Do Multiple Caregivers Speed up Language Acquisition?

L. ten Bosch[1]*, O. Räsänen*[4]*, J. Driesen*[3]*, G. Aimetti*[2]*, T. Altosaar*[4]*, L. Boves*[1]*, A. Corns*[1,2,3,4]

[1]Department of Linguistics, Radboud University Nijmegen, NL
[2]SPandH, University of Sheffield, UK; [3]ESAT, Catholic University of Leuven, Belgium
[4]Dept. Signal Processing and Acoustics, Helsinki University of Technology, Espoo, Finland
l.tenbosch@let.ru.nl

Abstract

In this paper we compare three different implementations of language learning to investigate the issue of speaker-dependent initial representations and subsequent generalization. These implementations are used in a comprehensive model of language acquisition under development in the FP6 FET project ACORNS. All algorithms are embedded in a cognitively and ecologically plausible framework, and perform the task of detecting word-like units without any lexical, phonetic, or phonological information. The results show that the computational approaches differ with respect to the extent they deal with unseen speakers, and how generalization depends on the variation observed during training.

Index Terms: Language acquisition, Computational modeling

1. Introduction

Language acquisition involves the discovery and representation of linguistic units from situated speech. There is evidence that infants start their language acquisition process by storing a large amount of acoustic/prosodic detail [3][4]. As a result, the 'early' representations would contain a large amount of speaker-dependent detail, which may impede the ability to recognize a 'known' word spoken by an unfamiliar speaker [6]. Thus, infants must learn to generalize speaker-dependent representations towards other speakers.

The discovery of word-like units is guided by cross-modal association (*word-referent pairing*). Infants receive *multimodal* stimuli: they hear speech in the context of tactile or visual information that is associated with the information in the auditory channel. Although for *individual* stimuli the relation between word and referent may be ambiguous, the accumulation of statistical evidence across many situational examples may facilitate the generalisation of acoustic representations [7].

In this paper we compare three computational approaches of language learning under development in the ACORNS project[1] with the aim to investigate the issue of speaker-dependent initial representations and subsequent generalisation.

The structure of this paper is as follows. In the next section, we will briefly describe the simulated learning situation. The following sections describe three learning methods, experiments and results. The final section contains a discussion and conclusion.

Research funded by the European Commission, under contract FP6-034362, in the ACORNS project, and partly by the Dutch Organisation for Scientific Research NWO.

[1]http://www.acorns-project.org

2. Learning

Each input stimulus in our model consists of an *auditory* part (a spoken utterance) in combination with an abstract *visual* representation of the concepts referred to in the speech signal. It is the task of the learner to find a relation between acoustic forms (word-like units) and the visual referent without any lexical, phonetic and phonological information.

Learning takes place in a communicative loop between the learner and a 'caregiver' [1]. The caregiver presents one multimodal stimulus to the learner. For input stimulus a structure discovery technique is applied to hypothesize new and/or adapt existing sound-reference pairs. While *learning*, the system 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.

3. Comparison of three learning methods

In ACORNS we are experimenting with different structure discovery approaches: Non-negative Matrix Factorization (NMF) [2] [8], Concept Matrices (CM) [5] and DP-Ngrams [9]. All approaches are incremental and are able to discover recurrent structure in speech signals and to associate audio and visual information. The exploration of different learning methods in parallel is motivated by the fact that neither theories nor experimental findings on language acquisition suggest a unique computational process or implementation. On the computational level the three approaches aim at the same task: the discovery of word-like units by building and updating representations of sound-reference pairs. The main conceptual difference is the way in which the step is taken from subsymbolic to symbolic processing. CM looks for recurrent patterns in sequences of discrete frame-based codebook labels, and so relies on symbolic processing at an early stage. DP-Ngrams operates primarily on the surface forms of the signals and postpones the symbolic processing until late in the word discovery process. NMF takes an intermediate position. Another difference between the approaches is how information from the speech signal is processed. Both CM and DP-Ngrams deal with the speech signal as the acoustic information evolves over time, while NMF takes the *entire* utterance as input to create an internal representation of the utterance and finds structure in the speech signal by a decomposition afterwards.

All methods start with the same MFCC-based frame-by-frame 10 ms-spaced vector representation of the speech signal. During learning, the internal representations are updated after each new multimodal stimulus. In all methods, the short- and long-term memory is initialised randomly, and the number of

concepts that are to be discovered during the entire training is not specified beforehand.

