A computational model of language acquisition: focus on word discovery

Louis ten Bosch[1], Hugo Van hamme[2], Lou Boves[1]

[1]Dept. Language and Speech, Radboud University Nijmegen, the Netherlands
[2]Dept. ESAT, Katholieke Universiteit Leuven, Belgium

`l.tenbosch@let.ru.nl`

Abstract

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

Index Terms: language acquisition, unsupervised word detection, computational modelling

1. Introduction

Healthy young infants perform language acquisition seemingly without any effort, but the large body of literature on cognition, language and memory shows that the underlying mechanisms are complex and far from completely understood (e.g. [9]). Undoubtedly, recognizing speech is tantamount to the mapping of continuous speech signals to discrete concepts, which we are used to think of as a sequence of word-like units. Infants learn to discover these units in speech without prior knowledge about lexical identities and despite the lack of clear word boundary cues in the signal. In order to do so, two related problems can be identified.

The first problem to be tackled by a young infant is the unimodal detection of words (word-like fragments) from utterances with mostly spontaneous connected speech. Newborns are not completely blank - they possess an auditory system that has been exposed in the pre-birth period to bandlimited sounds with the same type of rhythm and variation as speech. It appears that infants can identify their native language based on stress patterns very soon after birth. A few months old, they are able to segment words and distinguish between familiar and unfamiliar words based on stress patterns (whether or not the word actually means anything to the infant, e.g. [8]). An infant of six months old can distinguish vowels between the native and non-native phoneme space ([10]). And within 8 months, infants can segment words based on the statistical patterns in the observed phonotactics (e.g. [15], [9]). After 2 minutes of exposure, infants can then use the statistical properties of the co-occurrence of syllable-sized units to segment novel words.

These studies show that young infants are sensitive to the (statistical) structure in the speech signal on the level of phoneme or syllable-sized units. Evidently this does not necessarily imply that phoneme-like units are actually used to represent words in the internal (mental) lexicon. With respect to these lexical representations, both an episodic and abstractionist viewpoint are supported by experimental evidence. Episodic theories of speech perception assume that listeners store multiple entries, in the form of detailed perceptual traces ('episodes') ([2], [3]). In contrast, experimental data on e.g. perceptual learning in speech recognition are difficult to explain without hypothesizing more abstract phonological representations (features, phonemes or syllables) (for a discussion see [12]).

The second problem that a young infant is confronted with is the cross-modal pairing between a word and its potential referent. Parents, on average, may direct hundreds of utterances an hour to their children ([5]). That many words generate a potentially large number of ambiguities about possible word-referent associations. The behaviour of twelve to fourteen-month-old infants can adequately be modelled by evaluating the statistical evidence across many word-scene combinations ([17], [4]).

In this paper, we discuss a computational model based on both unimodal and cross-modal word learning. Its input and architecture are as much as possible motivated by cognitive plausibility. In order to simulate the stimuli a child receives from its environment, the input of the computational model consists of multimodal data with an 'audio' and a 'visual' modality. While the audio modality is represented by one utterance, the associated video modality is represented by a symbolic 'tag'. This tag is an abstract label associated with an object to which the utterance refers, and serves as a referent (so there is no real camera). The target word, its position in the utterance, and its acoustic/phonetic representation are unspecified, and it is up to the model to (statistically) determine the association between the word-like speech fragment and the referent.

The architecture and learning paradigm are different from conventional ASR. Instead, the model has similarities with the Cross-channel Early Lexical Learning (CELL) model [14]. It differs from CELL in that it does not assume that infants represent speech in the form of a lattice of pre-defined phonemes. While young children show sensitivity to phone-sized patterns in languages and sensitivity to native sounds after a few months, the current model does not assume that phones are available for use in the word detection task, thereby taking into account the distinction between 'perception units' and 'representation units' as discussed above. The model avoids the use of pre-existing representation for decoding the information in the input. Instead, the representations in the model gradually emerge from the multimodal stimuli that are presented. Actually, the cognitive plausibility of the model is an important consideration for its design.

In this paper, three experiments will be described. The first experiment shows that the learner is able to dynamically

Accepted after peer review of full paper

September 22 − 26, Brisbane Australia

build and adapt internal representations. The second experiment shows how the performance of the learner depends on the amount of speech data used for initialisation of the representations, and the amount of speech data used during the update of its internal representations. The third experiment describes how abstraction may be the result of grouping representations (based on structure in the representation space).

2. A computational model of word discovery

The model of language acquisition and speech communication that we are developing in the European project ACORNS [1] is based on four topics, viz. sensory front-end processing, memory access and organization, information discovery and learning, and interaction in a realistic environment. The model architecture (cf. fig. 1) is based on recent psycholinguistic research in speech and language processing ([7]).

