Modeling Sources of Uncertainty in Spoken Word Learning

Matthias Hofer (mhofer@mit.edu), Roger Levy (rplevy@mit.edu)
Department of Brain and Cognitive Sciences, 43 Vassar St
Cambridge, MA 02139 USA

Abstract

In order to successfully learn the meaning of novel words such as *bin* or *pin*, language learners must not only be able to perceive relevant differences in the speech signal but also learn mappings from words to referents. Prior work in native (Stager & Werker, 1997) and second (Pajak, Creel, & Levy, 2016) language acquisition has found that the ability to perceptually discriminate between words does not guarantee successful word learning. Put differently, learners fail to utilize knowledge that they can otherwise use in speech perception. To explore possible mechanisms accounting for this phenomenon, we present a probabilistic model capable of inferring a word's phonetic form and of integrating the result with prior label/referent representations in memory. By analyzing the reasons for successes and failures when fitting different versions of the model to experimental results from Pajak et al. (2016), we argue that a mechanism for spoken word learning needs to incorporate both perceptual uncertainty as well as additional, task-related sources of uncertainty to successfully account for the data.

Keywords: word learning, rational model, probabilistic inference, phonological similarity, speech representations

Introduction

From the perspective of a learner of English, successfully learning the meaning of novel words such as *bin* or *pin* requires the ability to perceptually discriminate between similar-sounding words. Creating distinct, non-overlapping representations of the input is necessary because the words need to be mapped onto different classes of referents. This requires perceptual sensitivities to the phonological contrasts critical for the discrimination (Pater, Stager, & Werker, 2004). In the case of *bin* and *pin* this contrast is along the voicing dimension (phonemes *b* and *p* differ in voice onset time). Studies in infant native-language (L1) acquisition have shown that while these perceptual abilities are present in 14 month old infants, they do not guarantee successful word learning (Stager & Werker, 1997; Pater et al., 2004). In a series of experiments conducted by Stager and Werker (1997), infants that were able to perceptually discriminate between two similar-sounding words, such as *dih* and *bih*, failed to utilize this knowledge during word learning (experiment 4). When first habituated to label/object pairings, infants did not reliably detect when the assignment of words to objects switched (experiment 1) and they failed to notice mispronunciations when an object that had before been introduced as *dih* was later referred to as *bih* (experiment 2). The authors interpreted their findings as infants being unable to attend to fine phonetic detail during word learning and argued that it constitutes a feature of linguistic development.

The same pattern of results has more recently been demonstrated for second-language (L2) learners. A study by Pajak et al. (2016) compared the performance of subjects of two different linguistic backgrounds in a perceptual discrimination and a word learning task. To create a situation paralleling that of L1 acquisition in infants, the researchers used a miniature language with word pairs at three levels of phonological similarity whose phonology, modeled after Polish, was novel and unfamiliar to all participants: Dissimilar words differed in multiple phonemes (e.g., *tala* / *kenna*); similar word pairs differed in one phoneme (e.g., *tala* / *taja*); and highly-similar words differed only along a single phonetic dimension, either in length (short vs. long, e.g., *tala* / *talla*) or place of articulation (alveolopalatal vs. retroflex, e.g., *gotɕa* / *gotʂa*). In order to examine the role of L1-specific differences in task performance, Pajak and Levy (2014) collected data from both Korean speakers, who are sensitive to length contrasts but not to the alveolopalatal vs. retroflex distinction, and from Mandarin speakers, who show the opposite pattern. To ensure that subjects were naive to the stimuli used in the experiment, the researchers tested separate groups of subjects on the perceptual discrimination and the word learning task.

