigate and quantify a specific characteristic of how word learning proceeds.

3.1.2 Training—Each auditory unit was initially connected to both visual units with a weight of zero. During each training trial, one auditory unit from each bank was activated along with the appropriate visual unit. For example, on a trial with *buk* as the auditory and visual stimulus, the third VOT unit (corresponding to a VOT of 0, a *b*) would be active, with the *buk* visual unit and the F0 and indexical unit for that speaker. This simulates the habituation trials during the Switch task.

Connection weights were adjusted using a variant of the Hebbian learning rule (e.g. Rumelhart & Zipser, 1985): when both a visual and an auditory unit were active, the weight between them increased; and weights decayed slightly on each epoch of training (to prevent them from growing indefinitely).

$$\Delta w_{ij} = \eta(a_i v_j - w_{ij}) \quad (1)$$

[1]We are simulating the difference between minimal pairs that differ in voicing here. While the Stager and Werker (1997) task used a place of articulation contrast, Rost and McMurray (2009) replicated this result with a voicing contrast (*buk* vs *puk;* see also Pater et al, 2004). We chose to model the voicing contrast because behavioral results from both protocols are available. Our approach predicts the same results from other contrasts (e.g. *bin* vs *din*).

[2]While we use F0 and indexical information as our non-contrastive cues in this model, we are not presuming that these are the only non-contrastive cues that are present to the infant, nor are we making any strong claims that these cues are preferentially encoded over other non-contrastive cues. Indeed, F0 alone has multiple dimensions along which it can vary (mean height, direction of change, range of change, etc). These cues are included as place holders for the cues that may vary in Rost and McMurray (2009, 2010), until more research illuminates precisely which cues are encoded at this age.

Here, w_{ij} is the associative strength between auditory unit *i* (a_i), and visual unit *j* (v_j), and η is the learning rate.

We simulated both the standard Switch task (henceforth SST) and Rost and McMurray's (2009; henceforth RM) variable training. In SST training, the activated VOT unit was perfectly correlated with the visual unit: the *b* VOT unit would always be activated along with the *buk* visual unit, while the *p* VOT unit always co-occurred with the *puk* visual unit. Meanwhile, the same F0 and indexical units were activated on every training trial. For example, F0 node 5 and indexical node 2 would be activated with every presentation, regardless of VOT and object. This corresponds to the Stager and Werker (1997) protocol, wherein phonological information consistently predicts visual stimuli, while the non-contrastive cues are unvarying across training trials for both words.

For the RM simulation, VOT unit activation was selected as in the SST simulation, but a random F0 and a random indexical unit were activated on each trial. F0 and indexical information had no correlation with the active visual node, and all 10 nodes were equally likely to be used. This corresponds to the protocol of Rost and McMurray (2009) as the phonological information (VOT) is predictive of the word, but non-contrastive cues were uncorrelated.

Training in both versions of the Switch task (e.g. Stager & Werker, 1997; Rost & McMurray, 2009) typically includes seven repetitions of any given auditory stimulus per trial. The infants in Rost and McMurray's (2009) study averaged approximately 18 trials, or just over 120 repetitions, to habituate. As such our model had 120 training epochs on each simulation.

3.1.3 Testing—At the completion of training the model was tested by presenting an auditory word (the VOT, F0 and indexical values), and computing the visual activation via the association strengths. Inputs were similar to training: in the RM simulation one of the representative VOT units was activated along with random F0 and indexical units; in the SST simulation, the VOT unit was activated along with the trained F0 and indexical units.

Activation of the visual layer was determined using the activations and connection weights from the auditory units. Each trial, the activations for both visual units were recorded. If the model heard *puk*, activation of the *buk* visual unit was treated as the mismatch activation, while activation of the *puk* visual unit was treated as the match activation. Visual unit activation is a measure of expectation: how much does the system expect to see a *buk* given the auditory input? This is consistent with current thinking about the role of expectation (or violation of expectation) in habituation (Sirois & Mareschal, 2004).

3.2 Results

We report results from a representative simulation using a learning rate of 0.01 and the parameters outlined above. We start with the RM simulation, as 14-month-olds succeed in learning minimal pairs with this design. Thus, for the purposes of calibrating how much visual activation corresponds to recognition (or surprise) it is the most relevant condition. After 120 epochs of training, the activation of the matching visual unit (0.41) exceeded activation of the mismatching unit (0.07). Thus, given the auditory input, the matching visual unit is more likely than the mismatching unit.

Assuming this difference is sufficient to drive the Rost and McMurray results, this means that an activation of 0.41 (the activation of the matching visual unit in this simulation) is above the recognition threshold, while 0.07 (the activation of the mismatching unit) falls

below this threshold, and so leads to dishabituation (with the current parameter settings). The recognition threshold for this model must be somewhere between 0.07 and 0.41.

In the SST simulation, after 120 epochs of training[3] we again see greater activation for the matching (1.05) than mismatching (0.70) target. Critically, however, *both* values are outside the range of thresholds in the RM model, as both exceed the matching value in that model. If the recognition threshold falls between 0.07 and 0.41, both the match and mismatch units will be recognized as matches to the auditory target, because both are activated above 0.41, so neither will produce surprise or dishabituation. Figure 3 presents this data; no matter where the threshold falls between the match and mismatch in the RM simulation, both SST targets will exceed this value.

Thus, as we described, co-activation of the competing words based on shared indexical and F0 cues is sufficient to drive failure in the SST. Similarly, variability is sufficient to block these erroneous associations. This accounts for the basic difference between Rost and McMurray (2009) and Stager and Werker (2009). However there are a number of further issues that need to be clarified to flesh out this basic account.

3.2.1 Non-minimal pair learning—In order to assess whether this same form of model could account for learning of non-minimal pair words (*lif* and *neem*), we ran an additional simulation, which included additional acoustic dimensions to code for words that differ by multiple phonemes. The architecture of this model is detailed in Appendix A. As can be seen in Figure 4, this simulation demonstrated similar patterns of performance for the RM stimuli and for *lif* and *neem*. Specifically, with this architecture, the RM model produced an activation threshold that must fall between 0.77 and 1.11. The SST model activated both the matching and mismatching units above this threshold (1.54 and 1.26, respectively). Critically, for *lif* and *neem*, the matching object received activation above the threshold (1.75), while the mismatching object's activation fell below the threshold (0.70). The model is able to recognize these non-minimal pair words, even without variability in non-contrastive information.

3.2.2 F0 Variability—In the above simulations, we used the same F0 value on every SST training trial. This matches Rost and McMurray's (2009) version of the SST, which used a single exemplar for all training trials. However, other labs have employed a version of the SST in which the same speaker produces multiple instances of the stimulus (e.g. Stager & Werker, 1997). It is likely that some small variability in F0 occurred between tokens, while indexical features of the speaker's voice remained the same. Empirically this was insufficient to drive learning (the infants did not notice the switch), but it is important to verify that our account shows the same result. Thus, we ran the SST simulation with moderate variability in F0.

F0 values were selected from a Gaussian distribution centered on a single value for the model (representing the speaker's typical pitch). The Gaussians had a standard deviation of 1, ensuring that while multiple pitch values occurred, they clustered closely around the speaker's typical F0. Training and testing were identical to the above simulations. As in the previous SST simulations, simulating small variability in F0 values yielded greater activation for the matching than the mismatching target (0.81 vs. 0.51), and both exceeded

[3]Our interpretation of the results depends on comparing across two models with the same training parameters. However, if infants habituate more rapidly during SST training because of the less variable (and thus potentially less interesting) stimuli, comparing across equal number of training epochs may be problematic. Rost and McMurray (2009, in press) tested this idea on their empirical data, and found that their training procedure elicited the same number of habituation trials as the SST (variable procedure: 18.3 trials; SST procedure: 18.4 trials). Thus, it does not appear that the variable stimuli drive shorter (or longer) habituation times, so comparing models with the same number of training trials is not only reasonable but preferable.

the matching activation for the RM simulation (0.41). Both are activated beyond the recognition threshold, so the model fails to notice the switch. As we increase this variability, both units remain above threshold until we use a purely random distribution, yet at each step activations decrease toward those of the successful RM values (see Figure 5). This suggests that variability across a few non-contrastive dimensions (e.g. F0 height, F0 direction and F0 range) may be sufficient to enable word learning even with a single speaker producing tokens.

3.2.3 Task Effects—Finally, we have not yet addressed why infants can succeed in single speaker training when tested with a different task, particularly the two-alternative preferential looking task used by Yoshida et al (2009). The SST simulations address this as well. While both visual units exceeded the recognition threshold defined by the RM simulation, the activation of the matching unit was greater than that of the mismatching unit. This suggests that while the auditory stimulus is accepted as a match for both visual items, the correct matching item still receives greater activation. The child will look more to this item because it is a *better* match, though both would be accepted as appropriate in the one-alternative Switch task.

Of course, these conclusions rely on our assumptions of how to tie activation levels to the task the infant is performing—essentially, a task demands account built on an associative core. In the Switch task, a threshold is appropriate, as the infant either looks or does not look at the target item – the activation level determines whether the visual item is a sufficiently good match to trigger recognition, and thus to show continued habituation. However, in the preferential looking task, the mapping is more complicated; here the infant is not merely deciding whether a visual stimulus matches the auditory input, but must decide which of two possible stimuli *better* matches this input. While both stimuli may be acceptable matches to the auditory word (both exceed the recognition threshold), the one with greater activation will be a better match, and so draw more looks. These differences in how activation maps onto behavior can explain why infants seem to exhibit word learning when tested in a preferential looking task even if they seem to fail to learn object labels when tested with one-alternative Switch tests (Samuelson, Schutte & Horst, 2009). When multiple items are on the screen simultaneously, the relevant operation driving the behavioral response is the comparison between their activation levels. When only a single item is on the screen, the relevant operation is comparing that unit's activation to some internal criterion or threshold. Thus, our model does not rule out task-demands—it embraces them. However, it is not general task demands (e.g. difficulty), but the way that the structure of a task maps onto the activated representations built from associative learning.

3.3 Generality of the Findings

At first glance, our explanation of the SST might seem an arbitrary fact about where the threshold is set. A concern for this and any computational demonstration is whether success is simply the result of a "sweet spot" in the parameter space—a uniquely suitable set of parameters. If different parameter settings yield different patterns of results, this weakens claims that the empirical behavior results from basic core principles governing these effects and may suggest limitations (Pitt, Kim, Navarro & Myung, 2006).

To test the generality of our account, we tested a large space of parameters, defined by the free parameters of the model and the architectural choices we made. This included variation in the number of training epochs; learning rate; how the input VOTs were chosen; initialization of the weights; how F0 and indexical values were selected during training (including the variable F0 manipulation detailed above); type of decay in the Hebbian

learning rule; and use of tuning curves in the auditory banks (for details see Appendix B and the Online Supplement).

