Dealing with uncertain input in word learning

Maarten Versteegh[†*], Louis ten Bosch[*], Lou Boves[*]
[*]Centre for Language Studies, Radboud University, Nijmegen, the Netherlands
[†]International Max Planck Research School for Language Sciences, Nijmegen, the Netherlands

***Abstract*—In this paper we investigate a computational model of word learning, that is embedded in a cognitively and ecologically plausible framework. Multi-modal stimuli from four different speakers form a varied source of experience. The model incorporates active learning, attention to a communicative setting and clarity of the visual scene. The model's ability to learn associations between speech utterances and visual concepts is evaluated during training to investigate the influence of active learning under conditions of uncertain input. The results show the importance of shared attention in word learning and the model's robustness against noise.**

***Index Terms*—1.1 computational neuroscience, 3.2 language development, 5.2 grounding of knowledge and representations, 6.1 language learning, 6.8 statistical learning**

I. Introduction

Language processing is arguably the most complex human cognitive capability. Speech production, speech perception and speech understanding are all processes that seem to be performed effortlessly, but on closer inspection appear to involve as yet ill-understood top-down and bottom-up processes.

Over the last few decades research in different disciplines such as psycholinguistics, linguistics, biology, neuroscience, psychology, phonetics and speech technology has resulted in the design of theories and models that account for *parts* of the chain that links the speaker's intentions and the listener's comprehension. However, we are still far from a coherent and comprehensive model or theory of speech communication and the way in which communication skills are learned.

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 signals. Werker and Curtis [1] presented a comprehensive model of human language acquisition, while Maloof and Michalski [2] focused on incremental learning. Infants learn to discover the word-like elements in speech without any prior knowledge about lexical identities, as shown by Saffran et al. [3] in experiments with artificial language learning. In another study Smith and Yu [4] show that word learning by young infants is enhanced by cross-modal, cross-situational learning, and that infants can discover associations from ambiguous combinations of audio and visual information. Roy and Pentland [5] focused on machine learning of words. Their Cross-channel Early Lexical Learning (CELL) model [5] is trained with audio recordings of play sessions between caregivers and seven-to-eleven-month-old infants. During these sessions, caregivers and infants played with toys from seven categories (balls, shoes, keys, cars, trucks, dogs, and horses). Pictures taken of each toy from various angles were used for building a visual model of each toy. CELL learned to discover words by listening to the speech, while simultaneously looking at the visual representations.

In the ACORNS project (www.acorns-project.org), a number of computational approaches have been investigated with the aim to model the emergence of word-like internal representations during the first stages of language acquisition. Similar to the CELL model, learning in ACORNS takes place within a communicative loop between a (virtual) caregiver and a (virtual) learner. The model is based on ecologically inspired stimuli, as well as a physiologically and psychologically motivated memory architecture.

The ACORNS model uses multimodal stimuli as input, and focuses on the discovery of word-like units from these stimuli. The ACORNS results [6] show that the representations in the model emerge from the multimodal stimuli. This is in line with growing evidence that speech and language skills are *emergent* capabilities of a developing communicative system [7], [8] and that the way in which linguistic patterns are stored and used during language acquisition changes constantly as these patterns become more numerous and fine-grained, and as the methods needed for processing the patterns correspondingly become more complex [1]. During learning, the learner hypothesizes and reinforces internal representations during the interaction with the caregiver.

The ACORNS results were based on an idealised learning situation. The learning agent is always presented with consistent and coherent stimuli and learns *passively*, i.e. without reinterpreting or overriding the presented stimuli based on prior experience.

This paper has two goals. First, we investigate the impact of making the learning agent gradually more autonomous and *active*. We implement this by giving the learner the opportunity to prefer its own interpretation of the meaning of a stimulus over the interpretation suggested by the caregiver. Second, we investigate the impact of making the input stimuli for the learner less crisp, by adding noise to the visual input and by allowing the learner to focus attention on different objects in the environment than those referred to by the caregiver. By doing so, we will investigate issues related to attention and grounding in a learning process that aims at simulating a cognitively plausible learning process.

