Modelling Novelty Preference in Word Learning

Maarten Versteegh[1,2], *Louis ten Bosch*[2], *Lou Boves*[2]

[1]International Max Planck Research School for Language Sciences, Nijmegen
[2]Centre for Language Studies, Radboud University Nijmegen, The Netherlands
m.versteegh@let.ru.nl, l.tenbosch@let.ru.nl, l.boves@let.ru.nl

Abstract

This paper investigates the effects of novel words on a cognitively plausible computational model of word learning. The model is first familiarized with a set of words, achieving high recognition scores and subsequently offered novel words for training. We show that the model is able to recognize the novel words as different from the previously seen words, based on a measure of novelty that we introduce. We then propose a procedure analogous to novelty preference in infants. Results from simulations of word learning show that adding this procedure to our model speeds up training and helps the model attain higher recognition rates.

Index Terms: language acquisition, word learning, computational modelling

1. Introduction

Extracting words from speech is an important part of human speech processing and plays a crucial role in child language acquisition. We define word learning as the development of multimodal pairings between patterns in the audio stream and referents in the environment. Children perform this task seemingly effortlessly, but in investigating the processes that govern this task we find several ill-understood processes.

Two of these processes are (1) statistical learning and (2) a preference for novel observations. The present paper studies the behaviour of a computational model of word learning that implements these two processes. We discuss them in turn.

First, research in the last years has shown that the ability of young children to process speech signals is at least partly based on the use of the statistical properties of the signal [1]. This ability may help infants discover suitable basic building blocks from a highly variable speech stream and eventually form meaningful combinations of these building blocks.

Infants have been shown to use statistical inference to discover matches between ambiguous combinations of auditory and visual information [2]. Studies like these show that infants learn word-referent mappings by inferring stochastic cross-modal and cross-situational associations.

The second phenomenon, infants' preference for novel or unexpected observations, has been reported in widely varying areas of development, from visual processing in newborn infants [3], to learning the sounds of a native language [4] to artificial grammar learning tasks performed by eight-month-olds [1]. The pervasive use of the Preferential Looking Paradigm in word learning studies attests to the role of novelty and familiarity in this area of development.

The exact neurological basis of novelty preference in developing children (and adults) is not yet completely understood, but it is assumed to be related to the developing memory structure of infants (for an overview, see [5]). A common interpretation of the role of infants' novelty preferences is that it is based on the orienting reflex [6]. Proponents of this interpretation propose that representations of a stimulus in working memory are compared with predictions about the stimulus based on long-term memory. If the two differ greatly, the stimulus has high information content given the learner's previous experience. The long-term memory representation of the stimulus is therefore underdeveloped with respect to the input representation, stimulating an update of the representation in memory.

In summary, we see in the developing infant two processes that help it learn words, statistical inference over cross-modal associations and a preference for novel observations.

In this paper we adapt an existing computational model of word learning to investigate how these processes may enable infants to efficiently learn words from speech. We hypothesize that the processes of statistical inference and novelty preference will help the model learn new words quickly, by providing mechanisms for detection and attention to new words. We will investigate this hypothesis by studying the performance of the model when it is first familiarized with a set of words and is then presented with words not previously encountered.

2. Computational Word Learning Model

2.1. Detection of words in speech

In the model under investigation, word representations are built by a computational method that discovers structure across sequences of stimuli, based on the Non-negative Matrix Factorization algorithm (NMF) [7]. NMF is a statistical machine learning algorithm that finds a decomposition of whole representations into their parts. In our application, the algorithm detects the basic building blocks and word-like units in a variable stream consisting of auditory and visual observations. This algorithm has successfully been applied to speech recognition tasks [8].

In the adaptation of NMF we discuss here, each multimodal stimulus, representing the low-level sensory information that is observed by an infant, is transformed into a feature vector and stored in an $n \times m$ data matrix $\mathbf{V}$. Each column of $\mathbf{V}$ contains n feature values of one of the observed m stimuli. The component parts of the observations are extracted by means of an approximate factorization of the matrix $\mathbf{V}$ into a product of two much smaller matrices $\mathbf{W}$ and $\mathbf{H}$, such that the dissimilarity between the observed matrix $\mathbf{V}$ and the reconstructed matrix $\mathbf{WH}$ is minimized with respect to the symmetrized Kullback-Leibler divergence D_{KL}, as investigated in [9] (equation 1, adapted from [7]).