Across these variants, 1600 simulations were tested with different parameter sets. While the raw activation values varied depending on the parameters used, each simulation could be classified in one of three ways. First, if the model failed to show more activation in the match than mismatch in the RM version, it was incapable of learning the words with the amount of training during typical habituation, and so is an invalid model. Second, the simulation may have succeeded at learning in the RM condition, but the RM-derived range of thresholds also predicted learning in the SST—a true failure to simulate the empirical data. Third, the simulation could pass on both criteria (as the above simulations did); this indicates successful simulation of the empirical data.

Figure 6 shows the results of this search. 86% of simulations were successful under this metric: the activation for matching words was higher than mismatching in both the RM and SST simulations; and, more importantly, in the SST simulation, both the match and mismatch visual units exceeded the RM activation in the match condition. 6% of simulations of models failed to learn the words entirely, and 8% learned the words but the RM threshold predicts success in the SST (see the Online Supplement for an analysis of the failures). Thus, of simulations that learned the words in the RM protocol, 91% showed the full empirical data pattern. The fact that the vast majority of simulations succeeded suggests that these results are fairly independent of specific parameter values – simple associative mechanisms robustly predict this behavioral pattern.

Finally, our account also robustly models the Yoshida et al (2009) results. 99% of the simulations we ran showed greater match activation than mismatch activation in the SST (Figure 6). This would lead to more looks to the match target in a two-alternative preferential looking design.

3.4 Discussion

These simulations demonstrate that simple associative learning principles can account for the differences in performance in the Switch task as a function of variability. The SST simulations learn to associate F0 and indexical values strongly with both visual items, because the same values were presented on every training trial. Presentation of a test item with these values activates both targets as the F0 and indexical values strongly predict both visual units. While the VOT units allow the matching visual unit's activation to exceed that of the mismatching unit, both are activated to a high level. The model strongly activates both units due to overspecification of lexical representations—they include both relevant details like VOT, and also irrelevant details like speaker voice (e.g. Goldinger, 1998).

In the RM simulations, the association between a visual unit and any specific F0 or indexical value never becomes strong. Each training trial causes different F0 and indexical units to increment their connections with the visual units; however all of these connections remain weak. Meanwhile, VOT is highly predictive of visual target identity, so these connections become strong. When testing, the matching visual unit will receive strong activation from the VOT unit and little activation from the F0 and indexical units. The mismatching unit will receive little activation from any auditory units. The system thereby expects (strongly activates) the matching object but not the mismatching one.

These simulations demonstrate that associative mechanisms can explain why adding variability to the training stimuli in the Switch task improves word learning. While such a manipulation appears to make the task more difficult by adding additional irrelevant information that the infant must filter out, our associative account shows how multi-talker

training can lead infants to better word learning by dispersing the connections between voice-specific information and lexical identity.

An interesting implication of this is that infants' failure to exhibit word learning in earlier Switch tasks may be because the word learning environment was designed in a non-optimal way once we realize that infants may know a lot less than we think (specifically: they do not know how to interpret talker voice). Presenting stimuli with very little variability may suggest to children that aspects of the speaker's voice, which do not vary across multiple repetitions of the word, should be part of lexical representations. By increasing variability in the speakers producing the words, infants can better identify which aspects of the acoustic information are meaningful.

Thus, infants' failure to demonstrate word learning may be dependent on using highly similar tokens across training and testing. During training, infants learn to associate both the phonological information and the speaker-specific information with the visual objects. If testing was done with a novel voice producing the items, infants may display better learning, as the phonological information would match what they had learned, while the non-contrastive information would not match either of the items. This represents another case in which adding difficulty to the Switch task might improve performance.

4.0 Further Simulations

In this section we extended these simulations show how two important findings can emerge from this simple associative account. First, we simulated the results of Rost and McMurray (2010) which demonstrate the locus of the variability effect. Only variability in *non-contrastive* information is sufficient to drive learning—variation in VOT alone does not work. Additionally, we expanded the account to simulate the mispronunciation effects of Ballem and Plunkett (2005), showing sensitivity to fine-grained phonetic information after single-speaker training, when testing with a slightly different task.

4.1 Variability and Specificity

Rost and McMurray (2010) showed that variation in VOT, with no variation in non-contrastive values, leads to failure to dishabituate on mismatch trials. They presented infants with a range of tokens from the same speaker, but with VOT values that varied between tokens in a bimodal distribution. In these experiments, infants failed to dishabituate on Switch trials. Conversely, when VOT was held constant, but talker-voice varied, infants learned the words.

To simulate this, we trained the same system with varying VOT values and unvarying F0 and indexical values. VOT values were selected from a bimodal Gaussian distribution. Each mode was centered on the prototypical category value, with a standard deviation of 0.9. For *b* VOTs, the distribution was centered on VOT unit 3, while *p* VOTs were centered on unit 8. As in the SST simulations above, and in Rost and McMurray (2010), a single F0 and indexical value were used in every training trial.

The results for this simulation mirrored those for the SST simulation. Specifically, the activation for the matching object, at 0.86, exceeded mismatch activation, at 0.74. Critically, both values exceeded the match activation in the RM simulation (0.41); both match and mismatch activations surpass the recognition threshold (Figure 7). The model recognized both match and mismatch items as somewhat appropriate visual targets, and so did not dishabituate to the mismatch, as seen in Rost and McMurray (2010). Interestingly, while the match activation exceeded the mismatch activation, the difference in activation values was smaller than in the SST simulation. This suggests that training with variable VOT but non-

variable F0 and indexical information might lead to greater impairment in two-alternative preferential looking testing.

Crucially, much like infants, it is not variability in general that drives good performance in this associative mode; rather it is the relative variability between relevant and irrelevant cues.

4.2 Mispronunciation effects

One initial interpretation of Stager and Werker's (1997) findings was the hypothesis that early lexical representations were underspecified and incapable of representing single-feature differences. Some of the most important evidence against this underspecification account has come from mispronunciation studies in which infants show reduced looking to the target (e.g. a picture of a *baby*) when it is mispronounced by a single feature (*vaby*) (Swingley & Aslin, 2002). Critically, Ballem and Plunkett (2005) showed this effect even with newly learned words: infants trained on the non-words *vope* and *tuke* are sensitive to the mispronunciation *fope*; that is, they treat *fope* differently than *vope*. Simulating this required minor architectural changes (since Ballem and Plunkett did not use minimal pairs). Given the importance of these findings in clarifying the nature of early lexical representations, it was important to examine these results in our framework. Thus, we incorporated additional input units to represent the training words used by Ballem and Plunkett (2005).

This simulation used a set of VOT units (representing *t* vs. *d*), a set of fricative voicing units (representing *f* vs. *v*) and a set of coda units (for coarsely coding offset information[4]). We included F0 and indexical units, and again used two visual units, now representing *tuke* and *vope* (Figure 8). On each training trial, the system was exposed to *tuke* or *vope* by activating the appropriate visual and auditory units, and the same F0 and indexical units for each exposure trial (since no studies have explored variability in this paradigm). Learning proceeded as in previous simulations.

The system was tested on the correctly pronounced target item (*vope*) and a mispronunciation (*fope*). As in Ballem & Plunkett (2005), this mispronunciation did not create another word, and the VOT and fricative voicing units activated for mispronunciation trials had not been presented during training. Visual activation of the *tuke* and *vope* units was determined based on activation values and connection strengths from training.

After 120 training trials, activation for *tuke* (1.54) exceeded activation for *vope* (0.63) when the correct *tuke* auditory units were activated. However, when a mispronounced *tuke* was presented (*duke*[5]), *tuke* still exceeded *vope*, but the magnitude of the difference decreased (*tuke:* 1.15; *vope:* 0.63; see Figure 9). When *vope* was presented, the pattern of activation mirrored that of *tuke*: *vope* was more active than *tuke*, and this effect was smaller when *fope* was heard. The decline in target activation when a mispronunciation is presented mirrors the behavioral findings of decreased fixation of the 4 Our use of a single localist unit to represent the entire coda is somewhat oversimplified. Further explorations with this architecture in our lab have confirmed that incorporating more distributed coda representations does not alter the results. 5 While duke is an English word, it should be noted that it is not a word in the model's lexicon, and is presumed to be absent from the

[4]Our use of a single localist unit to represent the entire coda is somewhat oversimplified. Further explorations with this architecture in our lab have confirmed that incorporating more distributed coda representations does not alter the results.

[5]*While duke is an English word, it should be noted that it is not a word in the model's lexicon, and is presumed to be absent from the vocabularies of 14-month-olds who are not members of the nobility.*

vocabularies of 14-month-olds who are not members of the nobility. target after a mispronunciation (Ballem & Plunkett, 2005).

4.3 Discussion

These two final simulations demonstrate that a range of effects seen in early word learning are compatible with an associative account. Variability in training items leads to improved ability to demonstrate learning in a Switch task in our model, much like in the empirical data. Further, our model emulates the selective benefit of variability in non-contrastive information over variability in contrastive dimensions. The model, like infants, is aided by variability in dimensions that are not criterial to phonological identity, while variability in VOT leads to failure in this task. As such, it appears that simple associative mechanisms are sufficient to explain why variability seems to help infants.

Additionally, these same mechanisms are effective in demonstrating why certain other tasks seem to yield better word learning, even with single-speaker training. Ballem and Plunkett (2005) found that children were sensitive to single feature mispronunciations of newly learned words when tested with a preferential looking task. Associative mechanisms predict this same pattern; even without variability in speaker-specific information, mismatching phonological information is noticed and can thereby lead to decreased looks to the target. While it is possible that variability in speaker during word learning in this task could magnify these effects, the effects seen with single speakers are not incompatible with our associative account.

5.0 General Discussion

We have demonstrated that basic learning processes can produce many of the behavioral results seen in low variability Switch task studies (Stager & Werker, 1997; Werker et al, 1998, 2002; Pater et al, 2004), high variability Switch studies (Rost & McMurray, 2009, 2010), and two-alternative preferential looking studies (Ballem & Plunkett, 2005; Yoshida et al, 2009). The critical link explaining why infant behavior varies so widely across these studies appears to be the structure of the acoustic variability in the training set. When infants are exposed to variable exemplars of words, their learning is focused on the consistent pieces of information – in this case, phonological information. However, when additional information is also highly consistent, as when a single speaker produces the words in a highly similar intonation, infants also associate this non-contrastive information with the object. This redundant information in the end reduces the contrast established by the phonetic cues.

Such an account suggests that despite evidence of adult-like abilities to discriminate between phonological categories (Kuhl et al, 1992; Werker & Tees, 1984), the task of phonological acquisition is not yet complete early in the second year. The additional task of phonological acquisition is to determine which dimensions meaningfully map on to words, and which are not useful for identifying words (Dietrich et al, 2007). To accomplish this second task, tracking the variability of different dimensions as they relate to object categories appears to be a highly effective mechanism for determining which cues are phonologically-relevant.

