Project no. 034362

ACORNS

Acquisition of COmmunication and RecogNition Skills

Instrument: STREP
Thematic Priority: IST/FET

D5.3
A system capable of rapidly learning a large vocabulary

Period covered: from 1 December 2008 to 30 November 2009
Due date: 1 Dec 2009
Submission date 27 Nov 2009
Revision 1.25 (final)

Start date of project: 1 December 2006 Duration: 36 months

Project coordinator name: Prof. Lou Boves
Project coordinator organisation name: Radboud University, Nijmegen

Project co-funded by the EC, FP6, 2002-2006

Dissemination level

PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	

CO	Confidential, only for members of the consortium (including the Commission Services)	

Review history			Reviewed by
v0.1	Oct 30, 2009	First draft version	Els den Os Roger Moore Lou Boves
v1.20	Nov 9	Updated draft version	
v1.21	Nov 12	Updated draft version	
v1.22	Nov 16	Updated draft version	
v1.23	Nov 20	Updated draft version	Els den Os
v1.24	Nov 24	Updated draft version	Lou Boves
v1.25 (final)	Nov 27	Final	

TABLE OF CONTENTS

TABLE OF CONTENTS....3

OVERVIEW MILESTONES AND DELIVERABLES IN WP5, YEAR 3:....5

Milestones....5
Deliverables....5

TASK DESCRIPTION IN WP5....6

Task description, with focus on the third year:....6

1 EXPERIMENTS PERFORMED IN YEAR 3....7

1.1 Background....7
1.2 Towards the third year....8
References....10

2 DO MULTIPLE CAREGIVERS SPEED UP LANGUAGE ACQUISITION? 11

2.1. Introduction....11
2.2. Learning....12
2.3. Comparison of three learning methods
....12
2.3.1. NMF....13
2.3.2. CM....13
2.3.3. DP-Ngrams....13
2.4. Experiments....14
2.4.1. Data....14
2.4.2. Results....15
2.5. Discussion and conclusion....18
2.6. References....19

3. ON THE REPRESENTATION OF THE VISUAL CHANNEL....20

3.1 Introduction....20
Reference....21
3.2 Feature and conceptual encodings....21
3.3 Variation across types and tokens....22
3.4 Multiple referent decoding and evaluation....23
3.5 Feature matrices: F and C....24
3.6 Database, training, evaluation and baseline....25
3.7 Using binary visual (semantic) features in combination with object recognition27
3.8 Using binary visual (semantic) features without object recognition....30
3.9 Unobserved features....31
3.10 Irrelevant features....32
3.11 Fuzzy features....33
3.12 Conclusions....33

References....34
Appendix A....35
4 DEVIATING FROM THE 'IDEAL' INTERACTION....37
4A LEARNING MEANINGFUL UNITS FROM MULTIMODAL INPUT – THE38
EFFECT OF INTERACTION STRATEGIES....38
4A.1. INTRODUCTION....38
4A.2. THE LEARNING FRAMEWORK....39
4A.2.1 Concept Matrices....40
4A.2.2 Dialogue....40
4A.3. INTERACTION STRATEGIES....41
4A.4. EXPERIMENTS....41
4A.5. DISCUSSION....43
4A.6. ACKNOWLEDGMENTS....44
4A.7. REFERENCES....44
4B DEVIATING FROM STRICT SUPERVISION DURING TRAINING....47
4B.1 Introduction....47
4B.2 Semi-supervised learning....47
4B.3 Results....50
4B.4 Conclusions....52
References....52
5 ALTERNATIVE FEATURES....54
5.1 Introduction....54
5.2 MMFCC....54
5.3 SA-MMFCC....55
Discussion SA-MMFCC....56
6 CONCLUSIONS....58

Overview milestones and deliverables in WP5, Year 3:

Milestones

M5.6	Due M27	Specification of the experiments in the third year
M5.7	M34	Complete implementation of learning system
M5.8	M34	Release of open source memory-prediction based learning software

Deliverables

D5.3	M36	System capable of rapidly learning large vocabulary

Task description in WP5

WP5 is a workpackage with four 'main tasks'. In the Technical Annex, these tasks are described as follows.

Task 5.1 Creation of a platform for learning in the memory-prediction framework
In this task we will create the basic software environment that is needed to integrate the modules produced in WP1 – WP4 and to conduct experiments with language learning. We will provide the part of the system that generates the agent's responses. The platform will come in two versions: one for off-line experiments, and one that can be used for demonstrations.
Task 5.2 Multimodal integration
This task is dedicated to the development of procedures and software for the integration of speech input and visual input for disambiguating spoken utterances and feedback that is equivalent to hugging.
Task 5.3 Architecture for interaction
In this task we will design and implement a fully operational system that can conduct a multimodal dialogue, using perception-action loops on several parallel levels. Loops at the lowest level cater for latency-free communicative responses, without the need for parsing the semantic contents of an utterance. At the highest level the system must be capable of conscious reasoning.
Task 5.4 Experiments with language learning
Three major experiments will be performed, corresponding to three stages of language learning. In the first stage the system will learn basic communicative behaviour, mainly to show that is can engage in interaction. In the second phase the system will acquire a basic vocabulary, resulting in the emergence of sub-word units. In the third experiment the system will learn a larger vocabulary and basic rules of syntax.

Task description, with focus on the third year:

Task 5.1 Creation of a platform for learning in the memory-prediction framework
From the experiments done in year 2, it appeared that the three major computational approaches (NMF, CM, and DP-ngrams) all had their own merit and advantages. In the December 2008 SAC meeting and the previous review, it was recommended to explore these computational approaches in parallel. **As a result, the third year** was devoted to creating the MATLAB platform in three instantiations, each specifically focussing on each of the three computational approaches. The resulting platform can be used in a 'stand-alone' version for off-line experiments – this version can also be used for demonstrations.
Task 5.2 Multimodal integration
This task has been addressed by exploiting the way in which information from the audio channel and the 'semantic-visual' channel can/must be integrated. The specific questions here are the encoding of the visual information and the manner in which realistic ambiguity can be added to the visual channel.
Task 5.3 Architecture for interaction
In year 1 and 2, all experiments were done in such a manner such that the caregiver and learner were involved in simple turntaking. Also the interaction was such that the learner accepted the ground truth from the caregiver, such that the learning was essentially supervised. **In the third year** we have done experiments that open up the set of options during the interaction, to decrease the level of supervision and to enable the learner to cope with her own certainty and uncertainty levels.
Task 5.4 Experiments with language learning
All the previous tasks are addressed in experiments based on the speech corpora that were recorded in the ACORNS project. In accordance with the recommendations of the SAC meeting and the previous review, the focus was not any more on the learning of ever more words, but **on understanding the processes involved in learning**.

1 Experiments performed in year 3

1.1 Background

The goals and design of the experiments in year 3 (Milestone M5.6 Due D27) were based on the experience and experimental results obtained in year 1 and 2, and the recommendations in the previous review in January 2009.

In year 1 and 2 of the project, the ACORNS models of language acquisition have been tested on utterances with increasing complexity. In the first year, we started with using utterances collected in the first ACORNS database. This database (in the ACORNS documents referred to as 'Y1-database') contains utterances of which the structure has been inspired by the properties of speech that is addressed to (very) young infants. Infant-directed speech differs in many aspects from adult-directed speech and is mainly characterized by a limited lexicon, a simple syntactic structure, repetitions, exaggerated prosodic patterns and a somewhat lower speaking rate (e.g. Kuhl, 2004).
In the 'Y1-database', each utterance contains only one target word. A 'target word' is a word that the learner (in our case, an algorithm) is supposed to detect and learn from being exposed to these utterances; in the Y1 database these target words were nouns and proper names. The set of target words has been inspired by the contents of Communicative Development Inventories (CDIs, Bates, Bretherton, & Snyder, 1988; Fenson, Dale, Reznick, Thal, Bates, Hartung, Pethick, & Reilly, 1993; Fenson, Dale, Reznick, Bates, Thal, & Pethick, 1994; Tomasello & Bates, 2001) that are available for several tens of languages.
The Y1 database is recorded in four languages (FIN, SWE, English and NL).

Typical examples of utterances (English) are

I see a truck.
There is Daddy.
The car is nice.

Several papers were published on the basis of these first year data and experiments (www.acorns-project.org). In general, the experiments in year 1 showed that each of the three computational approaches (NMF, CP-Ngrams, and Concept Matrices) were able to detect word-like units from multimodal stimuli that were composed of an audio part and an (abstract) 'visual' part. At the input side, audio and visual ('grounding') information were always coherent: The keyword in the audio part of the stimulus always referred to the object that was abstractly encoded in the visual modality. In the first year, the emphasis was on the use of these algorithms in batch mode.

In year 2, the database has been extended using utterances with a more complex syntactical structure, in which more than one target words can appear. Also the overall number of keywords was increased from 10 to 50, and the set of target words was extended to include action verbs and adjectives in addition to nouns and proper names. The database has been recorded in three languages: Dutch, English and Finnish. Typical sentences in the Y2 database are

There I see a green frog and a truck
Where is the red aeroplane?

In the construction of the sentences, semantics did not play a role. It was avoided to have semantic clashes such as

Mum likes a green red apple
I see a big small aeroplane and a frog

but implausible constructions such as *happy aeroplane* were allowed.

Also in the case of these more complex utterances, the ACORNS models were able to find recurrent word-like units from multimodal stimuli. While for the 'simple' utterances an accuracy of 97 percent and beyond could be attained, the performance on the more complex utterances was about 90-95 percent (these figures depend on the method and on the details of the evaluation.). In the second year, DP-Ngrams, CM and NMF became available in incremental mode.

1.2 Towards the third year

On the basis of experiments performed in the second year, a number of issues were brought up for investigation in the third year. These issues were also inspired by the Scientific Advisory Board (SAC) meetings that took place end of 2008, and by the recommendations of the reviewers during the second review in January 2009.
The discussions and scientific advice led us to move the WP5 focus away from the learning of 'ever more words', as the original deliverable title ('System capable of rapidly learning large vocabulary') suggests. Instead, we have shifted the focus to gaining insight about the learning processes themselves and the internal representations. This will be clear from the issues addressed.

The first issue is how the different computational approaches could be compared. The second issue is to what extent it is possible to relate model results to findings described in the literature on language acquisition. One of the findings, reported in Newman (2008), provided a good example of an empirical result that could be compared and contrasted with results obtained by computational simulation. Newman's statement is that young infants are better in recognizing novel speakers if they have been exposed to more different speakers earlier. This observation is closely related to the current debate about episodic and abstractionist processing of speech (see e.g. McQueen, 2007).
It was decided to investigate these two issues in a single experiment in which the three computational approaches NMF, CM, and DP-Ngrams were compared in their ability to reproduce Newman's results. This addresses Task 5.4. It was not the intention to see which algorithms performs 'better' than other algorithms. Instead, the comparison was meant to gain insight in the different types of behaviour, based on different learning principles. This experiment is described in section 2 (and is published in Interspeech 2009, ten Bosch et al, 2009).

The third issue, which was also raised by the reviewers, concerned the type of encoding of the visual information. This directly addressed task 5.2. Visual (or 'grounding') information is presented to the learner in combination with the auditory information. This encoding issue already surfaced from the experiments based on networks such as Self Organizing Maps (see Klein et al., 2008) and by word learning experiments involving multilingual input in year 1. The question is to what extent the

grounding information in the visual domain was too 'crisp' and to what extent this information could be made less invariant and therefore more realistic. In year 3 we investigated what types of low level ('sensory') or high level ('conceptual') variability could be used to make the visual input more ambiguous. The experiments addressed the question how different categories of objects can be distinguished and how individual tokens within a category can be recognized. Ideas underlying feature encoding, as well as experiments with different feature encodings are discussed in section 3.

The fourth issue relates to the way caregiver and learner interact with each other. In most ACORNS experiments, we have assumed that the caregiver always presents complete and consistent stimuli. Each stimulus consists of an audio part and a 'visual' (grounding) part. In the experiments so far, the learner takes each stimulus 'as it appears'. That is, the learner assumes the stimulus to be consistent, and does not doubt the consistency between the modalities in the stimulus.
Many scenarios can be designed that deviate from this idealized-world scenario.

- At the caregiver side: the caregiver may present a certain proportion of stimuli that are inconsistent, in addition to others that are consistent.
- At the learner's side, an internal confidence mechanism may be active such that if the confidence about a self-generated hypothesis exceeds a certain threshold θ, ($0 \leq \theta \leq 1$), the learner assumes that its own hypothesis is true and discards the information in the grounding part of the input stimulus. The self-generated hypothesis is kept in memory for later reuse. This means that the learning becomes less supervised.

