

Learning meaningful units from multimodal input – the effect of interaction strategies

Louis ten Bosch
Radboud University
Erasmusplein 1
Nijmegen, NL
+31243616069

l.tenbosch@let.ru.nl

Lou Boves
Radboud University
Erasmusplein 1
Nijmegen, NL
+31243616069

l.boves@let.ru.nl

Okko Räsänen
Helsinki University of Technology
Otakaari 5A
Espoo, Finland
+35894512499

okko.rasanen@tkk.fi

ABSTRACT

This paper describes a computational model of language acquisition based on meaningful interaction between an infant and its caregivers. Learning takes place in an interactive loop between (virtual) caregiver and (virtual) learner who only uses general and cognitively plausible learning strategies and who does not rely on unrealistic prior knowledge about linguistic categories. In this work, the model is used to study the effects of different attentional factors in learning of word-object pairing during learner-caregiver interaction.

Categories and Subject Descriptors

H.1.2 [Information Systems] User-Machine Systems – *Human Information Processing*; I.2.6 [Computing Methodologies] Learning – *Concept Learning, Knowledge Acquisition*; I.6.m Simulation and Modeling – *Miscellaneous*

General Terms

Algorithms, Human Factors, Theory.

Keywords

Interaction, learning, language acquisition

1. INTRODUCTION

Most (human) learning happens as a side effect of interaction, often between high- and lower-proficient participants. Language learning, which takes place through interaction between infant and caregivers, is a clear example. Caregivers are usually high-proficient users of the language that is learned by the infant.

Even if learning happens in a situation where a beginner interacts with one or more competent ‘agents’, several conditions must be distinguished. These conditions depend on the way in which errors that are made by the lower-proficient agent are corrected by the higher-proficient agent, and on the way the lower-proficient-agent is paying attention to the input from the higher-proficient agent. In language acquisition the caregivers may or may not

explicitly correct ‘errors’ of the infant, and the infant may or may not accept every sensory stimulus that it perceives as relevant. For example, an infant might hear an utterance from the caregiver, while at the same time not paying attention to exactly those objects referred to in that utterance. It will be evident that the way how and to what extent errors are corrected and to what extent information in a stimulus is processed will affect the eventual learning result and the shape of the learning curve. Literature on first language (L1) acquisition (see e.g. Kuhl, 2004; Houston & Jusczyk, 2000; Jusczyk & Aslin, 1995; Singh et al., 2004; Newman, 2008) suggests that young children are not very sensitive to systematic correction – but a recent longitudinal study suggests that word learning can be supported by subtle tuning by caregivers (Roy, 2009). For L2 acquisition, it is often assumed that error correction during language acquisition may affect the *rate* of learning; the *stages*, however, remain unaltered.

Given these findings, it is interesting to connect observations from language acquisition on the one hand with a study about the effect of interaction strategies on learning performance on the other hand. Since language acquisition is closely related to the detection of potentially meaningful units (words, word-like units), we can make a bridge by investigating the effects of different interaction strategies on the learning performance shown by a computational model of language acquisition which focuses on the detection of words.

In this paper, we explore this idea by investigating the impact of different learning strategies on the performance of a specific computational model. The model, developed in the ACORNS project (www.acorns-project.org), simulates language acquisition as a process in which infants learn associations between speech signals and objects or events in their environment. The model is extensively described in the literature (e.g. ten Bosch et al., 2009abc; Boves et al., 2007; see also Stouten et al., 2007; Van hamme, 2008; Klein et al., 2008), and is briefly summarized in section 2 for the sake of clarity. The model assumes that learning takes place through interaction between caregivers and learner. Thus, we need to define one or more interaction strategies. In section 3, we discuss possible strategies and investigate the effects on learning. Sections 4 and 5 present an experiment and contain a discussion, respectively.

Although the model was designed for simulating the discovery of meaningful speech units, it may be useful in a wider perspective for the study of internal learning models and possibly of user modeling and adaptation.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ICMI-MLMI'09 Workshop on Child, Computer and Interaction
November 5, 2009, Cambridge, MA, USA

Copyright 2009 ACM 978-1-60558-690-8/09/11 ...\$10.00.