Similarly to results from Stager and Werker (1997), the study found that the ability to perceptually discriminate similar-sounding words in the perceptual task did not successfully translate to the word learning task on a group level, nor did subjects' L1-specific perceptual advantages. Taken together, these findings suggest that the difficulty in learning similar-sounding words, especially during early lexical acquisition, is a general property of learning rather than a developmental stage (Pajak et al., 2016; Perfors & Dunbar, 2010). At present, little is known about the learning mechanisms that give rise to these difficulties. Here we seek to provide an account of such a mechanism in the form of a probabilistic model developed with the goal of reproducing the results from Pajak et al. (2016)'s original study. While we are not aware of any computational modeling work, the exists prior work addressing these issues on a conceptual level. The failure to utilize perceptual knowledge during word learning has previously been attributed to increased cognitive load (Werker, Fennell, Corcoran, & Stager, 2002; Stager & Werker, 1997). While discrimination only requires storage and comparison of perceptual input in phonological short-term memory, word learners must additionally attend to the referent stimulus and integrate label/referent information over the course of learning to infer probable associations. This lowers the resolution of auditory processing (Mattys & Palmer, 2015), which contributes to the failure of distinctly representing similar-sounding input. We will explore these verbal theories by analyzing which of the proposed components are necessary to account for the observed effects in the study by Pajak et al. (2016), which we will briefly describe in the next section.

Pajak et al. (2016)'s experiment

In a between-subjects design, ninety subjects, approximately equally split into speakers of Korean and Mandarin, participated in either a perceptual discrimination or a word learning task. Stimuli consisted of 16 bisyllabic consonant-vowel-consonant-vowel (CVCV) nonce words, split into similarity classes as described above.[1] Perceptual discrimination was tested in an ABX task. Subjects listened to three consecutive words, e.g., *talla*$_{[A]}$, *taja*$_{[B]}$, and *talla*$_{[X]}$, and had to decide which of the first two words sounded more similar to the last one. In the word learning task, each of the 16 labels was associated with a single referent in the form of a visual image and participants' goal was to learn which words were associated with which referents. The experiment consisted of four training blocks (each with 128 trials) and four interleaved testing blocks (each with 64 trials). In each trial, two pictures were presented side by side, corresponding to the referents for labels A and B. Similarly to the ABX task, subjects then heard a label X and had to decide which referent it belonged to. Error feedback was provided to the subjects during the training phase but not during the test phase. The stimulus triplets used in the discrimination task and in the test phase of the word learning task were identical, which made it possible to compare accuracy for triplets across the two tasks.

Computational Model

A computational analysis of spoken word learning must take into account the goal of the computation, the information available to the learner, and show how this information maps onto appropriate behavioral responses (Anderson, 1990). We suggest that learning novel words requires the learner to perform statistical inference on at least two distinct levels. While the ultimate goal of learning is to infer concepts, or label/referent associations, from a stream of observations, the learner must concurrently infer the label's phonetic form from the acoustic input, since it is not explicit in the speech signal. These two layers of inference give rise to a hierarchical probabilistic model, visually depicted in Figure 1, which we used to model spoken word learning and, using a variation of the generative process, model results from the perceptual discrimination tasks. Model behavior is influenced by three distinct factors: perceptual noise, which affects both discrimination and word learning when processing speech input, task-specific factors that lower the resolution of auditory representations of speech sounds during word learning (Mattys & Palmer, 2015), and overall memory capacity.

Formal characterization of the model

Each word and its corresponding referent define a concept, denoted c. To simplify our analysis, we assume that the referent stimulus is observed unambiguously. In our model, observing the referent stimulus is therefore identical to observing the concept directly, because of the one-to-one relation between referents and concepts. The a priori probability of choosing any concept is uniform. Corresponding labels are then sampled from the conditional probability $p(\mathbf{l}|c)$, whose probability mass is uniform across all possible labels in the language $\mathcal{L}$ and zero otherwise.

$$Pr(\mathbf{l}|c) = \begin{cases} \frac{1}{N} & \text{if } \mathbf{l} \in \mathcal{L} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Label $\mathbf{l}$ is a sequence of phonemes of the form $p_1p_2p_3p_4$ composed of pre-specified consonant and vowel primitives. Phonemes are represented mathematically as multivariate Gaussian distributions in one of two separate (phonetic) feature spaces, one for consonants and one for vowels. Following prior approaches to the representation of speech sounds (Richter, Feldman, Salgado, & Jansen, 2016; Bailey & Hahn, 2005), the feature dimensions of these phonetic spaces correspond to subsegmental features such as manner, place, length, and voicing, or to the first two vowel formants. Distributions are centered around a fixed category mean $\overline{\mu}[p_i]$, a vector of means indexed by the corresponding phoneme. Covariance matrices $\overline{\Sigma}_w[i]$, one shared across vowels and one across consonants, are indexed by i only (corresponding to whether the phoneme is of type C or V).

