INTERSPEECH 2010

Active word learning under uncertain input conditions

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

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

Abstract

In this paper we investigate a computational model of word learning that is cognitively plausible. The model is partly trained on incorrect form-referent pairings, modelling the input to a word-learning child that may contain such mismatches due to inattention to a joint communicative scene. We introduce a procedure of active learning, based on attested cognitive processes. We then show how this procedure can help overcome the unreliability of the input by detecting and correcting the mismatches by reliance on previously built up experience.

Index Terms: language acquisition, word learning, computational modelling

1. Introduction

Learning words from speech is an important part of human language processing and consequently plays an important role in language acquisition. It is a process that seems to be performed effortlessly, but on closer inspection is found to involve several ill-understood top-down and bottom-up cognitive processes.

Word learning, which we define as the development of pairings between patterns in the audio stream and referents in the environment, takes place under conditions of uncertainty for the learner. We distinguish two types of uncertainty. The first relates to the fact that the infant must discover suitable basic building blocks from a highly variable speech stream and eventually form meaningful combinations of these building blocks. Research in the last 15 years has shown that the ability of young learners to process speech signals is at least partly based on the use of statistical properties of the signal, as show by Saffran et al. [1] in experiments with artificial language learning.

The second type of uncertainty is due to possible inconsistencies in the matches between patterns in the speech stream and objects in the visual focus. Smith and Yu have shown that both adults [2] and young infants [3] use statistical inference to discover matches from ambiguous combinations of audio and visual information. Swingley and Fernald [4] reported that infants who are presented with a word that they know does not refer to the object they are looking at, will attempt to redirect their attention toward an object in the visual environment that *does* match the word. This shows that children are able to detect that form-referent pairings do not match their previous experience and that they will actively attempt to resolve these perceived mismatches by aligning a different visual referent with a given auditory form.

In summary, we see in children a set of processes that allow them to learn words while uncertain about both the composition of the patterns in the speech they hear and the reliability of the association of those patterns with objects they see.

In this paper we use an existing computational model of word learning to investigate how these processes may enable infants to accurately learn words from speech under uncertain conditions. We hypothesize that the described processes of detection and correction of mismatches may help to improve the word learning process, by providing a mechanism to overcome uncertainties in the input. We will investigate this hypothesis by investigating the performance of our model under conditions of uncertain input.

2. A model of active word learning

2.1. Detection of words in speech

In our model, word representations are built by a computational method that discovers structure across sequences of stimuli, using the Non-Negative Matrix Factorization algorithm (NMF), introduced by Lee and Seung [5]. NMF is a statistical learning algorithm, capable of solving the first type of uncertainty we described in section 1, i.e. the detection of basic building blocks and word-like units from a variable speech stream.

In our adaptation of NMF, an audio stream, representing low-level sensory information, is transformed into a feature vector and stored in an $n \times m$ database matrix V, each column of which contains n feature values of one of the observed m stimuli. The relevant structure is then extracted by means of an approximate factorization of the matrix V as a product of two much smaller matrices W and H, such that the dissimilarity between the observed matrix V and the reconstructed matrix $W \cdot H$ is minimized with respect to the symmetrized Kullback-Leibler divergence, as investigated in [6] (equation 1, adapted from [5]).

$$V_{ij} \approx (WH)_{ij} = \sum_{a=1}^{r} W_{ia} H_{aj} \tag{1}$$

Both W and H are internal to the learner. The r columns of W are the internal representations of the basic units that are being learned. Each column of H corresponds to a specific stimulus in V. The columns in H consist of the weights that must be applied to W such that a linear combination of basic units in W optimally approximates the stimuli. In most learning models building on NMF, the rank r of the factorization is chosen such that $(n+m)r \ll nm$, with the result that $W \cdot H$ forms a compression of the data in V.

Each stimulus is encoded as a single feature vector $\boldsymbol{y}$. This vector contains an audio part $\boldsymbol{y}^a$ (encoding the acoustic data in the stimulus) and a keyword part $\boldsymbol{y}^k$, which encodes the keyword that the utterance contains and is the target of learning.