5.1 The plausibility of overspecification

Our account of early word learning posits overspecification of lexical representations, such that acoustic information that is irrelevant to lexical identity is encoded as part of the word form. While at first such indiscriminate encoding appears to be a contentious claim, there is a sizable body of evidence supporting such an account. Thirteen-month olds associate non-

speech sounds (e.g. beeps and whistles) with pictures as though these sounds are appropriate object labels (Woodward & Hoyne, 1999; though Hirsh-Pasek, Golinkoff & Hollich, 2000, only found this with non-speech sounds made by humans, e.g. sighs). Further, there is evidence from segmentation tasks that at least as late as 10.5 months, infants encode affective cues along with word forms (Singh, Morgan & White, 2004) and at 7.5 months encode arbitrary variation in pitch (Singh, White & Morgan, 2008). When familiarized to words with a happy affect or a specific pitch value, infants show a preference for passages containing matching words only if they contain the same prosodic cues they heard during training.

The PRIMIR framework also suggests that overspecification could account for Switch task failures at 14 months (Werker & Curtin, 2005). Under this framework, infants encode all available information, and the indexical information overwhelms processing capacity, leading children to ignore phonological information. While this accords with our view of overspecification, this account differs from our approach in suggesting that children are ignoring phonological information; instead, we suggest that children use that information, but it is insufficient to overcome the wealth of indexical information, that they also use for word recognition. These talker cues suggest that the minimal pair words are the same, leading to the failure. Additionally, it is unclear how PRIMIR's account of overspecification could accommodate the benefit of multiple speaker training in the Switch task.

More directly, adults associate speaker-identity cues with lexical items (e.g. Bradlow, Nygaard & Pisoni, 1999; Creel, Aslin & Tanenhaus, 2008). Such associations are also reasonable assumptions for development – languages like Chinese use pitch phonologically, and many cues to speaker identity overlap with cues to phonation type (which is contrastive in languages like Green Mong; Andruski, 2006). Thus, children should not have innate constraints on which cues will be contrastive in their language; they must learn this to cope with the incredible diversity seen across phonologies of the world's languages. Our account presents a mechanism for learning which acoustic dimensions are most useful phonologically: the relative pattern of covariation between cues and words leads the child to down-weight non-contrastive cues and up-weight contrastive ones. Tracking variability in different acoustic dimensions is a highly effective tool in determining which dimensions should be attended. Crucially, this mechanism is intimately bound up with the emerging lexicon – variability is tracked with respect to specific words, and has its effect in the associations between specific acoustic cues and words.

Our assumption that the 14-month-old does not know which cues are phonemically relevant would seem to discount the learning already completed to this point. However, we are not arguing that 14-month-olds have not learned anything – they can clearly discriminate speech sounds (Eimas, Siqueland, Jusczyk & Vigorito, 1971; Werker & Tees, 1984), and may also be able to categorize them (although categorization is irrelevant to Switch task studies, since none have ever asked infants to generalize across different exemplars in a substantive way [e.g. to a new speaker]).

Indeed, in the aforementioned studies by Singh et al (2004, 2008), infants' ability to *generalize* across affects seems to slightly improve between 7.5 and 10.5 months. Thus, the ability to ignore this form of irrelevant variation is clearly developing as well. However, this has only been tested in a discrimination task using highly differentiable, well-known words (*bike* vs. *hat*), and so may not be sufficient for learning of minimal pairs at later ages. Thus while infants may have begun this process of dimensional weighting by 14 months, there is no reason to think it is complete. Indeed, Goldinger (1998) has shown that even into adulthood, speaker-specific information plays an important role in lexical encoding, especially for words that are heard infrequently. Moreover, if infants already knew which

dimensions to attend for word learning, it is unclear why they would learn better in the Rost and McMurray (2009, 2010) studies than in SST studies, and why learning (non-minimal) pairs of words with similar contrasts can improve learning of some contrasts (Thiessen, 2007, 2010). Natural learning settings may allow infants to learn minimal pairs (e.g. *ball* and *doll*), because of natural variability in non-linguistic information in these settings; the SST, however, plays to their weakness by reinforcing encoding of non-contrastive cues. Indeed while we have focused on auditory information as the source of false associations, it is possible that additional contextual cues could be encoded early in word learning as is often seen in memory tasks (Gooden & Baddeley, 1975).

5.2 Levels of representation

Our model assumes direct connections between acoustic information and lexical items, without intermediate (e.g. phonetic) representations. Such direct connections between acoustic information and object representations is consistent with exemplar models of speech perception and word recognition (Goldinger, 1998; Hawkins, 2003; Pierrehumbert, 2001), and with associated evidence that indexical information affects adult spoken word recognition (Bradlow et al, 1999; Creel et al, 2008; Goldinger, 1998). However, our model contrasts with pure exemplar models in two ways. First, we predict the pattern of statistical co-occurrence between cues and words eventually results in non-contrastive information (like F0 and indexical information) receiving significantly less weight than contrastive cues (though perhaps in a word-specific way). This can be accomplished by incorporating dimensional weighting into an exemplar approach (Kruschke, 1992, 2001; Nosofsky, 1986; Regier, 2005). While infant speech perception may start like a pure exemplar model, over the course of learning, the ability to (at least partially) abstract away from irrelevant cues emerges.

Second, and perhaps most importantly, individual exemplars are not stored in long term memory—rather, the weights of the network store an accumulation of their traces. Traditionally, exemplar approaches have confounded the encoding of speaker-specific information and the storage of individual lexical exemplars. Our approach shows that these two aspects can be separated; while our model does not need to store specific exemplars of spoken items, it remains sensitive to speaker-specific cues. By incorporating the exemplar-type sensitivity to speaker in a more prototype-style model, associative learning may offer an alternative account of such effects without the controversial claims regarding storage of huge sets of exemplars.

5.3 Other influences in early word learning

While our associative account can explain a fairly large set of the findings in word learning in 14-month olds, there are a number of other factors that likely play an additional role. For example, as the lexicon increases, infants become more able to learn minimal pair words even with single speakers (Werker et al, 2002). At some point infants become able to generalize their learning of which dimensions are meaningful to new words that they are learning. As currently instantiated, our model weights cues on a word-by-word basis – it cannot generalize this learning across words.

There are multiple ways one may be able to account for this. We earlier described perceptual learning mechanisms that weight dimensions as a whole (perhaps using the relative variability along each dimension as a source of information; Toscano & McMurray, 2010). This may operate alongside the associative mechanisms outlined here, and would need to work slower than these associative mechanisms to yield the right pattern of data. Such a mechanism is reasonable – it would not be too costly if infants committed to the wrong cue-weighting for a single word, while it would be substantially worse if an entire dimension

were ruled out prematurely. Thus, one would want associative forms of cue-weighting to work faster than perceptual learning.

Alternatively, Thiessen (2007) presents data that suggests that the distributional structure of the lexicon may guide learning through feedback mechanisms. When children learn very distinct words (*dawbow* and *tawgoo*), they were better able thereafter to learn the minimal pair (*daw* and *taw*). Upon hearing *daw*, children may partially activate the learned word *dawbow*. Feedback from this word could boost activation of the overlapping phonological information, thus helping draw attention to the relevant aspects of word identity. Even if irrelevant dimensions like talker voice were still being considered, this could help overcome it. Such a feedback mechanism may explain why children gradually become better able to learn minimal pair words. As phonological neighborhoods become more populated, a greater number of words will be co-activated during novel word learning, leading to a large degree of feedback suppressing irrelevant information. Indeed, Thiessen (2010) has shown just such an effect of neighborhood populating. While learning *dawbow* and *tawgoo* was sufficient to allow learning of *daw* and *taw*, this learning did not generalize to other vowel contexts (*dee* and *tee*). However, exposure to a wider array of vowel contexts before minimal pair learning helps infants generalize their learning to new contexts.

The lexical-feedback account could complement the associative mechanisms presented in our simulations. At its core, this account suggests that learners achieve dimensional weighting through links between acoustics and concepts. As more words are learned, the consistent weak association between non-contrastive information and lexical identity is more easily generalized to new words. A purely perceptual mechanism for dimensional weighting may have more difficulty accounting for such an effect, as it appears to depend on having formed links between acoustics and concepts.

Finally, word learning in the Switch task may not be an entirely auditory problem. Presentation of known objects with their names prior to Switch task training yields better word learning (Fennell & Waxman, 2010). This suggests the influence of higher level systems on word learning, though it is unclear what those might be in this work. The initial presentation may contribute to speaker familiarity or normalization affecting the way speaker cues are coded, or it may alter the attentional state of the infant changing the way activation thresholds map to responses. Regardless, such results may complement our own by suggesting that this associative core must be embedded in a richer system that supports language behavior and learning.

In this light, the associative mechanisms detailed in this paper are not meant to be an exhaustive model of word learning, but rather to represent one key component to the system. A number of features could be added to our model to make it a more complete account of word learning, including the addition of intermediate stages between the auditory and visual units (Schafer & Mareschal, 2001), dimensional attention weighting (Kruschke, 1992, 2001; Regier, 2005), competition between units, or feedback (McMurray et al, 2009). However, the primary goal of this study was to investigate the locus of variability effects in early word learning, not to provide a comprehensive model of word learning. From these simple simulations, it appears that the influence of variability in early word learning studies is related to how associative mechanisms link acoustic information to visual categories.

6.0 Conclusions

We have demonstrated how simple associative mechanisms, using basic learning principles, can account for the role of variability in early word learning, and explain differences in sensitivity to minimal phonetic differences exhibited by children in an array of tasks. 14-

month-olds fail in the SST not because they fail to encode contrastive cues, but because the lack of variability in the training exemplars leads to association of non-contrastive information with objects. Similarly, they succeed under multiple speaker conditions because variability in non-contrastive information blocks their ability to associate it with the words.

Our account, and the simulations that support it, suggests that differences in Switch task and preferential looking performance are not well described by resource limitations (Fennell & Werker, 2003; Werker et al, 1998; Werker & Curtin, 2005), variable constraint settings (Pater et al, 2004), lexical competition (Swingley & Aslin, 2007), phonological underspecification (Yoshida et al, 2009), or probabilistic phonological encoding (Yoshida et al, 2009). Instead, overspecification and the structure of word learning environment in these tasks lead to the variations in displayed abilities. Not only do we argue that phonetic cues are encoded in all these tasks, *but non-contrastive information is encoded as well*. As a result, these differences will emerge from basic associative mechanisms because of the incomplete nature of phonological understanding at 14 months coupled with tasks that exploit this incomplete development.

