Language acquisition: the emergence of words from multimodal input

Louis ten Bosch and Lou Boves

Dept Language and Speech, Radboud University Nijmegen, NL
{l.tenbosch,boves}@let.ru.nl
http://lands.let.ru.nl

Abstract. Young infants learn words by detecting patterns in the speech signal and by associating these patterns to stimuli provided by non-speech modalities (such as vision). In this paper, we discuss a computational model that is able to detect and build word-like representations on the basis of multimodal input data. Learning of words (and word-like entities) takes place within a communicative loop between a 'carer' and the 'learner'. Experiments carried out on three different European languages (Finnish, Swedish, and Dutch) show that a robust word representation can be learned in using approximately 50 acoustic tokens (examples) of that word. The model is inspired by the memory structure that is assumed functional for human speech processing.

Key words: language acquisition, word representation, learning

1 Introduction

Language processing is one of the most complex cognitive skills of humans. Infants seem to acquire this skill effortlessly, but the large body of literature on cognition, language and memory shows that we only begin to understand the processes involved. Understanding speech is tantamount to mapping continuous speech signals to discrete concepts, which we are used to think of as a sequence of word-like elements. Infants learn to discover word-like units in speech without prior knowledge about lexical identities and despite the lack of clear word boundary cues in the signal.

In this paper we propose a computational model of word discovery that is able to learn new words and that is a plausible analogy of the way in which infants acquire their native language. Specifically, the model should explain how babies can learn new words by using already stored representations that are continuously adapted on the basis of new speech input.

Strangely enough, today automatic speech recognizers are the most elaborate computational model of speech processing. Contrary to virtually all psycholinguistic models ASR is able to handle the entire chain from speech signal to a sequence of words. However, current ASR algorithms certainly cannot claim any cognitive or ecological plausibility. At the same time, ASR systems perform substantially worse than humans [8] [13]. It is widely assumed that for closing the

performance gap new ASR training and matching paradigms must be explored, preferably inspired by cognitive models of human speech processing and language acquisition. In this direction, several attempts have been made, e.g. learning of words [12], language acquisition [15], and incremental learning [9]. However, none of the models proposed in the literature are able to learn, adapt and generalize patterns quickly and effortlessly to recognize new variants of known words and novel words [5] [16].

In this paper, we propose an embodied computational model for language acquisition that has similarities with the Cross-channel Early Lexical Learning (CELL) model [12], but differs from CELL in that it does not assume that infants represent speech in the form of a lattice of pre-defined phonemes. The current model avoids the use of pre-existing representation for decoding the information in the input signals. Instead, the representations in the model emerge from the multimodal stimuli that are presented to the model.

The structure of this paper is as follows. In the next section, we will discuss the main components of the proposed computational model and the overall architecture of the ACORNS model, while the third section deals with experiments. The final section contains a discussion and our conclusions.

2 An embodied model of word discovery

The (partially) embodied model of language acquisition and speech communication that we are developing in the FET project ACORNS [2] will contain four sub-modules, viz. sensory front-end processing, memory access and organization, information discovery and learning, and interaction in a realistic environment.

Front-end processing: In the first step, the computational model converts sensory input signals into an internal representation which is used in subsequent sub-modules for learning new patterns and for recognizing known patterns. Front-end processing may include the conversion of input signals into representation such as the MFCCs which are used in conventional ASR.

Memory organization and access: Cognitive theories of memory distinguish at least three types of memory: a sensory store in which all information is captured only for a very short time (in the order of 2 seconds), a short-term memory (also called working memory) that holds representations of sensory inputs and serves as a processing system that is able to compare new sensory inputs to previously learned patterns that are retrieved from a long-term memory. The model makes use of these types of memory and stores, retrieves and updates internal representations. At the same time this memory model supports the Memory-Prediction Theory, which holds that intelligent action is based on memorized perception-action loops.

Information discovery and integration: In the Memory-Prediction Theory it is assumed that multilayered representations are formed in which structure at a lower level map to structures at a higher level (abstraction). In the experiments reported in this paper the abstraction method is based on Non-negative Matrix Factorization (NMF) [7] [4] [14]. NMF is member of a family of computa-

tional approaches that represent data in a (large) matrix and use linear algebra to decompose this matrix into smaller matrices. These smaller matrices contain the information in the original matrix in a more condensed and abstract form. There are close similarities with Latent Semantic Analysis, e.g. [1]. By using NMF processes such as abstractions receive a clear interpretation in terms of linear algebraic operations. NMF is a powerful tool for discovering structure in speech data [14]. In later stages of the ACORNS project we will also experiment with other structure discovery algorithms.

