Computational modeling of phonetic and lexical learning in early language acquisition:

existing models and future directions

Okko Räsänen

okko.rasanen@aalto.fi

Aalto University, School of Electrical Engineering, Department of Signal Processing and Acoustics

PO Box 13000, FI-00076 AALTO, FINLAND

Telephone: +358-9-470 22499

Fax: +358-9-460 224

Abstract

This work reviews a number of existing computational studies concentrated on the question of how spoken language can be learned from continuous speech in the absence of linguistically or phonetically motivated background knowledge, a situation faced by human infants when they first attempt to learn their native language. Specifically, the focus is on how phonetic categories and word-like units can be acquired purely on the basis of the statistical structure of speech signals, possibly aided by some articulatory or visual constraints. The outcomes and shortcomings of the existing work are reflected onto findings from experimental and theoretical studies. Finally, some of the open questions and possible future research directions related to the computational models of language acquisition are discussed.

Keywords: language acquisition, distributional learning, computer simulation, phonetic learning, lexical learning

1. Introduction

Normally developing human children acquire competence in their native language in a remarkable manner in the absence of explicit teaching. Towards the end of the first year of their life, infants are already sensitive to the fine details of the native phonetic contrasts [1] with decreased sensitivity to non-meaningful contrasts [2,3], understand several spoken words [4,5], and are on their way to the production of linguistically relevant speech acts (e.g., [4-6]). Although a great deal of knowledge regarding human first language acquisition (LA) has been accumulated, the underlying principles of the learning process are not well known. Unifying, generally accepted, and verified theories explaining language acquisition from birth to the linguistic competence of a normal five-year old child simply do not exist. Because language as a phenomenon is extremely complex by itself, an understanding of the big picture and verification of comprehensive theories requires integration and collaboration of research across a large number of disciplines [7].

One important part of the LA research is the development of computational models that can be used to test hypotheses regarding language learning: any theory or model related to LA should also endure computational implementation of the model so that the functionality of the model can be verified through simulations in realistic settings. However, implementation of a theoretical model typically leads to a number of issues: first of all, a theory may be useful in understanding the LA process and in the formulation of more specific research questions, but the theory may be too vague to be implemented as an algorithm (cf. Marr's levels of analysis; [8]). Another possibility is that the theory covers the computational aspects of the process, but is not implementable given the existing limitations in the hardware. However, possibly the largest issue is that, given the complexity of the phenomenon, no single model can address all aspects of the learning problem simultaneously. This

means that a large number of assumptions need to be made regarding the processes not studied in a given simulation, yielding very different results for different assumptions and having significant consequences to the ecological validity of the results. Ecological plausibility is also an issue with the data used in the experiments since no prerecorded data set can correspond to the sensory experience available to a human infant, not least because infants not only passively perceive but also actively affect their sensory experience by their own actions.

Finally, the question whether the proposed computational models can be actually realized in biological brains is not a trivial one. The idea in the majority of the models is to be able to explain aspects of LA at a phenomenal level and not to explain in detail how the brain executes the required computations. It is well known that there are multiple computationally equivalent implementations for any problem, making direct comparison of detailed neurophysiological data and computational models an ill-posed problem (e.g., connectionist networks can be used to approximate any non-linear functions; see, e.g., [9], for a neural network-based implementation of the dynamic time warping (DTW) algorithm). Finally, the memory requirements or the computational complexities of the algorithms can be usually cut down with approximations and heuristic solutions (e.g., [10]), making it difficult to judge whether a model scales up to real world learning conditions in the human brain (but see [11]). Together with the lack of knowledge on the principles of information processing in the brain, many researchers choose not to concentrate on the correspondence between their algorithms and their possible biological counterparts. Instead, the majority of the computational work attempts to reproduce results from the experimental studies or simply to study the extent of the distributional learning hypothesis using advanced machine learning techniques. Still, some aspects, such as incrementality versus batch learning (local vs. global criteria), temporal causality, and, e.g., the fidelity of the internal

representations for the entire history of sensory inputs, can be addressed when evaluating the plausibility of computational models.

Despite the evident challenges, mathematical and computational models can provide useful knowledge regarding the LA process. First of all, they can set statistical baselines to the *learnability of data* (what, at least, can be learned from the data with the given constraints and assumptions). Estimation of *upper limits of learning* can be also attempted (e.g., [12,13]), although reaching conclusive results without notable simplification from the real world settings is usually difficult. Moreover, in addition to *replicating behavioral findings*, computational models may also help to *formulate new behavioral hypotheses* to be verified and to better predict or understand the nature of different developmental disorders related to the language faculty. Finally, computational models able to demonstrate unsupervised learning capabilities are of interest to many other fields dealing with complex data, e.g. automatic speech recognition (ASR), robotics, intelligent data mining and analysis, and context-aware computing.