We investigate the effects of activity, attention, grounding and 'noise' in the form of their effects on the extent to which internal representations can be generalized. It has been shown that when children are exposed to more than one speaker, they generalize better to novel speakers [9]. Simulations in

978-1-4244-6901-7/10/$26.00 ©2010 IEEE

the ACORNS project have shown similar effects [10] for a passive learner and crisp stimuli. In this paper, we investigate a multi-speaker training situation in a setting where the learner is active, can focus attention on 'irrelevant' input and where inputs become increasingly fuzzy.

The paper is organized as follows. The next section describes the general learning model. Section III extends this model to incorporate active learning and introduces the variations that the learner's input is subjected to in order to make it more realistic. Sections IV and V contain the results from these experiments and a discussion.

II. Word learning model

A. Discovery of word-like units

In conventional pattern recognition systems the patterns to be recognized, as well as the primitive elements from which complex patterns can be formed, are defined a priori. By doing so, these systems sidestep the task of detecting suitable basic units – simply because these are pre-defined by the developer. In conventional ASR systems, for example, the basic units are usually phones, and all known words are represented as phone sequences, reducing the problem of word recognition in continuous audio to one of finding the sequences of discrete symbols that is most likely for a given signal.

However, the representations of words as sequences of phones, like beads on a string, does not capture the full complexity of spoken language [11]. In speech, including infant-directed speech, words are not separated by silences. Rather, words blend and merge at their boundaries. Thus, a word learning system that attempts to model a human infant, needs to discover the patterns in the continuous input stream that correspond to objects, attributes and events in the context. It has been shown that one of the mechanisms that young infants use in their language acquisition is the discovery of statistically recurrent patterns in speech signals [3], [12], [13].

In the ACORNS model, word representations are built by a computational method that is able to detect structure across sequences of multimodal stimuli [6]. The details of the internal representations depend to some extent on the specifics of the structure discovery technique [10]. In this paper we will focus on Non-Negative Matrix Factorization (NMF), introduced by Lee and Seung [14].

In our adaptation of NMF, low-level sensory information, obtained from multi-modal stimuli, 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 the symmetrized Kullback-Leibler divergence, as investigated in [15] (equation 1, adapted from [14]).

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

Both W and H are internal to the learner. The r columns of W are the internal representations of the basic units that are 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 approximate the stimuli. In learning models building on NMF, the rank r of the factorization is chosen such that $(n+m)r < nm$, 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 $\boldsymbol{y}^a$, which encodes the acoustic data in the stimulus, and a visual part $\boldsymbol{y}^k$, which encodes the concept that is present in the utterance.

The experiments reported here are based on an implementation of an *incremental* version of NMF, in which 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 has been obtained from input stimuli, the system can identify a concept from a speech utterances by making a reconstruction of the corresponding 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 visual component $\boldsymbol{y}^v$. We can estimate an encoding vector $\boldsymbol{h}$, based on the audio part of $\boldsymbol{y}$, by minimizing $\boldsymbol{y}^a \approx W^a \boldsymbol{h}$. The encoding vector $\boldsymbol{h}$ is used to reconstruct a visual vector by the equality $\hat{\boldsymbol{y}}^v = W^v \boldsymbol{h}$. This estimated reconstructed visual vector $\hat{\boldsymbol{y}}^v$ encodes the concept hypothesized by the system.

B. Active word learning

The passive learning model described in the previous section showed promising results in an idealized learning situation [10]. In this section we describe an *active* extension of the model that is able to deal with inconsistent and noisy input. We hypothesize that the active involvement of the learner may help in establishing robustness under these conditions.

Attention of the learner to the caregiver and the communicative scene is another important factor in the acquisition of language (cf. [18], [19]). We model the attention level of the learner, denoted by α, as the probability that the learner pays attention to the right object in word learning. With a probability of $(1-\alpha)$, the audio component of a stimulus is accompanied by the wrong visual component (chosen from a uniform distribution over the remaining possible visual concepts).

The *Clarity* of the communicative setting denotes the amount of noise degradation of the visual scene. This parameter, denoted by c, defines the ambiguity and fuzziness of the object space. Fuzzification is a two-step process. First multiplicative noise is added to the binary feature matrix, by multiplying the non-zero elements by random samples from a Gaussian distribution with $\mu = 1$ and $\sigma = (1-c)$. Next, zero

elements are replaced by random samples from an exponential distribution with probability density function f:

$$f(x) = \frac{1}{\frac{1}{2}(1-c)} e^{-x\frac{1}{\frac{1}{2}(1-c)}} \quad (2)$$

The parameter c governs the clarity of the visual scene; the clearer the setting, the less noise the visual input contains.