3.1. NMF

NMF represents input data in a (large) matrix V and uses linear algebra to decompose this matrix into smaller matrices W and H. W can be interpreted as representations of speech units; H contains the associated activations. Matrices W and H approximate the information in V in a (highly) condensed form. The number of columns in W (and rows in H) is equal to the number of different internal representations. The other dimension of W is specified by the dimension of the input. In our NMF-experiments an input utterance is coded in the form of counts of co-occurrences of Vector Quantization labels. 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 all NMF experiments. This allows us to represent utterances of arbitrary length in the form of a fixed-length acoustic vector. For NMF, the *visual* representation of the stimulus is appended to the acoustic part to obtain its full vectorial representation.

3.2. CM

The Concept Matrix (CM) approach [5] is a statistical method for weakly supervised pattern discovery from time-series input. During training, it builds statistical models for VQ-label pairs, using frequency of different label-pair co-occurrences at different time lags, and determines which of these pairs are characteristic for a specific concept (in the visual modality). Once the learner has seen time-series data in parallel with the visual information, the algorithm can be used to recognize new input.

Since the algorithm does not make a Markov assumption about the independence of subsequent states, but rather integrates information along the temporal dimension, it achieves high robustness against noise and variation in the input. For each concept, a separate co-occurrence matrix is created at each lag, and these concept-specific matrices are updated only in the presence of the corresponding tag in the visual input [5]. When recognising novel input, activation values of transitions occurring in the input at different lags are retrieved from co-occurrence matrices and added together for each frame, leading to a temporal activation curve for each learned concept. The concept with the highest activation is considered as a recognition hypothesis.

A code book of 150 labels (only statics) and lags ranging from 10 ms up to 250 ms was used in these experiments.

3.3. DP-Ngrams

The DP-Ngram approach detects repeating portions of the acoustic speech signal through a dynamic programming (DP) technique (cf. [9]), and finds word-like units by associating them to the visual information. DP is used for isolated word recognition by finding the shortest distance between an acoustic input and a set of templates. However, the current method uses an accumulative quality scoring mechanism to reveal repeating sub-portions of two acoustic signals, called local alignments. By means of a classical DP step, for each pair of utterances a matrix D is defined with local (frame-to-frame) distance scores. The distance is Euclidean. By applying a recurrence relation on D [9], local 'quality scores' are calculated such that a high local quality score corresponds with a long 'local alignment'. These stretches are interesting because they relate to potential candidates of recurrent 'words'. Frame insertion and deletion penalties are applied during this recurrence. Finally, the optimal local alignment is discovered by backtracking from the highest local 'quality score'. Multiple local alignments can be discovered by repeating this process.

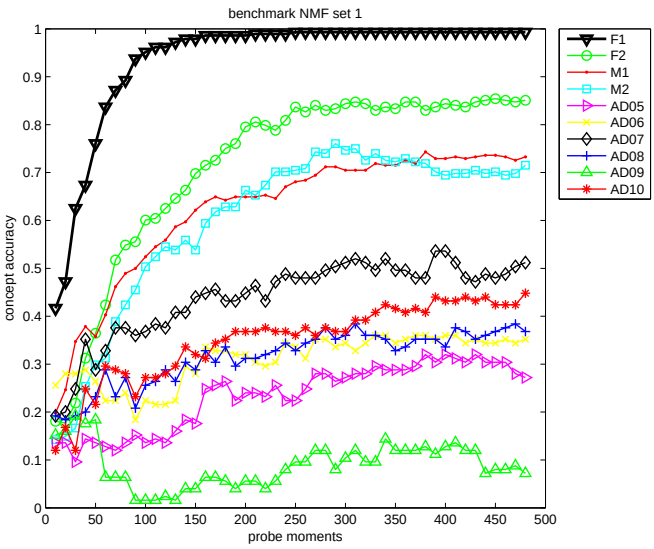

Figure 1: NMF, training set F1.

The internal representation of concepts are represented as a class of local alignments. Each class is constantly evolving with the accumulation of exemplar tokens, thus allowing the system to gradually become more robust to the variation.