Our use of basic learning principles to explain how variability interacts with and helps explain phonological abilities mirrors work in learning theory, which has explained many seemingly complex effects, like blocking, overshadowing and, learned irrelevance, as the product of simple learning rules (Wasserman & Miller, 1997 for review). The application of basic learning theory to dimensional weighting (or more specifically down-weighting of irrelevant dimensions) has its roots in work from the 1950s on *cue adaptation* or *cue neutrality* (Bush & Mosteller, 1951; Restle, 1955; Bourne & Restle, 1959). These studies showed that learning is dependent on the statistical relationships between stimuli, which are iterated over time. More recently, such findings have also been shown in statistical sequence learning (Gomez, 2002), segmentation (Bortfeld & Morgan, 2010), and visual categorization (Oakes et al, 1997; Quinn et al, 1993), where the variability in the category exemplars determines whether learning will take place. Similar associative mechanisms to those detailed in this paper may underlie all of these findings.

Perhaps the most important question raised by Stager and Werker (1997), however, does not concern learning. Rather, it is whether early lexical representations are underspecified or are sensitive to phoneme level contrasts. Our model proposes a third alternative. Early lexical representations are not underspecified – they are overspecified, including much irrelevant information. Yet associative learning, coupled with the fact that only contrastive cues consistently vary with word identity, appears to be sufficient to overcome this. As a result, it offers a compelling account of children's successes and failures in word learning.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank Ed Wasserman for helpful discussions on learning theory; Gwyn Rost for discussions on the Switch task and her work; and Katherine White and Marcus Galle for helpful comments on an earlier draft.

Appendix A: Modeling non-minimal pair learning

Our model of learning non-minimal pair words (*lif* and *neem*) without variability in speaker voice required minor changes to the model architecture. While *buk* and *puk* overlap in all but the onset phoneme, *lif* and *neem* differ at every position. As such, to more accurately

represent the differences between stimuli, it was necessary to use representations that included information on each position in the word.

To this end, we employed three banks of phonological input units. Each bank offered a coarse, localist phoneme encoding for one word position. While richer representations would more accurately match the acoustic differences between the words, we began this simulation with the simplest possible representations. Adding featural or cue-based phoneme representations would serve to further differentiate the items *lif* and *neem*, and so would not alter the results of our simulations.

The units in the onset phoneme bank could represent *l, n, b* or *p*, to allow the model to represent *lif, neem, buk* or *puk*. Similarly, the middle phoneme bank allowed representation of *I, i* and *u*, while the coda phoneme bank had values *f, m* and *k*. We included banks of F0 and indexical information as well.

Parameters

This model used the same base parameters as our initial SST and RM models. Specifically, units were initialized with connection weights of 0; the model was trained for 120 epochs at a learning rate of 0.01; units were localist, and thus did not include tuning curves; F0 and indexical information were uncorrelated with each other; and F0 values were either constant (in no-variability simulations) or completely random (in variability simulations). The same learning rule presented in the paper was used, in which connection weights were augmented when auditory and visual units were both activated, and all weights slightly decayed on each trial.

Training

To train the model, an auditory word was presented by activated the appropriate phonetic units and a single F0 and indexical unit. The model was trained in three conditions: learning *lif* and *neem* without variability in speaker voice (Werker et al, 1998); learning *buk* and *puk* without variability in speaker voice (Stager & Werker, 1997; Werker et al, 1998); and learning *buk* and *puk* with speaker variability (Rost & McMurray, 2009, 2010). In no-variability simulations, the same F0 and indexical units were activated on every trial. In variability simulations, a random F0 and indexical value was chosen on each trial. During each training trial, the word that was presented to the model was accompanied by its corresponding visual unit.

Testing

At the conclusion of training, the model was tested by activating an auditory word and calculating visual expectations. This calculation was performed by multiplying each activated unit by its connection strength to an output unit. The sum of all such inputs represented that output units activation. For each model, the visual unit matching the auditory word was considered the “matching” unit; the visual unit representing the other word trained for that model was the “mismatching” unit – that is, for the *lif* and *neem* simulation, if *lif* was presented auditorily, the visual *lif* unit represented the matching unit, while the visual *neem* unit represented the mismatching unit. The values presented in Figure 4 of the paper represent the average activation across both trials for the matching and mismatching unit.

Interpretation

As in the other simulations reported in this paper, interpretation of the results should be considered relative to the activations from the RM model simulating learning *buk* and *puk* with variable training exemplars. As we know this training methodology yields successful word learning, we can use the activations from this simulation as a guide to the activation threshold required for successful word recognition. Thus any words with activation below that of the mismatching unit from the RM simulation must be below the activation threshold, and so will not be recognized as matches to the auditory stimulus. Similarly, words with activation greater than that of the matching unit in the RM simulation exceed the recognition threshold, and so will produce recognition as matches to the auditory stimulus. As can be seen in Figure 4, the model predicts successful recognition of only matching words when learning *lif* and *neem*, while it accepts both matching and mismatching words as appropriate object labels when trained with *buk* and *puk* in the SST methodology.

Appendix B: Parameter space search

To explore the generality of the model's results, we manipulated the free parameters to explore how they affected model behavior. In particular, we wanted to show that the model's success at capturing the stimuli and task differences was not an artifact of any specific parameter settings. Thus, we varied the following factors, testing all possible combinations.

1. Pitch variability (2 levels, as described in **F0 Variability**): A) F0 in the SST model was constant, always using the same value. B) F0 in the SST model was selected from a Gaussian distribution centered around a typical value (node 5), with a standard deviation of 1.
2. Number of training epochs (5 levels): 30, 60, 90, 120, 180
3. Learning rate (5 levels): .001, .005, .01, .05, .1
4. Selection of VOT: during training of the RM model (2 types). A) During training, VOTs were chosen such that *b* was always node 3 and *p* was always node 8, as described in the main text. B) VOTs were chosen from a bimodal distribution such that *b* trials used a VOT selected from a Gaussian distribution centered at node 3, with a standard deviation of 0.9, while *p* trials' VOT was centered at node 8 (as in McMurray, Aslin & Toscano, in press). Selections outside the range of VOT values were set to extreme values (so selections less than 1 were set to VOT value 1, while selections greater than 10 were set to value 10). For testing, we used the prototypical VOT values (unit 3 for *b* trials, unit 8 for *p* trials).
5. Initial connection strengths (2 levels): A) Connection strengths were initialized to 0. B) Connection strengths were initialized to small random values between 0 and .1.
6. Pitch and timbre selection (2 levels): A) F0 and indexical values in the RM model were uncorrelated with each other. B) F0 and indexical values were perfectly correlated with one another (e.g. if F0 node 2 was active, indexical node 2 was active). This simulates their correspondence in natural speech.
7. Auditory units representation (2 levels): A) Localist units were used to represent the auditory information, such that a single auditory unit in each bank was activated. B) Tuning curves were used in the auditory banks, with multiple units active on each trial. An array of units was activated, with a peak of activation at the chosen target VOT node (from Parameter 3). The peak unit activation was set to 1,

while the remaining units' activations were determined by a Gaussian with a standard deviation of 0.3. During testing the same protocol was used to set auditory activation.

8. Decay in the learning rule (2 levels): We considered two variants of Hebbian learning, both of which varied in the way that weight decay was implemented. A) All weights decayed (Equation 1). B) Decay was based on anticorrelations. Connection strength decayed if the auditory unit was more active than the visual unit, according to the rule:

$$\Delta w_{ij} = \eta(a_i v_j(1 - w_{ij}) - (1 - v_j)a_i w_{ij}) \quad (2)$$

Across these seven factors, we examined all 1600 different possible combinations. 86% of models showed the behavioral pattern: greater match than mismatch activation in both RM and SST models, and greater mismatch SST activation than match RM activation. Of the remaining models, 6% failed to learn to discriminate the words in the RM model (primarily due to very small learning rates); and 8% showed greater RM match activation than SST mismatch activation. Among models that learned under the RM training regime, 91% showed the predicted pattern of results.

Many parameter settings did not have a major effect on model success. However, models with initial weights of 0 performed slightly better than those with initial random values (93% vs. 90%); varying VOTs in the RM improved performance (98% vs. 85%); and varying pitch in the SST impaired performance (99% vs. 84%). These effects are explored in more detail in the Online Supplement.

References

Andruski J. Tone clarity in mixed pitch/phonation-type tones. Journal of Phonetics. 2006; 34(3):388–404.

Aslin RN, Werker JF, Morgan JL. Innate phonetic boundaries revisited (L). The Journal of the Acoustical Society of America. 2002; 112(4):1257–1260. [PubMed: 12398431]

Atkins JE, Fiser J, Jacobs RA. Experience-dependent visual cue integration based on consistencies between visual and haptic percepts. Vision Research. 2001; 41(4):449–461. [PubMed: 11166048]

Ballem KD, Plunkett K. Phonological specificity in children at 1;2. Journal of Child Language. 2005; 32(1):159–173. [PubMed: 15779881]

Bortfeld H, Morgan JL. Is early word-form processing stress-full? How natural variability supports recognition. Cognitive Psychology. 2010; 60(4):241–266. [PubMed: 20159653]

Bourne LE, Restle F. Mathematical theory of concept identification. Psychological review. 1959; 66:278–296. [PubMed: 13803353]

Bradlow AR, Nygaard LC, Pisoni DB. Effects of talker, rate, and amplitude variation on recognition memory for spoken words. Perception & psychophysics. 1999; 61(2):206–219. [PubMed: 10089756]

Bush RR, Mosteller F. A model for stimulus generalization and discrimination. Psychological review. 1951; 58(6):413–423. [PubMed: 14900302]

Charles-Luce J, Luce PA. Similarity neighbourhoods of words in young children's lexicons. Journal of Child Language. 1990; 17(1):205–215. [PubMed: 2312642]

Charles-Luce J, Luce PA. An examination of similarity neighbourhoods in young children's receptive vocabularies. Journal of Child Language. 1995; 22(3):727–735. [PubMed: 8789521]

Coady JA, Aslin RN. Phonological neighbourhoods in the developing lexicon. Journal of child language. 2003; 30(2):441–469. [PubMed: 12846305]

Creel SC, Aslin RN, Tanenhaus MK. Heeding the voice of experience: The role of talker variation in lexical access. Cognition. 2008; 106(2):633–664. [PubMed: 17507006]

Dietrich C, Swingley D, Werker JF. Native language governs interpretation of salient speech sound differences at 18 months. Proceedings of the National Academy of Sciences. 2007; 104(41): 16027–16031.

