Modeling spoken language acquisition with a generic cognitive architecture for associative learning

Okko Räsänen[1], *Heikki Rasilo*[1], *Unto K. Laine*[1]

[1]Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland

okko.rasanen@aalto.fi, heikki.rasilo@aalto.fi, unto.laine@aalto.fi

Abstract

Human neo-cortex can be viewed as a modality invariant system for pattern discovery and associative learning. Similarly, research in the field of distributional learning suggests that much of human language acquisition can be explained by generic statistical learning mechanisms. The current paper argues that pattern processing capabilities of the human brain can be better understood if the process of early language acquisition is modeled using an entire cognitive architecture capable of unsupervised pattern discovery and associative learning. A high-level motivation and description for generic processing principles in such architecture are given, followed by examples of our current work in the field.

Index Terms: language acquisition, computational modeling, statistical learning, associative learning, multimodality, memory architectures

1. Introduction

The manner that human infants acquire their native language seems almost effortless. Instead of being explicitly taught, they learn to understand and produce speech through everyday interaction with other people in different contexts. Due to continuous linguistic exposure, children become able to understand speech in adverse acoustic conditions and despite different acoustic characteristics of different talkers. Moreover, they are able to fill in missing semantic and referential content of speech with the help of the context in which the communication takes place, and add tens of new words to their vocabularies on a daily basis.

The astonishing effectiveness of human learning becomes clear when one tries to build a machine able to understand spoken language. State-of-the-art automatic speech recognition (ASR) systems are based on the estimation of statistical correspondences between acoustic and textual representations. This calls for expert knowledge in phonetics and huge amount of work in preparing the speech material for the estimation of the system parameters. Still, the ASR systems perform well only on speech input that conforms to the acoustical and lexical content of the training material (see [1]). Novel words, grammatically incorrect constructions, background noise, and the paralinguistic aspects of everyday communication all cause major challenges to the system with a pre-defined set of capabilities. These shortcomings are not least due to the fact that ASR systems do not *understand* speech, but they are simply converting acoustic input into textual output using the given elementary units and their estimated correspondences in the both modalities.

Since a machine able to really understand spoken language is of interest from both practical and theoretical point of view, the question is how does one construct such a device? Given the complexity and notable differences of world's languages, it is evident that the language has to be *learned* through ever increasing experience. Since communication does not take place in a vacuum but the meanings of words only emerge through situated grounding, the system must also be able to understand its external and internal *context* and how the context is related to the *goals* and *needs* of the system. This means that learning must also take place outside the auditory domain, and that the meaning is ultimately coupled to the learned consequences of the auditory patterns. In order to accumulate experience, the system must be able to *interact* with the surrounding world in order to evaluate its behavior and proactively acquire knowledge of the relationship between sensory patterns and the external affordances. In other words, it can be argued that a system capable of full-scale language learning requires an entire cognitive architecture – an architecture that is able to learn the dependencies between its input and output modalities in a manner that the sensory patterns become predictive cues for the states of the surrounding world, thereby indirectly or directly guiding the actions of the learning system. Such architecture, if successful in explaining a variety of experimental data, would be a step towards an integrative theory of spoken language processing (see [2]).

The goal of the current paper is to illuminate the possibilities and challenges of studying language acquisition (LA) as a generic learning process of a cognitive system. The domain generality (versus language specificity) of the learning processes is inspired by the idea of neocortex as a universal pattern processing device [3] and motivated by the experimental findings in distributional learning that seem to point towards modality generic perceptual learning processes (e.g. [4]). Also due to the ongoing research of the authors, the focus is on the associative learning between patterns of different input and output modalities.

2. A general model for multimodal associative learning

The central characteristic of a cognitive architecture is the ability to learn meaningful patterns from sensory data without a priori knowledge of the patterns. This already poses two difficult questions: 1) *what is a pattern*, and 2) *where does the meaning emerge from*? According to our view, these questions can be approached from two different perspectives.