$$\mathbf{V}_{ij} \approx (\mathbf{WH})_{ij} = \sum_{a=1}^{r} \mathbf{W}_{ia}\mathbf{H}_{aj} \tag{1}$$

While $\mathbf{V}$ is the set of observations that the learner makes

from its surroundings, $\mathbf{W}$ and $\mathbf{H}$ are internal to the learner. The r columns of $\mathbf{W}$ are the internal representations of the basic units that compose the speech stream. The m columns of $\mathbf{H}$ correspond to the stimuli in $\mathbf{V}$. The matrix $\mathbf{W}$ contains a set of basis vectors, that can be interpreted as the extracted recurrent co-occurrences of speech units and concept units. Each column of $\mathbf{W}$ is an internal representation encoding a form-referent pair. In this sense $\mathbf{W}$ is the compressed representation (in memory) of the training stimuli. The columns in $\mathbf{H}$ consist of the weights that must be applied to $\mathbf{W}$ such that a linear combination of the basis vectors in $\mathbf{W}$ optimally approximates the stimuli. The rank r ($= 70$) of the factorization is chosen such that $(n + m)r \ll nm$ so that the matrix $(\mathbf{WH})$ forms a compression of the data in $\mathbf{V}$.

We use an *incremental* adaptation of NMF for reasons of cognitive plausibility. The data matrix $\mathbf{V}$ contains only the most recent stimuli and $\mathbf{W}$ is updated not on the entire dataset at once, but only on the partial matrix $\mathbf{V}$. This incremental approach has shown promising results in speech recognition [10].

Each stimulus is encoded as a single feature vector $\mathbf{x}$. This vector consists of a concatenation of an audio part $\mathbf{x}^a$, encoding the acoustic data in the stimulus, and a visual part $\mathbf{x}^k$, which denotes the keyword present in the audio part by 1-of-K coding and is the target of learning.

Once an initial estimate of the $\mathbf{W}$ matrix is obtained from some input utterances, the model can identify the keyword present in an audio file by the following procedure. Let $\mathbf{x}$ be a stimulus vector, consisting of $\mathbf{x}^a$ and $\mathbf{x}^k$, the auditory and visual component respectively. We use NMF to construct an encoding vector $\tilde{\mathbf{h}}$ based only on $\mathbf{x}^a$ such that it satisfies (2).

$$\tilde{\mathbf{h}} = \underset{\mathbf{h}}{\operatorname{argmin}}\ D_{KL}(\mathbf{x}^a, \mathbf{W}^a\mathbf{h}) \tag{2}$$

This vector $\tilde{\mathbf{h}}$ can then be used to reconstruct the visual vector by $\tilde{\mathbf{x}}^k = \mathbf{W}^k\tilde{\mathbf{h}}$. The corresponding keyword is indexed by the maximum of $\tilde{\mathbf{x}}^k$.

2.2. Novelty Preference

We propose a procedure for modelling infants' novelty preference as follows. Since we can estimate a keyword response vector $\tilde{\mathbf{x}}^k$ from an auditory vector $\mathbf{x}^a$, we can compute the amount of information with respect to the model's internal representations of a stimulus $\mathbf{x}$, by comparing the presented keyword vector $\mathbf{x}^k$ with the estimate $\tilde{\mathbf{x}}^k$. This comparison is analogous to that proposed in [6]. The amount of information $\mathcal{I}$ implicit in a stimulus is defined by the divergence in equation (3), based on the Jensen-Shannon divergence.

$$\mathcal{I}(\mathbf{x}) = \mathcal{H}(\frac{1}{2}(\mathbf{x}^k + \tilde{\mathbf{x}}^k)) - \frac{1}{2}(\mathcal{H}(\mathbf{x}^k) + \mathcal{H}(\tilde{\mathbf{x}}^k)) \tag{3}$$

where $\mathcal{H}$ is the Shannon entropy and all vectors are assumed to be normalized to sum to 1. The unit of $\mathcal{I}$ depends on the base of the logarithm in $\mathcal{H}$. We use the natural logarithm and so measure $\mathcal{I}$ in nats. Equation (3) is an extension to the well-known Kullback-Leibler divergence, with several additional properties desirable for our purposes, such as symmetry and finite-valuedness. The equation computes the number of nats needed to code $\tilde{\mathbf{x}}^k$ when using a code based on $\mathbf{x}^k$ and in this sense gives a measure of the amount of new information in $\mathbf{x}$ given the model's past experiences. In short, the higher $\mathcal{I}(\mathbf{x})$, the higher the *novelty* of $\mathbf{x}$ given the model's internal representations.