2. THE LEARNING FRAMEWORK

The model assumes a (virtual) learning environment in which a caregiver interacts with an infant. In each interaction cycle, the (virtual) caregiver presents a multimodal stimulus to the (virtual) learner. The learner processes this input and attempts to detect recurring auditory patterns in the speech signal and associate these acoustic elements to elements of the visual input. In this way, internal word representations are hypothesized and adapted during one training. To that end, the learner is able to extract features from the input signals, to encode and store the representations in its internal memory, to retrieve representations from its memory, and to produce a (virtual) response which is provided to the caregiver. After that, the next stimulus is presented to the learner.

In combination with the response (i.e. the hypothesis that a certain concept corresponds to the acoustic input), the learner can provide the confidence measure associated to that hypothesis. Each stimulus activates each of the internal representations according to the match between the signal and the internal model. Based on these internal activation scores, the learner can provide to the caregiver the confidence measure of a concept which indicates the level of certainty that the learner has about her response. The use of confidence measures opens the possibility of handling cases in which the stimulus is underspecified or inconsistent – for example, if the learner is sufficiently confident about a certain hypothesis, the learner may overwrite (or ignore) the information as present in the original stimulus, and instead believe in its own hypothesis. Used in this way, the confidence score is comparable to the way how humans (or infants) behave if they are use or not sure about their answer.

Considerable attention has been given to the cognitive plausibility of the design (*architecture*) of the model, especially concerning the data presentation (the input of speech and visual information), the data processing (Kuhl, 2004; Smith & Yu, 2008) and memory structure (Baddeley, 1986; Bar, 2007; see also Lewkowicz, 2002). The learner makes use of two basic principles that play a major role in language acquisition (e.g. Smith & Yu, 2008): detection of recurrent patterns in the speech signal, and cross-modal association between co-occurring acoustic and visual patterns (also called ‘form-referent pairing’). The learning starts without prior knowledge about speech – for example, the learner does not know about specific language-dependent sound inventories, nor does it know about words. Also the processing itself is not assumed to be speech specific or language specific – the learning algorithms are based on general cognitive principles (see also Thelen et al., 1995; Grabowski et al., 2007; Markovitch & Lewkowicz, 2004).

In ACORNS, we have designed and tested three different computational approaches for word detection from multimodal data: Non-Negative Matrix Factorisation (NMF, e.g. Van hamme, 2008), DP-Ngrams (Aimetti, 2009; Aimetti et al., 2009) and Concept Matrices (e.g. Räsänen et al., 2009). For the sake of clarity, one of the approaches, Concept Matrices (CM), will be discussed in more detail here.

2.1 Concept Matrices

CM is a technique able to find structure in data by *discovering* and *memorizing* associations between internal states of the learning system and multimodal external data. The input for the technique consists of a time series of discrete elements or sampled

spatial information to form one-dimensional sequences, and in the training phase, tags specifying some events associated with these sequences. These discrete elements may be based on e.g. the use of a vector quantization (VQ) codebook. The concept tags are discrete elements (in our case integer values) that represent invariant outputs of another perceptual modality than auditory perception. For example the tags may represent information from the visual or haptic modality (Räsänen et al., 2008, 2009).

In this way, CM is able to combine information from the combination of modalities to boost the detection of potentially meaningful patterns in one of these modalities. More generally, the method allows construction of statistical associations between different modalities. As mentioned above, this association is one of the key aspects in learning of meaning (by agents and humans). During *training*, when a label sequence s and a corresponding concept tag sequence c is presented, the algorithm starts to collect frequency data regarding the occurrences of label pairs in the sequence at specific temporal lags. This ‘bigram’ data is stored into histogram tables $T(l, c)$ specified by the lag l and c , i.e., a separate table exists for each tag at each lag, yielding a total of $N_l * N_c$ tables where N_c is the total number of all possible tags, and N_l denotes the number of used lags. This first step shares properties similar to those of the NMF-based HAC-model proposed by Van hamme (2008). In the next step, these tables T are normalized to an activation matrix $P(l, c)$ of size $N_q \times N_q$, where N_q is the size of the label codebook.