Interaction and communication: In order to simulate a learning environment, the learner is endowed with the intention to learn words in order to maximize the appreciation it receives from the carer. This is done by translating the appreciation from the carer into a strive for correctly responding to the carer, which in turn is interpreted as the optimization of the interpretation of the stimulus presented by the carer. This optimization involves the Kullback-Leibler distance between the input and output of NMF.

2.1 Architecture

The ACORNS architecture (cf. Fig. 1) is based on recent psycholinguistic research in speech and language processing [6].

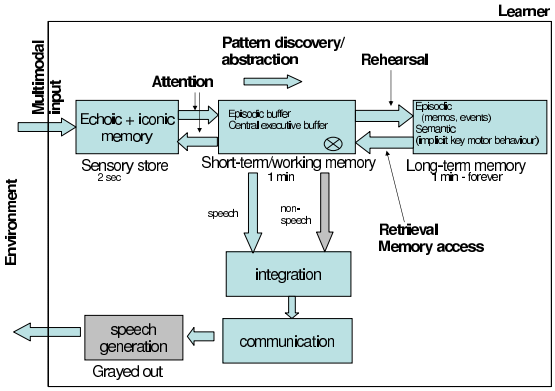

Fig. 1. Global architecture of the ACORNS system. Multimodal input is put into the sensory store. The sensory store, short-term/working memory and long term memory have different decay times. Two feedback loops are foreseen: one internal, governing the intrinsic processes and one external, in which the carer provides feedback to the model.

The learner receives multimodal input consisting of an audio stream (containing infant-directed or adult-directed speech) in combination with an abstraction of the visual modality (a visual 'tag'). This tag is provided in synchrony with

the speech signal. In this way, we simulate the presence of a visual sensory processing system. It is up to the learning agent to learn word-like entities from the repetitions in the audio signal and from cross-modally reoccurring systematic patterning.
In its current implementation, the model makes use of a simplified version of attention and rehearsal mechanisms. The attention and rehearsal mechanisms operate on representations stored in memory, and transform stored representations into possibly more abstract representations. We interpret attention as a process that reduces the part of the input stream that must be analyzed and is therefore indispensable to keep the computation load manageable, to reduce the storage into short-term (working) memory, and to reduce the ambiguity to be resolved during the search.

3 Experiments

A series of learning experiments has been conducted, inspired by phenomena observed in literature on language acquisition by young infants. In the first experiment, we have investigated the effect of a new speaker on the adaptation of already trained word representations. The arguments in favor of episodic representations used in psycholinguistics [3] suggest that different representation may be formed for different speakers. That means that representations that conflate episodes pertaining to several speakers corresponding to the same semantic object may only form on higher levels in the hierarchy. Another experiment deals with the effect of a new language (L2) on word representations that are trained on L1.

3.1 Material

For training and testing, three databases are available, in Dutch (NL), Finnish (FIN), and Swedish (SW). For each language we have utterances from 2 male and 2 female speakers. Each speaker utters 1000 sentences in two speech modes (adult-directed, ADS, and infant-directed, IDS), making a total of 2000 utterances per speaker. The set of 1000 sentences contains 10 repetitions of combinations of about 10 *target words* and 10 carrier phrases. (The content of the three databases differs in details that are not relevant for this discussion). The set of target words has primarily been chosen on the basis of literature on language acquisition.
For each utterance, the databases also contains meta-information in the form of a 'tag'. The tag represents abstract information and idealizes the input from other modalities. It translates to the presence or absence of vocabulary items in the audio stream. For example, the tag 'car' means that an object 'car' is referred to in the speech signal (and not that the *word* 'car' is pronounced). In the database, there is no information available about the words, phonetic content and position of words in the utterances.