The experiments reported here are based on an implementation of a cognitively plausible *incremental* version of NMF that has shown promising results in the field of speech recognition [7]. In this version W is updated each time a stimulus has been processed by the learner. This incremental approach allows the learner to decode (recognize) stimuli right from the start without the necessity to first collect stimuli in a V matrix. It also allows us to interpret the development of the internal W matrix as the dynamic result of an evolution across the training set.

Once an initial estimate of the W matrix is obtained from some input utterances $\boldsymbol{y}$, the system can identify keywords from audio files by reconstructing the visual part. This reconstruction is done as follows. Let $\boldsymbol{y}$ be a stimulus vector, consisting of an audio component $\boldsymbol{y}^a$ and a keyword component $\boldsymbol{y}^k$. We can estimate an encoding vector $\boldsymbol{h}$, based on the audio part of $\boldsymbol{y}$, by minimizing $\boldsymbol{y}^a \approx W^a \cdot \boldsymbol{h}$. The vector $\boldsymbol{h}$ indicates the linear combination of columns in W that best approximates y. We can then use $\boldsymbol{h}$ to reconstruct a keyword vector by using the equality $\hat{\boldsymbol{y}}^k = W^k \cdot \boldsymbol{h}$. In the estimated reconstructed keyword vector $\hat{\boldsymbol{y}}^k$ the index of the largest element indicates the keyword hypothesized by the system.

2.2. Active word learning

The findings described in section 1 suggest that human word learners possess two competencies to help overcome incorrect form-referent pairings in the input: First, the ability to detect these mismatches and second, the ability to correct them by finding a visual referent that matches the auditory form.

Both of these competencies rely on the child's ability to compare current input to the internal representations based on previous input. It is this comparison that forms the basis of our model of active learning. Once the learning model has constructed sufficiently strong internal representations such that it achieves some accuracy in word recognition, it can start to use these representations to judge the confidence it has in the correctness of form-referent pairings in further input. Active learning entails comparing the learner's own estimate of what referent (or keyword) conforms to a speech utterance in the input, to the referent that is actually presented with the utterance.

Thus, active learning is the process of detecting and correcting mismatches in the input, based on the comparison of stimuli with internal representations built from previous experience. We hypothesize that active learning may help in establishing robustness under conditions where the input associations of speech utterances with keywords are not always correct, i.e. where the input is *unreliable*. We formalize this unreliability as the probability that the audio component of an input stimulus is accompanied by the wrong keyword (chosen from a uniform distribution over the remaining possible keywords). This probability is denoted by λ.

If there is a strong match between the learner's own estimate of the referent and the presented referent, then the learner's confidence in the presented pairing will be high and it will update its internal representations accordingly. Conversely, if there is too strong a mismatch, the learner will actively attempt to 'shift focus' from the presented referent to one it estimates will fit better, as Swingley and Fernald [4] showed infants can do. As an example, if the child sees an apple and hears 'Look at the lion!' and detects the mismatch, it may decide to search for a referent conforming to its idea of 'lion' and learn to associate this with the speech.

We can formalize this notion of active learning as follows. First we define the confidence of the model in the correctness of a presented utterance-keyword (form-referent) pair. Let $\max(\boldsymbol{v})$ be the largest element of vector $\boldsymbol{v}$, $\mathrm{maxidx}(\boldsymbol{v})$ be its index and $\boldsymbol{v}_i$ indicate the i-th element of $\boldsymbol{v}$. Then the confidence of the model in the stimulus $\boldsymbol{y}$ is given by:

$$\mathrm{conf}(\boldsymbol{y}) = 1 - \frac{\max(\hat{\boldsymbol{y}}^k) - \hat{\boldsymbol{y}}^k_{\mathrm{maxidx}(\boldsymbol{y}^k)}}{\sum_{i=1}^{n} \hat{\boldsymbol{y}}^k_i} \qquad (2)$$

In words, the confidence of the learner in utterance-keyword pairing $\boldsymbol{y}$ is obtained by reconstructing a keyword vector $\hat{\boldsymbol{y}}^k$ from the utterance part of $\boldsymbol{y}$. The activation level of the estimated keyword is the maximal element of this vector $(\max(\hat{\boldsymbol{y}}^k))$. The presented stimulus indicates its keyword by $\mathrm{maxidx}(\boldsymbol{y}^k)$, so the activation level in the reconstructed vector of the presented keyword is $\hat{\boldsymbol{y}}^k_{\mathrm{maxidx}(\boldsymbol{y}^k)}$. The confidence, then, is 1 minus the normalized difference between the activation of the most activated keyword in the reconstruction and the activation level of the presented keyword in this reconstruction.