During *recognition*, the label transitions in a novel input sequence are used as weighted pointers to the activation matrices P . The activation level of a certain concept c at time t given a new input sequence s can then be computed by adding the probabilities of observing c according to the activation matrices P (see Räsänen et al., 2009, for mathematical details). This activation can be computed in parallel for all concepts in order to see what concept is most likely given the present acoustic input.

This procedure provides a temporally *local* activation estimate for each concept candidate. In many applications it is useful to examine the activation output in a larger temporal window since the events that are being recognized may spread over several subsequent time frames. One possible way by which good results were achieved is to apply a low-pass or median filter on all activation curves, in order to hypothesize a sequence of long-term winning concepts.

2.2 Dialogue

In the present implementation of the model interaction adheres to ‘ideal’ turn-taking behavior. By this we mean the following. In real life, natural turn taking between two human participants is characterized by a high number of interruptions, incomplete utterances, ungrammatical turns, and by specific discourse-dependent collaborative behaviour, such as mutual completion of a single phrase by the discourse participants. In contrast, ‘ideal’ turn-taking behavior as used here refers to interaction during which participants take turns without interruptions. The ‘ideal’ interaction is a sequence of single interaction cycles. Each interactive cycle consists of one stimulus from the (virtual) caregiver to the model, and the response of the model to the caregiver. The agents wait for the response of the other agent and do not interfere with each other’s process.

There is another difference between the interaction as used here and ‘natural interaction’. In the ‘ideal’ interaction, the auditory and visual input channels are always synchronized, while in a real interaction, the association between auditory information and visual information may be vague, asynchronous or even absent. Recent studies show that the form-referent pairing by young infants is supported by a consistent synchronized presentation of cross-modal information (Cogate et al., 2006), but that young infants are capable of making these associations also in cases where individual situations are more fuzzy (e.g. Smith & Yu, 2008 and references therein).

Despite and due to these simplifications, it is possible to investigate different interaction and learning strategies. These are described in more detail in section 3.

3. INTERACTION STRATEGIES

The simplest setting for the interaction between caregiver and infant is one in which it is assumed that the speech of the caregiver always refers to visible objects in the environment and the learner pays attention to those objects. Moreover, the learner assumes that the association between speech and visual representations in each multimodal stimulus is always ‘correct’. This ‘baseline’ strategy will be indicated as condition (strategy) A.

In a slightly more complex setting, the association between audio and visual input in the stimulus is always ‘correct’, yet the learner can make mistakes in the association; this setting is indicated as condition B. Condition B is more complex than condition A, since the learner can overrule the information that is presented during training on the basis of her own hypothesis.

The interaction complexity can be further increased when it can no longer be guaranteed that learner always looks at the objects referred to in the speech (condition C) or the learner looks at another object than the one referred to in the speech (condition D). Condition C is one in which the caregiver does not always provide *complete* multimodal stimuli, for example in the case of a single unimodal stimulus. Condition C is more difficult than condition B: in condition B it is left to the learner to hypothesize, while the stimulus itself is complete, correct and consistent; in condition C the learner is *forced* to hypothesize since not all stimuli are complete. Finally, condition D is the most challenging, because in this case stimuli may be misleading providing faulty information rather than just being incomplete.

Obviously, in conditions C and D the learner may or may not associate the speech with the ‘correct’ objects, depending on the confidence attached to such a cross-modal association. These settings in the learning and interaction strategies are strongly reminiscent of conditions used in *game theory* (e.g. Camerer, 2003).

4. EXPERIMENTS

In order to compare the different strategies on the learning result, we have conducted experiments with a fixed threshold for the confidence level in conditions B, C and D and a fixed proportion (20%) of non-ideal stimuli in settings C and D. Training and test sets were identical – the only difference between the experiments is the way the learner deals with the stimuli presented by the caregiver and the way in which the stimuli are presented to the learner.

