# On a computational model for language acquisition: modeling cross-speaker generalisation

Louis ten Bosch, Joris Driesen, Hugo Van hamme and Lou Boves

Dept Language and Speech, Radboud University Nijmegen, NL
ESAT, Katholieke Universiteit Leuven, Belgium
{l.tenbosch,boves}@let.ru.nl
{joris.driesen,hugo.vanhamme}@esat.kuleuven.be
http://lands.let.ru.nl

**Abstract.** The discovery of words by young infants involves two interrelated processes: (a) the detection of recurrent word-like acoustic patterns in the speech signal, and (b) cross-modal association between auditory and visual information. This paper describes experimental results obtained by a computational model that simulates these two processes. The model is able to build word-like representations on the basis of multimodal input data (stimuli) without the help of an a priori specified lexicon. Each input stimulus consists of a speech signal accompanied by an abstract visual representation of the concepts referred to in the speech signal. In this paper we investigate how internal representations generalize across speakers. In doing so, we also analyze the cognitive plausibility of the model.

**Key words:** language acquisition, word representation, learning

## 1 Introduction

Learning the sound patterns of new words is an essential part of learning language. In the literature on language learning, two fundamental processes are discussed that are needed to accomplish this task [13][20].

The first process is the word discovery and recognition from an acoustic point of view. Infants (until about six months old) are truly 'universal' learners without preference for a particular language or sound structure. Recent research strongly suggests that infants store a great deal of acoustic information about words [12][17]. For infants of 6-7 months old, the reliance on acoustic detail can even hamper the recognition of the same word when it is spoken by another speaker or in another speaking style ([11][19]). When infants are about one year old, they become better in generalizing across speaker and gender.

The second process in word discovery is cross-situational and cross-modal learning. Infants are mostly confronted with *multimodal* stimuli: they hear speech in the context of tactile or visual information that is associated with the information in the auditory channel (*word-referent pairs*). In *individual* word-scene

pairs, the relation between word and referent may be ambiguous, and it may only be revealed by accumulation of statistical evidence across many situational examples. [20] provides arguments for the hypothesis that young learners make use of *statistical* cues to associate acoustic patterns and referents.

How can word learning by young infants be modeled computationally? Currently, automatic speech recognizers (ASR) are the most elaborate computational models of speech recognition. Contrary to virtually all psycholinguistic models of speech processing, an ASR-based model is able to handle the entire chain from speech signal to a sequence of words. However, current ASR algorithms certainly cannot claim cognitive or ecological plausibility. Moreover, ASR systems perform substantially worse than humans [15] [21]. In [2][4][7] we have described a novel computational model that is designed in the framework of the ACORNS project [23]. The model bears some similarity with the CELL model developed by [18], but unlike CELL it does not assume (unrealistic) phonemic input representations for the speech signal. The ACORNS model is able to learn words in a cognitively plausible analogy of the way in which infants acquire their native language. Following [20], we model the detection of words by searching recurrent patterns in the speech modality, in such a way that hypothesized word-like units statistically correspond with visual information. In designing the model we have made an attempt to make it as cognitively plausible as possible with respect to processing and representations in memory. Our research pursues two goals: to better understand human speech recognition and to improve ASR by including essential knowledge from human speech recognition.

More often than not infants hear speech from a small number of speakers (their caregivers). This raises the question to what extent initial representations of word-like units are speaker dependent. Behavioral experiments suggest that infants do not readily generalize 'words' learned from one speaker to other speakers [17]. However, there are indications that recognizing other speakers is improved if infants learn from speech of several different speakers. In this paper, we investigate these issues by means of a number of simulation experiments.

The structure of this paper is as follows. In the next section, we will discuss the main components of the computational model. The following sections describe the experiment and results in more detail. The final section contains a discussion and conclusions.

## 2   Brief description of the model

An input stimulus in our model consists of a spoken utterance in combination with a visual representation of (some of) the objects referred to in the speech. For the time being, the visual representation of a referent (e.g. 'car') is represented by a low-dimensional vector containing visual features. This is provided in synchrony with the speech signal. The acoustic input consists of continuous speech; there is no word segmentation and there is no orthographic representation available for the learner. It is the task of the learner to find a coherent relation between acoustic forms (word-like units) and the referent.