Sensory front-end processing: In the first step, the computational model converts sensory input signals into an internal representation which is used in subsequent sub-modules for learning new patterns and for recognizing known patterns. The resulting representation includes the Mel-Frequency based cepstral coefficients as used in conventional ASR.

Memory organization and access: Cognitive theories of memory [7] distinguish at least three types of memory: a sensory store in which all sensory information is captured only for a very short time (in the order of 2 seconds), a short-term memory (working memory) and a long-term memory. The model makes use of these types of memory and stores (fig. 1). Multilayered representations are formed in which structures at a lower level map to structures at a higher level (abstraction, see below).

Information discovery and integration: In the experiments reported in this paper abstraction is based on Non-negative Matrix Factorization (NMF) [11] [6] [18]. NMF is member of a family of computational approaches that decompose a (large) matrix V into smaller matrices $W \cdot H$ such that the distance between V and $W \cdot H$ is minimised. Each 'episodic' column of V corresponds to an utterance and contains occurrence counts of specific acoustic events (defined via a VQ-codebook). The matrices W and H contain the information in the original matrix [6] [18]. By using NMF, a concept such as 'abstraction' receives a clear interpretation in terms of algebraic operations.

Interaction and communication: The carer model provides multimodal stimuli to the learner. Each stimulus consists of an utterance (infant-directed or adult-directed speech) in combination with a tag. The acoustic information is combined with the tag to form a high dimensional sparse representation, which is input for NMF. In order to simulate a learning environment, the learner is endowed with the intention to learn words. This is done by optimizing the appreciation from the carer, which in turn is interpreted as the optimization of the classification of the stimuli presented by the carer. To that end, W is used to decode the utterance (see below).

3. Experiments

3.1. Material

For training and testing, three databases are available (Dutch, Finnish, Swedish). For each language we have utterances from 2 male and 2 female speakers. In this paper we only report
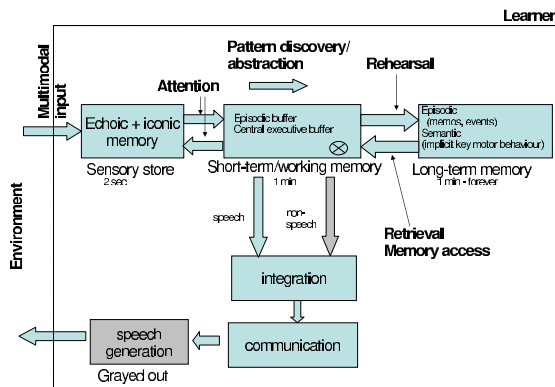

Figure 1: Global architecture of the ACORNS model.

about the Dutch database (the results on the other databases are comparable). Each speaker utters 1000 sentences in two speech modes (adult-directed, ADS, and infant-directed, IDS), making a total of 2000 utterances per speaker. The set of 1000 sentences contains 10 repetitions of combinations of the *target words* (a combination of nouns and proper names) and 10 carrier phrases. (The content of the three databases differs in details that are not relevant for this discussion). The set of target words has primarily been chosen on the basis of Child Development Inventories to include words that actually occur in real-life interactions. For each utterance, the databases also contain meta-information in the form of the 'visual' tag.

3.2. Procedure

A training run consists of interaction loops between the 'carer' (modelling an adult) and the 'learner' (modelling a very young child). At each interaction, the learner is confronted with a new (previously unobserved) multimodal stimulus from the carer. The learner then attempts to understand the audio part in terms of what it knows (that is, in terms of its internal representations trained so far). Its abstract reply is processed by the carer, who then continues the interaction by presenting the next stimulus. Once a multimodal stimulus has been processed by the learner, the information provided by this stimulus plays a role in subsequent updates of its internal representations and so sharpens or adapts these representations.

Basically, in the present setting, the NMF decomposition implies that the 'episodic' data matrix V is decomposed into $V \approx W \cdot H$ such that the KL-divergence is minimised:

$$D(V||W \cdot H) = -\sum_{ij} \left(V_{ij} \log(V_{ij}/(W \cdot H)_{ij}) \right) \quad (1)$$

in which the columns of W represent the representations on a level that is 'one level' more abstract than the level of the data in V. In the present experiment, the columns of W refer to words and word-like speech fragments.