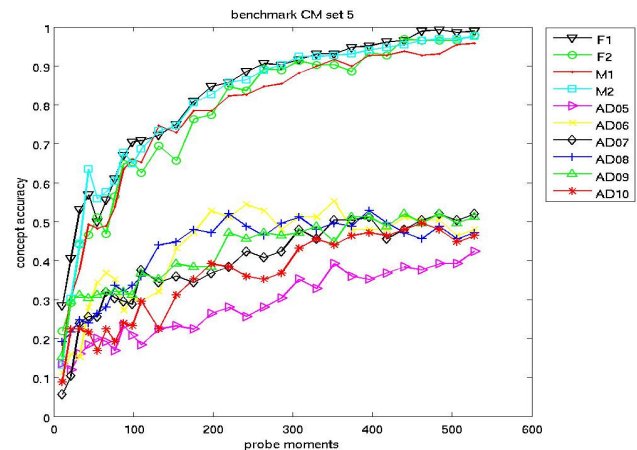


Figure 1. Results of the model using strategy A (condition A) on 10 different test sets (10 different speakers), using one fixed particular training set. There are ten learning curves - each curve is related to one of the test sets (i.e. one test speaker). One point (x, y) specifies the performance of the learner (y) on the test set after having observed x stimuli in the training set. (Results by Concept Matrix approach, Räsänen et al., 2009). As can be observed, four test speakers perform particularly well – these are exactly the speakers that are also present in the training set (indicated M1, M2, F1 and F2).

The training set consists of about 500 multimodal stimuli from four different speakers of (British) English: two male speakers (indicated M1 and M2) and two female speakers (F1 and F2). The number of target words (concepts) that are to be learned is 10 (so there are some 50 acoustic realizations for each of the concepts, about 12 per speaker).

Figure 1 presents a typical example of learning curves using strategy/condition A using this training set on 10 different test sets. Under the baseline condition A, the learner is able to discover associations between stretches in the speech signal and corresponding visual representations that are almost perfect after having processed some 500 interaction cycles (500 stimuli). However, the learned associations are highly speaker-dependent. When confronted with a new speaker (a speaker not earlier observed during training) the learner still makes a large number of errors. That can be seen in Figure 1: the 6 less performing speakers are those that are novel compared to the training set.

Figure 2 compares the use of different conditions A, B C and D on one of these 10 test sets, the test set associated to speaker M1 in fig 1. Therefore plot A (open circles) in figure 2 corresponds with the M1 plot in figure 1. As could be expected, among all conditions, condition A is the best with respect to learning rate and performance, and deviations from this condition A lead to a less favorable training. For example, for condition B performance starts lower but the eventual performance is comparable to condition A. For condition C, the learning rate is lower than condition B and performance drops significantly. An analysis of all the errors made shows that incomplete input stimuli are completed but at a price of introducing new errors, with no significant gain as net result. Condition D is worst: the learner makes about 30 percent errors, i.e. more than were in the input (20 percent). As could be expected, learning suffers more if the infant happens to focus on another object than the one referred to

in the speech utterance than when there is no visual object to accompany the speech.

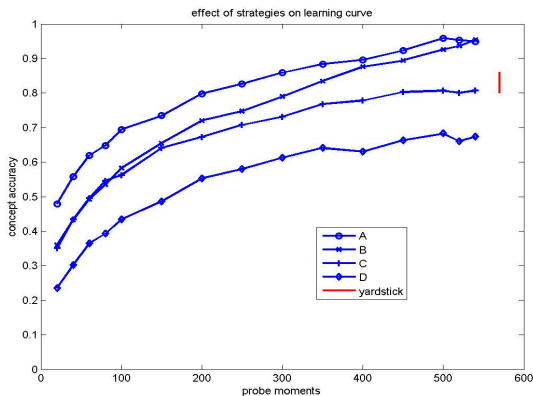


Figure 2. Comparable learning curves for conditions A, B, C, D. Significant differences are indicated by the red 'yardstick'.