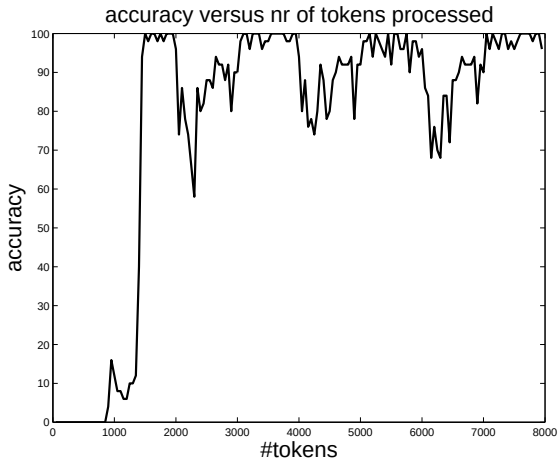

Fig. 2. (Dutch, speaker-blocked). Multimodal input data are presented blocked, speaker-by-speaker. Each time a new speaker starts (around number of tokens = 0, 2000, 4000, 6000), a drop in performance can be seen. Within about 1000 tokens (that is, approximately 100 tokens per word) the performance is back on its previous level. The decrease in performance is mainly due to different voice and speech characteristics which require an adaptation by the learning model.

3.2 Results

The result of the experiments are shown in figures in which the horizontal axis represents the number of utterances (tokens) presented during training. The vertical axis represents the accuracy of the learners replies. The accuracy is defined as the number of correct responses (defined by comparing the learners reply with the ground truth in the multimodal stimulus by the carer), divided by the total number of replies.

4 Discussion and conclusion

The computational model presented in this paper shows that learning relations between speech fragments and higher-order concepts can be accomplished with a general purpose pattern discovery technique. The performance of the learner depends on a number of factors - such as the ordering of the data (stimuli), the blocking per speaker, speaker changes, and multi-lingual training.

The learner is able to learn a limited set of concepts and classify a new stimulus in terms of one of these concepts. The learner needs a number of tokens before it can make a reliable representation. During the learning, it is able to gradually improve the quality of its internal representations, by minimizing the Kullback-Leibler distance between the observed data and the internal representations.

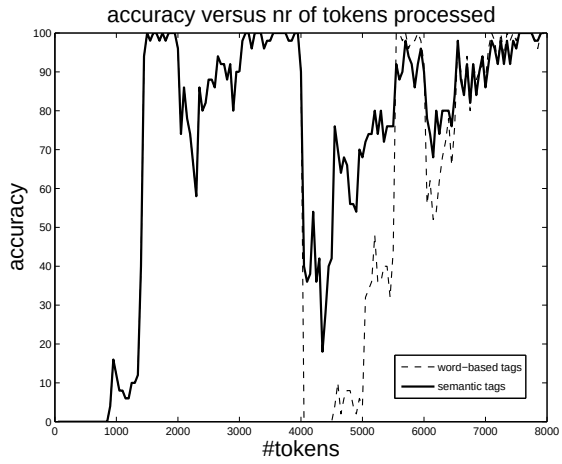

Fig. 3. (multilingual, speaker-blocked). Results of a multilingual experiment. First, two Dutch speakers are presented, then two Swedish speakers. The speakers are NL female, NL male, SWE female and SWE male. In the end the model is able to recognize Dutch and Swedish target words. The solid line represents the case were tags are language-dependent (word-based); the dashed line represents the ecologically more plausible case in which the tags are language-independent (semantic).

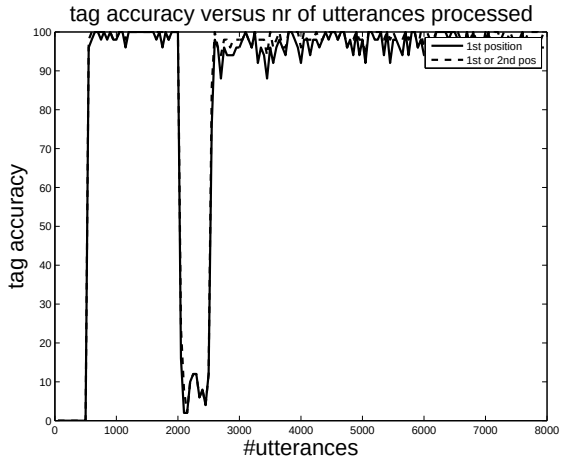

Fig. 4. The learner is first exposed to input from one Dutch speaker (the primary carer). After 2000 utterances from the primary carer, three other persons (2 males, 1 female) start interacting with the learner. The utterances from the new speakers are presented in random order. At first, the learner has difficulties adapting to new speakers, but it catches up after some 500 examples. The solid line indicates the accuracy of the 1-best reply of the learner; the dashed line shows the accuracy of the 2-best replies.