Conditioned on a choice for label $\mathbf{l}$, we can generate a sequence of phones $s_1s_2s_3s_4$, which can be seen as its discrete and noisy realizations of the label's phonemes:

$$Pr(s_i|p_i) = \mathcal{N}\big(\overline{\mu}[p_i], \overline{\Sigma}_w[i]\big) \tag{2}$$

$$\begin{aligned} Pr(\mathbf{s}|\mathbf{l}) &= Pr(s_1s_2s_3s_4|p_1p_2p_3p_4) \\ &= Pr(s_1|p_1)Pr(s_2|p_2)Pr(s_3|p_3)Pr(s_4|p_4) \end{aligned} \tag{3}$$

The covariance matrix that determines variability in the realization of speech sounds is specified through the following scalar-matrix-vector multiplication:

$$\overline{\Sigma}_{\text{task}}[i] = \alpha_{\text{task}}\mathbf{I}\overline{\mathbf{v}}[i]^{population} \tag{4}$$

The scalar α_{task} allows us to reflect task-specific sources of uncertainty (word learning vs. discrimination). We note that the parameter for word learning, α_w can be written as the product of a perceptual 'baseline' acuity parameter for the discrimination task times a constant factor: $\alpha_w = c\alpha_d$. Assuming pairwise independence across all feature dimensions, the covariance matrix is fully specified by its diagonal elements, encoded in the population-specific diagonal vector $\overline{\mathbf{v}}[i]^{population}$. Phonetic acuity along those feature dimensions is inversely proportional to variance: the higher phonetic acuity, the lower the variance. This allows us to model differences in L1 background (Korean vs. Mandarin) with respect to perceptual sensitivities along these feature dimensions. For example, for Korean speakers:

[1] The constraint that words always have four CVCV segments is a simplification for convenience. In principle, our model should be applicable to any set of phonemes and syllable structures. See Pajak et al. (2016) for the complete list of phonemes and stimuli used.

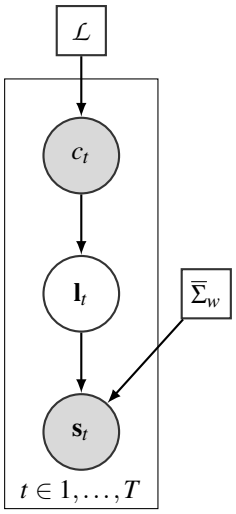


Figure 1: Graphical representation of the word learning model. Circles indicate random variables (variables shaded in gray are observed during learning); squares indicate fixed model parameters. To simplify our representation, the model does not include a referent node, which is deterministically generated by sampling from the concept.

$$\overline{\nu}[1]^{Korean} = \overline{\nu}[3]^{Korean} = \left(\tau_{F1_K}{}^{-1}\tau_{F2_K}{}^{-1}\right)^T$$
$$\overline{\nu}[2]^{Korean} = \overline{\nu}[4]^{Korean} = \left(\tau_{length_K}{}^{-1}\tau_{place_K}{}^{-1}\tau_{voicing_K}{}^{-1}\tau_{manner_K}{}^{-1}\right)^T$$

The vowel feature space consists of the first two formants F1 and F2. Consonant space consists of the dimensions voicing, place, manner, and length (Bailey & Hahn, 2005). All acuity parameters are set to 1 (corresponding to a unit Gaussian variance), except for τ_{length} and τ_{place}, which are population-specific free parameters in the model. As a simple approximation, means in $\overline{\mu}[p_i]$ are evenly spaced across perceptual space. Along each phonetic dimension, we defined a number of subsegmental features (e.g., 'voiced' and 'unvoiced' along the voicing dimension). Phonetic category means can then be written as the vectors composed of these features (the mean of phoneme *f*, for instance, is represented as [unvoiced, labial, fricative, short]). Although not fully accurate in its details, the coarse grained nature of this setup is sufficient with respect to the word pairs used in Pajak et al. (2016)'s experiment.[2]