4. Experiments

4.1. Data

In the experiments, training and test sets were carefully designed by selecting utterances from a database recorded in the ACORNS project [1]. All utterances have a simple syntax, similar to child-directed speech. The pool consists of 4000 English utterances spoken by two female (F1, F2) and two male (M1, M2) speakers (1000 utt/sp). Each of these utterances contains a single keyword, chosen from the following set: Angus, Ewan, bath, book, bottle, car, daddy, mummy, nappy, shoe and telephone. Each utterance is accompanied by an abstract symbolic tag (representing the information in the visual modality).

From this database, five different training sets have been created. These five different training sets are: F1, F1+F2, F1+M2, M1+M2, and F1+F2+M1+M2, the notation indicating the speakers present in the training set. The ordering of the stimuli (480 in F1, 520 in the others) within each training set was set up so that keywords would appear in a fixed and repeating order so as to produce a flat occurrence distribution. The number of examples per keyword in each training set was the same for each keyword and balanced per speaker. Each learning method (CM, NMF, DP-Ngrams) was applied to each of the five training sets. During learning, word representations were built, and after each 20 training stimuli the model was *probed* by measuring its accuracy on 10 different test sets: 4 test sets (F1, F2, M1, M2) containing held-out data from F1, F2, M1, and M2, and 6 sets from additional speakers (AD05, 06, 07, 08, 09, 10). There are no out-of-vocabulary words in the test sets.

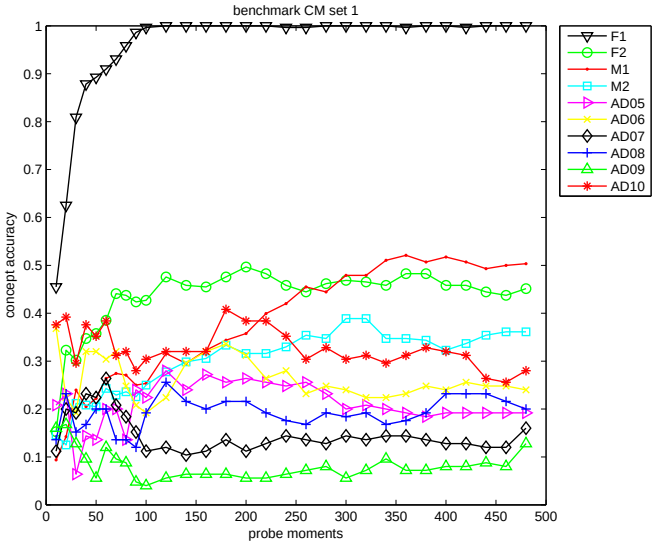

Figure 2: CM, training set F1.

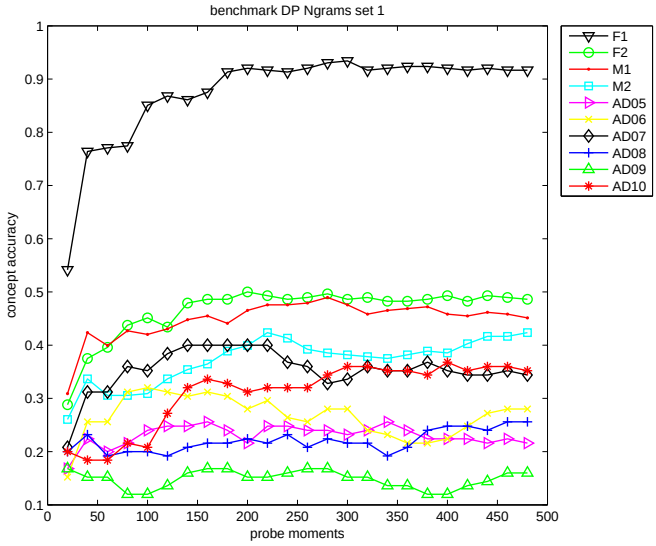

Figure 3: DP-Ngrams, training set F1.

Test sets did not overlap with any training set.

This set-up allows us to investigate the behaviour of the three different learning methods as a function of the variation present in training. We obtain 3 (number of methods) times 5 (number of training sets) times 24 (minimum probe moments during training) times 10 (number of test sets) (over 3600) accuracy measurements.