According to the first (traditional) view, and assuming a finite state space representation for a sensory signal, a pattern can be considered as a probabilistic construct of elementary events (or observations) that are dependent on each other in time and/or space. The dependency does not have to be deterministic, but above chance level probability of observing two or more elementary events in a specific configuration can already be

considered as a pattern. For example, an acoustic signal corresponding to a spoken word can be interpreted as a specific distribution of signal energy in time and frequency, analyzed up to a desired resolution. However, patterns discovered from a single data stream alone do not carry any meaning. According to the first view, the meaning of the pattern only emerges when the observed pattern is associated (*grounded*) to a jointly occurring contextual variable perceived through another modality or a variable representing an internal state of the system (e.g., active concepts in memory or current emotional state). For a spoken word such as "*a ball*", the contextual variable could correspond to the visual or haptic percept of a real ball. In other words, the pattern as such can be defined without the grounding component, but the meaning emerges only through the grounding process.

The inherent problem with the first viewpoint is that even though the quantification of statistical dependencies in time and frequency is possible, it is not possible to derive an "optimal" and finite set of distinct patterns (or categories) for a data set in isolation. The goodness of a representation is always measured with respect to the task or context against which the optimality of the patterns is reflected.

The second viewpoint argues that the patterns and their meanings are inherently intertwined so that there is no other without the other. According to this view, any processing beyond the learning of low-level sensory receptive fields always takes place in the context of multiple temporally proximate perceptions and mental states (memory, emotions) of the perceiver, and that this context affects the way how incoming sensory information from each modality is interpreted. This automatically attaches a set of multimodal associations to each percept and the elementary sensory events become bound together not only by their mutual co-occurrences but also by their *shared context*. In this case, the learning of pattern categories is no longer merely a question of bottom-up statistical clustering, but the categories are actually a function of the context: the sensory inputs belonging to the same pattern category are those that have equivalent predictions of the state of the world in other modalities, or equivalently, the current context defines the boundaries of a pattern category. In a sense, the idea of non-chance level dependency of elementary events in the first viewpoint is extended to allow these elementary events or states to occur across multiple input and output modalities of the system.

The obvious challenge with the latter viewpoint is that the estimation of all cross-modal dependencies through, e.g., normal joint probabilities is not possible due to the high dimensionality of the problem. Also, the direct associations between low-level sensory events (e.g., spectral features and visual receptive fields) may not be meaningful, but the useful dependency structure only emerges when at least one of the signal representations is sufficiently invariant to act as "labeling" for the remaining modalities (cf. visual tags in weakly-supervised LA simulations, e.g. [5]). How these abstracted representations can be learned from scratch is not clear, but one can hypothesize that the patterns in the sense of the first viewpoint may also allow the bootstrapping of the *cross-modal learning*.

In order to study the feasibility of both viewpoints, our goal is to develop a computational cognitive architecture where these theories can be tested. Since language (both spoken and written) is a good example of a complex system including patterns (acoustic, articulatory, written) at different scales (e.g., phones, words, phrases, letters, written words) and their meanings (the external and internal world of the perceiver), we use LA simulations to provide a solid experimental framework for unsupervised learning of meanings from data with a generic cognitive architecture. The basic framework is described in the following section.

3. Towards an interaction platform to study language acquisition with a cognitive architecture

LA is an excellent field to study unsupervised associative learning because of the several active modalities relevant in the learning situation, including visual and auditory information, articulatory motorics and emotional feedback. However, the earlier work on language acquisition modeling has specially focused on studying either learning of speech production (e.g., HABLAR, [6]; DIVA, [7]; Elija, [8]) or learning and grounding of words in weakly-supervised (e.g. [9]) or unsupervised (e.g. [10]) conditions. No unified work combining perceptual and motor learning have been carried out using an unsupervised learning paradigm; Elija comes closest, but for example its speech perception is strictly based on stored caregiver's exact reformulations of Elija's articulations using a dynamic time warping algorithm. We take a slightly different approach to develop a platform to study LA in multimodal environment. The idea is to include (but also control) the basic components of real-life LA that can be considered significantly important in the process, and to study how the learning process is guided or constrained by these factors. Figure 1 shows a schematic overview of our learning platform.