Word learning model In word learning, subjects engage in consecutive training and test blocks. Each training trial t consist of an observed label/referent pair $\{\mathbf{s}_t, c_t\}$. For simplicity we assume that learners discard the negative, second exemplar presented to them and only learn from the positive pairing. The learner's goal is to infer probable associations between referents c and labels $\mathbf{l}$, in other words, compute the posterior probability over labels given the observed stimulus and referent $Pr(\mathbf{l}_t|\mathbf{s}_t, c_t)$ according to:

$$Pr(\mathbf{l}|\mathbf{s}, c, \overline{\Sigma}_w) = \frac{Pr(\mathbf{s}|\mathbf{l}, \overline{\Sigma}_w)Pr(\mathbf{l}|c)}{\sum_{\mathbf{l}} Pr(\mathbf{s}|\mathbf{l}, \overline{\Sigma}_w)Pr(\mathbf{l}|c)} \tag{5}$$

The output of this computation is then used as a prior for the next trial. To model the difficulty of integrating multiple memory traces over time, one simple approach is to assume an upper bound on the number of memory traces that can be integrated, denoted m_c. We formalized this intuition by discarding samples from trials $t \geq m_c$ (no further updating of probabilistic representations occurs).

After computing $Pr(\mathbf{l}|\mathbf{s}_1,...,\mathbf{s}_T, c_1,...,c_T, \overline{\Sigma}_w)$ for the training block, in the test phase, the learner compares two alternative tuples $\{c_A, \mathbf{s}_X\}$ and $\{c_B, \mathbf{s}_X\}$ to assess which referent is more probable under $\mathbf{s}_X$. This is achieved by computing $Pr(c_A|\mathbf{s}_X, \overline{\Sigma}_w)$ and $Pr(c_B|\mathbf{s}_X, \overline{\Sigma}_w)$ by integrating over $\mathbf{l}$:

$$Pr(c|\mathbf{s}, \overline{\Sigma}_w) = \frac{\sum_{\mathbf{l}} Pr(\mathbf{s}|\mathbf{l}, \overline{\Sigma}_w)Pr(\mathbf{l}|c)Pr(c)}{\sum_c \sum_{\mathbf{l}} Pr(\mathbf{s}|\mathbf{l}, \overline{\Sigma}_w)Pr(\mathbf{l}|c)Pr(c)} \tag{6}$$

Discrimination model In the discrimination task, subjects perceive a stimulus triplet $\{\mathbf{s}_A, \mathbf{s}_B, \mathbf{s}_X\}$ and decide whether X is more similar to A or to B. We hypothesize that subjects use the generative process outlined above to judge similarity, where they independently determine the likelihood that the stimuli were sampled from two alternative generative models (Tenenbaum & Griffiths, 2001), described in the following:

$$Pr(\mathbf{s}_A, \mathbf{s}_B, \mathbf{s}_X|\mathbf{l}_1, \mathbf{l}_2) = \sum_{\mathbf{l}_1}\Big[Pr(\mathbf{s}_A|\mathbf{l}_1)Pr(\mathbf{s}_X|\mathbf{l}_1)Pr(\mathbf{l}_1)\Big]\sum_{\mathbf{l}_2}\Big[Pr(\mathbf{s}_B|\mathbf{l}_2)Pr(\mathbf{l}_2)\Big] \tag{7}$$

$$Pr(\mathbf{s}_A, \mathbf{s}_B, \mathbf{s}_X|\mathbf{l}_1, \mathbf{l}_2) = \sum_{\mathbf{l}_1}\Big[Pr(\mathbf{s}_A|\mathbf{l}_1)Pr(\mathbf{l}_1)\Big]\sum_{\mathbf{l}_2}\Big[Pr(\mathbf{s}_B|\mathbf{l}_2)Pr(\mathbf{s}_X|\mathbf{l}_2)Pr(\mathbf{l}_2)\Big] \tag{8}$$

The likelihood $Pr(\mathbf{s}|\mathbf{l})$ is the same as in Equation 3 but with covariance matrix $\overline{\Sigma}_d[i]$ specific to the perceptual discrimination task.