We have done experiments with the aim to investigate what happened when the interaction between caregiver and learner was modified towards less strict forms of supervision during training. This question addresses task 5.3, and is discussed in section 4. Section 4 consists of two related parts that both deal with deviations from the 'ideal' interactive setting:
4A Learning meaningful units from multimodal input – the effect of interaction strategies. Here we investigate four different interaction strategies between caregiver and learner. This section has been published in and presented at the WOrkshop for Child-Computer Interaction WOCCI-09 (ten Bosch, Boves & Räsänen, 2009).
4B Deviating from strict supervision during training. Here we show that a certain level of 'contrariness' at the learner's side helps to overcome inconsistencies in stimuli from the caregiver.

A final issue concerned the question what would happen in the case of realistic noise in the audio part. This task relates to both WP1 and WP5 . Since the computational approaches (NMF, CM, DP-Ngrams) differ with respect to how the acoustic features are handled during the learning, it is expected that they differ with respect to their robustness against background noise. It was expected that the novel features that were designed in WP1 show better performance in adverse conditions. WP1 designed two novel types of features ('modified'-MFCC and 'static adaptive'-MFCC); both these features were subject to various tests. Section 5 of this report briefly deals with the results. For all technical details regarding MFCC, the modified MFCCs (MMFCC) and the static-adaptive MMFCCs (SAMMFCC), the reader is referred to Deliverable D1.3.

References

Bates, E., Bretherton, I., & Snyder, L. (1988). From first words to grammar: individual differences and dissociable mechanisms. New York: Cambridge University Press. [Paperback edition issued 1991].

Fenson, L., Dale, P.S., Reznick, J.S., Thal, D., Bates, E., Hartung, J.P., Pethick, S., & Reilly, J.S. (1993). The MacArthur Communicative Development Inventories: User's Guide and Technical Manual. Baltimore : Paul H. Brokes Publishing Co.

Fenson, L., Dale, P., Reznick, J., Bates, E., Thal, D., & Pethick, S. (1994). Variability in early communicative development. Monographs of the Society for Research in Child Development, Serial No. 242, Vol. 59, No. 5.

Klein, M., Frank, S., van Jaarsveld, H., ten Bosch, L.F.M., & Boves, L. "Unsupervised learning of conceptual representations - a computational neural model", *Proc. 14th Annual Conference on Architectures and Mechanisms for Language Processing (AMLaP)*, 4-6 September 2008, Cambridge, UK

The MacArthur-Bates CDI. Online via http://www.sci.sdsu.edu/cdi/

McQueen, J. M. (2007). Eight questions about spoken-word recognition. In M. G. Gaskell (Ed.), The Oxford handbook of psycholinguistics (pp. 37-53). Oxford: Oxford University Press.

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

L.ten Bosch, O. Rasanen, J.Driesen, G. Aimetti, T. Altosaar, L. Boves, A.Corns. (2009) Do Multiple Garegivers Speed up Language Acquisition? *Proc. Interspeech* 2009. [pdf]

Louis ten Bosch, Lou Boves and Okko Rasanen "Learning meaningful units from multimodal input – the effect of interaction strategies", *Proc. Wocci2009*, [pdf]

Tomasello. M., & Bates, E., (Eds.). (2001). Language development: The essential readings. Oxford: Basil Blackwell.

2 Do Multiple Caregivers Speed up Language Acquisition?

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

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

2.1. 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 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.

2.2. 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.

2.3. 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.

2.3.1. 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, the deltas and delta-deltas) 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.

2.3.2. 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 cooccurrence 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.

2.3.3. 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.
The internal representations 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.

2.4. Experiments

2.4.1. 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 (denoted AD05, 06, 07, 08, 09, 10). There are no out-of-vocabulary words in the test sets.

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.

2.4.2. Results

Table I. Figure reference table

NMF, training set F1	Fig 1
DP-Ngrams, training set F1	Fig 2
CM, training set F1	Fig 3
NMF, full training set	Fig 4

DP-Ngrams, full training set	Fig 5
CM, full training set	Fig 6

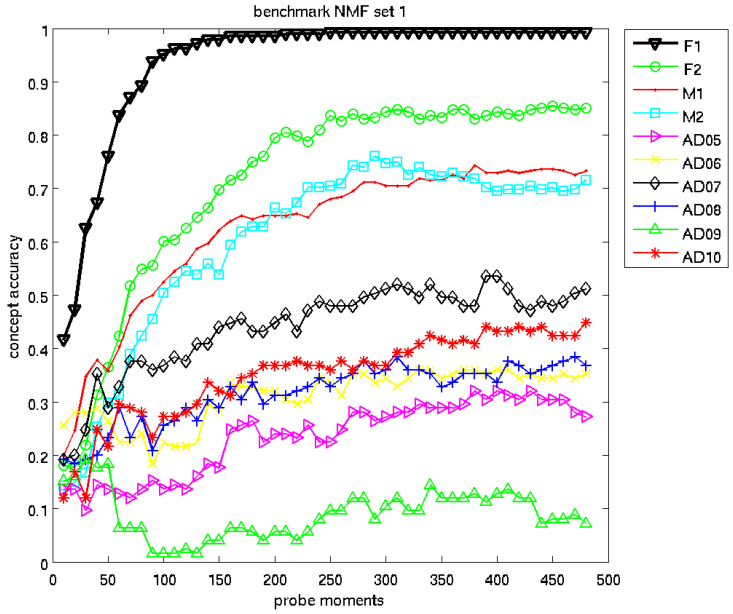

Fig 1. NMF. Training set F1.

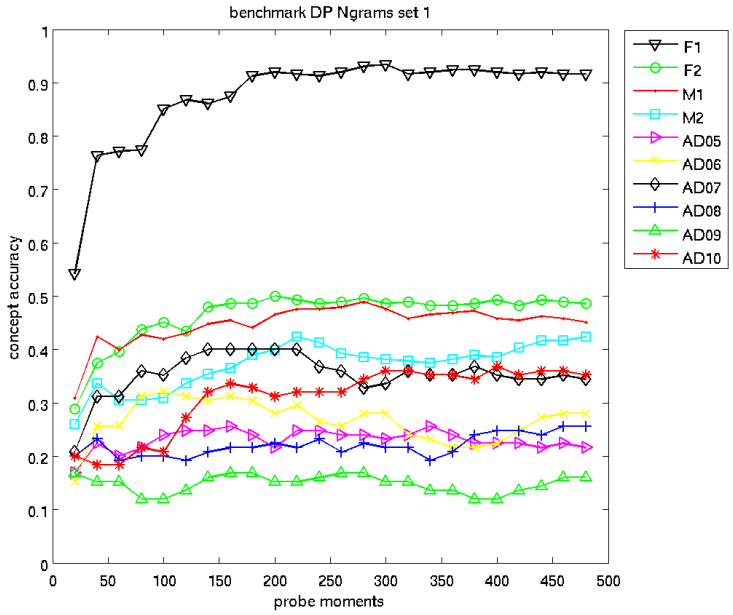

Fig 2. DP-Ngram, training set F1

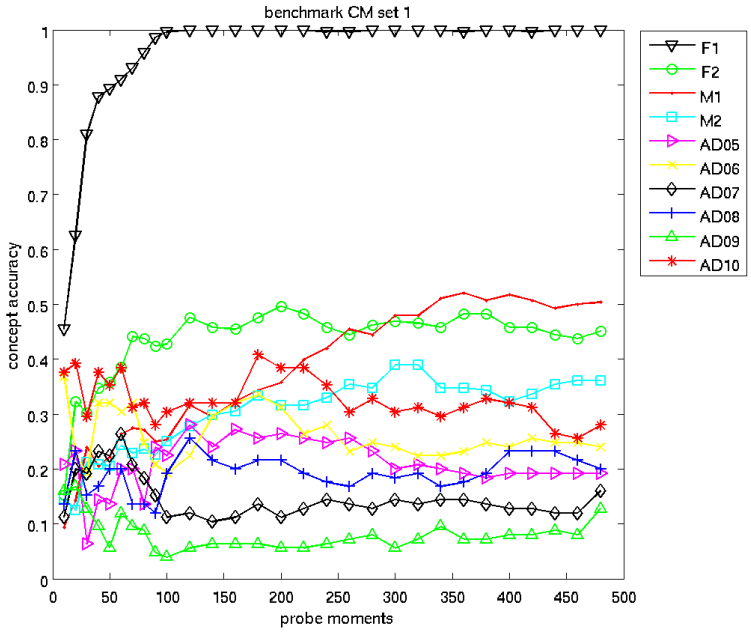

Fig 3. CM, training set F1

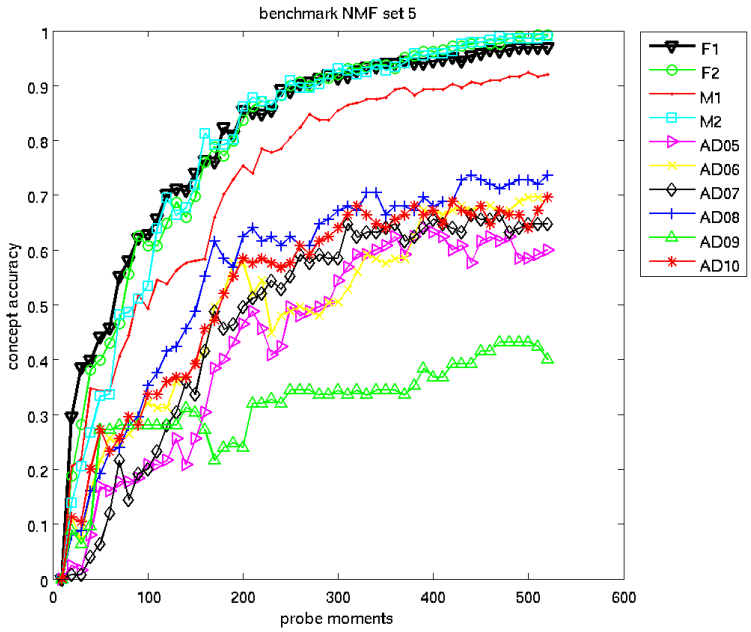

Fig 4. NMF, full set.

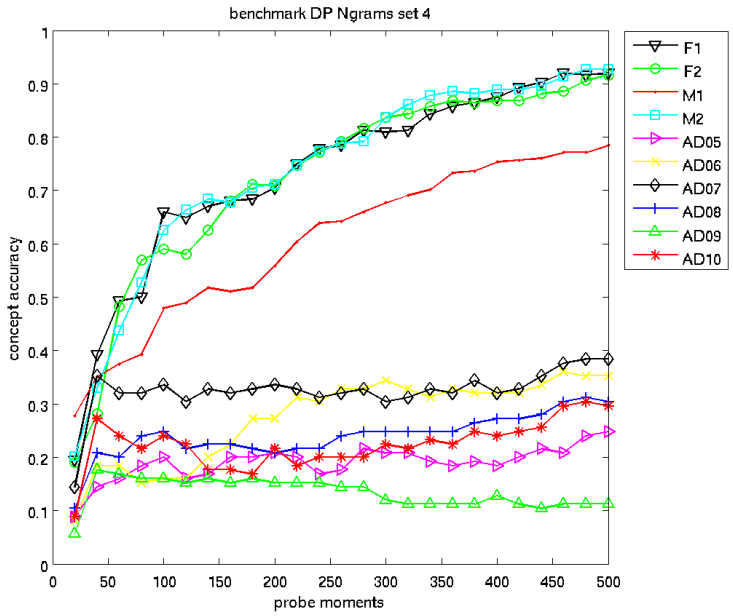

Fig 5, DP-Ngram, full set

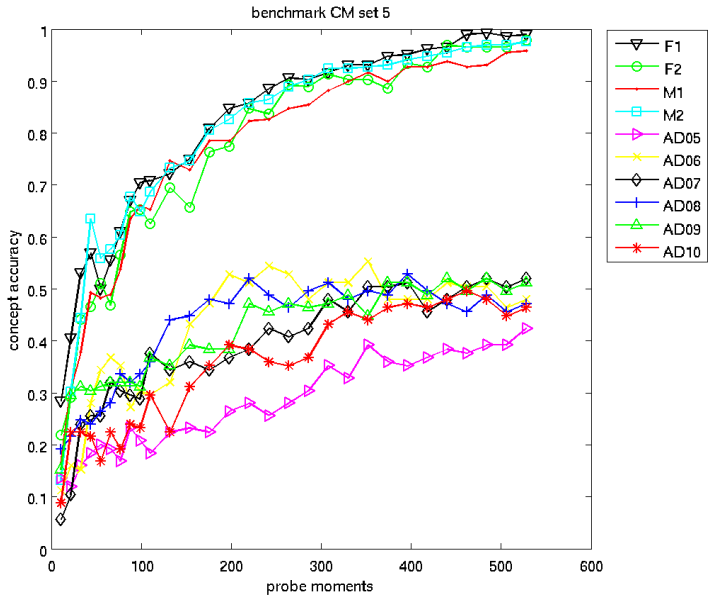

Fig 6. CM, full set

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, 2 and 3 show the results for NMF, DP-Ngrams and CM in the single-speaker training conditions, while figures 4, 5 and 6 show results for the full set (multi-speaker training condition). 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 Figs 1-3 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, $p < 0.01$). Furthermore, 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. For both CM and NMF, speakers 05 to 10 do significantly better on the full set compared to set F1 (t-test per speaker, $N = 480$, $p < 0.005$).