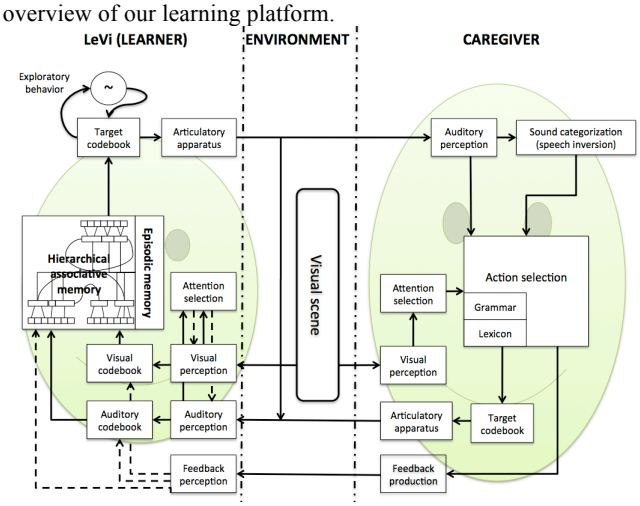

Figure 1: *A schematic of the basic learning platform*

The platform consists of three major components: the learning virtual infant agent, or LeVi, a caregiver, and a simulated environment in which the learning takes place. All signaling between the caregiver and the learner takes place through the simulated environment.

In the simplest case, **the environment** consists of a simulated visual scene with a number of objects and events with different visual characteristics. For enhanced ecological plausibility, the environment can be expanded to have, e.g., a true spatial dimensionality, specific transfer function characteristics for acoustic signals, additional background noise and more complex interaction between multiple agents.

LeVi is equipped with auditory and visual perception capabilities, and ability to follow caregiver's attention and to perceive emotional feedback (see [5]). In addition, visually salient articulatory gestures, or, e.g., orthographical representation of speech can be used as separate inputs to address specific research questions. LeVi's output modality consists of an articulatory apparatus that transforms articulatory gestures to speech. In addition to LeVi's pre-wired sensory and motor specific processing stages, it will be equipped with a general-purpose hierarchical associative memory and an episodic memory at the top of the memory hierarchy. All sensory channels and motor outputs are represented in the system using a universal coding by partitioning the parameter/feature space of each modality into a finite number of states and then coding the state of each modality with the unique labels assigned to these partitions. Both the state partitions (codebooks) and the patterns abstracted from the universal code are subject to learning. Our goal is to study different learning mechanisms and develop further LeVi's memory architecture.

The caregiver is also equipped with an articulatory apparatus similar to LeVi, but the sizes of the articulators and the fundamental frequencies of glottal excitation differ. In addition, the caregiver has pre-programmed knowledge of native speech sound production in articulatory terms and is able to interpret LeVi's speech in terms of intended (learned) articulatory targets. The caregiver is also able to attend specific objects in the simulated environments and construct grammatically correct sentences from the observations.

The key advantage of the platform in the study of unsupervised pattern discovery algorithms is in the increasing understanding of the involved processes without resorting to trivial synthetic test signals. For example, it is known that the sensory patterns in speech perception and motor patterns in speech production are connected to each other although their relationship is highly non-linear. This makes evaluation of the learned patterns and cross-modal dependencies feasible.

3.1. Basic principles of the learning process

In order to study whether LA can be simulated using a generic associative learning process, we have set up the following hypotheses to be verified or falsified: **i)** All input and output streams should be connected to each other in a hierarchical manner, allowing learning of predictive associations between any two modalities and across levels of hierarchy. **ii)** Principles of memory organization are similar for both sensory inputs and motor outputs. **iii)** Associative learning should not be only based on direct joint probabilities of sensory events and their abstractions, but each input also activates and updates the entire semantic network of the perceived patterns ("*priming*"). This allows the system to alleviate the sparseness of input statistics for high-level categorical associations. **iv)** The learning is driven simultaneously by bottom-up statistical structure of input streams and cross-stream dependencies (cf. Hebbian learning), and through feedback that induces plasticity in the memory structures. Feedback can be external (rewards or predictions of rewards, including caregiver feedback), or internal (failure of existing memory structures to predict the current situation; see, e.g. [11]). **v)** The semantic connectivity (synonymy; see, e.g. [12]) of internal representations is a function of the current external and internal context (or task) instead of being fixed. **vi)** The system will solve the relative importances of different input signals and their features in different situations by itself (cf. discussion in [13]). Naturally, the computational architecture must be flexible enough to support partitioning of input signals into multiple parallel representations at different scales.