Eimas PD, Siqueland ER, Jusczyk P, Vigorito J. Speech perception in infants. Science. 1971; 171(3968):303–306. [PubMed: 5538846]

Ernst MO, Banks MS. Humans integrate visual and haptic information in a statistically optimal fashion. Nature. 2002; 415(6870):429–433. [PubMed: 11807554]

Fennell CT, Waxman SR. What paradox? Referential cues allow for infant use of phonetic detail in word learning. Child Development. 2010; 81(5):1376–1383. [PubMed: 20840228]

Fennell CT, Werker JF. Early word learners' ability to access phonetic detail in well-known words. Language and speech. 2003; 46(2–3):245–264. [PubMed: 14748446]

Gliozzi V, Mayor J, Hu J-F, Plunkett K. Labels as features (not names) for infant categorization: A neurocomputational approach. Cognitive Science. 2009; 33(4):709–738. [PubMed: 21585482]

Goldinger SD. Echoes of echoes? an episodic theory of lexical access. Psychological review. 1998; 105(2):251–279. [PubMed: 9577239]

Gómez RL. Variability and detection of invariant structure. Psychological Science. 2002; 13(5):431–436. [PubMed: 12219809]

Gooden DR, Baddeley AD. Context-dependent memory in two natural environments: On land and underwater. British Journal of Psychology. 1975; 6(3):325–331.

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

Hirsh-Pasek, K.; Golinkoff, RM.; Hollich, G. An emergentist coalition model for word learning: Mapping words to objects is a property of the interaction of multiple cues. In: Golinkoff, RM., et al., editors. On becoming a word learner: A debate on lexical acquisition. Oxford, UK: Oxford University Press; 2000. p. 136-164.

Kruschke JK. Alcove: an exemplar-based connectionist model of category learning. Psychological review. 1992; 99(1):22–44. [PubMed: 1546117]

Kruschke JK. Toward a unified model of attention in associative learning. Journal of Mathematical Psychology. 2001; 45:812–863.

Kuhl PK, Williams KA, Lacerda F, Stevens KN, Lindblom B. Linguistic experience alters phonetic perception in infants by 6 months of age. Science. 1992; 255(5044):606–608. [PubMed: 1736364]

Lively SE, Logan JS, Pisoni DB. Training Japanese listeners to identify English /r/ and /l/ II: The role of phonetic environment and talker variability in learning new perceptual categories. Journal of the Acoustical Society of America. 1993; 94(3):1242–1255. [PubMed: 8408964]

MacWhinney, B. Competition and lexical categorization. In: Corrigan, R.; Eckman, F.; Noonan, M., editors. Linguistic categorization. New York: Benjamins; 1989. p. 195-242.Current Issues in Linguistic Theory 61.

Maye J, Werker JF, Gerken L. Infant sensitivity to distributional information can affect phonetic discrimination. Cognition. 2002; 82(3):B101–B111. [PubMed: 11747867]

McMurray, B.; Horst, J.; Toscano, J.; Samuelson, L. Towards an integration of connectionist learning and dynamical systems processing: case studies in speech and lexical development. In: Spencer, J.; Thomas, M.; McClelland, J., editors. Toward a Unified Theory of Development: Connectionism and Dynamic Systems Theory Re-Considered. London: Oxford University Press; 2009.

Metsala, JL.; Walley, AC. Spoken vocabulary growth and the segmental restructuring of lexical representations: Precursors to phonemic awareness and early reading ability. In: Metsala, JL.; Ehri, LC., editors. Word recognition in beginning literacy. Mahwah, NJ: Lawrence Erlbaum Associates; 1998. p. 89-120.

Nosofsky RM. Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology, General. 1986; 115(1):39–61. [PubMed: 2937873]

Oakes LM, Coppage DJ, Dingel A. By land or by sea: The role of perceptual similarity in infants' categorization of animals. Developmental Psychology. 1997; 33(3):396–407. [PubMed: 9149919]

Pater J, Stager CL, Werker JF. The lexical acquisition of phonological contrasts. Language. 2004; 80:361–379.

Peterson GE, Barney HL. Control methods used in a study of the vowels. Journal of the Acoustical Society of America. 1952; 24:175–184.

Pierrehumbert, JB. Exemplar dynamics: Word frequency, lenition and contrast. In: Bybee, J.; Hopper, P., editors. Frequency and the emergence of linguistic structure. Amsterdam: John Benjamins; 2001. p. 137-157.

Pitt MA, Kim W, Navarro DJ, Myung JI. Global model analysis by parameter space partitioning. Psychological Review. 2006; 113(1):57–83. [PubMed: 16478301]

Quinn PC, Eimas PD, Rosenkrantz SL. Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants. Perception. 1993; 22(4):463–475. [PubMed: 8378134]

Ranbom L, Connine CM. Lexical representation of phonological variation. Journal of Memory and Language. 2007; 57:273–298.

Regier T. The emergence of words: Attentional learning in form and meaning. Cognitive Science. 2005; 29:819–865. [PubMed: 21702796]

Restle F. A theory of discrimination learning. Psychological Review. 1955; 62(1):11–19. [PubMed: 14357523]

Rost GC, McMurray B. Speaker variability augments phonological processing in early word learning. Developmental Science. 2009; 12(2):339–349. [PubMed: 19143806]

Rost GC, McMurray B. Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning. Infancy. 2010; 15(6):608–635.

Rumelhart DE, Zipser D. Feature discovery by competitive learning. Cognitive Science. 1985; 9(1): 75–112.

Samuelson LK, Schutte AR, Horst JS. The dynamic nature of knowledge: Insights from a dynamic field model of children's novel noun generalization. Cognition. 2009; 110(3):322–345. [PubMed: 19131050]

Schafer G, Mareschal D. Modeling infant speech sound discrimination using simple associative networks. Infancy. 2001; 2(1):7–28.

Singh L, Morgan JL, White KS. Preference and processing: The role of speech affect in early spoken word recognition. Journal of Memory and Language. 2004; 51(2):173–189.

Singh L, White KS, Morgan JL. Building a word-form lexicon in the face of variable input: Influences of pitch and amplitude on early spoken word recognition. Language Learning and Development. 2008; 4(2):157–178.

Sirois S, Mareschal D. An interacting systems model of infant habituation. Journal of Cognitive Neuroscience. 2004; 16(8):1352–1362. [PubMed: 15509383]

Stager CL, Werker JF. Infants listen for more phonetic detail in speech perception than in word-learning tasks. Nature. 1997; 388(6640):381–382. [PubMed: 9237755]

Summerfield Q, Haggard M. On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. Journal of the Acoustical Society of America. 1977; 62(2): 435–448. [PubMed: 886081]

Swingley D, Aslin RN. Lexical neighborhoods and the word-form representations of 14-month-olds. Psychological Science. 2002; 13(5):480–484. [PubMed: 12219818]

Swingley D, Aslin RN. Lexical competition in young children's word learning. Cognitive Psychology. 2007; 54(2):99–132. [PubMed: 17054932]

Thiessen ED. The effect of distributional information on children's use of phonemic contrasts. Journal of Memory and Language. 2007; 56(1):16–34.

Thiessen, ED. Variability in lexical form facilitates children’s generalization of phonemic contrasts. Paper presented at the the 17th Biennial International Conference on Infant Studies; Baltimore, MD. 2010.

Toscano JC, McMurray B. Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. Cognitive Science. 2010; 34:434–464. [PubMed: 21339861]

Walley AC, Metsala JL, Garlock VM. Spoken vocabulary growth: Its role in the development of phoneme awareness and early reading ability. Reading and Writing. 2003; 16(1):5–20.

Wasserman EA, Miller RR. What's elementary about associative learning? Annual Review of Psychology. 1997; 48:573–607.

Werker JF, Cohen LB, Lloyd VL, Casasola M, Stager CL. Acquisition of word-object associations by 14-month-old infants. Developmental Psychology. 1998; 34(6):1289–1309. [PubMed: 9823513]

Werker JF, Curtin S. PRIMIR: A developmental framework of infant speech processing. Language Learning and Development. 2005; 1(2):197–234.

Werker JF, Fennell CT, Corcoran KM, Stager CL. Infants' ability to learn phonetically similar words: Effects of age and vocabulary size. Infancy. 2002; 3(1):1–30.

Werker JF, Tees RC. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. Infant Behavior and Development. 1984; 7:49–63.

Woodward AL, Hoyne KL. Infants' learning about words and sounds in relation to objects. Child Development. 1999; 70(1):65–77. [PubMed: 10191515]

Yeung HH, Werker JF. Learning words' sounds before learning how words sound: 9-month olds use distinct objects as cues to categorize speech information. Cognition. 2009; 113(2):234–243. [PubMed: 19765698]

Yoshida KA, Fennell CT, Swingley D, Werker JF. Fourteen-month-old infants learn similar-sounding words. Developmental Science. 2009; 12(3):412–418. [PubMed: 19371365]

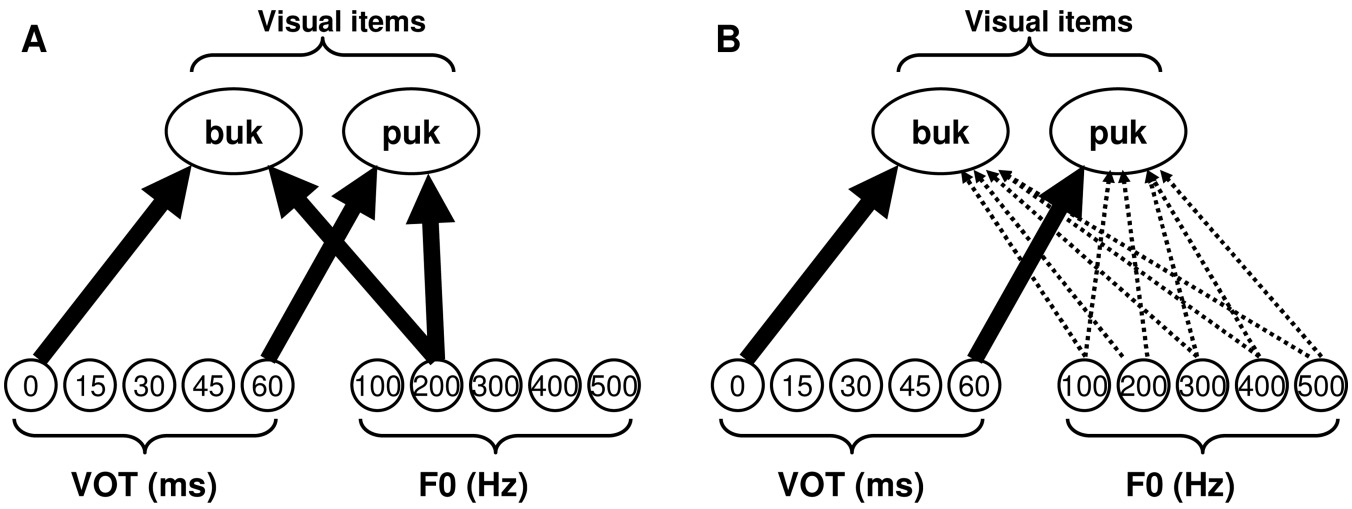

Figure 1.