We introduce a model parameter θ, a threshold which governs the amount of active learning that the system applies. If the confidence of the system in a presented stimulus is higher than this threshold, i.e. if there is a strong match between the presented keyword and the estimated keyword, the learner will update its internal representations with the presented association. If, on the other hand, the confidence is lower than θ, meaning that there is a too strong a mismatch between the presented and the estimated keyword, the learner will not associate the stimulus' keyword with the utterance, opting for the keyword it reconstructed instead. The threshold parameter θ in effect governs the amount of active learning that the system applies. The higher the threshold is, the more 'active' the system is in relying on its own representations.

In this paper, we investigate the behaviour of the model in conditions where the associations between speech utterances and keywords in the input to the learner are unreliable. This emulates directly the conditions of the experiments of Swingley and Fernald [4], but also has broader ecological validity. Attention of the learner to the caregiver and the communicative scene is an important factor in the acquisition of language (cf. [8, 9]). Unreliable associations in the input can be understood as modelling situations where the attention of the learner is on the 'wrong' part of the visual scene, thus risking associating the wrong object with an utterance.

3. Experiments

3.1. Data sets

In the experiments described in this section, the training set was designed by selecting utterances from a large database of human speech, recorded in the ACORNS project [10]. The utterances are all simple sentences with only elementary syntactic structure, consisting only of a main clause. This syntactic structure resembles the structure of child-directed language [11].

The training set consists of 450 utterances from a single speaker. Each of the utterances contains a single instance of one of the following 10 keywords: *mummy*, *looks*, *big*, *lion*, *see*, *bottle*, *square*, *cat*, *cow*, *fish*. The keywords are distributed evenly over the training set.

The nature of the NMF algorithm places restrictions on its input requirements that are not easily met by continuous audio recordings. Specifically, it demands that all input vectors are of equal length and contain only positive values. In order to comply with these specifications the utterances are coded as co-occurrence counts of Vector Quantization labels, as proposed by Van hamme [12]. The code book (150-150-100 for static MFCC, Δ and Δ^2) is trained on randomly selected feature vectors and is fixed throughout the experiments.

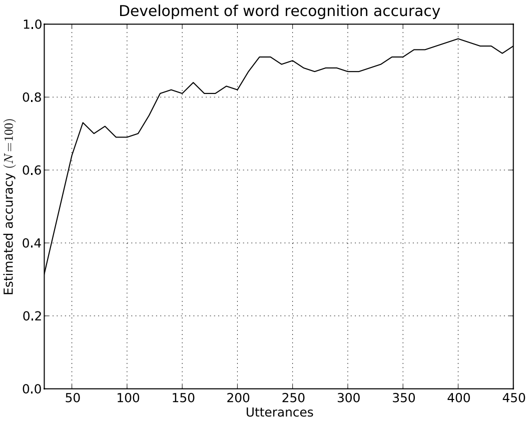

Figure 1: *Baseline of word learning experiment. The graph shows the development of accuracy of word recognition as the learning progresses as estimated from a held out test set* ($N = 100$)

Each speech utterance is associated with a binary keyword vector indicating which single keyword occurs in it.

3.2. Training and testing

In the training phase, the combined audio-keyword vectors are presented to the learning system incrementally. The training phase consists of two periods. In the first period, all presented associations are correct and the learner accepts them unconditionally ($\theta = 0, \lambda = 0$). This period spans 40 stimuli or 4 tokens of each keyword. In the second period, spanning 450 utterances, both θ and λ take values > 0, meaning that the input becomes unreliable and that the learner can learn actively. The first period of the learning phase is necessary for the system to build internal representations that are sufficiently strong to estimate keywords from audio signals.