Further, **vii)** Once set up, the only inputs or parameters affecting the system performance should arrive through established sensory channels, or through purely autonomous behavior. This means that different tasks (e.g., word learning and grammar acquisition) should be accomplished using the same overall system without further intervention. **viii)** There should be no a priori biases to use any specific input modalities or combinations of input modalities in specific tasks. These preferences should be universal across tasks or learned from experience. For example, the use of visual information to aid in speech perception should be based on the learned correlation between the visual and auditory modalities, and this learning is enabled by the fully interconnected modalities at different levels of hierarchy.

4. A learning example: acquisition of native phonetic categories in speech production

An early version of the learning platform has been utilized in studying phonetic category learning in infants [14]. The simulations were based on the findings that caregivers provide feedback to infant's babbling based on the adult-likeness of the productions, immediately shaping the infant's vocalizations [15], and that it is the caregivers who imitate their children when they hear vocalizations that can be interpreted as communicative acts [16]. By using these assumptions for caregiver-learner interaction, a simulation was run where the LeVi started to explore random canonical babbling, completely unaware of native language phonemes. By reinforcing the memory traces for articulatory productions that were associated with positive response from the caregiver, LeVi was able to converge to the native phone categories a priori known by the caregiver [14].

When the babbling became sufficiently close to adult-like speech (in terms of articulatory targets), the caregiver started to imitate LeVi. For example, "*amam*" produced by LeVi could be interpreted and imitated as "*mama*" by the caregiver. Mediated by the shared context of the communicative situation and cross-situational learning, LeVi was able to associate his own articulatory actions and acoustic speech output to the perceived and learned representations of caregiver speech, thereby also allowing the imitation on behalf of the infant [14].

After the learning process, LeVi was able to recognize and thus imitate caregiver's vowels almost perfectly (phone recognition rate 99.3 %) and consonants' place of articulation with an accuracy of 86.0 %. The results indicate that the highly non-linear inversion from caregiver's acoustic output to infant's articulation can be learned with satisfactory accuracy with a reasonably simple learning procedure when the language knowledge of the caregiver and interactive learning situation are properly modeled. When the phoneme categories have first obtained meanings in the articulatory domain during the rewarded babbling phase, the pattern processing device of the infant presumably begins associating the following reactions of the environment under these meaningful phoneme labels. When the caregiver's speech is later listened, and phoneme recognition happens directly based on the articulatory labeling, caregiver's speech can be easily imitated and new words added to the expressive vocabulary.

4.1. Future experiments and predicted advantages

Examples of the interaction between different modalities in language related tasks are numerous. McGurk effect confirms that the speech percept of a listened syllable can change due to watching a visual image of a person pronouncing a different syllable [17]. Listening to speech has been shown to activate motor areas in brain, suggesting that the motor system is directly coupled to the perception of acoustic speech stimuli (e.g. [18]).

The proposed learning platform enables further experiments with additional modalities and associations between different levels of hierarchy, allowing to study, e.g., the role of learned articulation in speech perception and formation of phonetic categories. Associations can be learned e.g. between the visual and acoustic domain to link caregiver's lip and jaw movements into either infant's or caregiver's acoustic production, or from physical objects into auditory representations of corresponding words.

Updating the associations across the whole semantic network during the learning process would have several advantages. As an example, let's imagine a situation where the virtual infant has learned to understand and imitate the speech of the mother, but has never heard speech of a male speaker. It is known that children are not always able to understand the speech of new previously unheard speakers (e.g. [19]).

Now let's imagine that the infant is able to identify visually a physical object of a ball. He has heard, and can link the auditory speech signal "ball" by the mother to the object, and is possibly able to pronounce the word himself. Now when a male speaker pays attention to the same object and pronounces "ball" with an unidentified voice, the associative network of the infant should be able to link the new vocalization to the object, but also to all the existing memory representations of the object. This way the new male vocalization would be linked to the mother's vocalization of the word and the phonemes, and further to the infant's articulation of the phonemes allowing the infant to rapidly adapt to new speakers.

In addition to speaker adaptation, the cognitive architecture would presumably enable robust language modeling without extensive training corpora through learning of synonymy or semantic connections between words, or using the most reliable acoustic or visual features for recognition depending on the observed background noise conditions.