The results show that the learning model can be used to investigate the effect of various learning schemes on the learning rate and eventual performance of the learner model. The results show that the model can support the study of alternative behavior during learning, of internal learning models and of improved user modeling and adaptation.

5. DISCUSSION

Although the model was designed for simulating the discovery of meaningful speech units, it may be useful in a wider perspective for the study of internal learning models and possibly of user modeling and adaptation.

We have shown that the baseline condition A is the best with respect to learning rate and performance. Deviations from this baseline condition lead to a less favorable training. For example, if the learner is less passive and is allowed to overrule information presented in the input (condition B), the learning curve starts lower than in condition A but the eventual performance is comparable to condition A. Apparently the learner has some problems with the bootstrapping of the learning, and probably with the internal evaluation of the 'confidence score' as well after having seen only a few data points.

For condition C, in which some of the information presented by the caregiver is incomplete, the learning rate is lower than condition A or B and the eventual performance drops significantly compared to condition A and B. And, as could be expected, condition D (deliberate inconsistencies in the input) is worst: the learner makes more errors than were in the input, implying that from an epigenetic point of view the training regime passed a critical boundary and has run into unstable regions in the learning state space (cf. Thelen & Smith, 1995).

The experiments show that the learning curve as well as the eventual learning performance significantly depend on the exact way how the caregiver and learner deal with the information and on the extent to which it is allowed to overwrite or ignore presented information.

In the future, it would be interesting to develop the learning platform further by incorporating simulations of actual consequences of communicative behavior instead of simple turn taking procedure ("correct", "wrong, try again"). It is clear that the result of a learning process depends on the way the information is presented by the teacher (in our experiments: caregiver), the way the learner deals with this information, and how errors made by the learner are handled by the caregiver. A rich set of strategies in the computational model would enable simulation studies where the needs and novelty seeking behavior of a learner would drive the learning process by itself instead of being dependent on 'passive' audiovisual perception. Behavioral consequences would "force" the learning algorithms to differentiate between perceptions that affect differently the state and rewards of the learner, whereas some other percepts in a specific context could be considered as equal. This way it is possible to study the development of categorical and semantic representations of the surrounding world.

Evidently, the design and implementation of such simulation platform in a plausible but yet flexible way is not a simple task. But the flexibility of computational models as a test bed for these and similar simulations is shown in this paper. Ultimately, the challenge is to derive useful information regarding real world learning processes, rather than building simulations where learning algorithms have very specific a-priori mechanisms for reverse engineering the expert designed learning environments.

6. ACKNOWLEDGMENTS

This research was funded by the European Commission, under contract number FP6-034362, in the ACORNS project, and by NWO, the Dutch organization for Scientific Research

7. REFERENCES

- [1] Aimetti, G. (2009). "Modelling early language acquisition skills: Towards a general statistical learning mechanism," in *Proc. of the Student Research Workshop at EACL, 2009*, pp. 1–9.
- [2] Aimetti, G., Moore, R.K., ten Bosch, L., Räsänen, O., and Laine, U. (2009). "Discovering keywords from cross-modal input: Ecological vs. engineering methods for enhancing acoustic repetitions," in *Proc. Interspeech*, Brighton, 2009.
- [3] Baddeley, A.D. (1986). *Working Memory*. Oxford: Clarendon Press, 1986.
- [4] Bar, M. (2007). "The pro-active brain: using analogies and associations to generate predictions," *TRENDS in Cognitive Science*, vol. 11, pp. 280–289, 2007.
- [5] ten Bosch, L., Van hamme, H., Boves, L., and Moore, R.K. (2009a). "A computational model of language acquisition: the emergence of words," *Fundamenta Informaticae*, vol. 90, pp. 229–249, 2009.
- [6] ten Bosch, L., Räsänen, O., Driesen, J., Aimetti, G., Altsaar, T., Boves, L., and Corns, A. (2009b). "Do multiple caregivers speed up language acquisition?" in *Proc. Interspeech*, Brighton, 2009.
- [7] ten Bosch, L., Driesen, J., Van hamme, H., and Boves, L. (2009c). "On a computational model for language