After every 10 stimuli presented in the second period of the training phase, the system is tested for accuracy on a separate set of utterances ($N = 100$). The test set consists of held-out data from the same speaker as in the training set. The keywords occur evenly in the test set. During testing, training is halted so that the internal representations in W are not updated on the test set. The accuracy of the model is estimated as follows. Based on the audio part of an input stimulus, the model reconstructs the keyword part. The accuracy on a given test set is estimated by comparing the reconstructed keyword vector with the original keyword vector for every item in the test set.

4. Results

Section 2 introduced the basic word learning model. In our extended model it corresponds to a setting with passive learning ($\theta = 0$) and completely reliable input ($\lambda = 0$). Since this combination of parameter settings is conceptually closest to an idealized learning situation, it will serve as our baseline.

Fig. 1 shows the development of the accuracy of word recognition of the baseline. Denoting the estimated accuracy after training on n utterances by $\hat{a}_n$, we note that the estimated final accuracy is $\hat{a}_{450} \approx .94$ and convergence to this level occurs around 350 utterances of training.

Fig. 2 depicts the development of the accuracy of word recognition for partly unreliable input, i.e. $\lambda = 0.4$ and for three levels of active learning ($\theta = 0, \theta = 0.5$ and $\theta = 0.9$). The figure shows that without active learning ($\theta = 0$), the model is still able to learn from uncertain input, improving significantly from $\hat{a}_{40} = 0.41$ to $\hat{a}_{450} = 0.74$ (McNemar's Test, $p \ll 0.01$). Note that this improvement over training is significantly less pronounced than the improvement over training shown in the baseline graph in figure 1 ($p \ll 0.01$).

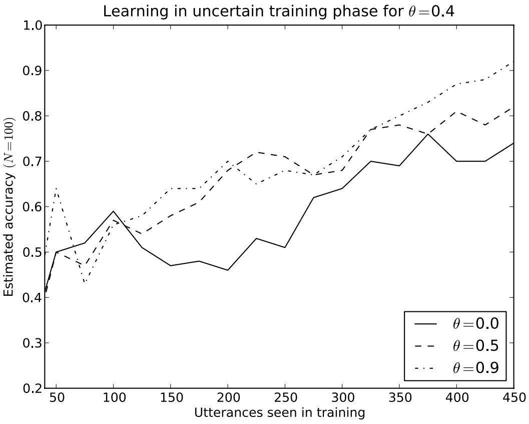

Figure 2: *Effects of active learning from uncertain input. The figure depicts the development of accuracy of word recognition for three levels of active learning (θ), showing that performance improves with increasing θ.*

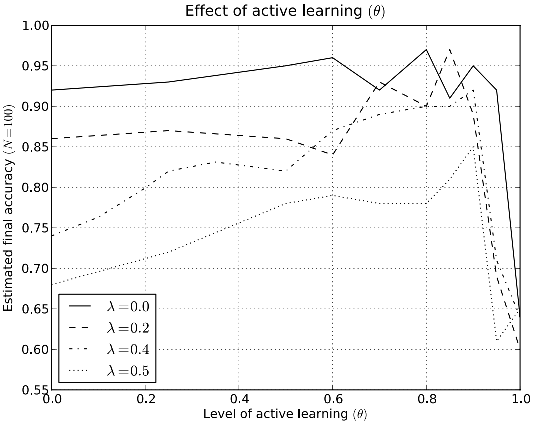

Figure 3: *Effects of active learning on final performance for different levels of λ. Accuracy is estimated after training on 450 uncertain utterances.*

Accuracy improvement during training is significantly higher when the model learns actively ($\theta = 0.5$ and $\theta = 0.9$) as also shown in figure 2. Final accuracy estimations are 0.82 and 0.9 respectively. In summary, these graphs display the increase in performance with increased θ for $\lambda = 0.4$.

Fig. 3 gives an overview of the effects of active learning on final performance after training under different levels of uncertainty ($\lambda = 0.0, \lambda = 0.2, \lambda = 0.4, \lambda = 0.5$). From this figure we observe the following.