A) Results of learning with non-varying pitch values (e.g. Stager & Werker, 1997). VOT values become strongly predictive of visual target identity, while trained F0 values become strongly associated with both visual targets. Note: this figure only presents F0 data for simplicity; indexical values would be expected to mirror F0 data. B) Results of learning with variable F0 values (e.g. Rost & McMurray, 2009). VOT values are strongly predictive of visual target identity, while F0 values are very weakly and diffusely related to visual targets.

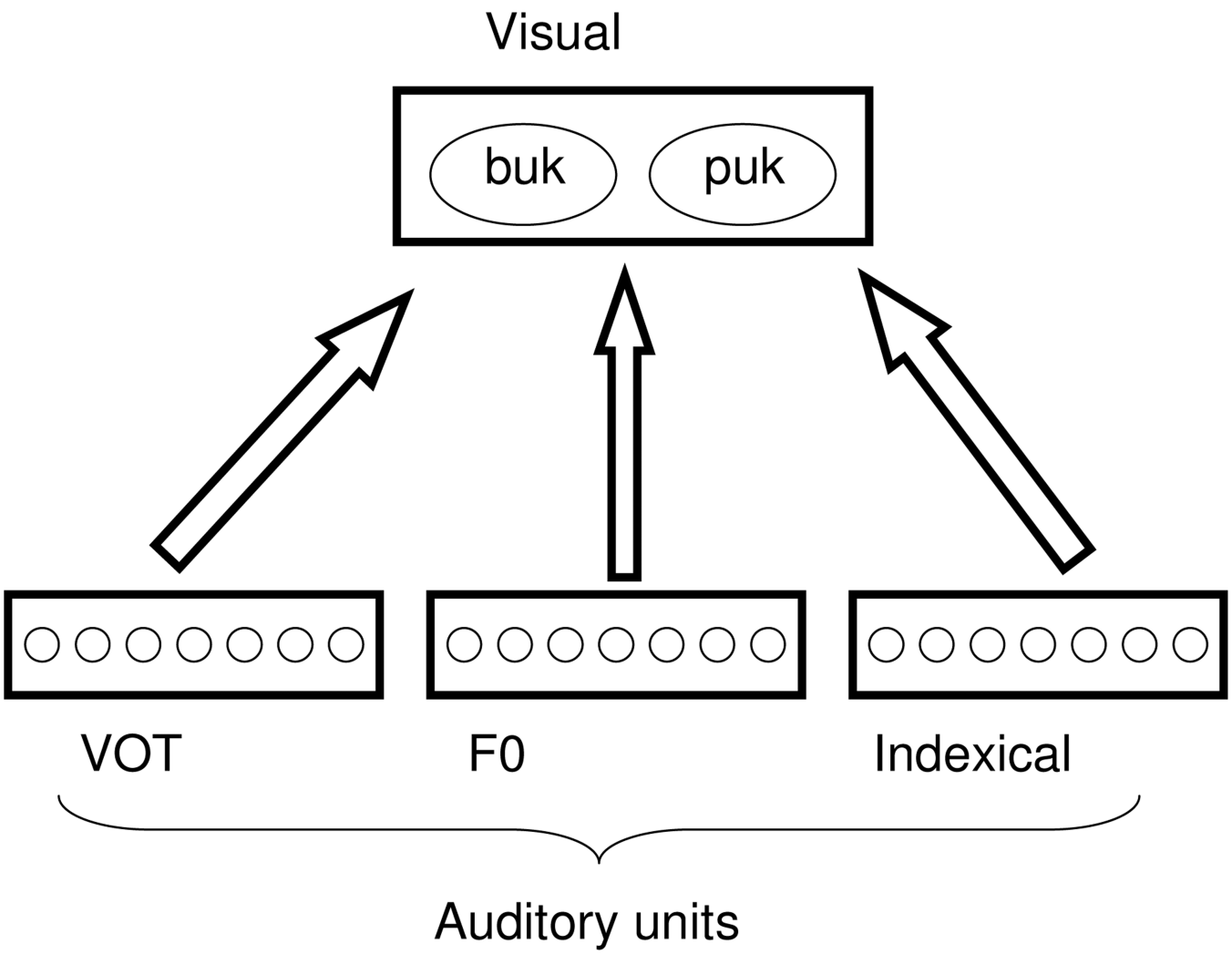

Figure 2.
Schematic of the SST and RM models.

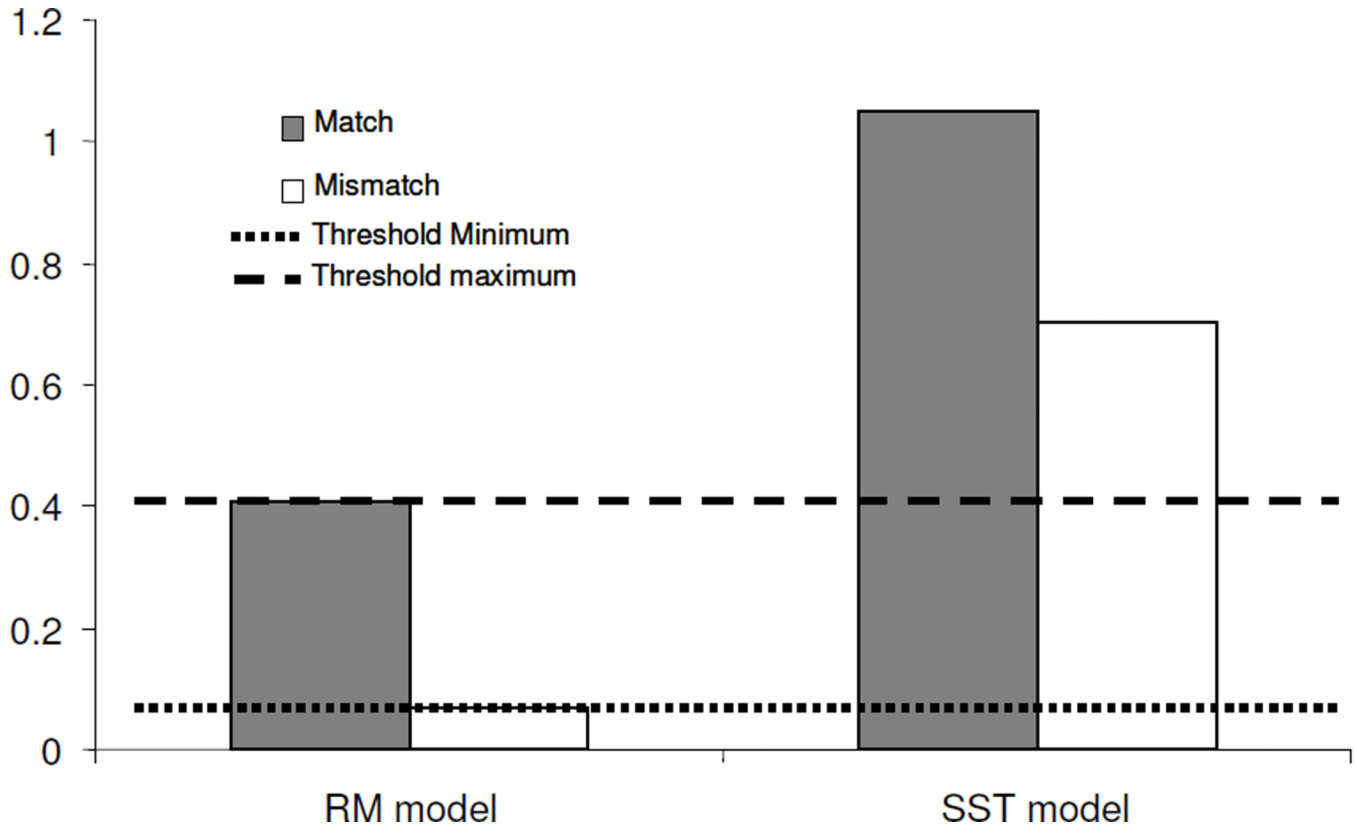

Figure 3.
Expectation values generated from testing the RM and SST models. The recognition threshold must fall between the activation of the match and mismatch items in the RM model; both match and mismatch activation in the SST will exceed any threshold between these values.

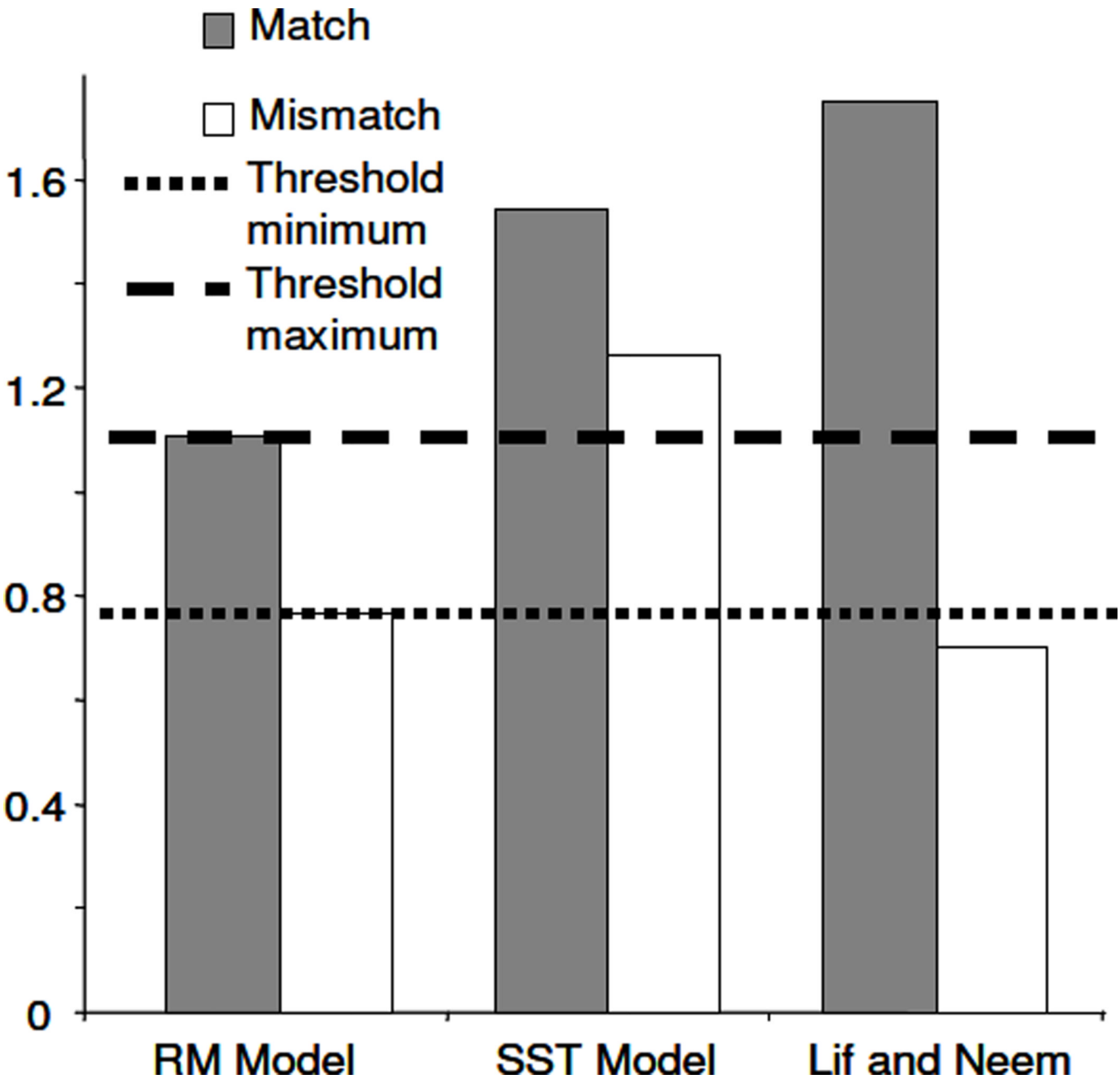

Figure 4.