The parameters α and c cover the uncertainties in the input that the learner is confronted with. The model attempts to overcome these uncertainties by learning *actively*. Active learning in this study entails comparing new input stimuli to the model's internal representations obtained from previous input and basing the decision on how to update these internal representations on the result of this comparison.

The procedure for active learning in our model is as follows. First, for each new input stimulus, the model builds a reconstruction of the visual part of the stimulus on the basis of its internal representations. Second, the model estimates how confident it is of this reconstruction by determining how unambiguous the concept associated with this reconstruction is. Finally, on the basis of this confidence, the model decides whether to accept the association of the audio utterance with the concept from the input stimulus, or to associate the audio utterance with its own reconstruction instead.

This procedure is formalized as follows. First we define the notion of confidence in a reconstruction on the basis of input stimulus $\boldsymbol{y}$. Let $\max(\boldsymbol{v})$ be the maximal element of vector $\boldsymbol{v}$ and $\max_2(\boldsymbol{v})$ the second maximal element. Then the confidence of the model in reconstruction $\hat{\boldsymbol{y}}$ with n elements is given by:

$$\text{conf}(\hat{\boldsymbol{y}}) = \frac{\max(\hat{\boldsymbol{y}}) - \max_2(\hat{\boldsymbol{y}})}{\sum_{i=1}^{n} \hat{\boldsymbol{y}}_{\boldsymbol{i}}} \quad (3)$$

In words, the confidence of the model in a reconstruction is the normalized difference between the value of the model's first and second guesses of the concept present in the audio utterance. If the values of the guesses are close together, the confidence is low, if the first guess 'stands out' from the second guess, the confidence of the model in its reconstruction is high.

Second, we introduce a model parameter θ, which denotes a threshold that governs the amount of active learning that the model applies. If the confidence of the model in the reconstruction it makes on the basis of a presented stimulus is higher than the threshold θ, it will associate the reconstructed visual vector with the audio utterance instead of the presented visual vector. If it is lower, it will simply accept the presented stimulus. Lower values for θ thus indicate more active learning and vice versa.

In our definition, active learning comprises the process of detecting and correcting uncertainties 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 visual scenes may not always be correct due to inattention of the learner or where the visual scene is unclear to due noise degradation.

III. Experiments

A. Data sets

In the experiments described in this section, the training set was designed by selecting utterances from a large database recorded for the ACORNS project [6]. The utterances are all simple sentences, with only elementary syntactic structure, devoid of embeddings and consisting only of a main clause. This structure resembles that of child directed speech [16].

The training set consists of 520 utterances from four different speakers, 130 utterances from each speaker. The recordings of the speakers, two male (M1,M2) and two female (F1,F2), occur in the order F1-M1-F2-M2. Each of the utterances contains a single instance of one of the following keywords: Angus, Ewan, bath, book, bottle, car, daddy, mummy, nappy, shoe, telephone. The keywords are distributed evenly over the training set and over the speakers. Each utterance is associated with a boolean-valued visual vector indicating which of the keywords occurs in it.

Nmf 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 in the form of co-occurrences of Vector Quantization labels, as proposed by Van hamme [17]. 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.

B. Training and testing

During training, the combined audio-visual vectors are presented to the learning system in incremental fashion. After every 10 stimuli, the model is probed for accuracy on four separate fixed test sets ($N = 30$). The test sets consist of held-out data from the four speakers, with again evenly distributed keywords. During testing, training is halted, meaning that the internal representations W are fixed.

The accuracy of the model is estimated as follows. Based on the audio part of an input stimulus, the model reconstructs the visual part. The accuracy on a given test set is estimated by comparing the reconstructed visual vector with the visual vector from the stimulus for every item in the test set.

IV. Results

Section II introduced the basic ACORNS model. In our extended model it corresponds to a setting with a passive, attentive learner and clear visual input. Since this combination of parameter-settings is conceptually closest to previous work in the ACORNS project, it will serve as our baseline.

Fig.1 shows the accuracy as estimated on the held-out sets during a training run. The horizontal axis specifies the number of learning utterances that the model has processed so far. The vertical axis shows the estimated accuracy. The five curves relate to the four test sets, plus an average, denoted 'Total'. The four vertical lines indicate the points in the run at which a new speaker is introduced.