2.5. 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 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.

2.6. 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 wordreferent 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.

3. On the representation of the visual channel

This section consists of two parts. Subsections 3.1–3.4 (by Louis ten Bosch, Michael Klein, and Hugo Van hamme), discuss ideas about possible realistic representations for the information presented in the visual channel. The emphasis is on how one can go beyond the crisp, invariant visual feature encoding in a realistic and plausible manner. The following subsections (from 3.5 onward, by Hugo Van hamme) present and evaluate a number of NMF-based experiments with different visual feature representations.

3.1 Introduction

An issue that came up during the first year of ACORNS, in the multi-lingual experiments concerns the nature and impact of the symbolic, invariant ('crisp') character of the visual tags that were presented to the leaner during training. As a discussion issue, it came back during the second review.

The invariant symbolic nature of the keyword tags, as used in the early ACORNS experiments, had the following consequences:

- In the database, an utterance such as '*There I see a car*' was accompanied by the tag 'car'. In reality, an utterance '*There I see a car*' relates to an extra-linguistic referent 'car', more specifically a specific TOKEN of the type 'car'. So the first-year encoding assumed that *each* token of the type 'car' was represented by the same abstract invariant code 'car'. By construction, between-token variation was not modelled and could not be dealt with. For the modelling of learning processes this is obviously a disadvantage: Variation in the input is needed in order to be able to learn which visual features are relevant for recognizing a referent in the context.
- At the learner's side, the option of observing invariant tags theoretically allows the learner to 'cheat': The mere fact of observing a new symbolic tag will already tell the learner to create a representation for a new target word, thereby entirely ignoring the audio part.

The way from an invariant symbolic tag towards a more realistic, plausible and grounded representation is not evident and has far-reaching consequences for theory and algorithms. The exact encoding of the visual 'grounding' channel (that is presented to the leaner in parallel with the audio information in the speech channel) is important for almost all aspects of computational modelling and the concepts underlying the model. One of the questions is to what extent learning based on multimodal input is facilitated by the amount of information that is available in the parallel channel(s). This is an interesting issue since the language acquisition literature suggests that linking information is an issue *in itself*. For example, Saffran, Werker, Werner (2006) and several other recent studies show that when it comes to learning words, infants must learn to relate sounds with objects, and that this linking involves a substantial cognitive effort. This 'cognitive effort' is evidenced by experiments: There is massive evidence that tasks that only involve auditory discrimination are tapping a process that is less complicated than tasks that involve lexical tasks (and so involve linking between referents and sounds). It has been proposed that this corresponds to the difference between 'phonetic' and

'phonological' representations of words (the representation discontinuity hypothesis; Saffran et al, 2006, p 89).

Also in the ACORNS tasks carried out in the first two years, the learning of associations between acoustic and visual information on the one hand and the learning of recurrent acoustical patterns (units) are different aspects of the same learning process. Detecting recurrent units is possible without cross-modal association; however, cross-modal association facilitates the detection of recurrent units and of course also facilitates grounding.

Reference

Saffran, J.R., Werker, J., & Werner, L. (2006). The infant's auditory world: Hearing, speech, and the beginnings of language. In R. Siegler and D. Kuhn (Eds.), Handbook of Child Development. New York: Wiley (p.58-108).

3.2 Feature and conceptual encodings

We need to explain what we mean by making tags less invariant (more fuzzy, here sometimes referred to as 'descrispification'). There are multiple ways to introduce fuzziness in the visual domain. The following options represent two extreme positions among the gamma of alternatives.

a. **Feature encoding** (also called **visual feature encoding**). A referent is represented by a vector of binary features, where *each* component reflects *one* visual (or 'semantic') property of the referent.
b. **Conceptual** encoding (also called **canonical encoding)** A referent is represented by a vector that contains activations for a number of referents that were learned before.

A feature encoding might look like:

	legs	moving	round	eating	alive	soft	Etc
Car	0	1	0	0	0	0	Etc

while the conceptual representation for the same referent may look like:

	apple	aeroplane	daddy	truck	car	Etc
Car	0.08	0.12	0.01	0.58	0.98	Etc

Apart from the mathematical differences, both encodings also differ in the way they represent the real world. Feature encoding represents an encoding which is close to the lower-level visual or semantic properties of objects. Examples of such properties are 'roundness', 'color', 'moving', 'have legs'. Using feature encoding, one assumes that these 'raw' sensory data can be input for the learning approach. So, feature coding implies the assumption that there is no single unit representing the meaning of a word – instead, the grounding vector refers to a bundle of properties.
In contrast, the conceptual encoding assumes that the grounding information must be provided on a more cognitive level. In this case, the individual components of the

grounding vector do not refer to individual low-level properties of the referent, but to activations or similarities between the referent and other referents.

Both encoding systems can be considered as extensions of the crisp invariant tag coding: The use of crisp invariant symbolic tags is a special case of both systems. For example it is equivalent to the use of a unique conceptual encoding for each keyword. With crisp tags, the input is equivalent to a vector with one active unit, e.g. '0 0 0 1 0 0 … 0' with the fourth unit representing a certain concept such as 'car', or, alternatively, with a vector that has a non-zero activation for only one concept.
Both encoding systems are extensions of the crisp invariant coding, and they can handle variation. However, the systems differ with respect to how exactly variation between types and tokens can be handled. This is topic of the next subsection.

3.3 Variation across types and tokens

Both encodings deal in different ways with the between-type distance. That is already an improvement compared to the crisp tags. In the case of crisp tags, the between-type distances between (for example) 'man' and 'woman' and between 'man' and 'car' are basically the same. By using a suitable feature or conceptual encoding, the between-type distance for 'man'-'woman' may be smaller than for 'man'-'car' (but still the same for each *token* (instance) of 'man' or 'car'). In that sense, the (visual) feature encoding and the conceptual encoding are already more realistic than the crisp tags.

What about between-*token* variation? Both encodings can handle this, albeit in different ways.
The *feature* encoding can handle between-token variation by appending token-specific binary feature vectors. This appended pattern may be a random pattern. In such a case the feature encoding of a referent consists of a type-specific fixed part, augmented with a token-specific random part (indicated ***bold italic***)

	legs	moving	round	eating	alive	soft	Etc	***augm***	***augm***	***augm***	***Etc***
Car1	0	1	0	0	0	0	Etc	***0***	***1***	***1***	***Etc***
Car2	0	1	0	0	0	0	Etc	***1***	***1***	***0***	***Etc***

Introducing between-token variation in the case of visual features ('augmented binary visual feature encoding')

The *conceptual* encoding can handle between-token variation by adding random variations to the same type-specific conceptual encoding. In such a case, the conceptual encoding of a referent might for example consist of a sum of a type-specific fixed conceptual vector and a token-specific random vector.

	apple	airplane	Daddy	truck	car	Etc
Car1	0.083	0.114	0.007	0.56	0.979	Etc
Car2	0.076	0.125	0.008	0.59	0.991	Etc

Introducing between-token variation in the case of conceptual features ('fuzzy' conceptual encoding)

Between-type and between-token variation makes it harder for the algorithms to know whether two visually presented objects do indeed pertain to the same concept with the

same name, making it harder to know whether a new internal representation is required or not. This will be shown in the following sections, in which experiments are described in which different feature encodings are used. Subsections 3.5-3.8 deal with different experiments using different encodings without between-token variation. Between-token variation is introduced in 3.9-3.11. For example, the augmented binary visual feature encoding is presented in section 3.10, the 'fuzzy' conceptual encoding in section 3.11.

The use of these encodings can be motivated on a linguistic basis by the fact that under- and over-generalisation can be modelled by these encodings. For example, in language acquisition children sometimes under-generalize, i.e. they don't apply the word they know for an object if it seems significantly different from the tokens that we previously classified as that object (i.e. they might not call a white rose a rose, because so far they have only seen red roses). Also over-generalisation (children often name somewhat similar objects (e.g. a cat and a dog) with the same word) can be adequately dealt with by the proper use of either feature or conceptual encoding.

An example of feature encoding in which over-generalisation (cat → cat, dog 2 → cat) might appear is presented below.

	legs	moving	Animal	hard	alive	barks	Etc
Cat	1	1	1	0	1	0	Etc
Dog 1	1	1	1	0	1	1	Etc
Dog 2	1	1	1	0	1	0	Etc

3.4 Multiple referent decoding and evaluation

When one wants to apply feature or conceptual encodings, there are at least two additional issues to solve.

1. Above, we discussed the encoding of a *single* referent. How to encode an utterance in which several different referents occur, such as in 'There I *see* a *green duck* with an *aeroplane'* (all target words italicized)?
2. How to evaluate the learner? If the learner is reconstructing certain feature vectors or conceptual vectors, it is likely that these hypotheses are not exactly the same as what has been presented during training. What is the best metric to be used?

Multiple referent encoding

A sentence in the Y2 database may contain more than one (of 50) keyword. The encoding of a single keyword is straightforward, but the encoding of a scene in which multiple keywords make sense is not. For example, the encoding of 'There I *see* a *green duck* with an *aeroplane'* depends on the way how the word-based encoding for 'see', 'green', 'duck' and 'aeroplane' are combined into one single visual feature vector or matrix. In most experiments, we opted for the addition of feature vectors. That is, if a feature is present in more than one keyword, the corresponding values are

added. This addition option is used in the experiments described in the following subsections.
Another option (which is *not* tested here) would be to apply other ways to combine features, different from straightforward addition, for example based on visual scene analysis.
This multiple referent encoding issue addresses both the concept of the encoding as well as the way in which the back-end learning algorithm deals with this encoding.

Evaluation
The evaluation of tests in which complex encodings are used may become complicated. To explain this, we take NMF-based learning as a starting point. Of the two different encodings, a relatively simple case is the use of the high-level conceptual (canonical) features. In this case, the model estimates (reconstructs) the conceptual features from the acoustic part of the test stimulus. During the evaluation, the 'ground truth' (given in the original stimulus) can be compared to the learner's reconstruction. One may e.g. apply a 'winner takes all' strategy on the learner's hypothesis – this makes sense since the individual components refer to similarities of the presented referent with other referents.
In the case of the lower-level visual feature encodings, however, the model will attempt to reconstruct these low-level features (from the acoustic ones). What is a good metric to measure the (dis)similarity between the ground truth and the reconstruction as hypothesized by the learner in this case? There is no obvious relation between distances in the 'visual property domain' and the distances in the 'concept domain'. In other words, although one might measure the Euclidean (or other) distance between the true visual features and the estimated ones, it is difficult to interpret this distance in terms of the error rates that we are used to deal with (such as concept error rate, tag error rate). Moreover, we are not interested in the learner's capability to reconstruct individual properties of the referent (if that would make sense at all, based on audio) – instead, we are interested in the referent itself. This issue is addressed in more detail in subsections 3.5-3.8.

3.5 Feature matrices: F and C

As observed above, by using visual feature encoding one represents an object by a set of characteristics or *features* which hold or do not hold. A particular instance of an apple may have the features "round", "green", "eatable" and does not have the features "black", "four-legged" and "furry".
In this case, it is no longer possible to determine that an object is present from the observation of a single feature, e.g. the feature "round" may apply to the objects "ball" and "apple".

In order to perform experiments, we need explicit definitions for both visual features and conceptual features. For the ACORNS Y2 database, we defined a set of 64 binary visual/semantic features, listed in appendix A (at the end of this section), resulting in a feature matrix **F** (of size 64x50 = #features x #words). This matrix defines features for the 50 words occurring in the Y2 UK database. The visual-feature experiments will be using this feature matrix **F**.

For the conceptual/canonical representation, we will denote the grounding matrix by **C**, i.e. a matrix with 50 rows that has a "1" in entry *i,j* if the *i*-th keyword occurs in the *j*-th utterance and zeros elsewhere.

3.6 Database, training, evaluation and baseline

As already mentioned, the Y2 UK database has a 50 keyword vocabulary. These words can refer to objects, actions, colours or size; here they will be referred to as *keywords*. There are 9821 utterances for training and 3268 for testing. Each training utterance contains between 1 and 4 keywords (1994 utterances contain 1 keyword; 695 utterances contain 2 keywords; 3819 utterances contain 3 keywords and 3313 utterances contain 4 keywords). A similar distribution is observed on the test set. The number of keyword occurrences in the test data is shown in Figure 1.