All graphs show a distinct peak in performance around $\theta = 0.9$, after which performance drops off sharply. Up until this point, we observe that higher levels of active learning coincide with higher levels of final performance for all values of λ (including $\lambda = 0$). Secondly, for the condition $\lambda = 0$, we observe that levels of θ up to 0.9 do not decrease the performance substantially relative to $\theta = 0$. This is important, because if the input is in fact reliable, any lack of confidence in the presented associations incurs the risk of incorrectly rejecting stimuli. Fig. 3 shows that this is not the case in our model.

The performance as θ approaches 1 under the different levels of λ is understood from this same perspective. At $\theta = 1$ every association in the input is rejected in favour of a reconstruction based on internal representations. This means that the model cannot learn effectively from correct inputs and projects its errors in its internal representations.

5. Discussion and Conclusions

This paper set out to investigate the role of active learning under conditions of uncertainty in a recent computational model of word learning. We formally defined uncertainty as stochastic mismatches in form-referent pairings in the input to the learning algorithm. Active learning was implemented by allowing the learner to override the presented keyword if its confidence in the pairing with an audio signal was lower than a threshold. The detection and correction of mismatches by choosing a different referent if deemed necessary models the process of gaze shifting under inconsistent input described by Swingley and Fernald [4].

The effects of active learning were quantified by measuring the word recognition accuracy. We investigated our hypothesis that active learning helps the learner overcome uncertainty in the form-referent pairings in the input.

The results described in section 4 lead us to the following conclusions. First, we observe that higher levels of unreliability of the input decrease the final performance of the model. This ties in with the posed centrality of attention sharing in language acquisition [9]. However, even with higher levels of uncertainty, the model still performs above chance and is able to learn throughout the uncertain phase of training by relying on its past experience.

Second, the results show that the learning model performs better under unreliable input if it learns actively. This confirms our hypothesis that actively detecting and correcting mismatches in the input by relying on previous experience helps the learner gain a higher accuracy of word recognition.

Third, the less reliable the input, the greater the improvement gained from active learning. Since unreliability is defined stochastically over all input stimuli, the model can still improve its performance when it is does not learn actively, although these improvements will be marginal. When the unreliability increases, learning inactively becomes less effective. If the model learns actively, however, it can better counteract the unreliability of the input and gain better performance.

Fourth, the model never performs substantially worse when it learns actively instead of passively (up to a certain level of active learning). This is an important result, since it shows that the active learning strategy is viable even when the input is completely reliable. When it is not known how reliable the input will be (what the level of λ will be), active learning is a good default strategy, guaranteed to achieve the best performance given the level of reliability.

In summary, we have shown how a model of active learning based on cognitively plausible processes that infants are known to apply can achieve good performance when trained with uncertain input. This result ties in with the observation that a corrective procedure that is able to overcome attentive mismatches between an utterance from a caregiver and an object in the communicative scene, can be an important part of word learning in human language acquisition.

For future research this model can be extended to have a dynamic and variable level of active learning depending on the strength of the internal representations. In the current paper we model the infant's burgeoning word recognition capacity as a system that is trained just sufficiently to gain some accuracy in recognition but has not reached convergence yet. In a future extension, the role of the first, 'clean', phase of training could be studied to bring the model closer to cognitively and ecologically plausible learning situations.

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] C. Yu and L. Smith, "Rapid word learning under uncertainty via cross-situational statistics," *Psychological Science*, vol. 18, pp. 414–420, 2007.

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

[4] D. Swingley and A. Fernald, "Recognition of words referring to present and absent objects by 24-month-olds," *Journal of Memory and Language*, vol. 46, pp. 39–56, 2002.

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

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

[7] V. Stouten, K. Demuynck, and H. Van hamme, "Automatically learning the units of speech by non-negative matrix factorisation," in *Proceedings Interspeech*, 2007.

[8] L. Smith, "How to learn words: an associative crane," in *Breaking the word learning barrier*, R. Golinkoff and K. Hirsh-Pasek, Eds. Oxford: Oxford University Press, 2000.

[9] M. Tomasello, *Constructing a language – a usage-based theory of language acquisition*. Harvard University Press, 2003.

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

[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.