5. Conclusions

The change from study of task specific machine learning setups towards a general learning architecture allows the analysis of how much of language acquisition can be explained by non-speech specific statistical learning mechanisms (given the physiological constraints of auditory perception and speech production). If successful, this methodology for associative learning will be also applicable to other domains outside language dealing with automatic analysis of data streams. A notable advantage in this type of generic approach is that the simulations addressing different aspects of LA are directly compatible, as the same framework is supposed to describe these multiple phenomena in parallel. Only when the assumptions of domain generic learning processes fail, one should start looking for specially tailored solutions for specific aspects of LA.

Naturally, the approach does also come with a number of drawbacks. For example, the ecological plausibility of the simulations is highly affected by the used interaction framework and complexity of the simulated environment. Also, evaluation of the system performance becomes more difficult when moving away from strictly defined machine learning tasks towards more natural interaction framework.[1]

6. References

[1] Lippmann, R., "Speech recognition by machines and humans", *Speech Communication*, 22, 1-15, 1997.

[2] Moore, R. K., "Spoken language processing: Piecing together the puzzle", *Speech Communication*, 49, 418-435, 2007.

[3] Mountcastle, V.: An Organizing Principle for Cerebral Function: The Unit Model and the Distributed System. In *The Mindful Brain,* (Edelman G. M. & Mountcastle V., eds.), Cambridge, MA: MIT Press, 1978.

[4] Saffran, J., Johnson, E., Aslin, R. and Newport, E., "Statistical learning of tone sequences by human infants and adults", *Cognition*, 70, 27-52, 1999.

[5] Räsänen, O., "Non-auditory cognitive capabilities in computational modeling of early language acquisition", accepted for publication in Proc. Interspeech'2012, Portland, Oregon, 2012.

[6] Markey, K. L., The sensorimotor foundations of phonology: a computational model of early childhood articulatory and phonetic development, Ph.D. Thesis, University of Colorado, Boulder, 1994.

[7] Guenther, F., "Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production", *Psychological Review*, 102(3), 594-621, 1995.

[8] Howard I. S. and Messum P, "Modeling the development of pronunciation in infant speech acquisition." *Motor Control* 15(1), 85-117, 2011.

[9] Boves, L., ten Bosch, L. and Moore, R. K., "ACORNS – towards computational modeling of communication and recognition skills", Proc. ICCI-2007, 349-356, 2007.

[10] Räsänen, O.: "A computational model of word segmentation from continuous speech using transitional probabilities of atomic acoustic events", *Cognition*, 120, 149-176, 2011.

[11] Schultz, W., "Predictive Reward Signal of Dopamine Neurons", *J. Neurophysiol.*, 80, 1-27, 1998.

[12] Räsänen, O., "Context induced merging of synonymous word models in computational modeling of early language acquisition", Proc. ICASSP'2012, Kyoto, Japan, pp. 5037-5040, 2012.

[13] Scharenborg, O., "Reaching over the gap: A review of efforts to link human and automatic speech recognition research", *Speech Communication*, 49, 336-347, 2007.

[14] Rasilo, H., Räsänen, O. and Laine, U., "Feedback and imitation by caregiver guides a virtual child to learn native phonemes and the skill of speech inversion", in preparation, 2012.

[15] Goldstein, M. H. and Schwade, J. A., "Social Feedback to Infants' Babbling Facilitates Rapid Phonological Learning", *Psychological Science*, 19(5), 515-523, 2008.

[16] Gros-Louis, J., West, M. J., Goldstein, M. H., and King, A. P., "Mothers provide differential feedback to infants' prelinguistic sounds", *International Journal of Behavioral Development*, 30(6), 509–516, 2006.

[17] McGurk, H. and MacDonald, J., "Hearing lips and seeing voices", *Nature,* 264, 746–748, 1976.

[18] Wilson, S. M., Saygin, A. P., Sereno, M. I. and Iacoboni, M., "Listening to speech activates motor areas involved in speech production", *Nature Neuroscience*, 7, 701–702, 2004.

[19] Houston, D. M. and Jusczyk, P. W., "The role of talker-specific information in word segmentation by infants", *Journal of Experimental Psychology,* 26(5), 1570-1582, 2000.

[1] This work was partially funded by Nokia Research Center, the Graduate School in Electronics, Telecommunications and Automation (GETA), Nokia Foundation, TES, and KAUTE.