Figure 1: number of keyword occurrences in the test set.

We use the Histogram of Acoustic Co-occurrence acoustic representation with the NMF learning method (Van hamme, 2008). The acoustical features used for learning are the WP1 MFCCs, which are quantized with codebooks of sizes of 150 (static), 350 (velocity) and 200 (acceleration)[1], each with lags 2, 5 and 9. The number of co-occurrence features is 555,000.
In the experiments described below, the NMF uses 75 internal representations (so W and $\mathbf{W}_g$ have 75 columns). The classical NMF equation reads:

$$\begin{bmatrix} \mathbf{G} \\ \mathbf{V} \end{bmatrix} \approx \begin{bmatrix} \mathbf{W}_g \\ \mathbf{W} \end{bmatrix} \mathbf{H} \qquad (1)$$

Hence, there are about 50 representations that should be associated with the keywords, while (only) 25 representations (i.e. the remaining columns of **W**) can be used for the words in the carrier sentences and for deviations from the linear model. The matrix **G** contains the feature values for each utterance (column), **V** are the acoustic features. **W**, $\mathbf{W}_g$ and **H** are estimated on the training set.

For the baseline, we learn from the canonical features, i.e. **G** are the canonical features for each utterance (column), i.e. **G** = **C**. For initialization with canonical features, $\mathbf{W}_g$ is initialized as a diagonal matrix (it is non-square: 50x*R*) plus some random noise. The diagonal structure directs the NMF (which is sensitive to local extrema) towards a solution in which each keyword is associated to exactly one column in **W**. The noise is added to $\mathbf{W}_g$ such that better solutions can be found where a keyword would have multiple columns in **W** (internal representations).

The random initialization may lead to different solutions as there is no guarantee that the multiplicative NMF update algorithm leads to the global optimum of the Kullback-Leibler (KL)-divergence. Therefore, in the experiments reported on in this section we always perform 5 independent trials and report the lowest and the mean error rate.

For evaluation of the baseline, only the acoustics are assumed to be available, i.e. using only the acoustics **V** and the trained model **W**, we determine **H** with a "one sided NMF" (convex problem) and determine the word activations as $\mathbf{A} = \mathbf{W}_g\mathbf{H}$. Per column of **A**, we now have the activation levels of each word and we threshold these with a **fixed**, utterance and word-independent threshold θ. If the activation is above the threshold, we conclude that the word was present in the sentence, else it is absent. As θ increases, there will be more missed detections and fewer false alarms. Since the threshold may depend on parameters such as codebook sizes, training data etc., we

[1] Following Meng's experiments done in Leuven, this is known to be suboptimal for accuracy.

factor out θ by plotting the DET-curve (i.e. false alarm rate vs. missed detection rate – see also ACORNS deliverable D4.1) and always take the equal error rate (EER) point as the operating point. We then report EER in % for all experiments. (Notice that this operating point does NOT lead to equal insertion and deletion rate, since a word is more often not present than present).
These baseline results are presented in Table I in the lines "Canonical". The first line uses the initialization method described above. The second line uses a fully random $\mathbf{W}_g$. We notice that both initialization methods lead to comparable error rates.

As already observed in subsection 3.4, the evaluation of tests with low-level visual ('semantic') features is not that trivial. Since the grounding matrix **G** will now contain the semantic features, we would be tempted to reconstruct and evaluate on each of the 64 features and not on the 50 keywords, which is very different from acoustic recognition of an object. The question asked would then be "did you hear a word that has the feature *round* ?" instead of the question used in the baseline "did you hear the word *apple* ?". To proceed with evaluation at the keyword level instead of at the feature level, we observe that the internal structures do contain word activations in **H**, though there may be more internal representations (75) than keywords (50). These additional representations model words from the carrier phrases as well as deviations from the linear model (i.e. an observed **V** has a higher rank than is implied in **V** = **W** **H**).
In order to make evaluation at the keyword level possible, we estimate a new matrix **Q** such that **C**=**Q** **H** on the training data, i.e. a transformation that maps the internal representations directly onto the intended keywords. The matrix **Q** 'peeks' into the brain of the learner. In the simplest of cases, the method builds a single internal representation for each keyword and the peek matrix **Q** has (modulo permutations) a diagonal structure (beware: it is not square). It is also possible that a keyword received multiple internal representations, in which case the peek matrix **Q** will combine their activations to a single word activation. Inspection of this matrix reveals if the training was successful in the sense that every keyword or canonical feature has at least one internal representation. If a single representation is used for two different keywords, they are bound to be confused and lead to errors.
During the test, we can compute the activations of the representations **H** from the acoustics, then multiply with the peek matrix **Q** to estimate which keywords these activations correspond to. Moreover, we can construct the keyword activation matrix **Q** **H** on the test data and construct a DET-curve, similar to the case of conceptual/canonical features, and also compute the EER on this curve. This way, the error rates become comparable to the baseline.

3.7 Using binary visual (semantic) features in combination with *object recognition*

The *binary visual (semantic) features* are obtained by multiplying the *conceptual* (*canonical) features* **C** with the feature matrix **F**, i.e. **G** = **F** **C**.
This linear model is an ideal case for NMF since the linearity applies to both the grounding and the acoustic part of the data. Whether this is realistic, is debatable. In subsection 3.4 we addressed the question how to combine features in the case of multiple referents. Do features add when they are present multiple times? This is reasonable if we assume that the learning system recognizes *L* separate objects with their features. Adding the feature values of the separate objects is a possible way for NMF to deal with multiple objects in the visual scene. Alternatively, methods could be designed that use the *L* feature vectors separately instead of adding them.

Initialization. The **W**-matrix (acoustic part) is always initialized randomly. In the case of the high level conceptual/canonical features, $\mathbf{W}_g$ was initialized as a diagonal matrix plus some random noise. This diagonal initialisation not a good choice for the present situation, as we might direct the NMF towards a solution where **W**-columns would be associated to features rather than keywords. We therefore consider three types of initialization:

1) Initialization with **F** (plus some random noise). This is a "cheating" experiment since actually **F** is unknown to the learner. Started sufficiently close, we expect to find the global optimum of the NMF.
2) Random initialization for $\mathbf{W}_g$.
3) Using *singling out*. Singling out means that we first try to recognize simple sub-scenes from more complex scenes. Relating this method to human learning, it is as if we ignore complex scenes and first try to make sense of the simple scenes. This is typically what we do in learning: we start by using simple language to our children if we present them with learning stimuli: caretakers will simplify their language and not read from the Financial Times. Mathematically, this is done as follows: given **G**, we try to remove scenes that can be written as an additive combination of other scenes. To reduce computational load, we first remove doubles in the columns of **G**, which only makes sense for integer-valued **G**. For noisy data, there will never be an exact match, so we work with a threshold. The following pseudo-code is applied:

```
θ initialized to some small value.
repeat
   for all columns g_k of G
      if KL(g_k, G_k a)< θ, remove column k from G
         (where a≥0 and G_k is G with its kth column removed)
multiply θ with 1.3 (or something like that)
until size(G,2)≤R
```

In the third line of Table I, we observe that the semantic features are more difficult to handle than the canonical features, even if initialized with **F**.

Feature stream weights. In past experiments with canonical features, we have found that the weight of the semantic vs. acoustic stream has only a minor effect on the accuracy. This is questioned here, since we are concerned about converging to internal representations that reflect features rather than keywords. For normalization, a first weight is applied such that $\sum_{i,j} \mathbf{V}_{ij}$ is 1 for both streams. Then we multiply the acoustic stream with the weight indicated in Table I. We notice that – at least for semantic features – the weight of the semantic stream is to be large enough. With too strong a weight on the acoustics, we fall back to completely blind (unsupervised) acoustic pattern discovery. This worked ok for small vocabulary (11 digits) (Stouten *et al.*, 2008), but for the larger vocabulary where not all words correspond to a tag, it does not seem to work as well.

Notice that initialization with **F** cannot be applied to real-life data, since **F** is unknown. The *singling out* method, which addresses this issue, seems to be effective here and leads to the same recognition rate as the canonical features (last line of table).

Method	acoust weight	Trial					Best	Avg.
		Trial 1	Trial 2	Trial 3	Trial 4	Trial 5		
Canonical, initialization with identity	1	2,19	2,09	2,16	2,10	2,08	2,08	2,12
Canonical, random initialization	1	2,18	2,35	2,03	2,35	2,61	2,03	2,30
Init with **F**	0,01	3,11	3,14	2,96	2,99	2,95	2,95	3,03
Init with **F**	1	6,16	4,69	5,14	5,09	6,13	4,69	5,44
Init with **F**	100	23,8	23,5	22,5	23,4	24,4	22,5	23,52
Random init	1,E-04	7,47	5,87	6,86	7,15	6,29	5,87	6,73
Random init	0,01	7,45	6,49	5,92	7,30	6,61	5,92	6,75
Random init	1	7,26	9,66	7,81	9,85	7,40	7,26	8,40
Random init	100	24,4	23,7	23,3	23,8	23,4	23,3	23,7
Only acoustics	N/A	23,5	22,9	23,3	23,8	23,9	22,9	23,5
Singling out	0,01	2,88	2,54	2,86	2,95	2,62	2,54	2,77

Table 1: Equal Error Rate in % on the ACORNS Y2 UK database. Detection of 50 words.

Inspection of the internal representations. The 'peek' **Q**-matrix introduced above allows to inspect the mapping between internal representations and keywords. Its columns are always permuted to obtain a maximally diagonal structure. Internal representations are on the abscissa, keywords on the ordinate.

1) Canonical features: random versus identity initialization
Random initialization only marginally increases EER. Comparing Fig 2 and 3, we see that some keywords receive two internal representations. Representation 70, however, models two words ("mummy" and "like") and might be the cause of the increased error rate.

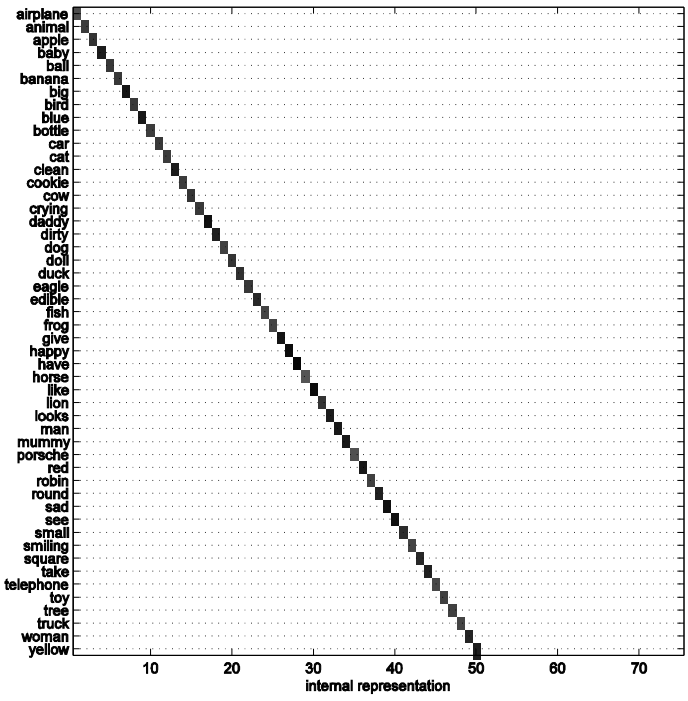

Figure 2: with canonical semantic features and "identity" initialization, a nice diagonal structure is obtained, i.e. all keywords have one internal representations. Other internal representations are used to model non-keywords and approximations.

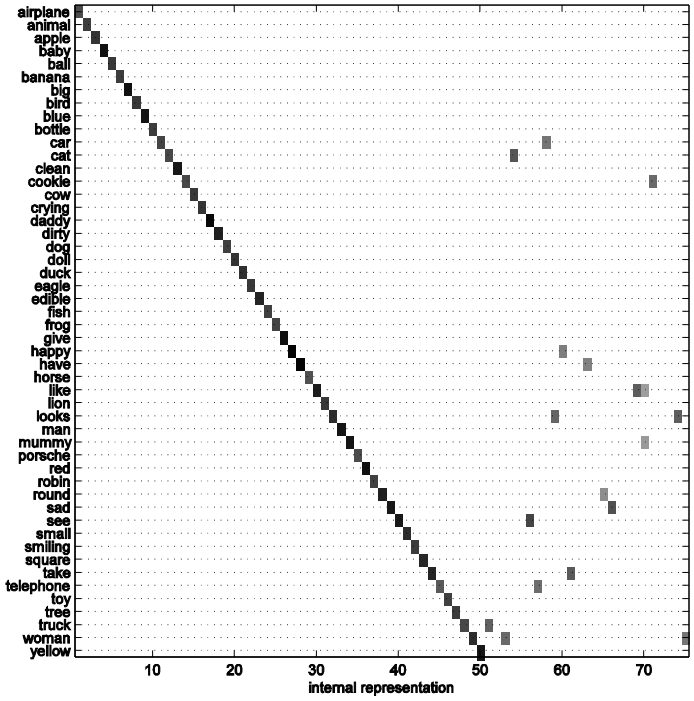

Figure 3: with canonical features and random initialization, some words get more than one internal representation.