Response probability Both experimental paradigms use a two-alternative forced choice task (2-AFC) to assess subjects' knowledge. Subject either compare two posterior probabilities over concepts given labels (word learning task) or the likelihoods that the stimulus triple was generated by one of two alternative generative models (discrimination task). In

[2] In particular, the distance between dissimilar phonemes in feature space is large because their means differ in multiple units across multiple dimensions. The distance between highly-similar phonemes on the other hand is small since they are only one unit apart along a single dimension.

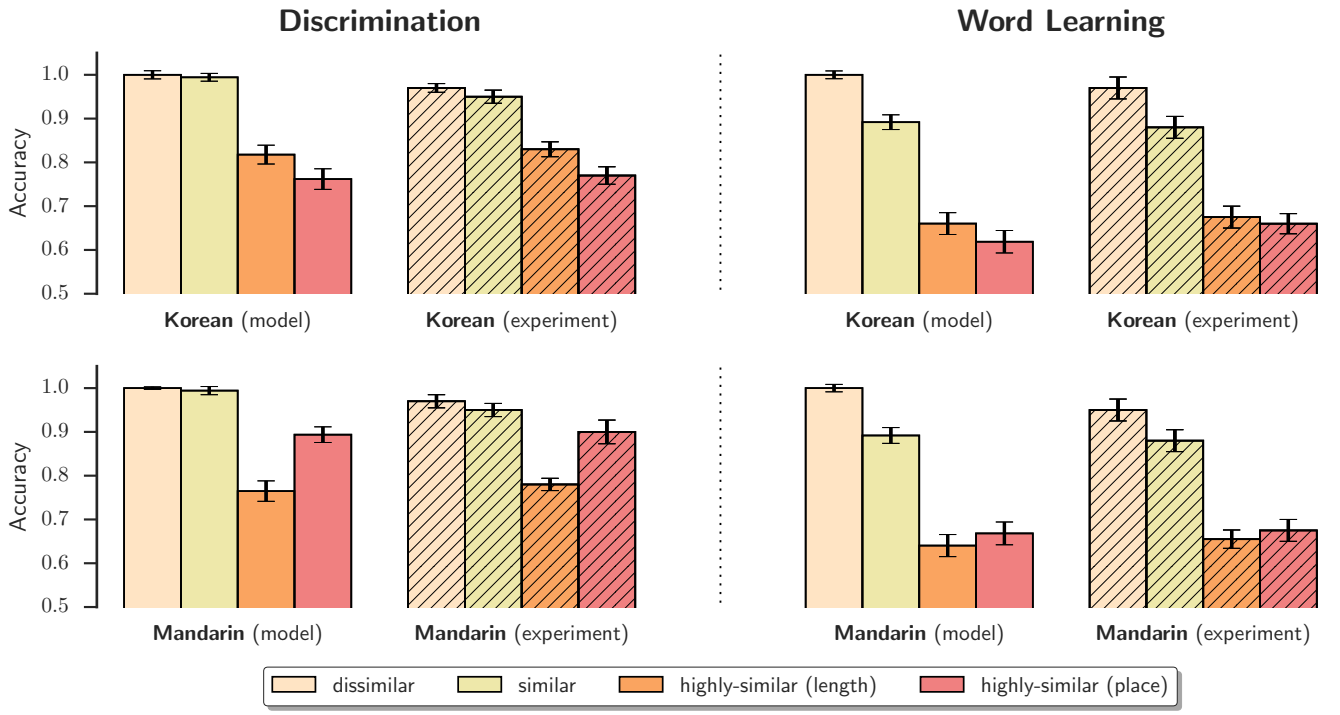


Figure 2: Comparison of the M_+A_+ model to experimental results for the L1 *Korean* population (top) and for the L1 *Mandarin* population (bottom). Error bars are standard errors. Accuracy scores are percentage correct in the discrimination task and during the test phase of the word learning task.

both cases, choices are modeled using a Bernoulli distribution and the probability of choosing one alternative over the other is computed using Luce's choice rule (Luce, 1959). Response parameter β controls the stochasticity of responses.