The learning is incremental and takes place in a communicative loop between the learner and the 'caregiver'. An interaction takes place as follows. The (virtual) 'caregiver' presents one multimodal stimulus to the (virtual) learner. After the learner receives the input stimulus, structure discovery techniques are applied to hypothesize new and/or adapt existing internal representations for word-like units. This process is based on the stimulus and the information stored in the learner's memory (see e.g. [2][4][5][7] for more detail). During *training*, the learner uses *both* modalities of an input stimulus. In the *test*, only the auditory part of the stimulus is processed, and the learner responds with the hypothesized concept(s) that match(es) best with the utterance.

In the ACORNS project we are experimenting with several different structure discovery techniques [23]. In this paper we only report results obtained with Nonnegative Matrix Factorization (NMF) [10][14][22]. NMF is member of a family of computational approaches that represent input data in a (large) observation matrix $V$ and uses linear algebra to decompose $V$ into much smaller matrices $W$ and $H$ such that $V \approx W \cdot H$. After this decomposition, $W$ can be interpreted as a set of representations of speech units, while $H$ contains the associated internal activations. In combination, $W$ and $H$ represent the information in $V$ in a more condensed form ('reconstruction'). The number of columns in $W$ (and rows in $H$) is equal to the number of different internal representations. This number is a model parameter (see next section). The other dimension of $W$ is specified by the dimension of the input. In our experiments, an input utterance is coded in the form of counts of co-occurrences of Vector Quantization labels. This allows us to represent utterances of arbitrary length in the form of a fixed-length vector (this is the acoustic part of the stimulus). The *visual* representation of the stimulus is appended to the acoustic part to obtain its full vectorial representation.

NMF is reminiscent of Latent Semantic Analysis, e.g. [1]. It can be shown that NMF provides a clear interpretation of concepts such as *abstraction* in terms of linear algebraic operations [23]. Prior to a training, both $W$ and $H$ are initialized randomly, and during training both matrices are updated on the basis of the stimuli. In its conventional form NMF is a technique for decomposing matrices that comprise a large number (possible all) of input stimuli [14]. To make the approach more cognitively plausible, the decomposition algorithm has been adapted such that the influence of past stimuli decays exponentially over time. The decay is determined by a model parameter $\gamma$ (set to 0.99 – the closer to 1, the smaller the forgetting decay). Apart from details, the adapted NMF-update has the following form ($\Theta$ denoting an auxiliary matrix) ([8], cf [14]):

$$
\begin{aligned}
W_{new} &= W_{old} \cdot ((V/W_{old}H_{old}) * H_{old}^{tr}) + \gamma\Theta_{old} \\
\Theta_{new} &= W_{new} \cdot ((V/W_{new}H_{new}) * H_{new}^{tr}) + \gamma\Theta_{old}
\end{aligned}
\tag{1}
$$

It is left to the learner to decide how many different internal representations will be built. Also, we exercise no control over the 'contents' of the internal representations: they may represent phrases, words, sub-word units, etc. Input utterances are not kept in memory after the update of the $W$ matrix is complete. Therefore, in psycholinguistic terms our present model is not purely episodic.

However, it is possible to mix abstractionist and episodic representations by deferring the update of $W$ until after a certain number of input utterances have been received.

The virtual learner is endowed with the *intention* to learn words in order to maximize the appreciation (s)he receives from the caregiver. Here it is assumed that appreciation is a function of the proportion of utterances that are correctly understood. In addition, the learner tries to minimize the stress between the internal representations of speech units and the representation of a new utterance. This optimization boils down to the minimization of the Kullback-Leibler distance between the original input $V$ and the reconstructed $\tilde{V} = WH$ [10].

## 3    Experiment

In this section we describe the design of the experiments with which we tested the hypothesis that if the observed variation during training is small, the ability to generalize will be small as well, while a larger amount of variation observed during training will lead to a larger ability to generalize.