2) Semantic features

At acoustic weight of 0.01, the cases of **F**-initialization, random initialization and singling out initialization are shown. With **F**-initialization, we notice that "ball" and "bottle" get a common representation, as well as "Porsche" and "airplane", which is bound to cause increased error rates over canonical features. "Porsche" does not have its own representation and leads to recognition errors.

Figure 4: F-initialization. "Porsche" gets no representation. "Bottle" is cluttered by "ball", which still has its own representation.

Figure 5: Semantic features with random initialization. Confusions caused by semantic similarity are apparent.

With singling out, some confusions still remain ("airplane"/"Porsche"; "bottle"/"ball"), but the EER is surprisingly good despite the fact that these words *do* occur in the test.

Figure 6: confusions for the "singling out" approach.

3.8 Using binary visual (semantic) features without *object recognition*

In the previous section, we assumed that the features of all objects were observable separately. Then we processed the semantic stream by adding up all features. In the present case, we assume that we dispose of a video stream in which we can detect features, but they are not grouped into objects. The feature is present or not in the image as a whole, i.e. there is no weighting for multiple occurrences. Mathematically speaking, the original semantic feature matrix **G** is now replaced by the binary matrix **G**>0.5.

For the high-level conceptual/canonical features, there does not seem to be a significant impact of this deviation from linearity (see first 4 rows of Table 2). With our data, the canonical case is actually the same as the previous section: every keyword occurs at most once in every utterance, so in both cases, **G** contains at most one '1' per utterance.

For the low-level visual/semantic features, the situation is different: a given feature can apply to multiple keywords in a sentence and now **G** of the previous section did contain values greater than 1. After mapping these to 1, the linearity in equation (1) is broken. For the model, it is like introducing noise.

From Table 2, we observe that the use of these low-level visual semantic features, even with the 'cheating' initialization using **F**, does impact the error rate negatively. The word confusions are apparent from Fig 7. We see that keywords like "apple" and "ball" have a common internal representation, as well as "cat" and "dog". Singling out does not work as well as before (see last 2 rows of Table 2 and Figure 8 where we observe many keyword pairs that receive a common representation), but is still below the error rates obtained with random initialization *with* visual object recognition.

By this comparison, we conclude that not having high-level object-related features makes the task much harder for the NMF-algorithm. This has to do both with the deviation from linearity in the visual stream as well as with convergence issues.

Method	objects	acoust weight	Trial					Best	Avg.
			Trial 1	Trial 2	Trial 3	Trial 4	Trial 5		
Canonical, initialization with identity	yes	1	2,19	2,09	2,16	2,10	2,08	2,08	2,12
Canonical, init with identity	no	1	2,08	2,20	2,08	2,16	2,13	2,08	2,13
Canonical, random initialization	yes	1	2,18	2,35	2,03	2,35	2,61	2,03	2,30
Canonical, random init	no	1	2,14	2,27	2,89	3,09	2,24	2,14	2,53
Semantic, init with **F**	yes	0,01	3,11	3,14	2,96	2,99	2,95	2,95	3,03
Semantic, init with **F**	no	0,01	5,18	3,94	4,05	4,25	4,95	3,94	4,47
Semantic, singling out	yes	0,01	2,88	2,54	2,86	2,95	2,62	2,54	2,77
Semantic, singling out	no	0,01	6,76	7,25	6,57	7,32	6,82	6,57	6,94

Table 2: Impact on EER of having a visual stream that recognizes objects (objects="yes") and associates the visual features to these, or having a visual stream that does not distinguish objects (objects="no") and detects the presence or absence of features in the scene as a whole.

Figure 7: word confusions when the vocabulary is acquired in a setting where features are detected in the whole scene and NOT per object. Initialization with F.

Figure 8: word confusions in the same case as Fig 7, but initialization with *singling out*.

In the next three subsections, experiments are presented in which the visual encoding is modified. In 2.9 we use unobserved ('obscured') featyres, while in 3.10 and 1.11 we deal with various types of variation.

3.9 Unobserved features

In realistic scenario's, it may happen that some features cannot be observed. For instance, a color cannot be seen when there is not enough light, objects may be occluded so we cannot determine their shape correctly or we may not know if an animal can fly, for it sits on the ground presently. To simulate occlusion, we assume that objects can be identified in the visual stream and we randomly decrease the count in **G** by one. The total number of decrements in **G** is a parameter in this experiment. Table 3 shows that with proper initialization, the method can withstand quite a lot of missing features. The error rate increases smoothly with the amount of noise and does not really break down at a certain noise level. The initialization with singling out produces a slightly higher error rate than the ideal initialization using the matrix **F**. The solution seems acceptable but is not perfect.

Method	acoust weight	Trial					Best	Avg.
		Trial 1	Trial 2	Trial 3	Trial 4	Trial 5		
Canonical, initialization with identity	1	2,19	2,09	2,16	2,10	2,08	2,08	2,12
Canonical, random initialization	1	2,18	2,35	2,03	2,35	2,61	2,03	2,30
Init with **F**, 0 % noise	0,01	3,11	3,14	2,96	2,99	2,95	2,95	3,03
Init with **F**, 5 % noise	0,01	4,75	4,65	4,45	4,39	4,59	4,39	4,57
Init with **F**, 10 % noise	0,01	5,75	5,39	6,00	5,72	5,68	5,39	5,71
Init with **F**, 15 % noise	0,01	6,57	6,97	6,67	6,86	5,75	5,75	6,56
Init with **F**, 20 % noise	0,01	7,72	7,65	8,19	7,93	8,56	7,65	8,01
Init with **F**, 25 % noise	0,01	8,84	8,76	8,91	9,09	9,50	8,76	9,02
Singling out, 0% noise	0,01	2,88	2,54	2,86	2,95	2,62	2,54	2,77
Singling out, 5% noise	0,01	5,98	6,08	5,51	5,58	5,83	5,51	5,80
Singling out, 10% noise	0,01	7,53	7,16	7,07	7,48	6,64	6,64	7,18
Singling out, 15% noise	0,01	9,6	8,97	8,68	9,3	8,11	8,11	8,93
Singling out, 20% noise	0,01	9,57	10,3	9,92	9,92	9,81	9,57	9,90

Table 3: the impact of unobserved features. "x % noise" means that x% of the per-object visual features are unobserved.

We did not conduct experiments with the opposite noise type where some features are falsely detected, for this seems to be less relevant.

3.10 Irrelevant features

In real scenario's objects usually have a lot of features which are irrelevant for recognizing the object as such. For instance, the color is not a relevant feature for a chair for its basic property 'on-sit-able'.

We simulate this by adding L random rows to the low-level visual semantic features. However, we should select the distribution of the noise we add carefully, i.e. the number of features that are present/absent in this noisy stream should be realistic. To ensure that these added rows of **G** have the same sparsity as the relevant rows of **G**, as we draw them from **G** and randomly permute along the columns. The impact of noise rows is shown in Table 4. Remember that there are 64 relevant visual features. We again see a gradual ('graceful') degradation of the recognition accuracy. However, in any realistic scenario, the number of irrelevant features will be higher than what is simulated here and context-dependent feature selection mechanisms are required. Again, we see that *singling out* adds to the error rate (last row) and is not capable of fully solving the initialization problem for the NMF.

Method	# irrelev. features	Trial					Best	Avg.
		Trial 1	Trial 2	Trial 3	Trial 4	Trial 5		
Init with **F**	0	3,11	3,14	2,96	2,99	2,95	2,95	3,03
Init with **F**	1	3,38	2,94	3,51	3,38	3,53	2,94	3,35
Init with **F**	2	3,67	3,19	3,61	3,12	3,07	3,07	3,33
Init with **F**	5	3,89	3,44	3,69	4,84	4,24	3,44	4,02
Init with **F**	10	4,48	3,67	3,36	5,06	3,61	3,36	4,04
Init with **F**	20	4,41	4,51	4,41	4,51	5,50	4,41	4,67
Init with **F**	50	7,31	7,72	8,28	8,83	7,72	7,31	7,97
Init with **F**	100	9,05	8,49	8,69	9,97	8,22	8,22	8,88
Singling out	0	2,88	2,54	2,86	2,95	2,62	2,54	2,77
Singling out	1	3,64	3,44	4,07	3,49	4,08	3,44	3,74
Singling out	2	2,99	3,62	3,23	3,23	3,26	2,99	3,27
Singling out	5	7,08	7,84	8,73	7,06	8,38	7,06	7,82
Singling out	10	9,41	8,98	11,69	9,47	9,04	8,98	9,72
Singling out	20	9,91	10,34	10,16	9,68	10,18	9,68	10,05

Table 4: The impact on EER (in %) of irrelevant features.

3.11 Fuzzy features.

In reality, we only have noisy observations of features. Features such as transparency, softness, noisiness, … are valid to some degree. Here, the features are not binary any more, but real-valued numbers. To simulate this, we add multiplicative noise to the non-zero values of **G**. Non-zero values are multiplied with |x|, where x is normally distributed with mean 1 and standard deviation sigma. Zero elements are replaced by an exponential distribution

$$P(x) = \frac{1}{\mu} e^{-x/\mu}$$

which has mean μ and variance μ^2, hence MSE of $2\mu^2$. Table 5 shows how this type of noise degrades performance. Again, we observe that quite a lot of noise can be allowed.

Method	zero's μ	nonzero's σ	Trial					Best	Avg.
			Trial 1	Trial 2	Trial 3	Trial 4	Trial 5		
Init with **F**	0	0,0	3,11	3,14	2,96	2,99	2,95	2,95	3,03
Init with **F**	0,05	0,1	2,77	2,77	3,30	3,07	3,03	2,77	2,99
Init with **F**	0,05	0,5	7,15	7,18	6,50	7,35	7,40	6,50	7,12
Init with **F**	0,01	0,1	3,25	3,41	3,28	3,63	3,59	3,25	3,43
Init with **F**	0,01	0,2	3,16	3,89	3,55	3,32	3,37	3,16	3,46
Init with **F**	0,2	0,1	5,74	4,73	4,21	4,36	5,73	4,21	4,95
Init with **F**	0,2	0,5	9,03	8,91	9,20	8,97	8,07	8,07	8,84
Init with **F**	0,2	1,0	10,9	10,4	10,2	10,6	10,2	10,2	10,5
Init with **F**	0,5	1,0	15,4	15,0	15,4	15,2	15,2	15,0	15,2

Table 5: the impact of fuzzy feature observation.

3.12 Conclusions

When learning from visual features instead of conceptual/canonical features, the NMF algorithm is not always able to find the correct mapping between keywords and internal representations. This is a problem of local extrema in the NMF cost function: if initialized close to the solution, a valid solution can be found. The technique of "singling out" was proposed to produce a better initialization, but the method is not always satisfactory.

To simulate the fact that visual information is 'noisy' in practical learning situations, we added various types of noise to the data. We observe that the NMF can handle noisy input and degrades **gracefully** as more noise is added. It is advisable to filter inputs to alleviate the impact of the introduced between-token variation.

References

Hugo Van hamme (2008). "HAC-models: a Novel Approach to Continuous Speech Recognition", In *Proc. International Conference on Spoken Language Processing*, pages 2554-2557, Brisbane, Australia, September 2008.

Veronique Stouten, Kris Demuynck and Hugo Van hamme (2008), "Discovering Phone Patterns in Spoken Utterances by Non-negative Matrix Factorisation*", IEEE Signal Processing Letters*, volume 15, pages 131-134, 2008.