4.2. Results

For each learning method, the results show a clear tendency. For the sake of clarity, we have summarized the results in figures that represent the major findings and concentrate on F1 and F1+F2+M1+M2 (referred to as the 'full' set). Figure 1 and 2 show the results for NMF and CM on the F1, while figures 4 and 5 show results for the full set. Along the horizontal axes, the probe moments are specified. The 10 curves relate to the 10 test sets (across all figures they have the same symbols). The vertical axes show the concept accuracy. In Fig 1 and 2 we clearly see that the test speaker F1 profits from the fact that she is the single speaker in the training set F1. The methods however differ in detail how they handle the other nine speakers. NMF is significantly better than CM for F2, M1, M2 in the F1 training case (t-test, $N = 480$, $p \ll 0.01$). Furthermore, when we compare figs. 1 to 4 and 2 to 5 we observe that speakers F2, M1, M2 profit from full training in both cases, while F1 does not deteriorate.

In general, the 6 *additional* speakers that do not play a role in training also profit from the speaker variation during training: all their eventual scores are significantly better than in case of the F1-training. In general, NMF seems more sensitive to differences between speakers than CM appears: in all NMF-results the variation across speakers is larger than for CM.

Results are summarized in table 1: for both CM and NMF, speakers 05 to 10 do significantly better on the full set compared to set F1 (t-test per speaker, $N = 480$, $p \ll 0.005$).

DP-Ngram for learning from speaker F1 (cf. Fig 3) shows accuracies that are comparable to CM (table 1, columns 3 and 6). For 5 out of 10 speakers, DP-Ngrams outperforms CM (t-tests, $N = 480$, $p = 0.05$), while the opposite is true for the other speakers.

Table 1: Final results of the three learning approaches, for the 10 different test speakers. 'Full' refers to the training set F1+F2+M1+M2.

sp	NMF F1	NMF full	CM F1	CM full	DP-N F1
F1	0.99	0.96	1.00	0.98	0.91
F2	0.85	0.99	0.45	0.97	0.48
M1	0.73	0.92	0.50	0.95	0.45
M2	0.71	0.98	0.36	0.97	0.42
05	0.27	0.60	0.19	0.42	0.22
06	0.35	0.69	0.24	0.48	0.28
07	0.51	0.64	0.16	0.52	0.34
08	0.36	0.73	0.20	0.47	0.25
09	0.07	0.40	0.12	0.51	0.16
10	0.44	0.69	0.28	0.46	0.35

5. Discussion and conclusion

During language acquisition infants must learn to ignore perceptible but irrelevant detail in speech. Learning to understand other speakers than the primary caregivers (in most cases mother and father) is essentially related to learning to ignore these irrelevant aspects in the speech signal. It is argued that the variability in the input helps infants recognize which aspects are important and which can be ignored. As children gain more linguistic experience, they begin to learn which detail is relevant for distinguishing words, supporting the recognition of novel speakers [6].

All three learning approaches presented here show substantial differences between a one-speaker and multi-speaker training condition for new speakers. The approaches differ with respect to how information from new speakers is integrated into the internal models. Learning must find a balance between adaptation on the one hand and long-term accuracy on the other. From an ASR-standpoint these results seem straightforward: in

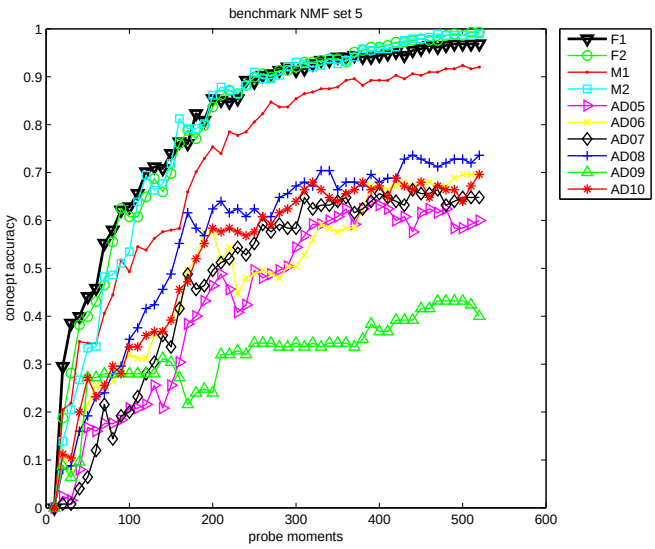

Figure 4: NMF, on the full training set F1+F2+M1+M2.