During the training, it is up to the learner when and how often to update W. Evidently, in the general case the updated W will not be radically different from the previous W. Furthermore, cognitive plausibility of the model is enhanced by stipulating that W should not be updated on the entire collection of observed stimuli, but rather on the recently observed ones only (possibly with a forgetting factor that smooths the way forgetting operates). In the present learning algorithm, the *recency*
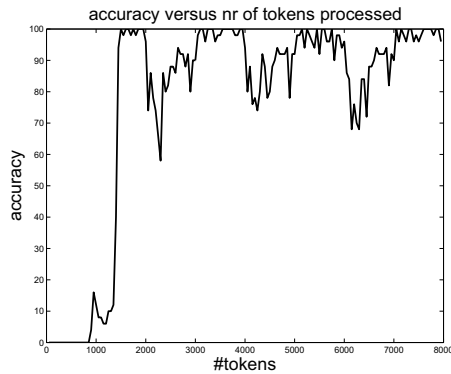
accuracy versus nr of tokens processed

Figure 2: (Dutch, speaker-blocked). Multimodal Dutch input data are presented blocked, speaker-by-speaker. The four speakers (female, male, female, male) start at number of tokens = 0, 2000, 4000, 6000.

is implemented by using the memory length (ml) with refers to the number of recently observed stimuli that is used for updating the internal representations.

3.3. Experiment 1

This experiment aims at showing that the learner is able to dynamically build and adapt internal representations via NMF. The training is based on all 8000 utterances, presented blocked by speaker, random within speaker-block. In total there are 13 different target words. The result is displayed in figure 2. The horizontal axis represents the number of utterances (tokens) presented during training. The vertical axis represents the accuracy of replied answers. The *accuracy* is defined as the number of correct responses (the learners' reply is correct if it is identical to the tag provided by the carer), divided by the total number of replies. To better monitor the 'instantaneous' accuracy, the plot shows the average accuracy obtained over the most recent 50 utterances. Each time a new speaker starts (around number of tokens = 2000, 4000, 6000), a drop in performance can be seen. This drop is mainly due to the different voice and speech characteristics of the new speaker which require an adaptation of the internal representations. The learner is able to catch up within about 1000 tokens (that is, approx. 70 tokens per word).

3.4. Experiment 2

The second experiment specifically focuses on the learning performance as a function of cognitively relevant parameters: the number of stimuli (denoted: nsbt) presented before the actual initialisation of the internal representations take place, and the *memory length* (denoted: ml) that indicates how many recent utterances are used to update the internal representations. The used database is a subdatabase consisting of 200 utterances by one female speaker, followed by a random ordering of 1800 utterances from other speakers. Figures 3 and 4 show the performance by varying nsbt and ml. It appears that delaying the initialisation to 500 utterances yields a slightly better performance on short term, but that this advantage tends to vanish in the end (fig. 3). In contrast, variation of ml (fig. 4) has a drastic influence on the performance. The results point at the relevance of the accessibility of recently preceived stimuli for the eventual performance.

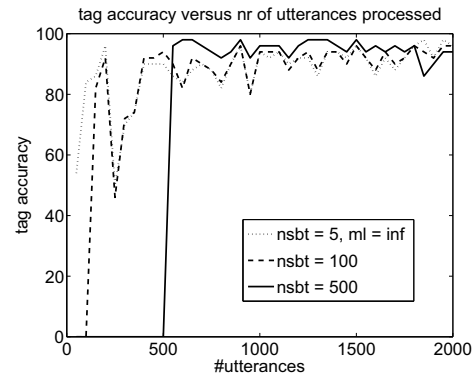tag accuracy versus nr of utterances processed

Figure 3: Performance as a function of the number of stimuli presented before training the internal representations. Delaying the initialisation leads to slightly better performance on short term.

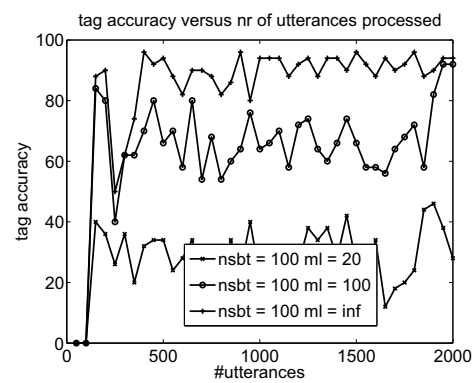tag accuracy versus nr of utterances processed

Figure 4: Performance for different values (20, 100, and inf) of the memory length (ml) used in the NMF update. The value 'inf' means that the entire past is taken into account.