Appendix A

Binary semantic features for the 50-word vocabulary from the Y2 UK database, version of March 13, 2009.

airplane:	0	0	0	0	0	0	0	1	0	1	0	1	0	1	1	0	0	0	0	1	1	0	0	1	0	1	0	1	0	1	1	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
animal:	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	1	0	0	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
apple:	0	0	0	0	0	0	1	0	0	1	0	1	0	1	0	1	0	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
baby:	0	0	0	1	0	1	0	0	1	0	0	1	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ball:	0	0	0	0	0	0	1	0	0	1	0	1	0	1	1	0	0	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
banana:	0	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	0	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
big:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
bird:	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
blue:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0
bottle:	0	0	0	0	0	0	0	1	0	1	0	1	0	1	1	0	0	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
car:	0	0	0	0	0	0	0	1	0	1	0	1	0	1	1	0	0	0	0	1	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
cat:	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	1	1	0	0	1	0	1	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
clean:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
cookie:	0	0	0	0	0	0	1	0	0	1	0	1	0	1	1	0	0	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
cow:	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	1	1	0	0	1	0	1	0	1	1	0	0	1	1	0	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
crying:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
daddy:	1	0	1	0	1	0	0	0	1	0	0	1	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
dirty:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
dog:	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	1	1	0	0	1	0	1	0	1	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
doll:	0	0	0	0	0	0	0	0	1	0	0	1	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
duck:	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	1	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
eagle:	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	1	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
eatable:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
fish:	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
frog:	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	1	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
give:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0
happy:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
have:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
horse:	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	1	1	0	0	1	0	1	0	1	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
like:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lion:	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	1	1	0	0	1	0	1	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
look_at:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
man:	1	0	0	0	1	0	0	0	1	0	0	1	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
mummy:	0	1	1	0	1	0	0	0	1	0	0	1	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
porsche:	0	0	0	0	0	0	0	1	0	1	0	1	0	1	1	0	0	0	0	1	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
red:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
robin:	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	1	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
round:	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sad:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
see:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
small:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
smiling:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
square:	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
take:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
telefone:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
toy:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
tree:	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
truck:	0	0	0	0	0	0	0	1	0	1	0	1	0	1	1	0	0	0	0	1	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
woman:	0	1	0	0	1	0	0	0	1	0	0	1	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
yellow:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0

4 Deviating from the 'ideal' interaction

The following two sections (4A and 4B) are narrowly related: both deal with experiments in which the interaction between caregiver and learner deviates from the 'ideal' interaction. 'Ideal' refers to the situation in most ACORNS experiments: the caregiver presents consistent stimuli (maybe noisified), and the learner takes each stimulus 'as it appears'. That is, the learner assumes the stimulus to be consistent, and does not doubt the consistency between the modalities in the stimulus.
One may deviate from this idealized-world scenario in several ways.

- At the caregiver side: the caregiver may present a certain proportion of stimuli that are inconsistent, in addition to others that are consistent. This compromises the belief of the learner that all stimuli are consistent.
- The learner may take a much more active role. For example, an internal confidence mechanism may be active such that the learner might assume her OWN hypothesis being true, irrespective of what was presented in the stimulus. This directly means that the learning becomes less supervised.

Section 4 consists of two related parts that both deal with deviations from the 'ideal' interactive setting:
4A Learning meaningful units from multimodal input – the effect of interaction strategies. Here we investigate four different interaction strategies between caregiver and learner. The reason to keep this section separate from 4B is that this text has been published in and presented at the WOrkshop for Child-Computer Interaction WOCCI-09 (ten Bosch, Boves & Räsänen, 2009).
4B Deviating from strict supervision during training. Here we show that a certain level of 'contrariness' at the learner's side helps to overcome inconsistencies in stimuli from the caregiver.

4A Learning meaningful units from multimodal input – the effect of interaction strategies

Louis ten Bosch, Lou Boves, Okko Räsänen

(This paper has been published in WOCCI09)

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 a (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 paring 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

4A.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.

4A.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.

4A.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 *Nl*Nc* tables where *Nc* is the total number of all possible tags, and *Nl* 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 *Nq* x *Nq*, where *Nq* 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.

4A.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.

4A.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 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 then 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).

4A.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.

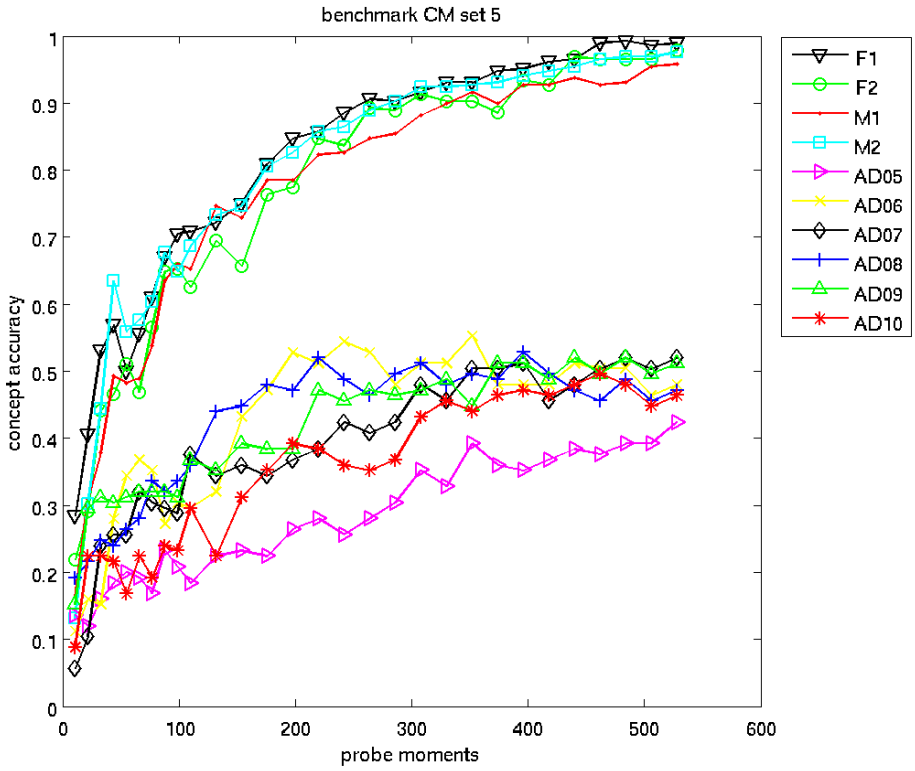

Figure 1. Results of the model using strategy A (condition A) on 10 different test sets (10 different speakers), using one fixed particular training set. There are ten learning curves - each curve is related to one of the test sets (i.e. one test speaker). One point (x, y) specifies the performance of 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.

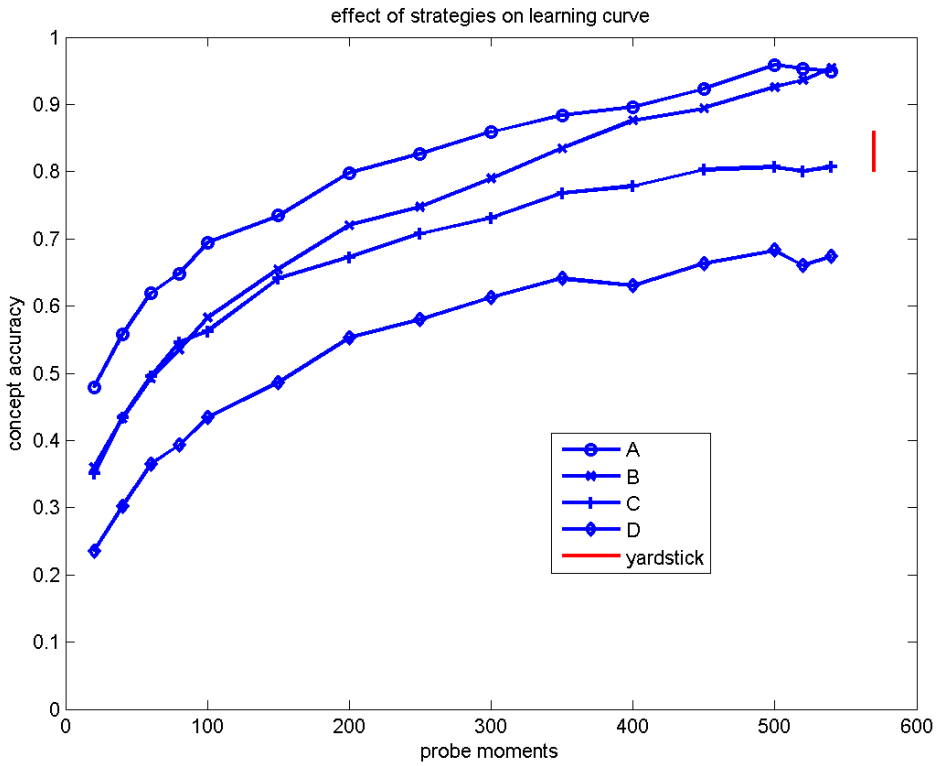

Figure 2. Comparable learning curves for conditions A, B, C, D. Significant differences are indicated by the red 'yard stick'.

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 behaviour during learning, of internal learning models and of improved user modeling and adaptation.

4A.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 that 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 behaviour 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.

4A.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

4A.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., Altosaar, 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 Altosaar T. (2008). "Computational language acquisition by statistical bottom-up processing", *Proc. Interspeech'08*, pp. 1980-1983, 2008.

[22] Räsänen, O., Laine U.K., and Altosaar 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.

4B Deviating from strict supervision during training

4B.1 Introduction

In section 3, we discussed the various scenarios if one 'noisifies' the visual channel, while section 4A presented the effect of different learning conditions on the performance of the learner. Both studies show that, as soon as one deviates from the situation without variation, with consistent stimuli, and with a passively believing learner, the performance of the learner deteriorates, almost always in a graceful manner.
In section 4B, we will investigate in more detail what happens in the caregiver-learner interaction, with a focus on the relation between *'proportion inconsistent stimuli'* (as presented by the caregiver) and the amount of *contrariness* (implemented by using internal confidence measures at the learner's side). By doing so, section 4B also addresses the issue of supervision during learning. Using hindsight we conclude that in *most* experiments done so far we have investigated what is essentially supervised learning. This issue was also discussed during the January 2009 review. In most experiments the correct visual tag was appended to the speech stimulus after it was first presented to the learner, and the learner believed the presented stimulus 'as is' (as ground truth), even if the learner's hypothesis was wrong. That is:

- The caregiver presents coherent, complete, true combinations of auditory and visual information in each stimulus, or noisified versions thereof
- The learner takes these inputs 'as is'. The stimulus is incorporated into the learning mechanism, without doubt about the internal (in)consistency of the stimulus.

Especially the second point represents an unrealistic setting for learning, and it would be interesting to investigate situations that deviate from this 'blind-belief' ('blind-trust') situation. The relevance within the ACORNS project to investigate other types than strictly supervised learning relates to the fact that language learning is actually only very mildly supervised.

Another motivation to investigate semi-supervised learning is the role of the learner during learning. In the experiments presented so far, the role is (quite) passive in the sense of 'taking everything for granted'. This is not realistic. It is very likely that the role of any learning system must be more 'pro-active' - especially in the case where the truth or consistency of learning stimuli cannot be taken for granted. In a realistic teacher-pupil scenario, and probably also in a caregiver-infant situation, the level of suspicion at the learner's side will probably rise if the teacher makes occasional errors. That means that the communication between teacher and pupil, and between caregiver and infant, is a dynamic process: 'if I start doubting what you say, I increase my contrariness levels and may become more cocksure'. This mutual balancing between 'belief' and 'suspicion' is probably a basic mechanism in many, if not all, forms of human-human communication.

4B.2 Semi-supervised learning

In this section we will investigate a route towards semi-supervised learning. The experimental results suggest the computational modelling of the 'belief-suspicion balance' at the learner's side. The following two directions are explored:

- At the caregiver side: the caregiver may present stimuli that are inconsistent. The probability of presenting an inconsistent stimulus is denoted p. ($0 \leq p \leq 1$). Compared to the

noisified visual presentations in section 3, the inconsistencies are very drastic: either the the stimulus is entirely consistent (i.e. no noise or variation), or it is entirely inconsistent.

- At the learner's side, an internal confidence mechanism is implemented such that if the confidence about a certain hypothesis exceeds a certain threshold θ ($0 \leq \theta \leq 1$), the learner assumes that its **own** hypothesis is true. In that case, the learner **ignores** the information in the grounding section of the presented input stimulus. Instead, the own hypothesis is put in short term memory for later reuse, and so is assumed to be of value for all later internal updates.

The situation is depicted in Figure 1. The 'ideal' situation is the one combining 'coherent and correct' with 'always believe'. The parameter p ($0 \leq p \leq 1$) models deviations along the vertical axis; θ ($0 \leq \theta \leq 1$) models deviations along the horizontal axis. The vertical arrow left refers to the probability of presenting inconsistent audio-visual pairs to the learner. At the caregiver's side, the value p=0 is represented by the left upper option, higher values are referred to by the option 'incomplete or incorrect'. Horizontally, the choices are determined by θ at the learner's side. A value of θ =1 corresponds to the middle column; the lower θ, the lower the learner's internal threshold to believe herself and the more 'cocksure' the learner will be.