The goal of the current paper is to review the existing work on the computational modeling of language acquisition in order to provide a synthesis of the existing knowledge on the field. The review specifically focuses on computational studies related to the learning of phonetic categories and words from *continuous speech* without assuming any a priori linguistic knowledge, i.e. on the acquisition of the very first building blocks upon which later stages of language learning can rely on. We adopt the distributional approach to LA in this work, trying to understand the language learning process as a statistical inference problem with as few assumptions regarding innate linguistic abilities as possible (see [14]). Since no model has been successful in acquiring a phonemic or orthographic representation of auditory speech (see also the discussion in [15]), the work related to computational models dealing with word learning or grammar induction from phonetic transcription or the orthographic layer are

excluded from this work, but are discussed in depth elsewhere, e.g. in [16-18]. On the same basis, models of adult word perception such as TRACE [19] or Shortlist [20] are not reviewed since they do not explain how the representations used in the models come into being during the early human development, but the reader is recommended [21] for an overview. This also excludes models that use a deterministic mapping from linguistically motivated representations to articulatory or acoustic features in their analysis (e.g., [22]). Also, the relationship between human and machine speech perception is covered in the special issue of Speech Communication "*Bridging the Gap between Human and Automatic Speech Recognition*" [23] and is not discussed here. Finally, although central to all languages, syntax and syntactic learning are not covered here since the existing work on computational modeling of LA has been purely focused on the search for fundamental units of speech perception. It has been assumed that the syntax can be learned only after the learner is able to interpret incoming speech signals as sequences of categorical lexical elements, and no existing model has been able to integrate the acquisition of these basic building blocks and their syntactic relationships under a single computational framework.

The paper is organized as follows: the next section discusses the basic theoretical background and a number of challenges related to the computational models of LA from continuous speech. The third section reviews a number of studies attempting to explain how categorical perception of phonetic contrasts may emerge on the basis of distributional information in speech signals. In the fourth section, we discuss how words and word semantics can be learned using unsupervised and associative learning mechanisms, and the final section draws together the current findings and their shortcomings. A separate treatment of syntactic learning is beyond the scope of the current review, but the focus will be on how even the most basic representations of spoken language can be learned – the representations upon which later linguistic learning can then take place.

2. The theoretical background of computational models of LA

The theoretical background of the computational models of LA can be understood from the perspective of two theories of LA that try to integrate the existing findings on early language acquisition: the native language magnet theory expanded (NLM-e; [24]) and the PRIMIR framework of language acquisition [25]. Although the theories do not claim fully explicit sequential ordering of developmental stages, their main connotations are as follows: the NLM-e states that language learning starts by learning the distributional properties of native speech sounds, leading to enhanced phonetic perception of native contrasts. Once phonetic perception has achieved a sufficient proficiency level, words can be segmented and learned based on the sequential organization of perceived phonetic units. PRIMIR, on the other hand, states that the organization of the language faculty is driven by the acquisition of word forms directly from the acoustic surface properties of speech signals (or on the "general perceptual plane") that combine both phonetical and indexical features. Later, once sufficiently many lexical tokens have been memorized, the learner is able to discover similar sub-word patterns across different word tokens, giving outset to the sub-word level organization and perception of language (phonemes). Phoneme representation of spoken language then enables fast accumulation of new vocabulary since the learned tokens automatically generalize to their acoustic variants through phonemic encoding. However, PRIMIR also maintains that the distributional properties of native phonetic units affect the way that spoken words are represented in the general perceptual plane, but this representation does not yield proper generalizations across different contexts and talkers without the help from lexical learning.

The distinction of the learning order of words and subword units between PRIMIR and NLM-e also divides the computational models into two basic categories: to those models where the phonetic system is learned before words (cf. NLM-e) and those where proto-lexical items are learned before the phonetic system (cf. PRIMIR). Despite the intuition that knowledge of sub-word units such as phones

or syllables must precede word learning because they are the basic building blocks of words, the answer to the question of representational learning order is not obvious. As Peter Jusczyk [26] wrote,

"One potential problem with using characterizations of the mature state to guide research about the initial state is that it may lead one to assume that the elementary units that yield the best description for the adult's knowledge function as elementary units during acquisition of the knowledge… … to the extent that a description of the adult state of knowledge of the sound patterns of the language is best captured by assuming phonemic representations, we have to provide an explanation of how these representations develop in the course of language acquisition."