### 3.1    Data

A pre-recorded database of Dutch utterances was applied for constructing specific training and test sets ([4]). It consists of utterances with a simple syntactic structure, in analogy to infant-directed speech. For this experiment, the data was narrowed down to utterances with exactly one concept (target word). For example, one of the stimuli is 'daar is een auto' (English: 'there is a car'), in which 'car' is one of the target words. In total, there are 13 different target words, all inspired on the basis of child language inventories [24].

### 3.2    Procedure

Table 1 presents a summary of the experiments that have been carried out. Each row represents a separate experiment. The column 'training' indicates which speaker(s) are used by the model for learning the internal word representations.

**Table 1.** Overview of the experiments on generalization across speakers.

| experiment | training (600) | held-out test (400) | accuracy (%) | conf. intervals (5% 1%) |
|---|---|---|---|---|
| 1 | female 1 (F1) | F1 | 99–100 | - - - - |
| 2 | male 1 (M1) | M1 | 99–100 | - - - - |
| 3 | F1 | F2 | 85.25 | 82.33–88.17 81.13–89.37 |
| 4 | M1 | M2 | 87.30 | 84.56–90.04 83.43–91.17 |
| 5 | F1 | M1 | 61.80 | 56.24–65.90 55.36–66.76 |
| 6 | M1 | F1 | 67.10 | 63.57–70.82 61.51–72.44 |
| 7 | F1+F2 (600) | F1 | 89.00 | 86.43–91.57 85.36–92.64 |
| 8 | F1+F2 (1200) | F1 | 95.25 | 93.50–97.00 92.78–97.72 |

The column 'test' indicates the test speaker(s). Each experiment consists of one incremental training starting with blank memory, using 600 stimuli presented in randomized order. Each test set consists of 400 stimuli. Experiments 1 and 2 assume a single speaker as the primary caregiver for training *and* test. The other experiments represent different possibilities for speaker generalization. Experiments 3 and 4 deal with cross-speaker, within-gender generalization. Experiments 5 and 6 show cross-gender performance. The final two experiments 7 and 8 can be compared to experiment 1 and 3, but investigate the effect of mixing in other speakers during training. In experiment 7, the 600 training stimuli are a randomized combination of 300 training stimuli from F1 and another 300 from F2. In experiment 8, the training consisted of the *entire* combined set from F1 and F2 (1200 utterances). The number of columns in $W$ was set to 70, which is more than enough to build speaker-dependent representations of the 10 concepts presented to the model.
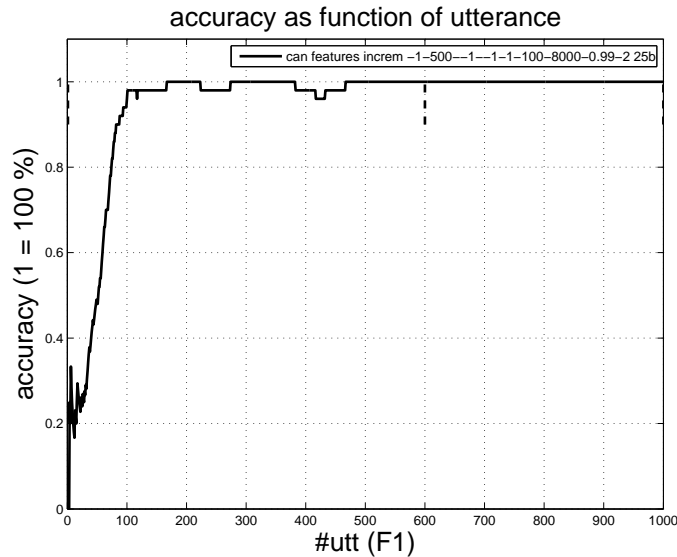


**Fig. 1.** Example of the learning performance of the computational model. The stimuli are taken from a single (female) speaker (F1). The horizontal axis indicates the stimulus index, while the accuracy of the learner (in terms of fraction of correctly recognized key words) is displayed along the vertical axis. The first 600 utterances are processed in incremental training mode, and the remaining 400 utterances are tested without any further training. Within about 200 utterances, that is, about 20–25 tokens per key word, the learner is able to build useful word representations.