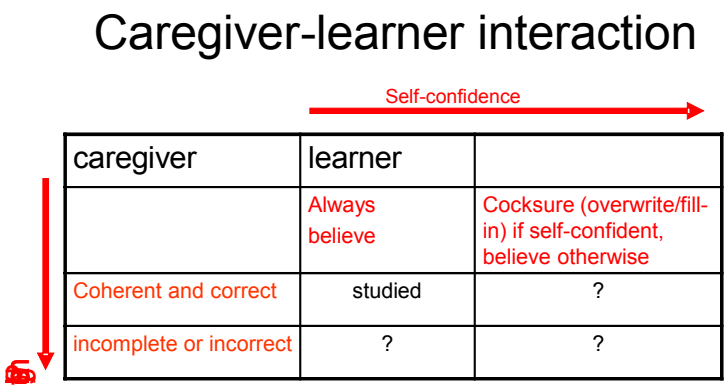

Figure 1. The parameter p models the vertical axis; θ models the horizontal axis. The vertical arrow left refers to the probability of presenting inconsistent audio-visual pairs to the learner. At the caregiver's side, the value p=0 is represented by the left upper option, higher values are referred to by the option 'incomplete or incorrect'. Horizontally, the choices are determined by θ at the learner's side. A value of θ =1 corresponds to the middle column; the lower θ, the more 'cocksure' is the learner.

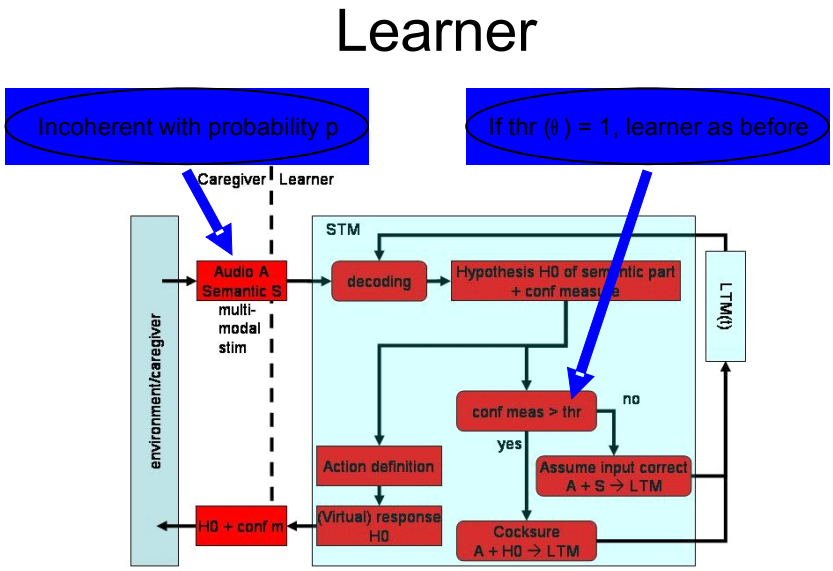

Figure 2. This figure shows how the learner is extended for the experiments described here. The input stimulus is recognized on the basis of its acoustic (audio) part. Next, the confidence is evaluated on the basis of all activations of the internal representations. In the following step, this confidence value is compared with a threshold ('thr', θ) – this comparison determines the behaviour of the learner, in particular what is stored in the learner's Long Term Memory. If the threshold has a value of 1, the learner assumes that all stimuli are correct. If the threshold equals 0, the learner is cocksure for all stimuli.

Semi-supervised learning means that the learner must be able to overrule the 'ground truth' as presented in the stimuli, and so the learner plays a more 'active' role than during fully supervised training. This more active role may vary from 'filling in' missing information to 'overruling' the information in the presented stimulus. In effect, this experiment exploits the possible types of interaction between caregiver and learner in such a way that the experiments done in the first ACORNS year can be considered as a special case (i.e. large θ).
The changes made in the learner as depicted in Figure 2. The learner does not passively assume any more that inputs are 'correct'. Instead, it uses an internal confidence measure to quantify its own belief in the labelling of an unseen stimulus, and if the learner is sure enough (of the internal confidence exceeds a certain threshold), it assumes its OWN hypothesis to be correct. The learner's internal threshold is denoted 'thr' (= θ, the value of this threshold can be set by the experimenter). The confidence measure that the learner evaluates for each input stimulus, in combination with this threshold, determines how the learner proceeds after having recognized the stimulus. In concreto, the steps are as follows:

- During test, a novel stimulus is presented to the learner
- The audio part of the stimulus is used for recognition, in combination with the internal representations. A hypothesized conceptual feature vector is reconstructed by the learner. In combination with the reconstructed conceptual decoding, a confidence measure *c* is evaluated. This value is defined by $c = (a_N - a_{N-1})/\Sigma a_i$, where a_N and a_{N-1} denote the confidence attached to maximum and second best element in the reconstructed visual feature vector, respectively, and Σa_i represents the sum of all elements of this feature vector. Figure 3 shows the effect of this particular definition of confidence on the performance of the learner: the larger the learner's confidence value, the more likely it is

that the recognition was actually correct. The lower black curve corresponds to a confidence threshold of 0 and therefore includes all stimuli. By selecting the stimuli by increasing the condfidence threshold (0.04 and 0.08 refer to the middle red and upper purple curves, respectively), the performance increases, showing that the more confident the learner was, the larger the likelihood of the hypothesis being correct.

- In the current implementation, the learner uses this internal confidence measure *c* in the following way. If the confidence measure *c* exceeds a certain (experimenter defined) threshold θ (i.e. $c > \theta$), the learner assumes its own hypothesis is correct (no matter the label in the actual stimulus). In that case, the hypothesized label is assumed to be correct and (in combination with the audio part) stored in Long Term Memory (LTM). Otherwise the learner assumes the information in the stimulus is correct, and the stimulus is stored in LTM.

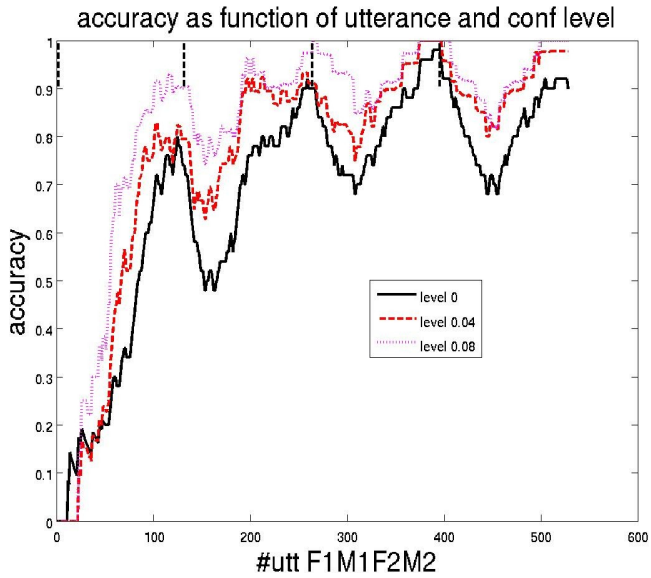

Figure 3. This figure shows the relevance of the chosen implementation of the internal confidence measure c. The three curves shown here (black, red and purple) all refer to the same training run. The black curve is the original plot of the accuracy of the learner, varying over time. The red curve is based on the same data, but restricted to all recognition results with $c > 0.04$. The purple curve displays the data for which $c > 0.08$. The higher the confidence of a certain hypothesis, the higher the likelihood of this hypothesis being correct.

4B.3 Results

Each combination of (p, θ) leads to a specific performance of the learner. This is shown in figure 4 (left and right). This figure shows a contour plot of the performance of the learner as a function of *p* (in the figure referred to as the noise in transmission, along the horizontal axis) and θ (confidence threshold, along the vertical axis). The 'ideal' training condition is in the upper left part of the plot. The more the training condition deviates from this position, the worse the learner's performance (in general). Interestingly, the effect of both parameters is different: while in one direction one can detect a 'graceful' degradation, the effect in the other dimension is very drastic. It is interesting to see that a certain amount of 'contrariness' does not deteriorate the performance. On the contrary: for certain amounts of incoherence in the stimuli (i.e. along one vertical slice) a certain level of contrariness at the learner's side even *improves* the learner performance.

Figure 4 further suggests that the learner is robust against inconsistency in the input as long as the confidence threshold is low enough. Even a value of $p = 0.5$ (audio-referent pairing is incorrect in 1 out of 2 stimuli) may lead to a correctness of 0.75. In this direction, degradation is graceful
The effect along the vertical dimension is made clearer in figure 5. This figure displays a section of the contour plot shown in figure 4, right panel, along the vertical slice $p = 0.25$ (this means that the audio-referent pairing in 1 out of 4 stimuli is incorrect). At the right hand side, θ=1 and the learner assumes all stimuli are coherent. According to figure 5, this leads to less optimal performance than the situation in which the contrariness is higher (that is, lower values of θ), with an optimum for θ around 0.1. Interestingly, this means that the learner can improve its own performance, compared to the fully-believing case, by being much more cocksure and only take the stimuli for granted in those cases where the learner itself is not sufficiently sure.
Another interesting fact is that in this dimension, degradation is not so graceful any more. If the learner is too 'cocksure' (θ → 0), the learning radically breaks down.

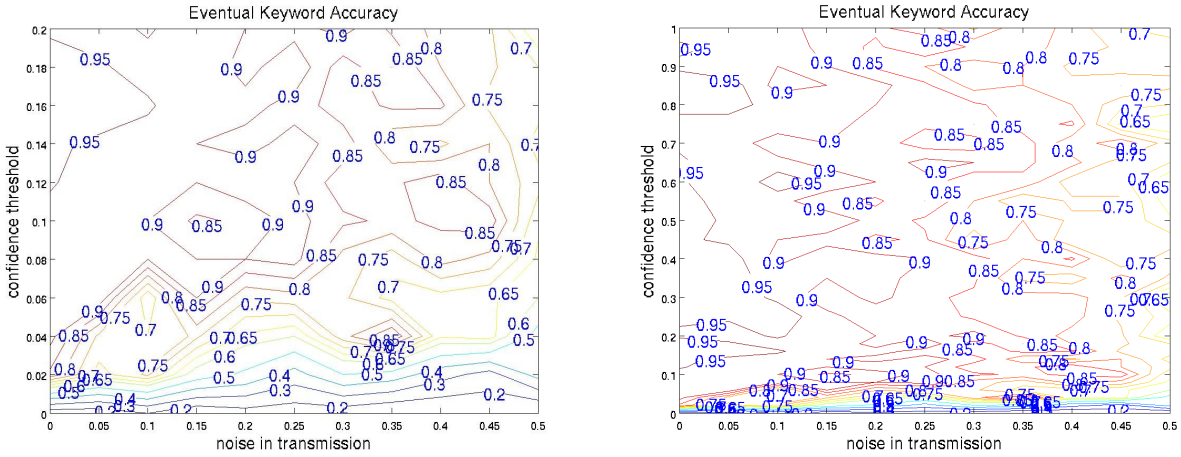

Figure 4. Performance of the learner, as a function of the probability p (horizontal axis) and the confidence threshold θ (vertical axis). Small grid (left) and large grid (right) A contour plot of the performance of the learner as a function of p (in the figure referred to as the noise in transmission, along the horizontal axis) and the confidence threshold □□along the vertical axis□□. The contour plot suggests that the learner is robust against inconsistency in the input as long as the confidence threshold is low enough.

Figure 5. A cross section of the contour plot along the vertical slice p = 0.25 (this means that the audio-referent pairing in 1 out of 4 stimuli is incorrect). At the right hand side, θ=1 and the learner assumes all stimuli are coherent. Evidently this leads to less optimal performance than the situation in which the contrariness is higher (that is, lower values of θ), with an optimum for θ=0.1. This means that the learner can improve, compared to the fully-believing case, by being much more cocksure and only take the stimuli for granted in those cases where the learner itself is not sufficiently sure. If the learner is too 'cocksure' (θ → 0) the learning breaks down.

Figure 4 and 5 have been based on a grid computing in which various values of (p, θ) were explored. Each combination of (p, θ) in the square [0, 1] x [0, 1] leads to an entire learning curve.

4B.4 Conclusions

This section discusses the effect of deviations from an 'ideal' setting of caregiver and learner on the performance of the learner. An experiment was done by systematically investigating the effect of two parameters. The results show that the performance of the learner can gain from a certain amount of contrariness at the learner's side. It is the computational equivalent of a 'belief-suspicion' balance between caregiver and learner. It is to be investigated how this contrariness level can be adjusted in an automatic way during training.
Interestingly, the experiment also shows various degrees of graceful degradation. The degradation of the learner's performance is graceful along one dimension (the inconsistency in the input). This is in line with the findings reported in section 2B. However, along the other dimension, degradation might not be graceful, and looks actually rather catastrophic. Both the graceful and less graceful form of degradation are shown in figure 4 (right).