Equipped with equation (3) as a measure of novelty, we can implement a *novelty preference* as follows. As described above, the model's internal representations are updated after every observation. At each update step, the parameter γ determines the ratio between the weight of past observations and the current observation. Higher levels of γ effect a bias toward reusing past observations. We implement a preference for novel observations by making this ratio γ for an observation $\mathbf{x}$ dependent on the novelty of $\mathbf{x}$, as expressed by $\mathcal{I}(\mathbf{x})$.

The function to express this dependence should be decreasing, bounded and differentiable. We use the generalized logistic function in equation (4). Here, L and H are the lower and higher asymptote respectively, M is the horizontal shift and α is the growth rate of the function. L and H are set to empirically determined values 0.9 and 0.999, M is set to the median of $\mathcal{I}(\mathbf{x})$, 0.49, and α is set to be sufficiently steep (> 20).

$$\gamma(\mathbf{x}) = H + \frac{L - H}{1 + \exp(-\alpha(\mathcal{I}(\mathbf{x}) - M))} \tag{4}$$

By defining a dependence of the weight of an observation on its novelty we have implemented a procedure for novelty preference that is analogous to that described in section 1. The procedure first measures the novelty of an observation by comparing its prediction about that observation with the observation itself. In the second step, the model decides how much weight to give to this observation, giving more weight (preferring) to novel observations than to observations that match its expectations. This means that novel observations lead to stronger adaptation of the internal representations. As the internal representations become more similar to the stimuli with continued exposure, the weight given to the novel observation will gradually diminish. In short, we have extended our model to include a novelty preference as described in [6].

3. Experiments

3.1. Data sets

In the experiments described in this section, the training sets were designed by selecting utterances from a large database of simulated infant-directed speech, recorded for the ACORNS project [8]. The utterances are all simple sentences, consisting only of a main clause. This elementary structure, as well as the pronunciation clarity, resemble the form of child-directed speech [11]. The natural variation in pronunciation of the keywords creates a cognitively plausible source of uncertainty in the audio domain.

The training set is divided in two parts. The first part, training set A, consists of 200 English utterances from a single female speaker. Each of the utterances contains a single instance of one of the following ten keywords: *sad*, *yellow*, *square*, *give*, *duck*, *ball*, *banana*, *fish*, *cow* and *cat*. The keywords are distributed evenly over the training set.

The second part of the training set, part B, consists of 280 utterances spoken by the same speaker, but in this set each utterance contains one of *fourteen* keywords. This training set contains the keywords of set A extended with four new words: *telephone*, *lion*, *woman* and *apple*.

The speech utterances are coded in the form of co-occurrences of Vector Quantization labels, as proposed by [12]. The code book (150-150-100 for static MFCC, Δ and Δ^2) is trained on randomly selected feature vectors from the training set and is fixed throughout the experiments. This coding ensures that the audio part of the stimulus is a fixed-length vector

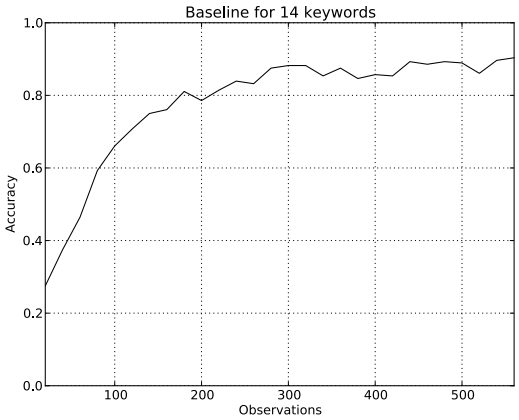

Figure 1: Baseline performance of the word learning model for fourteen keywords. The graph shows the development of recognition accuracy as the learning progresses as estimated from a held out set.