3.5. Experiment 3

The aim of the final experiment is to show how abstraction may follow as a result of competition between crowded collections of representations on a lower level. For eaxmple, can speaker-dependent word representations be grouped in such a way that the common characteristics of these representations combine into one higher-level word representation? We used the Dutch database in which the tags have been extended to contain both the original tag *and* the identity of the speaker. This means that the new task of the learner is to find the association between a triplet {word, reference, speaker}. After training, each word now has four different representations (one for each speaker). We investigated the initialisation and update of the internal representations in more detail. All representations are one-to-one with columns in W after each NMF update step. The metric of the vector space in which these columns reside is defined by the *symmetrised* Kullback-Leibler divergence. This means that for any vector pair (v_1, v_2) the distance $KL(v_2, v_2)$ can be used as a dissimilarity measure, resulting in a KL-distance matrix M_{KL}. A 10-means clustering using M_{KL} then yields 10 groups (such that each group contains one or more word-speaker representations). Eventually, these groups appear to cluster the word representations as shown by table 1. Since the between-group variance Σ_b increases while
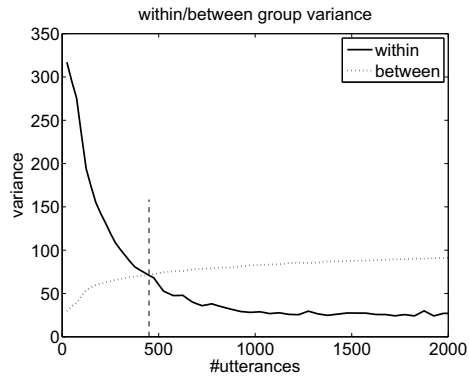
Figure 5: The evolution of the between and within group variance, based on M_{KL}.

group 1	$auto_1$, $auto_2$, $auto_3$, $auto_4$, $otto_2$
group 2	$otto_1$, $otto_3$, $otto_4$
group 3	$luier_1$, $luier_2$, $luier_3$
group 4	$mama_1$, $mama_2$, $mama_3$, $mama_4$, $papa_4$
…	…
group 10	bad_1, bad_2, bad_3

Table 1: Overview of abstract groups of representations defined after clustering low level (word-speaker) representations. The indices refer to one of the four Dutch speakers.

the average within-group variance Σ_w decreases (fig. 5), groups statistically emerge from the entire set of representations, which indicates that NMF is able to group speaker-dependent word representations towards one single abstract representation. To enforce abstractions of this type, the entire set of representations must be sufficiently large to have Σ_w/Σ_b falling below a threshold.

4. Discussion and conclusions

The computational model presented shows that learning relations between speech fragments and referents can be accomplished with a general purpose pattern discovery technique. For the recognition of 10 words, the performance of the learner is about 97%, which can be reached after having processed about 30-50 tokens per word. The learner is able to gradually improve the quality of its internal representations, by minimizing the Kullback-Leiber divergence between the observed data and the internal representations. Furthermore, at speaker changes, the internal representations are adapted to accommodate the new speaker characteristics (fig. 2).

The performance of the learner depends on the number of utterances that are used to initialise the internal representations. This effect is small (fig. 3). However, the effect of the memory length (the number of recent utterances used for the NMF updates) is substantial (fig. 4). This shows that the way representations are constructed and updated can be improved by better selecting those utterances that are critical for the learning, for example by focussing on those utterances that may help in better shaping the class boundaries. Furthermore, NMF is able to group speaker-dependent word representations to form a more abstract general-word representation (fig. 5).

One of the research lines that will be pursued in the near future deals with the effect of corrective feedback on the learn-

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

5. References

[1] Boves, L., ten Bosch, L. and Moore R. (2007). ACORNS - towards computational modeling of communication and recognition skills. Proceedings IEEE-ICCI 2007.

[2] Goldinger, S.D. (1998). Echoes of echoes? An episodic theory of lexical access. Psychological Review 105: 251–279

[3] Goldinger, S.D. (2000). The Role of Perceptual Episodes in Lexical Processing. In: SWAP-2000, 155–158.

[4] Graf Estes, K., Evans, J.L., Alibali, M.W., and Saffran, J.R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. Psychological Science. 18(3): 254–260

[5] Hart, B., and Risley, T. (1995). Meaningful differences in everyday experience of young American children. Baltimore: Paul Brookes Publishing Co.

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

[7] Jones, D.M., Hughes, R.W. and Macken, W.J. (2006) Perceptual organization masquerading as phonological storage: Further support for a perceptual-gestural view of short-term memory, J. Memory and Language 54, 265 281.

[8] Jusczyk, P.W. (1999) How infants begin to extract words from speech. TRENDS in Cognitive Science, 3: 323–328.

[9] Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. Nature Reviews Neuroscience, 5, 831–843.

[10] Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., and Iverson, P. (2006). Infants show facilitation for native language phonetic perception between 6 and 12 months. Developmental Science, 9, 13–21.

[11] Lee, D.D., and Seung, H.S. (2001). Algorithms for non-negative matrix factorization. In Advances in Neural Information Processing Systems 13, 2001.

[12] James M. McQueen, Anne Cutler, Dennis Norris (2006). Phonological Abstraction in the Mental Lexicon. Cognitive Science 30 (2006), 1113–1126.

[13] Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. Cognition, Vol. 52, 1994, pp. 189–234.

[14] Roy, D.K. and Pentland, A.P. (2002). Learning words from sights and sounds: a computational model. Cognitive Science, 26: 113–146.

[15] Saffran J.R., Newport E.L., Aslin R.N. (1996). Word segmentation: the role of distributional cues. J Mem Lang 35:606-621.

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

[18] 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, 131–134.