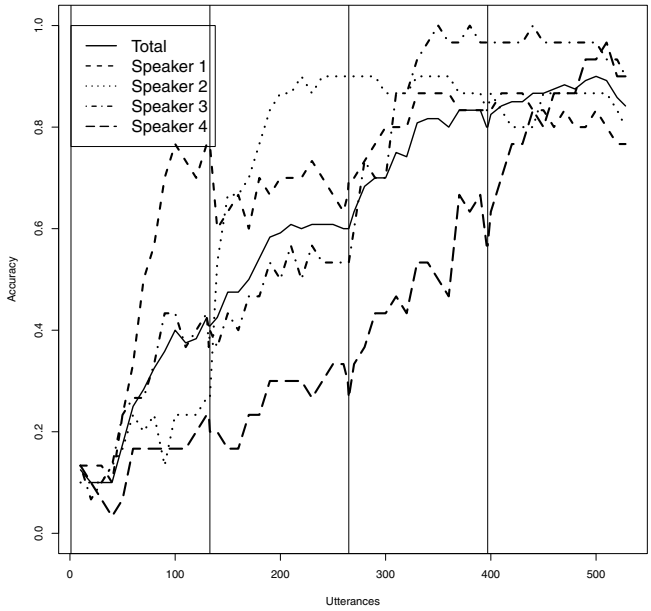

Fig. 1. Baseline simulation ($\theta = 1$, $\alpha = 1$, $c = 1$)

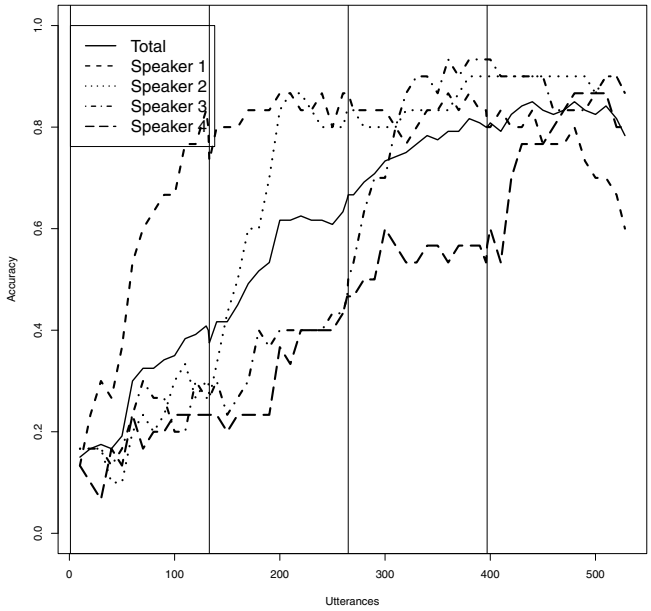

Fig. 2. Simulation with unclear visual input, showing similar tendencies as in Fig.1 ($\theta = 1$, $\alpha = 1$, $c = 0.5$)

TABLE I
SUMMARY OF BASELINE INDIVIDUAL ACCURACY JUMPS IN FIRST 50 UTTERANCES AFTER THE INTRODUCTION OF A NEW SPEAKER, CORRESPONDING TO FIG.1

Speaker	Accuracy difference	p-value
1	0.22	$\ll 0.01$
2	0.51	$\ll 0.01$
3	0.27	0.006
4	0.17	0.009

The baseline shows clear tendencies that were not discovered in earlier work. In Fig.1 we observe that the model gains considerably from experience with the first speaker as it trains on sentences from the following speakers. In fact, the curve corresponding to speaker 4 increases gradually, even though none of his utterances are used for learning during the first three quarters of the run. The biggest gains in accuracy on a specific speaker, however, predictably occur during training on that speaker. Large 'jumps' in accuracy are made during exposure to the first 50 utterance from a new speaker, as summarized in table I.

From Fig.1 it can also be seen that the performance of the model on previous speakers does not drop drastically when a new speaker is introduced in the training. Even the average performance shows no negative impact from speaker changes. This finding has two implications. First, the model is robust against sudden extreme variations in the input, even the introduction of new speakers of a different sex. Second, and related, the internal representations in the model are not simply blends of information from the different speakers, like in single-gaussian HMMs. Rather, the process of updating the representations is non-destructive in the sense that previously gained experience is not lost, but added to.