Results

We fitted the model to aggregate subject data from Pajak et al. (2016)'s word learning and perceptual discrimination tasks. Free model parameters included task-specific phonetic acuity for word learning (α_w) and for discrimination (α_d), four population-specific acuity parameters (τ_{length_K}, τ_{length_M}, τ_{place_K}, τ_{place_M}), the response parameter (β), as well as the memory capacity parameter (m_c).

Table 1: Results from fitting four versions of our model to the experimental data. Fits are quantified using the product of RMSEs to the word learning and discrimination data across the two speaker populations.

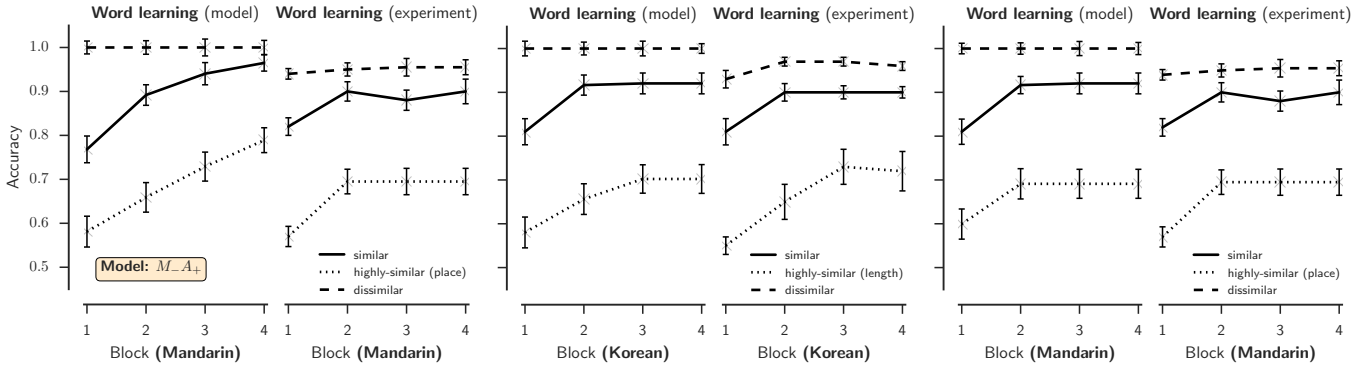
Model	n.p.	w.l.	w.l.b.	disc.	all
M_-A_-	6	0.089	0.138	0.179	0.407
M_-A_+	7	0.029	0.140	0.017	0.186
M_+A_-	7	0.076	0.098	0.172	0.347
M_+A_+	8	0.025	0.099	0.018	0.142

To assess which model components are necessary to account for the experimental data, we fitted four alternative versions of the model that were composed out of two binary factors: the presence ($+$) or absence ($-$) of memory constraints m_c (M) and the presence of task-specific, non-perceptual uncertainty in the form of separate ($+$) or shared ($-$) perceptual acuity parameters across tasks (A), where in the case of shared parameters: $\alpha_w = \alpha_d$. We also considered separate response rule parameters β for word learning and discrimination but found that the improvements were only minimal.

The models were fitted to the data by minimizing the product of six separate error terms. For each group of L1 speakers (Mandarin vs. Korean), we calculated the root mean squared error (RMSE) between model predictions and experimental results, resulting in three separate error terms for (i) overall discrimination accuracy across trial types [disc.], (ii) overall word learning accuracy across trial types [w.l.], and (iii) word learning accuracy across blocks and trial types [w.l.b.]. For each of the four models, Table 1 shows the RMSE for these three scores (averaged across Korean and Mandarin speakers) and their sums [all]. Column [n.p.] indicates the fitted model's number of free parameters.

The necessity of separate acuity parameters

Pajak et al. (2016)'s main finding was that the difference in accuracy between tasks was mediated by similarity. In other words, performance takes a greater hit from increased perceptual similarity in the word learning task compared to discrimination. The study also found that L1-specific perceptual



(a) M_-A_+ model (*Mandarin* speakers) (b) M_+A_+ model (*Korean* speakers) (c) M_+A_+ model (*Mandarin* speakers)

Figure 3: Figure (a) shows how the M_-A_+ model fails to account for the time course of learning. Figure (b) and (c) show comparisons of the M_+A_+ model to L1 *Korean* speakers and L1 *Mandarin* speakers. Error bars indicate standard errors.