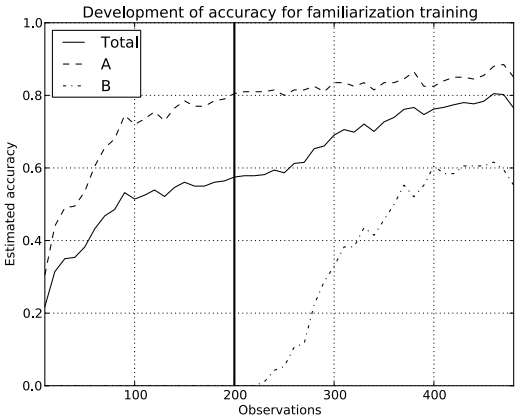

Figure 2: Effect of familiarization training. The graph shows the development of recognition if the model is first familiarized with ten keywords (set A). The vertical line marks the point where the model is presented with words from set B.

of non-negative reals. The visual part of each stimulus is a 1-of-K coding of the keyword present in the auditory part.

3.2. Training and testing

In the training phase, the concatenated audio-keyword vectors are presented to the learning system one by one, as we use an incremental learning method. The training phase consists of two parts, corresponding to the two training sets described in the previous section.

After every twenty observations presented to the model, it is tested on a separate set of 280 stimuli. The test set consists of held-out data from the same speaker that produced the training sets. The keywords again occur evenly in the test set. During testing, training is halted, so that the model's internal representations in $\mathbf{W}$ are not updated on the test set. The test set also consists of two parts. Test set A contains the ten keywords present in training set A, while B contains only the novel keywords in training set B. We report on the results on these test sets separately, so we can investigate the effects of the addition of new words to the input after the model has been familiarized with a set of words.

4. Results

Section 2.1 introduced the basic word learning model, without the novelty preference procedure in place. Figure 1, which will serve as our baseline, shows the development of accuracy of word recognition if all fourteen keyword classes are available from the beginning of training. After training on the full training set, recognition accuracy is around 0.91 and the model converges to this level after about 300 observations.

Figure 2 shows the development of accuracy if the training set is split into sets A and B, as described in 3, but the novelty preference procedure is not in effect. The model is first familiarized on the keywords in set A and only after 200 observations are the keywords in set B introduced intro training. The vertical line marks this transition point. The figure depicts the overall accuracy on all fourteen keywords, as well as separate accuracy measures on sets A and B.

While the model attains recognition accuracy scores comparable to those in Figure 1 on set A, the performance on set B is significantly worse, thereby also affecting the total accuracy score on the combination of sets A and B. The low scores on set B are due to the model's tendency to map novel keywords onto existing representations.

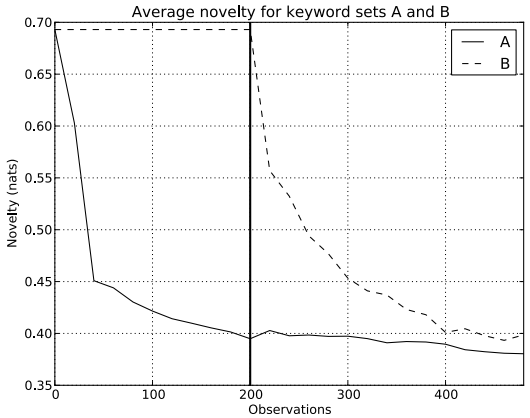

Figure 3: Development of the model's estimate of the novelty of observations in set A and B. The vertical line marks the point where the model is presented with words from set B.

Figure 3 shows the average novelty of observations in set A and B as a function of the number of observations. The data are drawn from the same simulation run as Figure 2, so the novelty preference procedure is not yet in effect. For both sets, the novelty decreases rapidly as the model has more observations to learn from. When only a few observations of a keyword are made, each observation is still highly different from the model's prediction about that observation, and so the novelty is high.

We see that after 200 observations, the average novelty for set B, which was maximal before, starts to drop as the model becomes more familiar with the keywords in this class. The initially high values for the keywords in set B support the assumption in our novelty preference procedure that the measure

of novelty introduced in equation (3) provides a good basis for detecting novel classes of observations.

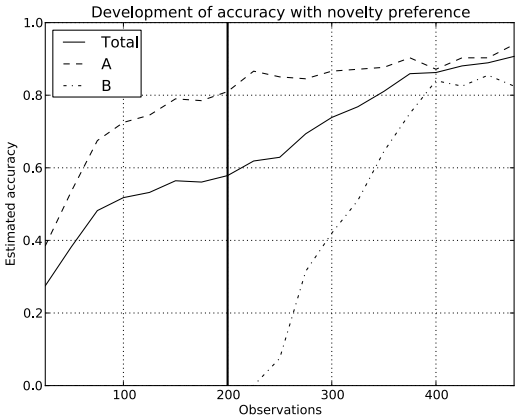

Figure 4: The effect of introducing novelty preference. Training is identical to that in figure 2.