Fig.2 shows the effect of introducing fuzzy visual features into the training. As shown in Fig.2 the observations made in Fig.1 hold up under the pressure of this noise. The eventual average performance is comparable, but there are some differences in the accuracy curves of the training runs. Apparently, the unclarity of the visual features affects the rate of convergence of the individual test sets. Specifically, the model is less able to take advantage of previous experience in recognizing the utterances from speaker 4.

In general, the model behaves well under conditions of unclarity, showing gradual, but very slow degradation of the final performance. The observation from Fig.2 that convergence for individual speakers is slower holds across the board. As the visual scenes become less clear, the system needs more utterances to reach the final accuracy rate. Table II illustrates this point for $\theta = 0.5$, $\alpha = 1.0$.

Fig.3 illustrates the effect of low attention levels on the learning curves. In this particular run of the simulation, α was set to 0.5, meaning that for each stimulus in the input, there was a 50 percent chance that the visual component was switched to one corresponding to a keyword not occurring in the audio component of the stimulus.

Average final performance for this run is significantly lower (McNemar's test of proportion, $p \ll 0.01$) than the baseline. Especially striking is the bad performance on utterances from speaker 1 throughout the training run. This indicates a persistent trend in runs with lower attention levels; the amount of data needed to reach better-than-chance performance increases as the attention levels are lower. The accuracy on test sets of unseen speakers profits less from seen speakers compared to

TABLE II
CLARITY IN THE VISUAL SCENE VERSUS ESTIMATED FINAL AVERAGE ACCURACY ON THE ENTIRE HELD-OUT SET, FOR $\theta = 0.5$ AND $\alpha = 1.0$. THIS TABLE SHOWS THAT THE MODEL DEGRADES ONLY GRADUALLY UNDER NOISE. FROM $c = 0.6$ DOWNWARDS THE DIFFERENCES BETWEEN SUBSEQUENT VALUES ARE SIGNIFICANT ($p < 0.01$).

Clarity	Accuracy
1.0	0.85
0.9	0.89
0.8	0.87
0.7	0.89
0.6	0.88
0.5	0.82
0.4	0.78
0.3	0.71
0.2	0.69
0.1	0.62
0.0	0.55

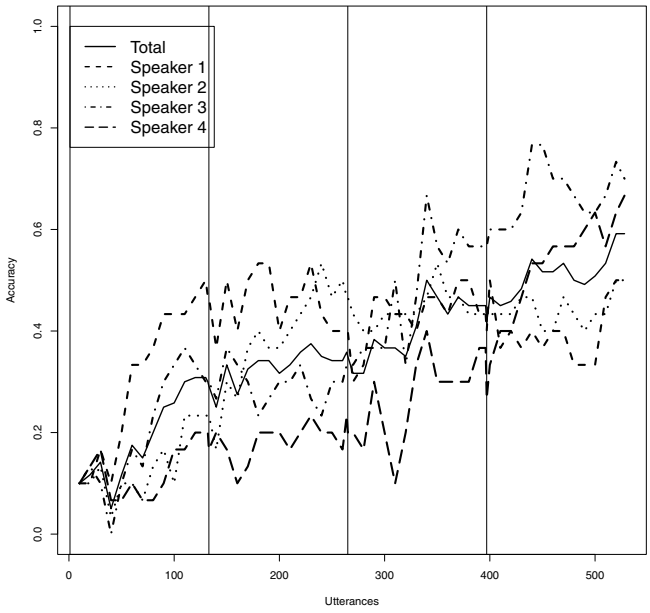

Fig. 3. Simulation with inattentive learner ($\theta = 1$, $\alpha = 0.5$, $c = 1$)

the baseline.

Fig.4 shows the curves of another inattentive learning situation, this time with a high setting for active learning. It illustrates the effect of active learning. Final performance is slightly higher under the passive learning condition depicted in Fig.3 (McNemar test, difference $= 0.025, p \ll 0.01$), but the learning curves are very different. Active learning tends to increase the number of utterances needed to get the learning 'off the ground', i.e. showing an upward trend in accuracy. When the learner is active, past experience becomes more important. Especially in conditions of inconsistent input, the learner will have a harder time establishing a base it can trust.