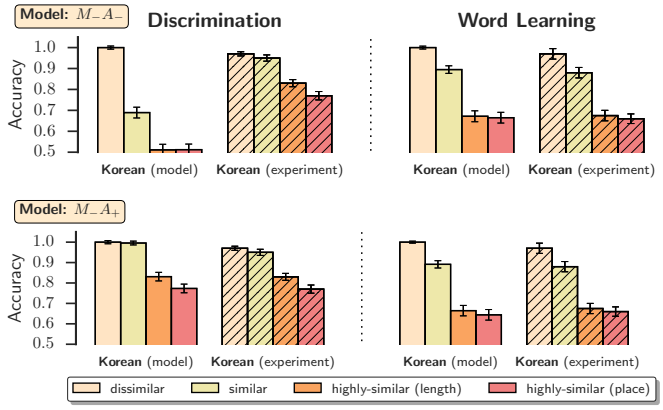


Figure 4: Results for the M_-A_- model (top) and for the M_-A_+ model (bottom; both *Korean* only; results are qualitatively similar for Mandarin speakers)

advantages that are manifest in the discrimination task cannot be utilized in word learning. Model variants M_-A_- and M_+A_-, missing the additional acuity parameter α_w, were not able to account for these observations. Figure 4 (top) illustrates this point by showing overall results for Korean speakers. One key observation for the M_-A_- model is that, when sharing a single perceptual acuity parameter between the two tasks, the original pattern of findings reverses and the discrimination model performs worse than the word learning model.[3] The reasons for this are twofold. All other things being equal, there is more uncertainty in the generative process for discrimination (see Equation 7 and 8) than there is in word learning. In discrimination, subjects need to infer the phonetic form of three auditory stimuli, compared to a single stimulus in word learning. Moreover, the word learning model can profit from additional information in the form of label/referent representations, gradually sharpening over the course of learning. The fact that perceptual uncertainty alone (in the form of a shared acuity parameter across models) cannot account for the superiority of discrimination performance over word learning suggests that word learning is influenced by additional sources of uncertainty. The bottom of Figure 4, which depicts results for the M_-A_- model, illustrates that adding an additional phonological acuity parameter that is specific to word learning is sufficient to account for both discrimination as well as word learning results.

Accounting for the time course of word learning

In Pajak et al. (2016), word learning performance only improved over a certain number of trials, resulting in learning curves to asymptote after roughly the second learning block. Reproducing this pattern while simultaneously accounting for the results presented above was an important aspect of our modeling efforts. While model M_-A_+ provides an almost ideal fit to aggregate results from discrimination and word learning (see Figure 4), underlining the importance of incorporating acuity factor A into the model, it fails to capture the time course of learning (see Figure 3a).

The only model that fit the entire range of empirical findings was the M_+A_+ model. Figure 2 shows that the model is a good fit, both qualitatively as well as quantitatively, to aggregate accuracy scores for both Korean and Mandarin L1 speakers. In particular, simulated data successfully reproduce the lack of L1-specific advantages in word learning compared to discrimination. Figures 3b and 3c show that the added memory constraint allows the model to better account for the shape of the learning curve. Models without this component are not able to reproduce this pattern. Table 1 further shows that adding such memory constraints also slightly improves fits to the aggregate word learning data [w.l.] compared to the same model where they are absent.

[3]Depending on how model fits are quantified it may also be possible to fit the discrimination results very well but overestimate accuracy on the word learning task. Critically, however, discrimination accuracy will always be lower than word learning accuracy in the M_-A_- model.

Discussion

Recent work by Pajak et al. (2016) suggests that the difficulty of learning label/referent associations for similar-sounding words is a general feature of learning rather than a developmental stage unique to infancy. In working towards a computational theory that could account for this phenomenon, we developed a probabilistic model capable of learning label/referent pairs while at the same time inferring the label's phonetic form. We fitted and compared four versions of the model to data from Pajak et al. (2016), contrasting different factors that are thought to influence performance.

