Unsupervised detection of words – questioning the relevance of segmentation

Louis ten Bosch, Hugo Van hamme, Lou Boves
Overview of this talk

• Background of this research
• Computational model of language acquisition
  – Focus on word detection
• Architecture
• Experiments
• Conclusion
Background

- European FP6 FET project ACORNS
  - Acquisition of Communication and Recognition Skills
- Dec 2006-Dec 2009
- Partners: universities/research institutes
- Aim: to design and develop a computational model for language acquisition, with focus on word detection
Language acquisition

• Babies and young infants perform a remarkable task:
  – Learn concepts, words and phonemes from spontaneous multimodal input
  – Bottom-up process + bottom-up expectation
  – Mildly supervised
  – On average 8 words/day during 10++ years
Input & output

• The **input** for the model will consist of
  – audio signals, in combination with
  – symbolic representations of the environmental context to which the audio signals refer

• **Output:**
  – detected auditory forms
  – possibly with some confidence value
Model

• Can perceive sensory input
• Has memory (sensory store, STM+LTM)
  – Can store representations
  – Empty at start of training
• Is able to communicate
  – Can output something
• Has a drive to learn
  – Wants to explain its input with what it currently knows
Multimodal sensory data

- Echoic + iconic memory
- Sensory store

Attention

- Episodic buffer
- Central executive buffer
- Short-term/working memory

Data flow and processes

- Rehearsal/retrieval
  - Episodic (memos, events)
  - Semantic

Abstract planning of output (communication)

Long-term memory

- 1 min - forever

2 sec

1 min

Output ("words", gazing)
Comparator – evaluating fulfilment of intrinsic need

ACORNS functional blocks in feedback loop

Multimodal input

Carer

Internal sensors

Estimation of intrinsic fulfilment of needs

Recognized output

Action generation

Behaviour (abstract)

External behaviour (“gazing time”)

Learning Agent

[quality of a parse perplexity]
Learning drive

• Any learning system needs something that is optimized. Here two aspects:
  – the ‘external’ appreciation it receives from its ‘carer’
    • Approval, disproval
  – its internal ‘happiness’ about to what extent input signals can be explained by using its internal representations, e.g.
    • Quality/completeness of a parse
    • Approximation of a decomposition

ACORNS Aalborg June 2008
Loop in more detail

Stimulus selection

Stimulus list

Exp design

Database

Stimulus processing /learning

Carer’s appreciation

Response LA

Stimulus processing /learning

Audio+abstr

Flag + conf

Audio+abstr

ACORNS Aalborg June 2008
Word discovery (e.g. Van hamme, 2007)

# utterances = # representations

positive numbers = positive numbers

‘raw’dat a matrix = base vectors

weights
Non-negative matrix factorisation

- NMF (Hoyer, 2004)
- One of modern techniques for structure mining

- Store all ‘episodes/traces’ in a (big) matrix $X$
- Apply methods from linear algebra to decompose $X = WH$
  - all components $\geq 0$
  - convex hull
- Interpret the columns of $W$ as associated to finer-grained representations
  - E.g. sentences $\rightarrow$ ‘words’ $\rightarrow$ phones
NMF basis weights
decomposition

utterance → matrix → M

w2, w3, w6
Segmentation lost

utterance

matrix

NMF

weights

b2

w2

w3

w6

2x

b3

b6

ACORNS Aalborg June 2008
Multimodal information in database

Utterances

Abstract labels

Aa

“”

Aa

Bb

Aa

ACORNS Aalborg June 2008
Multimodal information in database

Utterances

Abstract labels

Aa

“”

Aa

No information about Actual word form Phonetic properties Location in utterance

Bb

Aa

ACORNS Aalborg June 2008
Experiment

• System that is able to recognize some 15 words
  – including ecologically plausible words
• By building internal representations of these words
• Input: continuous speech with the characteristics of infant-directed speech (‘parentese’)
Claim: Segmentation

- Is not required in speech database for learning
- Is not used in training/updating the internal representations
- Is not provided as output of the learning algorithm

- But can be obtained from the decoding
Input data, experiments

• Database
  – FIN, SWE, NL, (UK)
  – 2m+2f per language
  – IDS, ADS
  – Small number of target words (13)

• Experiments
  – Mono-lingual, multilingual input
  – Batch mode/Incremental learning
NL, random ordering of stimuli

Accuracy versus nr of tokens processed

ACORNS Aalborg June 2008
NL, speaker blocked

accuracy versus nr of tokens processed

New speaker at 2000, 4000, 6000

ACORNS Aalborg June 2008
Assessment and truth
Word activation over time

![Graph showing word activation over time for telefoon (38)]
Learner can ‘correct’ the carer

**Left:** correct tags provided by carer

**Right:** same data but with noisified tags. Of two tags, 20 percent is incorrect

Of these incorrect tags, 89 percent is corrected by the carer

The increase in errors on correct tags is 4.2 percent

ACORNS Aalborg June 2008
Conclusions

• For the detection of word-like entities from multimodal input,
  – Information about segmentation is not required in speech database for learning
• Segmentation is not used in training/updating the internal representations
• Segmentation is not part of the output of the learning algorithm

• Segmentation can be obtained from the decoding
Upcoming experiments

- Detection of words from larger speech databases (\textit{hundreds of words})
- Detection of \textit{sub-word units}
- Use of \textit{semantic} features in combination with audio-based features (MFCC)
  - Weighting, under- and over-representations
- Focus (\textit{attention})
Semantic features

• Together with auditory signals, visual features are presented to the learner

• Carer and learner see a common scene
  – Objects + relations between objects

• car \rightarrow +movable, +wheels

• chair \rightarrow +legs, +on-sit-able
  – Details to be investigated
thank you for your attention

ACORNS Aalborg June 2008