A second point illustrated by Fig.4, is that active learning tends to smooth out the 'jumps' in accuracy that were so characteristic of the curves in the baseline condition. Since input is not taken at face value when the learner is active, it

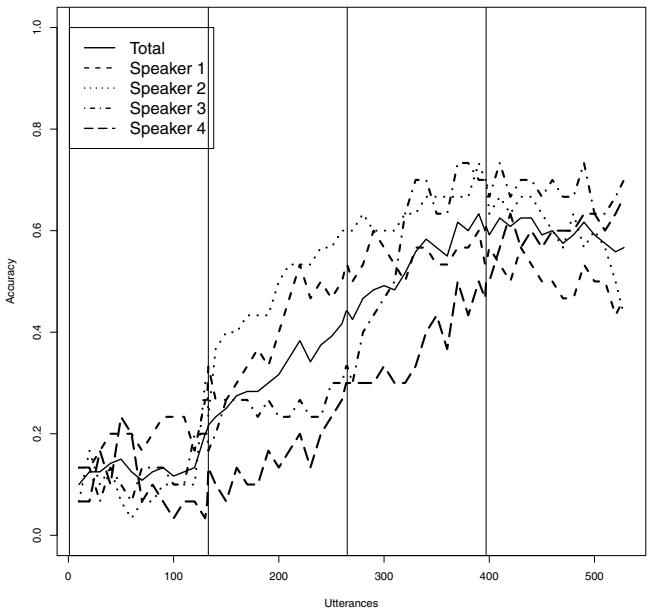

Fig. 4. Simulation with inattentive, but active learner, showing the effect of active learning ($\theta = 0.3$, $\alpha = 0.5$, $c = 1.0$)

also does not have the chance to make the great increases in accuracy that a passive learner has. We were not able to find an effect of learning activity on the degradation of performance under noisy conditions.

V. DISCUSSION

This paper set out to investigate the roles of attention, visual clarity and active learning in an existing computational model of word learning. We formally defined these three concepts in terms of a learning algorithm that had previously shown good results on word learning, but that had only been tested under ideal conditions. Attention and visual clarity were defined as stochastic changes to the visual part of multimodal input to the learning algorithm. Active learning was implemented by allowing the learner to override the visual grounding of an input if its confidence in its own interpretation exceeded a threshold, a refinement to the learning algorithm that factored in previous experience in deciding the establishment of associations between visual and auditory patterns.

The effects of the three factors on learning performance were investigated by measuring the accuracy with which the learner could recognize a fixed set of test stimuli as the learning process proceeded. By doing so, we were able to investigate the conditions under which the claim of [9] that interacting with more caregivers helps infants to generalize internal representations more easily and more effectively to novel speakers.

Word and concept learning is complicated by potential ambiguities in the communicative setting. Learner and caregiver may focus on different particulars of their surroundings, leading the learner to form an association between an object

and a sound pattern that was not intended by the caregiver. The visual scene may also contain inherent ambiguities, i.e. contain overlapping objects, indiscernible features etc. Due to these uncertainties a learner may profit from overriding the apparent visual grounding of speech stimuli if she is confident about her own interpretation of a stimulus.

Our results show that the specific learning algorithm investigated in the paper breaks down when the settings of one of the three factors exceeds some extreme value. However, the algorithm shows graceful degradation while the settings of the three factors move to the break-down point. Although a full exploration of the interactions between the three factors remains for future research, the data presented in this paper allow us to draw a number of interesting conclusions.

The results show that the learning model profits from its experience with previous speakers in processing new speakers. The more speakers the learner has been exposed to, the better a novel speaker's utterances are recognized. This confirms a finding reported in [10], but on the basis of our present results, we are able to extend this analysis and show the developments in terms of learning curves. Regarding the goals we set in section I, we come to the following conclusions.

First, we see that speakers that are no longer observed are not entirely forgotten. While the baseline curve does show a drop in the estimated accuracy on speaker 1's held-out set when the input switches to speaker 2, this drop is slight, compared to the gained accuracy. The retainment of information about older speakers is an important finding because the training presents the speakers in blocks and the NMF algorithm we used is incremental. This means that the model is able to remember aspects of speaker characteristics. As Fig.1 shows, there is a slight decrease, especially visible for speaker 1, after input from this speaker stops coming in. While not significant (McNemar's test between performance at utterance 130 and 528, $p = 0.34$), this decrease might indicate that this 'remembering' is not entirely perfect or robust.