We found that, besides structural differences in the way the generative models for word learning and for discrimination are set up, a single multiplicative factor operating on perceptual uncertainty was sufficient to account for the major differences between perceptual discrimination and word learning. A second additional factor, representing long-term memory constraints, was only necessary to account for the time course of learning.

Task-related sources of uncertainty

Conceptually, the acuity factor combines sources of uncertainty unique to the word learning task, such as attention to the referent stimulus and the encoding of label/referent exemplars over the course of learning. An interpretation broadly consistent with our model and with previous work (Stager & Werker, 1997; Mattys & Palmer, 2015) is that, although originating from post-perceptual sources, the locus of this added uncertainty is perception itself, operating through lowering attention to phonetic detail. On this view, the model's discrimination acuity parameter can be interpreted as representing various sources of perceptual uncertainty, ranging from the transduction of the speech signal at the periphery to phoneme recognition. Word learning-specific sources of uncertainty can be interpreted as a multiplicative factor on perceptual uncertainty, which, multiplied together, constitute the model's word learning acuity parameter. This added factor also accounts for the finding that L1-specific perceptual advantages cannot be utilized in word learning. The overlap of highly-similar word pairs in perceptual feature space is so large that potential advantages along the length and place feature dimensions are washed out.

Another important insight comes from models that lack this separate acuity parameter, which suggest that the discrimination is actually harder than word learning. This makes sense when considering that the generative model for the discrimination task must infer the phonetic form of three stimuli instead of a single stimulus. As a consequence, perceptual uncertainty affects the discrimination task more severely. In the absence of other factors that independently operate on the generative model for word learning, this leads to relative performance benefits in the word learning task.

Memory constraints

While distinguishing between two major sources of uncertainty might alone be sufficient to account for time-averaged results, it is not enough to account for the time course of learning in Pajak et al. (2016), which showed that learning stagnates after the second training block. The fact that these performance deficits are specific to word learning suggests that they are due to memory-related processes. We found that incorporating capacity constraints in the form of an upper bound on memory was necessary to fully account for the observed effects.

Conclusion

Our model is a first step in addressing the question of what are the factors that make the learning of similar-sounding words hard. In particular, the model is consistent with the original explanation given by Stager and Werker (1997); Werker et al. (2002). According to this view, word learning is an inherently hard information processing problem and the difficulties of learning similar-sounding words are a consequence of optimally distributing limited resources across the perceptual and memory-related processes involved in learning.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.

Bailey, T. M., & Hahn, U. (2005). Phoneme similarity and confusability. *Journal of Memory and Language*, *52*(3), 339–362.

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

Mattys, S. L., & Palmer, S. D. (2015). Divided attention disrupts perceptual encoding during speech recognition. *The Journal of the Acoustical Society of America*, *137*(3), 1464–1472.

Pajak, B., Creel, S. C., & Levy, R. (2016). Difficulty in learning similar-sounding words: a developmental stage or a general property of learning? *J Exp Psychol Learn Mem Cogn.*, *42*(9), 1277-99.

Pajak, B., & Levy, R. (2014). The role of abstraction in non-native speech perception. *Journal of phonetics*, *46*, 147–160.

Pater, J., Stager, C., & Werker, J. (2004). The perceptual acquisition of phonological contrasts. *Language*, 384–402.

Perfors, A., & Dunbar, D. (2010). Phonetic training makes word learning easier. *Congnitive Science Proceedings, 2010*.

Richter, C., Feldman, N. H., Salgado, H., & Jansen, A. (2016). A framework for evaluating speech representations. *Congnitive Science Proceedings, 2016*.

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, *388*(6640), 381–382.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and bayesian inference. *Behavioral and brain sciences*, *24*(04), 629–640.

Werker, J. F., Fennell, C. T., Corcoran, K. M., & Stager, C. L. (2002). Infants' ability to learn phonetically similar words: Effects of age and vocabulary size. *Infancy*, *3*(1), 1–30.