Special session proposal
From sounds to words: Modelling cognitive architecture for the speech sound acquisition

The study of human phonological development can provide insight into the modelling of cognitive learning architecture, and hence developmental robotics, as well as vice versa. As evidence mounts that speech sound can be learnt prior to semantics or syntax, efforts to computationally simulate such domain-general learning mechanisms have started to emerge. Developmental robotics offers an ideal platform to implement their models on.

Our three papers all demonstrate a language-independent learning models of speech units (phones, syllables or phonological words). They represent different stages in the cognitive maturity under investigation and are put in the ascending order. Howard and Messum's paper is motivated biologically and starts from the acquisition of basic phonetic units. Van Hamme looks at pre-phonemic acoustic perception and explores the effectiveness of recurrence-based learning. Finally Sato et al. examine, assuming phonemes, the role of syllables in word learning by using a robot's 'babbling'. All involve an important shared feature: a multi-modal approach to acquisition, making use of perception-production or cross-modal learning loop: another good reason to work with roboticists.

The purpose of the special session is to provide a forum where such computationally oriented speech acquisition researchers can present their work and discuss the future directions of this exciting emerging area of research. As important an objective is to disseminate this line of work to roboticists and researchers of other neighbouring disciplines (such as the cognitive and neural sciences) so that it can be extended and applied by these researchers.

Modelling unsupervised and caregiver tutored development of a young childs pronunciation

Ian S. Howard, Computational and Biological Learning Laboratory, University of Cambridge, U.K.
Piers Messum, Centre for Human Communication, University College London, UK.

We model learning to pronounce using Elija, a computational model of infant speech acquisition. Elija has a speech production capability based on a modified Maeda articulatory synthesizer and a perceptive system that is based on an auditory filterbank. Elijas representation of motor actions are akin to the gestural score used in the Task

Dynamics model and movement of his articulators between targets is implemented by assuming 2nd order dynamics that follow critically damped trajectories. Elija first discovers sounds in an unsupervised manner by finding underlying target motor patterns in a process formulated as an optimization problem. In an interactive phase, the natural reactions (that is, reformulations) Elija receives from a caregiver in response to the sounds he previously discovered are used; 1) first to reinforce those speech sounds that are appropriate for L1 and 2) to learn equivalence relations between his vocal actions and the caregivers speech. Here we show that Elija is able to discover simple vowel and consonant articulations and generate syllables with a repertoire that exhibits a wide distribution of both vowel and consonant qualities. Using caregiver reformulations, we then show that Elija progresses to generating sounds specific to the caregivers language L1. We then demonstrate that using the association between caregiver reformulations and his productions, Elija can then parse input speech and learn the names of objects by imitation.

Phonetic analysis of a computational model for vocabulary acquisition from auditory inputs

Hugo Van hamme
Katholieke Universiteit Leuven, dept. ESAT, Belgium

Psycho-acoustic experiments have shown that statistical regularities of sounds (phones) play an important role in the acquisition of vocabularies by infants. In earlier work a method for modelling early word acquisition was proposed, in which Non-negative Matrix Factorization (NMF) was used to exploit phone transition statistics. Deployed in an unsupervised way, NMF merely discovers word-sized recurring acoustic patterns that can later be detected in the acoustic inputs. In this method perceived acoustic signals are mapped to a histogram, i.e. feature counts observed over the course of an utterance. The features concern event co-occurrence, e.g. the occurrence of a phone pair or the co-occurrence of two prototypical spectra. The feature representation is therefore named Histogram of Acoustic Co-occurrence (HAC).

Cross-modal learning is achieved by augmenting the feature set with detectors of events or objects in other modalities such as vision. The feature set from one modality then acts as supervisory information for another modality, enabling multi-modal representations to be learnt and features in one modality to be generated based on observations in another. An 'idealized modality' is then considered, with perfect detectors for keywords present in training utterances. Keyword recognition has now become a special case of cross-modal learning.

In this work, results are reported on a data set with 21 keywords in Dutch, without exploiting phonemic knowledge in the NMF model. The emerging models are subsequently analyzed phonetically and lexically.

Discovery of words through babbling of syllables

Yo Sato, Caroline Lyon, Hagen Lehman and Chrystopher Nehaniv
Adaptive Systems Group, University of Hertfordshire, Hatfield, U.K.

Syllables have long been argued to be a unit which children build on to acquire phonological lexicon, but psycholinguistic evidence has been inconclusive. It *is* well-established,

however, that infants start their speech with 'canonical babbling', which takes a syllabic form and is gradually developed into more word-like sequences (Vihman et al 2009). We therefore examine this question: if infants *did* syllabify their input and 'test' their tentative segmentations ('*syllable candidates*'), would this improve the performance of word discovery? We explore it with a human-robot interaction experiment, which allows for incorporating another factor: positive reinforcents.

The learner syllabifies a sequence of phonemes online, and we use only one statistic, transitional probability of phonemes (to which Saffran et al 1996 famously showed infants are sensitive), for deciding on the preferred segmentation.

In our experiment the participants are asked to 'teach' shapes drawn on a cube to an infant-like humanoid robot and to 'praise' it when it utters a word. The robot, equipped with the syllabifier, responds to the participant by uttering a sequence of syllable candidates, according to the frequencies of the candidates in what it has previously heard. When it then recognises a reinforcement expression (e.g. 'Well done'), that particular sequence is registered as a word and gets a statistical reward.

We compare the results with other popular (non-interactive) methods of word segmentation both in performance and efficiency.