Second, we can conclude that a decrease in clarity leads to a gradual degradation in performance, as shown by tableII. This indicates that, given enough data, the model can overcome even fairly large amounts of noise in the visual channel. This provides an indication that this model indeed possesses 'life-like' properties, as learning under noisy conditions is more the rule than the exception in infant word-learning. Moreover, the characteristic 'jumps' observed in the baseline are retained up to high levels of noise, showing that the character of learning does not change under noisy conditions.

Third, lower levels of attention are more harmful than lower levels of clarity. This ties in with the posed centrality of attention sharing in language acquisition [19]. However, even with low levels of attention, the model performs well above chance. Paying attention to only 50 percent of the stimuli does not mean a 50 percent drop in performance compared to the baseline. The parameter θ has its effect on the learning curve, but, surprisingly, barely on the final performance. The learning curves show increasing variance as attention levels lower.

Finally, active learning does not affect final performance either, but does provide stability to the learning curve, making it more gradual and smoothing out 'jumps' in both directions. These effects will be investigated further.

Future research will include investigations into the interaction between caregiver and learner to examine the role of different types of feedback. A relatively unexplored area is the question of the role of multi-modal information in grounding the internal representations in learning algorithms like the one presented here. We intend to explore this role more in depth in future work. We will further elaborate on the relations between variations and inconsistencies in the input stimuli and cognitively plausible learning situations.

ACKNOWLEDGEMENTS

The authors thank Hugo Van hamme for the procedure for obtaining fuzzy visual features. The research of Maarten Versteegh and Louis ten Bosch is supported by grant number 360-70-350 from the Dutch Science Organisation NWO.

REFERENCES

[1] J. Werker and S. Curtis, "Primir: a developmental framework for infant speech processing," *Language learning and development*, vol. 1, pp. 197–234, 2005.
[2] M. Maloof and R. Michalski, "Incremental learning with partial instance memory," *Artificial Intelligence*, vol. 154, pp. 95–126, 2004.
[3] J. Safran, R. Aslin, and E. Newport, "Statistical learning by 8-month-olds," *Science*, vol. 274, pp. 1926–1928, 1996.
[4] L. Smith and C. Yu, "Infants rapidly learn word-referent mapping via cross-situational statistics," *Cognition*, vol. 106, pp. 333–338, 2008.
[5] D. Roy and A. Pentland, "Learning words from sights and sounds: a computational model," *Cognitive Science*, vol. 26, pp. 113–146, 2002.
[6] 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.
[7] S. Johnson, *Emergence*. New York: Scribner, 2002.
[8] B. MacWhinney, "Models of the emergence of language," *Annual Review of Psychology*, vol. 49, pp. 199–227, 1998.
[9] B. Barker and R. Newman, "Listen to your mother! The role of talker familiarity in infant streaming," *Cognition*, vol. 94, pp. B46–B53, 2004.
[10] L. ten Bosch, O. Räsänen, J. Driesen, G. Aimetti, T. Altosaar, L. Boves, and A. Corns, "Do multiple caregivers speed up language acquisition?" in *Proceedings of Interspeech 2009*, 2009.
[11] M. Ostendorf, "Moving beyond the 'beads-on-a-string' model of speech," in *Proceedings IEEE ASRU-99*, Keystone, Colorado, 1999.
[12] C. Snow and C. Ferguson, *Talking to children: language input and acquisition*. Cambridge, New York: Cambridge University Press, 1977.
[13] E. Newport and R. Aslin, "Learning at a distance: I. statistical learning of non-adjacent dependencies," *Cognitive Psychology*, vol. 48, pp. 127–162, 2004.
[14] D. Lee and S. Seung, "Learning the parts of object by non-negative matrix factorization," *Nature*, vol. 40, pp. 788–791, 1999.
[15] P. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.
[16] J. van de Weijer, "Language input for word discovery," Ph.D. dissertation, Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands, 1998.
[17] H. Van hamme, "HAC-models: a Novel Approach to Continuous Speech Recognition," in *Proceedings Interspeech 2008*, Brisbane, Australia, 2008.
[18] 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.
[19] M. Tomasello, *Constructing a language – a usage-based theory of language acquisition*. Harvard University Press, 2003.