A second characteristic of the learner's behavior is the adjustment to a new speaker. As soon as a new speaker starts interacting with the learner, the internal representations are adapted to accommodate the speaker characteristics. Moreover, the learner reuses already stored representations whenever possible. This is particularly clear in the multilingual experiment based on the semantic tags (fig 4).
The speech database contains infant- and adult directed speech. This distinction is now not used, but other experiments have shown that the learner is able to distinguish these styles.
The computational model illustrates the relevance of various issues that are known to play a role in (models of) human speech processing. One of these issues is how words get activated (and to what extent), the second with the way how competition may act during the word search. In the current model, the activation of lexical items is separated from the actual competition. This is similar to Shortlist, one of the widely used computational models for human word processing [11]. Shortlist is a two-stage model in which activation of words by incoming speech input is separated from competition between the activated words. Other than Shortlist, however, the current model plays out the entire lexicon, while in Shortlist the network in which competition plays a role is constructed from only those words supported by the input. In the current model, competition is not explicitly implemented. Instead, it emerges from the parallel search among multiple candidates. This is in line with earlier findings e.g. obtained with another model of human word processing TRACE [10]. TRACE showed that competition is not a necessary consequence of parallel processing.

One of the research lines that will be pursued in the near future deals with the mechanisms that underly the emergence of words as a function of utterance-based training. In a genuine communicative setting, the learner must be able to learn not only from the presented multimodal stimuli, but also from the feedback that she receives from the carer. This will open the possibility of investigating the effect of corrective feedback on the learning process in more detail than is possible now. This will both enhance the ecological and cognitive plausibility of the computational model.

The second research line that will be exploited is also directly related to the cognitive plausibility. This research line deals with the use of *semantically related* tags that are presented to the learner in combination with the speech signal. In the current interaction model, the tags represent high-level references to objects that the learner receives and processes with 100 percent certainty. We aim at a model of a learner that receives multimodal input (speech and semantic tags) in such a way that the construction and adaptation of new representations is entirely controlled by the learner's internal learning mechanisms.

Acknowledgments. This research was funded in part by the European Commission, under contract number FP6-034362, in the ACORNS project (www.-acorns-project.org)

References

1. Bellegarda, J. R. (2000) Exploiting Latent Semantic Information for Statistical Language Modeling. Proc. IEEE, Vol. 88: 1279-1296..
2. Boves, L., ten Bosch, L. and Moore R. (2007). ACORNS - towards computational modeling of communication and recognition skills. Proceedings IEEE-ICCI 2007.
3. Goldinger, S.D. (1998). Echoes of echoes? An episodic theory of lexical access. Psychological Review 105: 251-279
4. Hoyer, P.O. (2004) Non-negative matrix factorization with sparseness constraints, Journal of Machine Learning Research, 5, 1457-1469.
5. Johnson, S. (2002) Emergence. New York: Scribner.
6. Jones, D.M., Hughes, R.W. and Macken, W.J. (2006) Perceptual organization masquerading as phonological storage: Further support for a perceptual-gestural view of short-term memory, J. Memory and Language 54, 265 281.
7. Lee, D.D., and Seung, H.S. (2001). Algorithms for non-negative matrix factorization. In Advances in Neural Information Processing Systems 13, 2001.
8. Lippmann, R. (1997) Speech Recognition by Human and Machines. Speech Communication, 22: 1-14.
9. Maloof, M.A., Michalski, R.S. (2004). Incremental learning with partial instance memory. Artificial intelligence 154, 95-126.
10. McClelland, J. L. and Elman, J. L. (1986). The TRACE model of speech perception. Cognitive Psychology, Vol. 18, 1986, pp. 1-86.
11. Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. Cognition, Vol. 52, 1994, pp. 189-234.
12. Roy, D.K. and Pentland, A.P. (2002) Learning words from sights and sounds: a computational model. Cognitive Science, 26: 113-146.
13. Sroka, J. J. and Braida, L. D. (2005) Human and machine consonant recognition, Speech Communication: 44, 401- 423.
14. Stouten, V., Demuynck, K., Van hamme, H. (2007). Automatically Learning the Units of Speech by Non-negative Matrix Factorisation. Interspeech 2007, Antwerp, Belgium.
15. Werker, J.F. and Curtis, S. (2005) PRIMIR: a developmental framework for of infant speech processing. Language Learning and Development, 1: 197-234.
16. Werker, J.F. and Yeung, H.H. (2005) Infant speech perception bootstraps word learning. TRENDS in Cognitive Science, 9: 519-527.