References

Roger K. Moore, Louis ten Bosch "Modelling Vocabulary Growth from Birth to Young Adulthood", *Proc. Interspeech 2009* [pdf]

5 Alternative features

5.1 Introduction

One of the tasks in WP1 is the design and test of features that deviate from the standard features that conventional speech decoding algorithms use (MFCCs). In WP1, two new types, MMFCC and SA-MMFCC features, were designed in such a way that they optimally encode acoustic details in the signal. The differences between the novel features is basically the dimension of the resulting feature vector and the way how the features are selected from a constrained search space. In the ACORNS approach, computational models of the human hearing system were involved for optimizing the features. In experiments, the novel features are then contrasted with the more conventional MFCCs.
The hypothesis is that once features are tuned towards the human hearing system, they are principally better equipped to capture details from the speech signal that appear relevant for upstream processing of human speech across many conditions. For example, it was expected that certain feature types could outperform MFCCs in adverse conditions. In total, three feature sets have been compared in various noise conditions:
(a) the classical MFCC-based features (dim 39)
(b) modified MFCC (MMFCC) features (dim 39)
(c) features specifically found by using auditory model optimisation SA-MMFCC (dim 51)

For the description of these features we refer to deliverable D1.3. In this deliverable it is described how the features are constructed and how the computational model of the human auditory system in involved in the improvement of the features. The deliverable also mentions the various tests to show that these features indeed make sense for improving the recognition rates on specific test sets in specific conditions. These tests were based on the conventional HMM-ASR framework

The question remains to what extent MMFCC and SA-MMFCC could improve results when they are combined with other back-ends, such as the three ACORNS learning algorithms.

5.2 MMFCC

In order to test the performance of MMFCC and to contrast these results with the classical MFCCs, a large-scale experiment was done with the following independent experimental factors:

- training sets (5: varying in speakers)
- test sets (10: one speaker per test set)
- noise type (3 or more: clean, white20 (20 dB SNR), white10, factory, …)
- channel normalisation (2: CMVN, no-CMVN)
- feature type (2: MFCC, MMFCC)
- learning algorithm (2: NMF, CM (DP-Ngrams))

This experiment provided a substantial amount of results for at least 1200 different combinations of independent factors. The result of each combination was reflected in one single number (the performance of the learner on the entire test set).

The results can be summarized as follows.

The modified MFCC (MMFCC) features perform significantly better than MFCC in almost all cases (independent of the computational method NMF, DP-ngrams, and CM). In absolute terms, the improvement in accuracy varies between 3 and 8 percent.

In 4 percent of all comparable cases, MMFCC does NOT lead to improvement of the measured accuracy compared to MFCC. This effect is attributable to the random initialisation effects that are inherent in all training runs. The same training rerun may yield a (slightly) different performance, with a certain underlying statistical distribution. Since each training run also depends on (random) initialisations, it can happen that the MMFCC variant turns out at the low side and the MFCC variant at the high side of this distribution. Under the assumption of this effect being normally distributed, the probability of this 'flip' happening is estimated between 3 and 5 percent.
The amount of improvement depends in general on the type of background noise applied and whether cepstral mean and variance normalisation (CMVN) is used or not. In case of stationary noise (white 10 dB SNR, white 20 dB SNR) the gain is higher than for non stationary noise. Also CMVN helps the improvement between MMFCC and MFCC, with an average of 1.9 percent absolute.

5.3 SA-MMFCC

In order to test the performance of SA-MMFCC and to contrast these results with the classical MFCCs and with MMFCCs, a small-scale experiment was done with the following independent experimental factors:

- training sets (5: varying in speakers)
- test sets (10: one speaker per test set)
- noise type (3 or more: clean, white20 (20 dB SNR), white10, factory, …)
- one learning algorithm: DP-Ngrams

The SA-MMFCC features have only been tested in combination with DP-ngrams – it is not combined with NMF and CM. The reason for this was that building reasonably optimal code books, such that a fair comparison is possible, for NMF en CM is prohibitive: the dimension of the MFCC and MMFCC vectors is 39; the dimension of the SA-MFCC vector is 51. The design of a good codebook allows a few degrees of freedom (number of codewords, the sampling accuracy, the way in which multiple streams are weighted and combined), and therefore the creation of fair and comparable codebooks on features with an entirely different type of content and statistical character is not evident. For that reason, it was decided to limit the SA-MFCC tests to DP-Ngrams, the only method of the three that could be tested right away without codebook.

The results obtained for SA-MFCC appeared less promising, for reasons that became clear only after having analysed the results of these preliminary experiments. In these experiments, it appeared that by using DP-grams, the results show significant deterioration compared to both MFCC and MMFCC. One of the results is plotted in Figure 1

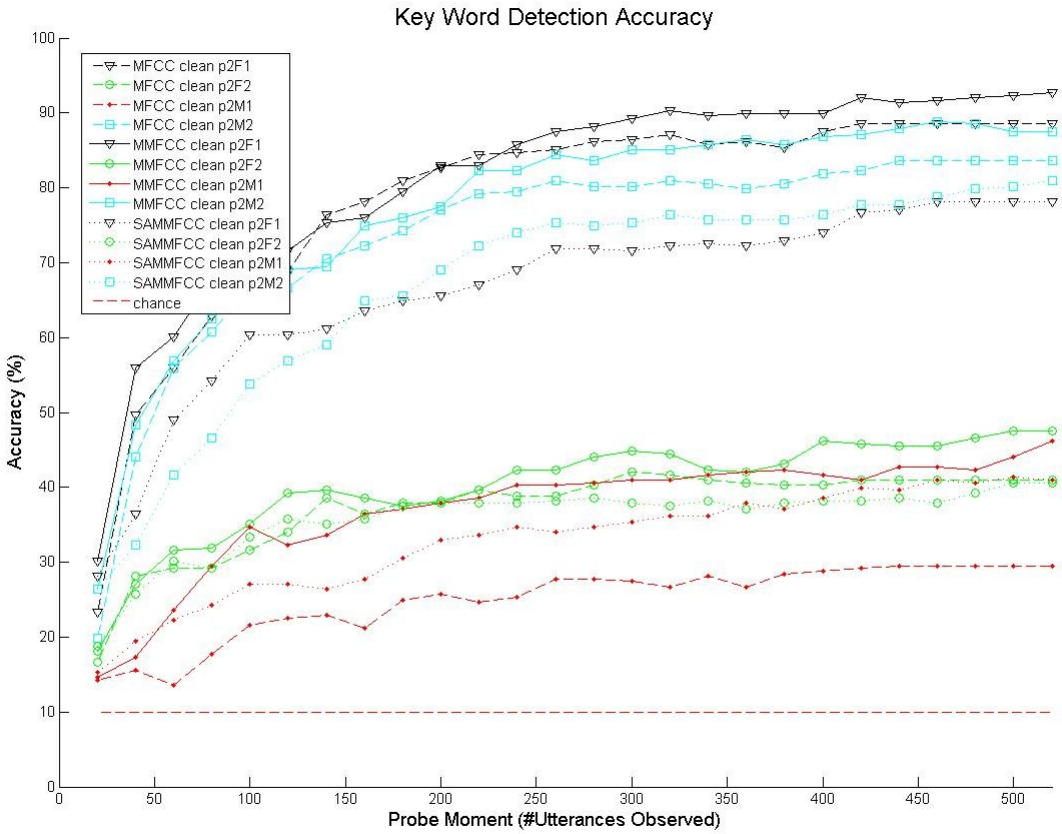

*Figure 1. The figure shows the results for MFCC, MMFCC and SA-MFCC in **clean** condition, for four different speakers F1, F2, M1 and M2. The computational approach is DP-ngrams. The figure shows that MMFCC clearly outperforms MFCC, but the results with SA-MFCC are substantially worse (as can be seen by e.g. comparing the plots with the downward triangle for speaker F1). Graph made by Guy Aimetti.*

Discussion SA-MMFCC

Prior to the ACORNS data, the novel features MMFCC and SA-MFCC have been tested on other data and other recognition tasks (Deliverable 1.3). In the case of phone decoding tasks (on TIMIT) and word recognition tasks on Aurora-2, these modified features (both MMFCC and AS-MFCC) gave better results than MFCC with an HMM-ASR system (Hidden Markov Toolkit, HTK) as recognition back end.

Given these results, it appears that these HMM-ASR-based results cannot be directly generalised to the DP-Ngrams approach in ACORNS. So far, all results point to the issue of feature correlation between the input features as the most likely hampering effect. The HMM-ASR-based improvements were obtained with a Gaussian mixture per state. Since DP-Ngrams uses Euclidean distances, the rather low SA-MFCC results may be due to correlation in the SA-MMFCC features. The amount might be low enough to be adequately coped with by HTK Gaussian mixtures but too high to be captured by the L2 norm in DP-ngrams.

In general, a mixture of Gaussians is able to cope with correlated data, by explicitly modelling the correlations between the features of the input vectors using a weighted sum of diagonal covariance Gaussians. In total, this means that pretests with an HMM-ASR system using Gaussian mixtures may obscure the usefulness of the novel features in DP-ngrams, and that a de-correlation step must be applied between the feature extraction and the recognition back end.

Another possible cause is the non-gaussianeity of the SA-MMFCC features compared to the MFCC and the MMFCC features.

These effects will be investigated in subsequent research. Within the ACORNS project, there is hardly time left for investigating the precise effect of correlation. Within the ACORNS time frame, only a few limited small-scale tests could be performed due to the late availability of the novel SA-MMFCC features.

6 Conclusions

In section 2, we have seen how the different computational approaches could be compared, and to what extent it was possible to relate model results to findings described in the literature on language acquisition. One of the literature findings, reported in Newman (2008), provided a good example of an empirical result that could be compared and contrasted with results obtained by computational simulation. Newman's statement is that young infants are better in recognizing novel speakers if they have been exposed to more different speakers earlier. This observation is narrowly related to the current debate about episodic and abstractionist processing of speech (see e.g. McQueen, 2007) and therefore opens the discussion to what extent episodically-based algorithms can/should deal with abstraction and generalisation.
The experiment was a opportunity to contrast the three computational approaches NMF, CM, and DP-Ngrams, which was one of the recommendations of the reviewers. It was explicitly not the intention to see which algorithms performs 'better' than other algorithms. Instead, the comparison was meant to gain insight in the different types of behaviour, based on different learning principles.

In section 3, different alternatives for the visual encoding are discussed and NMF-based experiments are reported in which the visual channel is noisified in a number of different ways. When learning from visual features instead of canonical/conceptual features, the NMF algorithm is not always able to find the correct mapping between keywords and internal representations. This is a problem of local extrema in the NMF cost function: if initialized close to the solution, a valid solution can be found. The technique of "singling out" was proposed to produce a better initialization, but the method is not always satisfactory.
To simulate the fact that visual information is noisy in practical learning situations, we added various types of noise to the data. We observe that the NMF can handle noisy input and degrades **gracefully** as more noise is added. It is advisable to filter inputs to alleviate the impact of the noise at the learner's side.

Section 4 (4A and 4B) discusses the way caregiver and learner interact which each other. In most ACORNS experiments, we have assumed that the caregiver always present complete and consistent stimuli. Each stimulus consists of an audio part and a 'visual' (grounding) part. In the experiments so far, the learner takes each stimulus 'as it appears'. That is, the learner assumes the stimulus to be consistent, and does not doubt the consistency between the modalities within the stimulus (learning condition A)
We have done experiments with the aim to investigate what would happen if the interaction between caregiver and learner was modified towards less strict forms of supervision during training. 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.
A certain contrariness at the learner's side can help to overcome the inconsistency that is present within the input stimuli. It was also shown that degradations of the learner's performance are graceful along the dimension of inconsistency in the input (a higher inconsistency leads to a graceful degradation of the learner's performance). However, the change in learner's contrariness levels does not lead to graceful degradation of the performance. Along this dimension, a rather catastrophic degradation can be observed.

Section 5 briefly addressed the question what would happen in the case of realistic noise in the audio part when we use novel features. This task does relate to WP1 (features) in combination with WP5 (experiments). Since the computational approaches (NMF, CM, DP-Ngrams) differ with

respect to how episodic information is dealt with during the learning, it is expected that they differ with respect to their robustness against background noise. In theory, episodic approaches (such as DP-Ngrams) will deal with noise in another way than less-episodic approaches. Moreover, it was expected that the novel features that were designed in WP1 show better performance in adverse conditions. WP1 designed two novel types of features (‘modified’-MFCC and ‘static adaptive’-MFCC); both these features were subject to various tests.
It appears that the modified MFCC features (MMFCC) outperforms MFCC, but that SA-MMFCC scores lower than both MFCC and MMFCC for DP-Ngrams. The discussion focuses on the amount of correlation in the features, and non-normality within the features and the way how HTK and DP-Ngrams deal with correction in different ways.