Expectation values generated from the lif and neem model. When trained on buk and puk with the SST methodology, both units are activated above the recognition threshold from the RM buk and puk simulations. However, for lif and neem, only the matching unit exceeds the recognition threshold, while the mismatching unit falls below the threshold.

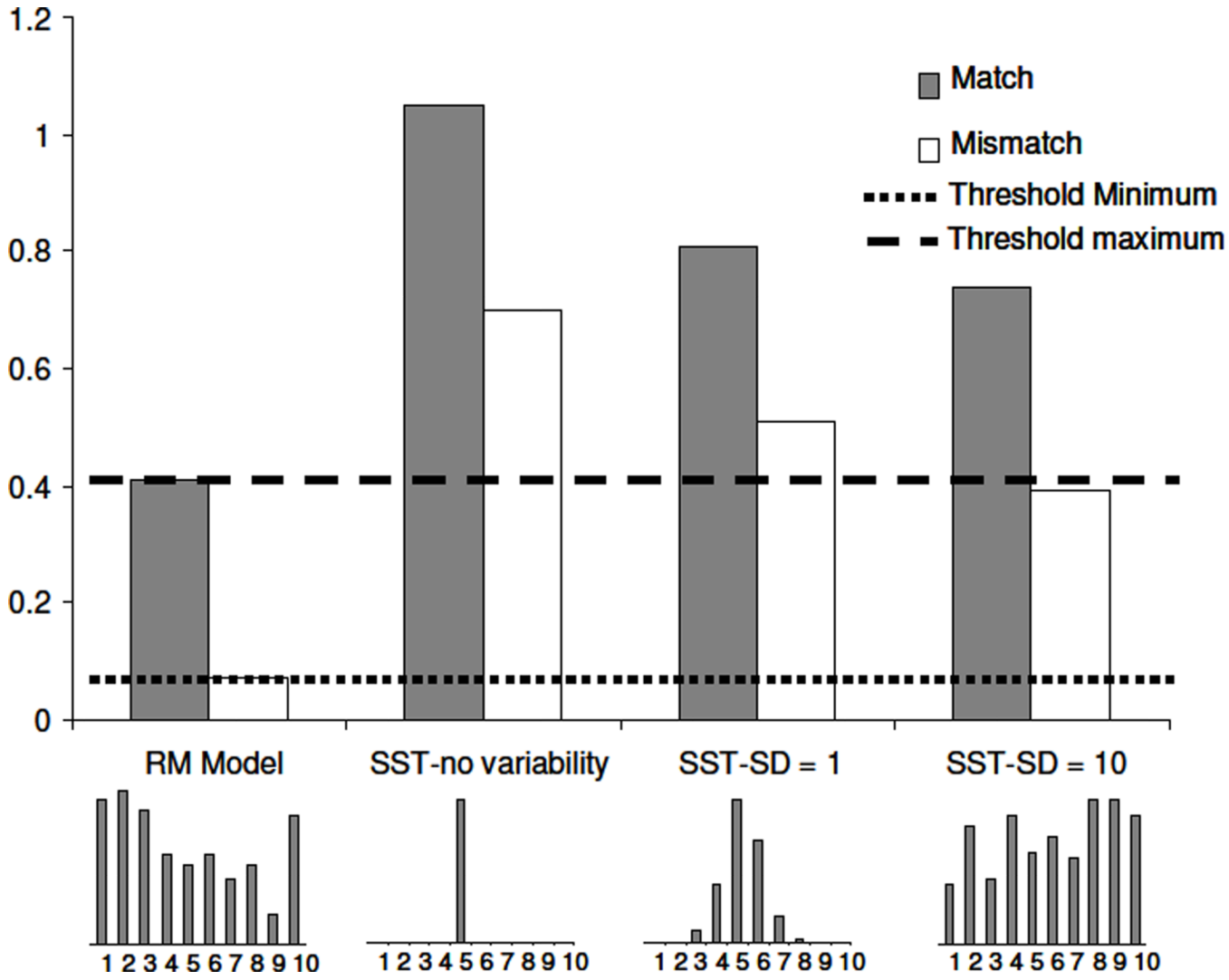

Figure 5.

Expectation values generated by the RM model, compared against various iterations of the SST model with different variability in F0. The lower panel displays histograms of the distribution of F0 values in the different conditions. In the SST model with no F0 variability and with small F0 variability, both the match and mismatch expectation values exceed the match value for the RM model. When a large amount of F0 variability was added, the SST mismatch activation began to enter thethreshold range from the RM simulation.

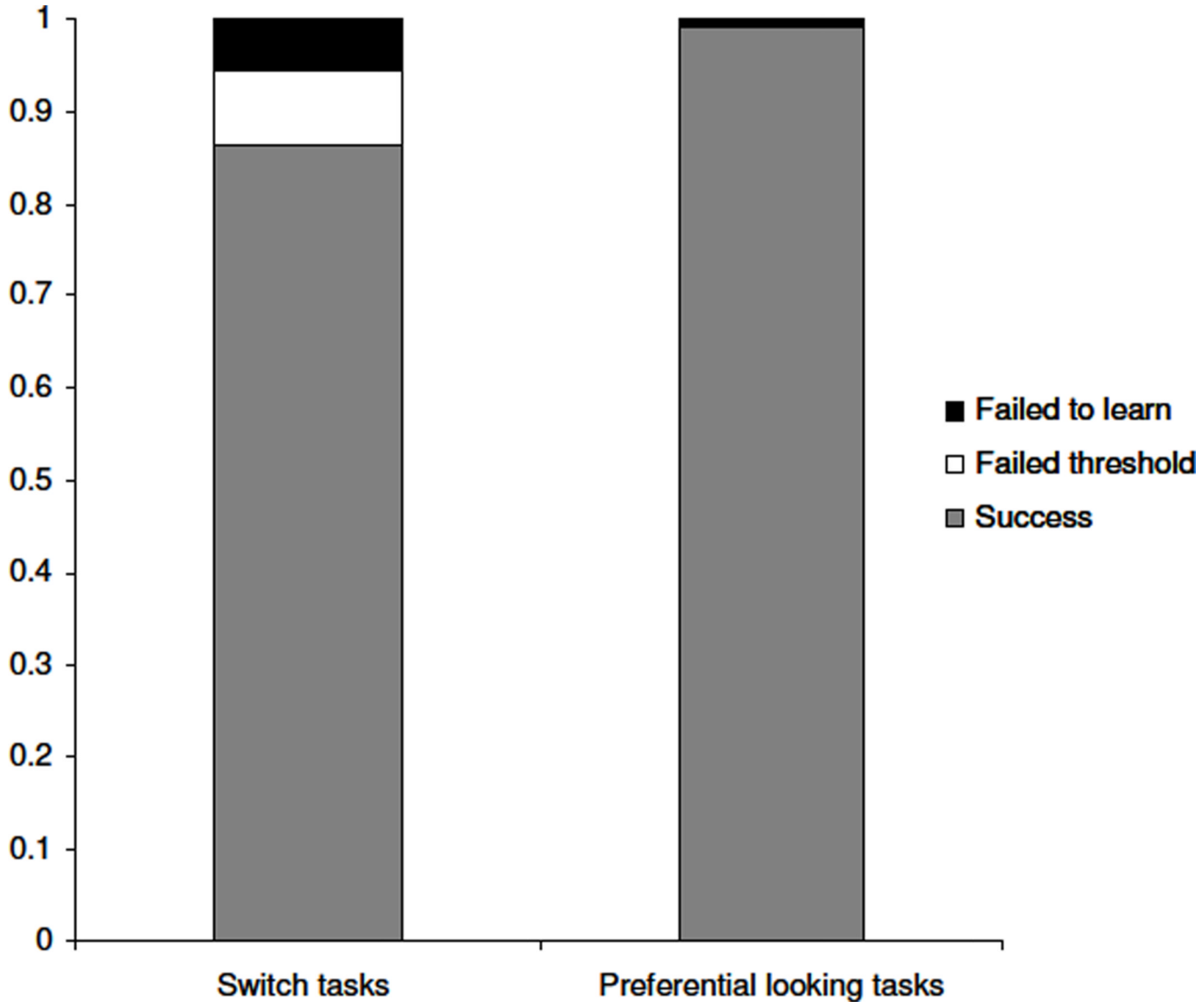

Figure 6.

Results from parameter manipulations. The vast majority (86%) of models showed the empirical pattern of results, in which match was greater than mismatch activation for both RM and SST models; and both match and mismatch activation for the SST model were greater than match activation for the RM model. Of models that failed, 6% failed because they were incapable of learning the items at all (often due to very small learning rates); 8% failed because the threshold for learning in the RM model also predicted learning in the SST model.