## 4    Results

The results are shown in the column 'accuracy' of table 1. The accuracy is given as the percentage correctly recognized keywords in a test with 400 utterances. The confidence intervals (based on the student's t-test and a binomially distributed success-rate, both for the $p = 0.05$ level and the $p = 0.01$ level) are provided in the last column.

Not surprisingly, the model performs best in case of speaker-matched training-test conditions. This is true for both the female and the male speaker. Rows 1 and 2 in table 1 show that speaker-dependent learning is almost perfect. Rows 3 and 4 show that generalization to a new speaker of the same gender, however, already provides significantly worse results on the $p = 0.01$ level. Generalization across gender (experiments 5 and 6) is significantly worse than the performance in the same gender cases.

Experiments 7 and 8 show that learning is affected by the number of training speakers, and not necessarily in a positive way. If the model learns from speakers F1 and F2, recognition of F1 deteriorates compared to the situation (in experiment 1) where there is only one training speaker. The accuracy of 89% in experiment 7 cannot be attributed to lack of training data. Fig. 1 shows how learning proceeds for speaker F1 if she is the only person from whom the model learns. It can be seen that recognition performance is already close to 100% after some 200 utterances. In the presence of speaker F2 that performance level is not even reached after learning from 600 utterances.

## 5    Discussion

Most of the results in table 1 are unsurprising when considered from an Automatic Speech Recognition (ASR) point-of-view. The table shows that test results are best for within-speaker conditions, and deteriorate for the different-speaker same-gender condition, and are worst for different-gender conditions. However, we did not perform a conventional ASR training-test experiment. It was left to the learner to figure out how many different representations to build from the input utterances, and how to link these to the keywords. In our experiments the word-referent pairs were quite strongly tied. While this may seem to enhance learning, it might actually have had the opposite effect. From experiments 7 and 8 it appears that the learner tried to build speaker independent representations for the referents and that this slows down learning and results in sub-optimal performance.

A possible explanation for this behavior of the learner is the use of memory in the present NMF-based structure discovery approach. In all experiments reported above, a 70-column $W$ matrix was initialized randomly (more than 5 times the number of concepts to be learned). The close to perfect recognition performance on *independent* test sets observed in experiments 1 and 2 shows that the ratio between the number of concepts to be learned and the number of representations that can be built is not the culprit. More likely, the interference between two speakers is due to the fact that the learner did not try

to build (seemingly inefficient) speaker dependent representations, which would then only (much) later be reorganized into more abstract speaker independent representations.

For adults, the recognition of new speakers seems easy, but the underlying learning and adaptation processes are not well understood. The literature on language acquisition provides evidence that children must *learn to adapt*: they start storing a great deal of phonetic detail and gradually learn to generalize towards new speakers when they are between 7 and 12 months old. Recent theories in psycholinguistics assume that adaptation (by adults) can probably be best explained by hybrid models in which both so-called 'episodes' (detailed acoustic representations) and abstractions play a role ([9][16]). If we would exactly understand the processes involved, the 'gap' between human and automatic speech recognition could be narrowed in many realistic conditions.

When humans adapt to a new speaker, this adaptation process does not come at the cost of 'forgetting' older information that appeared to be useful in the past. Instead, learning to understand new speakers involves adapting existing internal representations if this does not destroy older information, and by creating additional representations if the new input differs too much from what was already in memory. In our model this can be accomplished if we change the present update procedure: instead of updating the $W$ matrix after every utterance the learner should estimate the degree of fit between the new utterance and the internal representations. Only if that fit is good enough, the update should proceed. In all other cases a new representation should be added to the $W$ matrix. However, implementing such a conditional update procedure is not straightforward. So far, we have not been able to define a reliable distance measure for the fit between an utterance and the internal representations. In [6] we have shown that reorganization of a large set of over-detailed representations (formed by enforcing the learner to create fully speaker dependent representations) can be implemented by means of clustering procedures.