Figure 4 shows the effect of introducing the novelty preference procedure into our model. The increased weight allotted to novel observations helps the model to learn internal representations corresponding to the novel keywords in set B more quickly. This leads to higher recognition accuracy for this class, and so to a higher overall accuracy on all fourteen keywords.

5. Discussion and conclusions

This paper set out to investigate the role of novelty preference in a recent computational model of word learning. We formally defined novelty as the amount of information that an observation represents given the model's previous observations. We implemented a novelty preference procedure by assigning extra weight to those observations that the model considered novel.

The effects of this novelty preference procedure were quantified by measuring the model's word recognition accuracy. We investigated our hypothesis that a preference for novelty helps the model to learn novel keywords.

The results described in section 4 lead us to the following conclusions. First, we observe that the measure of novelty allows the model to distinguish newly introduced keywords from previously observed ones. This means that this measure is adequate, i.e. it reflects our intuitions about what novelty means and is of practical use for the learner as it provides a base upon which it can decide to give more importance to novel observations.

Second, the results show that the novelty preference procedure causes the model to attain higher recognition accuracy scores after the introduction of new keywords. This confirms our hypothesis that adapting the weight of observations dependent on their novelty helps the model to learn new word classes after a familiarization phase.

Third, we note that the novelty preference procedure does not negatively affect the model's performance on the initial set of keywords. This means that this procedure is a good default strategy that ensures high performance even if the training set is not unevenly distributed.

In summary, we have shown how a model of novelty preference based on cognitively plausible processes can achieve good performance when trained on unevenly distributed input. This leads us to conjecture that novelty preference in infants is a process that can help infants deal with the variable and uneven nature of language acquisition and learning in general.

For future research this model can be extended to investigate different distributions of keywords over the training set, for example by deliberately overtraining the model on a small number of keywords and subsequently presenting novel words. In addition, we note that our implementation of the novelty preference procedure is stated in general terms and is not dependent on the precise implementation of the word learning model. Other models with similar input/output specifications may also benefit from this procedure, as it provides a general and principled means for data point weighing in the face of unevenly distributed training examples.

6. Acknowledgements

The research of Maarten Versteegh and Louis ten Bosch is supported by grant number 360-70-350 from the Dutch Science Organisation NWO.

7. References

[1] J. Saffran, R. Aslin, and E. Newport, "Statistical learning by 8-month-olds," *Science*, vol. 274, pp. 1926–1928, 1996.

[2] L. Smith and C. Yu, "Infants rapidly learn word-referent mappings via cross-situational statistics," *Cognition*, vol. 106, pp. 333–338, 2008.

[3] A. Slater, V. Morison, and D. Rose, "Locus of habituation in the human newborn," *Perception*, vol. 12, pp. 593–598, 1983.

[4] P. Jusczyk, *The Discovery of Spoken Language*. MIT Press, 1997.

[5] O. Pascalis and M. de Haan, "Recognition memory and novelty preference: what model?" in *Progress in Infancy Research, Volume 3*, 2003, pp. 95–120.

[6] E. Sokolov, *Perception and the conditioned reflex*. Oxford, UK: Pergamon, 1963.

[7] D. Lee and S. Seung, "Learning the parts of object by non-negative matrix factorization," *Nature*, vol. 40, pp. 788–791, 1999.

[8] L. ten Bosch, H. V. hamme, L. Boves, and R. Moore, "A computational model of language acquisition: the emergence of words," *Fundamentae Informaticae*, pp. 229–249, 2009.

[9] P. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[10] J. Driesen, L. ten Bosch, and H. V. hamme, "Adaptive non-negative matrix factorization in a computational model of language acquisition," in *Proceedings Interspeech 2009*, 2009, pp. 1731–1734.

[11] J. van de Weijer, "Language input for word discovery," Ph.D. dissertation, Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands, 1998.

[12] H. Van hamme, "HAC-models: a Novel Approach to Continuous Speech Recognition," in *Proceedings Interspeech 2008*, Brisbane, Australia, 2008.