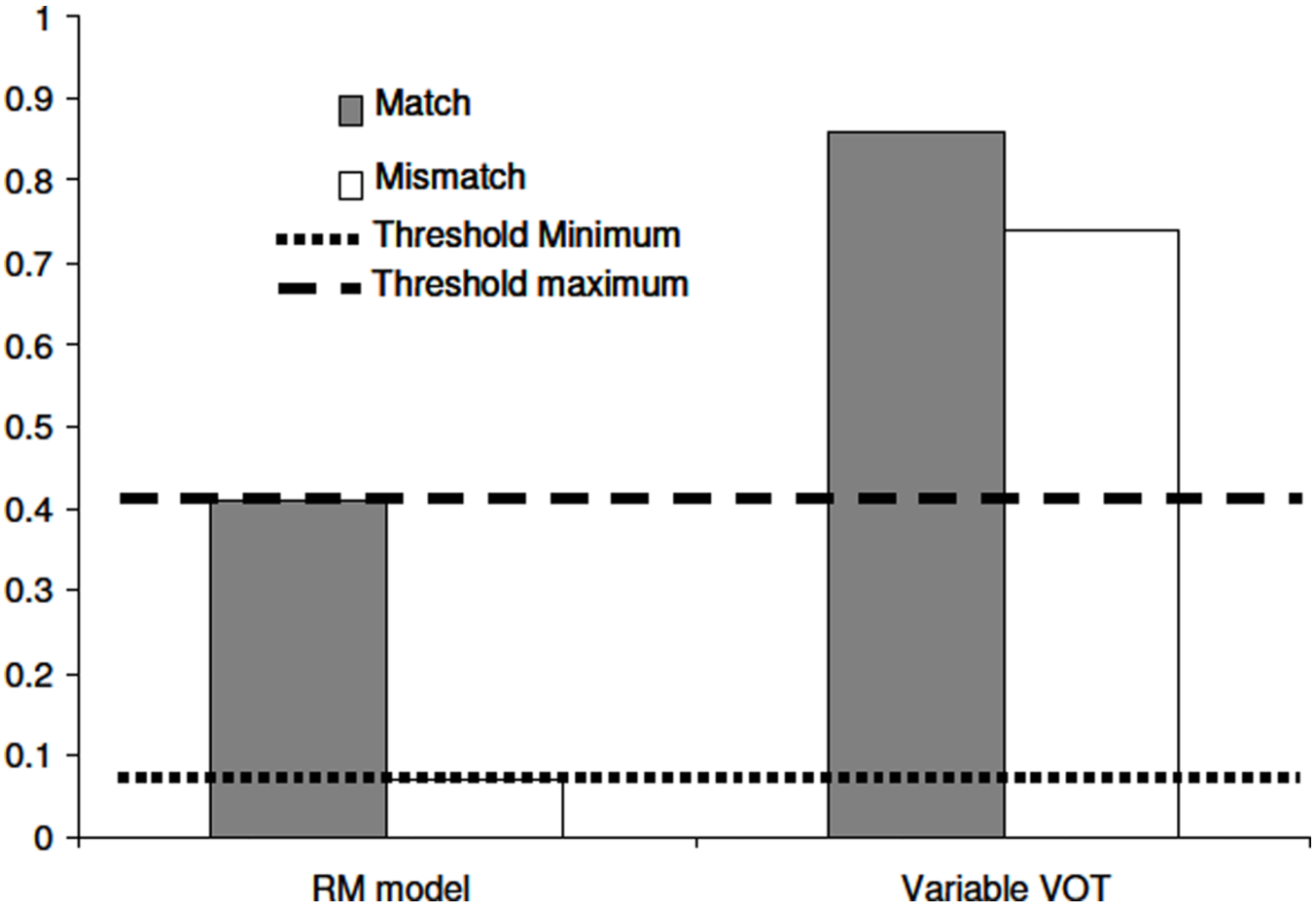

Figure 7.
Expectation values generated from simulations of Rost and McMurray (2010) compared to expectation values from the RM model. In the Rost and McMurray (2010) simulations, F0 and indexical values do not vary while VOT around prototypical values; both match and mismatch activations in this simulation will exceed the recognition threshold for the RM model.

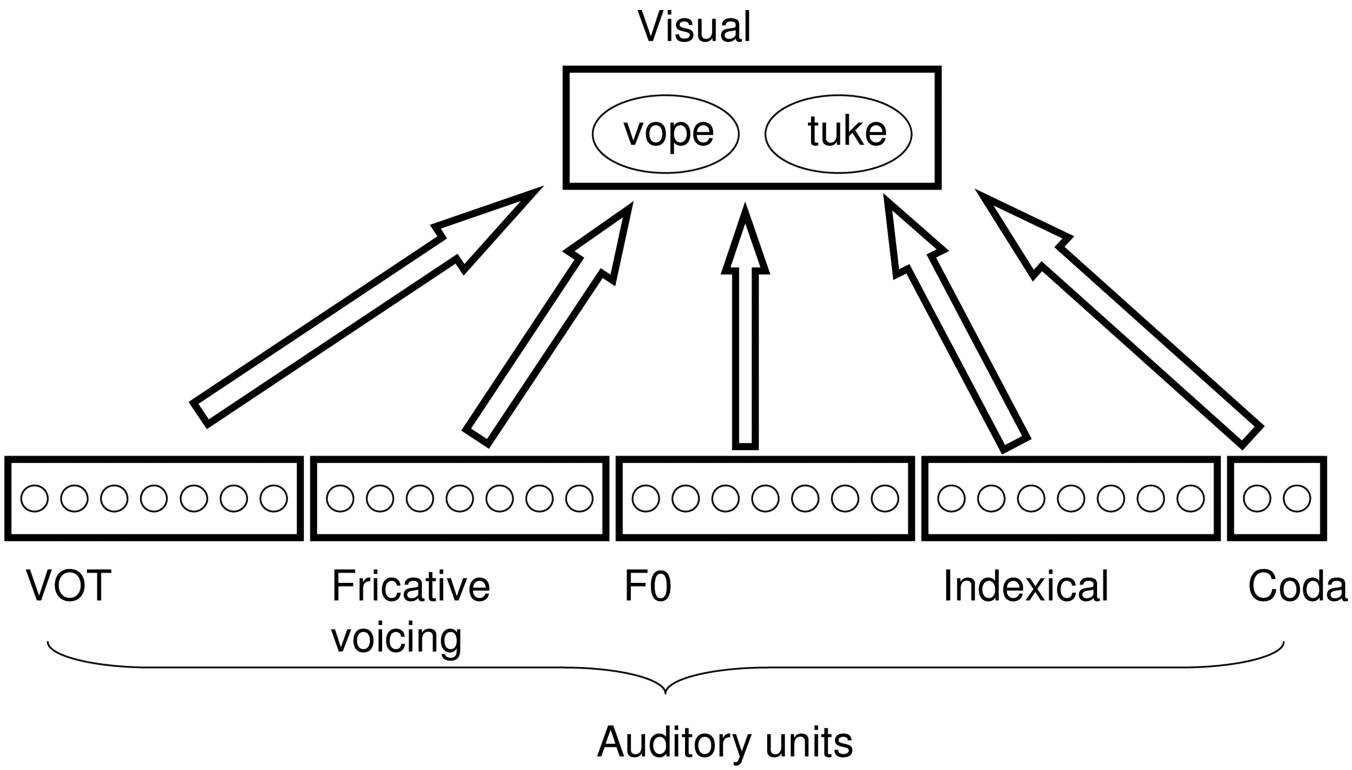

Figure 8.
Schematic of the model used to simulate Ballem and Plunkett (2005).

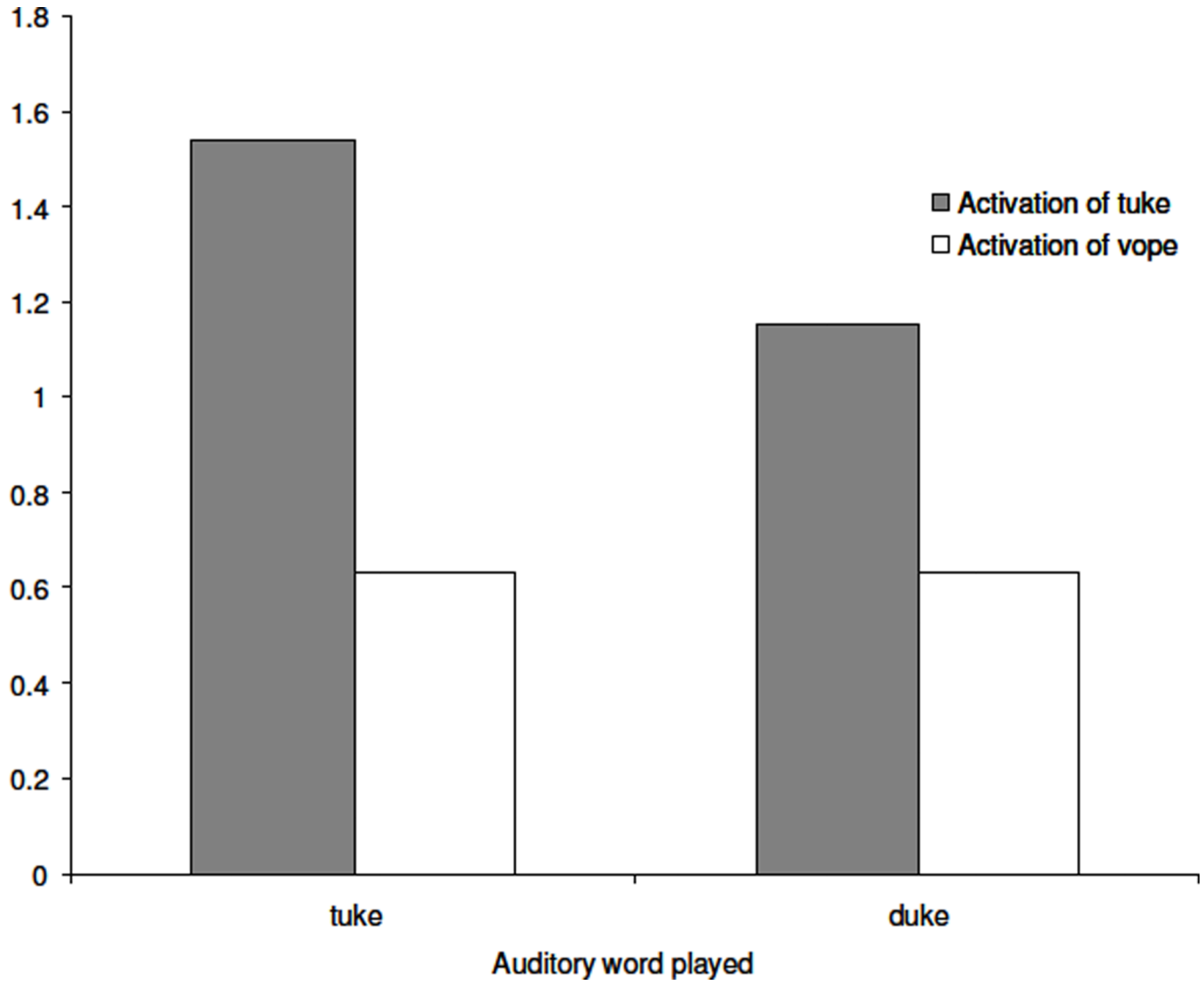

Figure 9.
Results of simulations of Ballem and Plunkett (2005). When presenting a correctly pronounced target, activation of the target exceeds that of the competitor: the model correctly activates the correct target. When a mispronunciation is presented, this difference decreased, representing children's decreased looking preference for mispronounced targets.