So far, the ACORNS project has been more successful in applying knowledge from ASR and machine learning to elucidating models of human speech processing. However, we are confident that the underlying idea, viz. that the learning system should build its representations on the basis of its experience with input data rather than building models of pre-defined units, will result in automatic speech recognition systems that will rival human performance.

# References

1. Bellegarda, J. R. (2000). Exploiting Latent Semantic Information for Statistical Language Modeling. Proc. IEEE, Vol. 88: 1279-1296.
2. Van hamme, H. (2008). Integration of Asynchronous Knowledge Sources in a Novel Speech Recognition Framework, ISCA ITRW, Speech Analysis and Processing for Knowledge Discovery.

3. ten Bosch, L., Van hamme, H., Boves, L. (2008). Unsupervised detection of words questioning the relevance of segmentation", ISCA ITRW, Speech Analysis and Processing for Knowledge Discovery.
4. ten Bosch, L., Boves, L. (2008). Language acquisition: the emergence of words from multimodal input, in Sojka, P., Hork, A., Kopecek, I. & Pala, K. (Eds.) Text, Speech and Dialogue, 11th Intern. Conference, TSD 2008, Brno, pp. 261–268.
5. ten Bosch, L., Van hamme, H., Boves, L. (2008). Discovery of words: Towards a computational model of language acquisition, in: France Mihelic and Janez Zibert (Eds.) Speech Recogition: Technologies and Applications, Vienna: I-Tech Education and Publishing KG, pp. 205–224.
6. ten Bosch, L., Van hamme, H. and Boves, L. (2008). A computational model of language acquisition: focus on word discovery, Proc. Interspeech 2008, pp. 2570-2573.
7. Boves, L., ten Bosch, L. and Moore R. (2007). ACORNS - towards computational modeling of communication and recognition skills. Proceedings IEEE-ICCI 2007.
8. Driesen, J. and Van hamme, H. personal communication.
9. Goldinger, S.D. (1998). Echoes of echoes? An episodic theory of lexical access. Psychological Review 105: 251-279
10. Hoyer, P.O. (2004) Non-negative matrix factorization with sparseness constraints, Journal of Machine Learning Research, 5, 1457-1469.
11. Houston, D.M., & Jusczyk, P.W. (2000). The role of talker-specific information in word segmentation by infants. Journal of Experimental Psychology: Human Perception & Performance, 26, 1570 1582.
12. Jusczyk, P.W., & Aslin, R.N. (1995). Infants detection of the sound patterns of words in fluent speech. Cognitive Psychology, 29, 123.
13. Kuhl, P.K. (2004) Early language acquisition: cracking the speech code. Nat. Rev. Neuroscience, 5: 831-843.
14. Lee, D.D., and Seung, H.S. (2001). Algorithms for non-negative matrix factorization. In Advances in Neural Information Processing Systems 13, 2001.
15. Lippmann, R. (1997) Speech Recognition by Human and Machines. Speech Communication, 22: 1-14.
16. McQueen, J.M., Cutler, A. & Norris, D. (2006). Phonological abstraction in the mental lexicon. Cognitive Science, 30, 1113–1126.
17. Newman, R.S. (2008). The level of detail in infants' word learning. Current directions in Psychological Science. Vol. 17 (3). 229–232.
18. Roy, D.K. and Pentland, A.P. (2002) Learning words from sights and sounds: a computational model. Cognitive Science, 26: 113-146.
19. Singh, L., Morgan, J.L., & White, K.S. (2004). Preference and processing: The role of speech affect in early spoken word recognition. Journal of Memory and Language, 51, 173189.
20. Smith, L., Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. Cognition 106 (2008) 15581568.
21. Sroka, J. J. and Braida, L. D. (2005) Human and machine consonant recognition, Speech Communication: 44, 401- 423.
22. Stouten, V., Demuynck, K., Van hamme, H. (2007). Automatically Learning the Units of Speech by Non-negative Matrix Factorisation. Interspeech 2007, Antwerp, Belgium.
23. http://www.acorns-project.org
24. http://www.sci.sdsu.edu/cdi/