- acquisition: modeling cross-speaker generalisation,” in *Proc. Text Speech and Dialogue*, Plzen, 2009.
- [8] Boves, L., ten Bosch, L., and Moore R.K. (2007). “ACORNS - towards computational modeling of communication and recognition skills,” in *Proc. IEEE-ICCI*, 2007.
- [9] Camerer, C.F. (2003). *Behavioral game theory: experiments in strategic interaction*, Princeton University Press.
- [10] Gogate, L.J., Bolzani, L.H, and Betancourt, E.A. (2006). “Attention to maternal multimodal naming by 6- to 8-month-old infants and learning of word-object relations,” *Infancy*, vol. 9(3), pp. 259–288, 2006.
- [11] Grabowski, L., Luciw, M., and Weng, J. (2007). “A system for epigenetic concept development through autonomous associative learning,” in *IEEE 6th International Conference on Development and Learning*, 2007, pp. 175–180.
- [12] Van hamme, H. (2008). “HAC-models: a novel approach to continuous speech recognition,” in *Proc. Interspeech*, Brisbane, 2008.
- [13] Houston, D., and Jusczyk, P. (2000). “The role of talker-specific information in word segmentation by infants,” *Journal of Experimental Psychology. Human Perception and Performance*, vol. 26, pp. 1570–1582, 2000.
- [14] Hoyer, P. (2004). “Non-negative matrix factorization with sparseness constraints,” *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.
- [15] Jusczyk, P. and Aslin, R. (1995). “Infants detection of the sound patterns of words in fluent speech,” *Cognitive Psychology*, vol. 29, pp. 1–23, 1995.
- [16] Lewkowicz, D. and Lickliter, R. (2002). *Conceptions of development: Lessons from the laboratory*. New York: Psychological Press, 2002.
- [17] Markovitch, S. and Lewkowicz, D. (2004), “U-shaped functions: Artifact or hallmark of development?” *Journal of Cognition and Development*, vol. 5(1), pp. 113–118, 2004.
- [18] Klein, M., Frank, S., van Jaarsveld, H., ten Bosch, L., and Boves, L. (2008). “Unsupervised learning of conceptual representations - a computational neural model,” in *Proc. 14th Annual Conference on Architectures and Mechanisms for Language Processing (AMLaP)*, Cambridge, UK, 2008.
- [19] Kuhl, P.K. (2004). Early language acquisition: cracking the speech code. *Nature Reviews: Neuroscience*, Vol. 5, pp. 831–843.
- [20] Newman, R. (2008). “The level of detail in infants’ word learning,” *Current directions in Psychological Science*, vol. 17, pp. 229–232, 2008.
- [21] Räsänen, O., Laine, U. K., and Altsaar T. (2008). “Computational language acquisition by statistical bottom-up processing,” *Proc. Interspeech’08*, pp. 1980-1983, 2008.
- [22] Räsänen, O., Laine U.K., and Altsaar T. (2009). “A noise robust method for pattern discovery in quantized time series: the concept matrix approach,” in *Proc. Interspeech*, Brighton, 2009.
- [23] Roy, D. (2009). New horizons in the study of child language acquisition. Keynote at *Interspeech 2009*, Brighton, UK.
- [24] Singh, L., Morgan, J., and White, K. (2004). “Preference and processing: the role of speech affect in early spoken word recognition,” *Journal of Memory and Language*, vol. 51, pp. 173–189, 2004.
- [25] Smith, L. and Yu, C. (2008). “Infants rapidly learn word-referent mappings via cross-situational statistics,” *Cognition*, vol. 106, pp. 1558–1568, 2008.
- [26] Stouten, V. Demuynck, K., and Van hamme, H. (2007). “Automatically learning the units of speech by non-negative matrix factorisation,” in *Proc Interspeech*, Antwerp, 2007.
- [27] Thelen, E., and Smith, L. (1995). “A dynamic systems approach to development of cognition and action,” *Journal of Cognitive Neuroscience*, vol. 7(4), pp. 512–514, 1995.