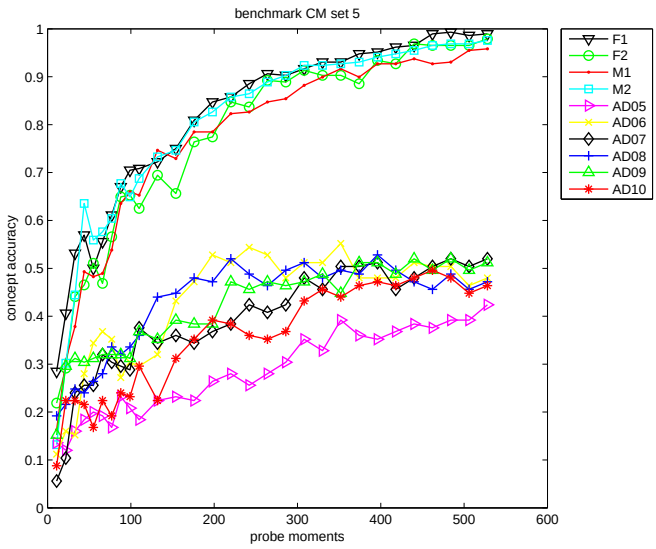

Figure 5: CM, training set F1+F2+M1+M2.

ASR multi-speaker training usually shows better results on new speakers. However, in ASR the training is always supervised and based on pre-existing knowledge about words and speech sounds. In our model the learner must discover sound-reference pairs without prior knowledge that would conflict with the requirement that learning must be plausible from a cognitive perspective. For example, in the case of NMF, new information could be redistributed across multiple columns of the W-matrix or dealt with by adapting just one specific W-column. That means that new information is *not* necessarily 'blended into' the existing internal model.

In summary, all learning approaches show the same tendency which supports the finding from behavioural experiments that a multi-speaker training condition helps to recognize speech from novel speakers. The approaches differ with respect to the degree the training speakers deteriorate. In the case of CM, none of the training speakers does significantly sacrifice in the end (fig. 5).

Conceptually, all three approaches have their own merit to be investigated in more detail. DP-Ngrams is a method able to hypothesize word-like units by strengthening internal representations on the basis of straightforward alignments between stretches of speech in different utterances. NMF needs the entire utterance to build a representation of the speech signal, but provides a powerful scheme in which bottom-up *and* top-down information in a multi-level hierarchy can be dealt with in a coherent framework. CM has an open architecture where the processes and internal representations are easily analyzable, and the internal representations actually predict input in the temporal domain.

Perhaps not surprisingly, our results with respect to the putative advantage of learning from multiple speakers for the recognition of new speakers are not completely conclusive. Our data suggest that learning from a speaker of a certain gender enhances performance for other speakers of the same gender, but that there may still be substantial differences between speakers of the same gender. It is still not very well understood how differences between speakers are best quantified.

In future work we will investigate learning schemes in which novel inputs may not cause the most similar existing internal representations to adapt; rather, additional representations can be built, which afterwards may or may not be merged with other representations that have the same semantic reference. Here, it is especially interesting to investigate the processing of new (out-of-vocabulary) words.

6. References

[1] ten Bosch, L., Van hamme, H., Boves, L., Moore, R.K. (2009). A computational model of language acquisition: the emergence of words, *Fundamenta Informaticae*, Vol. 90, pp. 229–249.

[2] Hoyer, P.O. (2004). Non-negative matrix factorization with sparseness constraints, *Journal of Machine Learning Research*, 5, 1457–1469.

[3] Jusczyk, P.W., & Aslin, R.N. (1995). Infants detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, 29, 1-23.

[4] Kuhl, P.K. (2004). Early language acquisition: cracking the speech code. *Nat. Rev. Neuroscience*, 5: 831–843.

[5] Räsänen O., Laine U.K. & Altosaar T. (2009). A noise robust method for pattern discovery in quantized time series: the concept matrix approach. *Proc. Interspeech 2009*, Brighton, England.

[6] Newman, R.S. (2008). The level of detail in infants' word learning. *Current directions in Psychological Science*, Vol. 17, 229–232.

[7] Smith, L., Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106, 1558-1568.

[8] Stouten, V., Demuynck, K., Van hamme, H. (2007). Automatically Learning the Units of Speech by Non-negative Matrix Factorisation. *Proc. Interspeech 2007*, Antwerp, Belgium.

[9] Aimetti, G. A. (2009). Modelling Early Language Acquisition Skills: Towards a General Statistical Learning Mechanism. *Proc. of the Student Research Workshop at EACL 2009*, pp. 1–9.