Researchers in speech sciences have struggled for decades in order to find comprehensive descriptions for the mapping from variable speech sounds of the acoustic domain to the invariant and abstract linguistic units such as phones or even phonemes that can be placed serially to construct larger linguistic units such as words [27]. Despite tremendous amount of work on this issue, the basic problem always seems to be that the variability in the acoustic tokens cannot be captured into segmental models that assume independence of a phone from the preceding and following phones, making accurate categorization of phone-sized units impossible when they are isolated from their context. The standard solution to get away with the difficulties at the segmental level is to extend the units to be context sensitive by making their characteristics dependent on the neighboring phones. Another possibility is to use lexical memory for disambiguation of difficult segments by first retrieving the most likely word, given the sequence of initial phone hypotheses, and then seeing which phones (or phonemes) correspond to the ambiguous segments (e.g., TRACE, [20]; but see also [28]). Although feasible for segment disambiguation, both of these approaches are not compatible with the idea that speech perception consists of the perception of sequences of independent units that are realized as sequences of phones. Otherwise the surface structure of speech should enable this type of serial

segmentation into the discrete building blocks despite the variation introduced by coarticulation. If a phonemic system of sequential discrete elements exists, at least it seems that there is no direct access to it from the surface structure of speech.

The second issue from the perspective of language learning is that, even before the categorization of speech sounds into a finite set of categories, the discovery of the phone-like segments themselves is problematic in the absence of a priori knowledge of their structure. While it has been proposed that humans and other primates are capable of primitive segmentation of a continuous acoustic stream into acoustically coherent segments, namely, basic-cuts [29,30], the correspondence of these units with linguistically motivated phones is not direct. Several diverse computational methods for blind segmentation of continuous speech into phone-like units have been proposed [31-36] and they all systematically fall short of ideal performance if manually performed phonetic transcription is used as a reference. What is common to these methods is that they analyze changes in spectral content of the speech signal and hypothesize phone boundaries at points of notable discontinuity in the spectrum. While this type of chunking of the speech signal can detect approximately 70-80% of phone boundaries (with ±20 ms accuracy), many of the phone transitions are still detected with very low accuracy. For example, the overall quality of the segmentation is too low to be directly utilized as a front-end processing before feature extraction in ASR-systems (see also [37]). Only when context-sensitive phone models are imposed in a top-down manner and taught to the segmentation algorithm using pre-recorded speech data in supervised training paradigms, the segmentation algorithms reach segmentation performance that starts to converge with the definitions of phone boundaries (e.g., [38-40]). In this case the segmentation models are essentially built manually upon the criteria that are also used to evaluate their performance.

There is also notable evidence that human listeners do not only pay attention to the sequential evolution of phonetic units, but store detailed supra-segmental and episodic acoustic information regarding speech tokens. Variables such as talker and speaking style characteristics have been shown to affect speech perception performance (e.g., [41]). Young infants' representations of words appear to be holistic and contain information regarding not only phonetic, but also indexical and stress information related to the word forms [42-44]. Infants as old as 14 months also fail to discriminate phonetic contrasts in otherwise similar novel words when learning names of external referential objects [45].

Due to the inability of the phonemic/segmental view of speech perception to explain the acoustic mapping problem and the effects of suprasegmental acoustic details on adult speech perception, contemporary views have emerged that question the entire existence of segmental phonemes as fundamental units of speech perception [27,41,46]. These views are also supported by the detailed analyses of pronunciation errors in young children that point towards suprasegmental or even word level representations of produced words instead of phonologically motivated encoding of word forms (e.g., [47], see also [48] for an overview).

Even if the phonemic representation of language is present in our minds and used to code and decode linguistic messages, the problem is that the mapping from acoustic signals to phonemic representations cannot be easily learned directly from continuous speech by simply analyzing distributional properties of acoustic events without support from some additional source of information. Even if the phone segmentation would succeed with perfect accuracy, the speech sounds from a number of different talkers do not neatly group into clusters of phonetic categories in terms of their acoustic features, but largely overlap in the acoustic space. This is demonstrated in the work of Feldman et al. [12] where Bayesian modeling (clustering) with theoretically well-justified mechanisms for learning was used to learn phonetic categories of American English vowels from the formant data

of Hillenbrand et al. [49]. Despite the fact that the formant frequencies were estimated from isolated productions of the vowels instead of continuous speech, classification of the segments into correct phone categories was far from perfect. Only when support from the lexical layer was utilized in order to perform context-sensitive classification, the categorization of the segments became successful [12].

Given all the considerations above, it is not obvious that the infants would learn their native language by *first* acquiring a fully functional phonetic system of the language and only then start learning words as sequences of phones. As for the phonemes, some sort of proto-lexical layer becomes almost necessary as long as the phonemes are defined as the smallest units of language that contrast between two words. However, the evidence is not conclusive. First of all, the above discussion does not take into account the fact that human infants are not only equipped with auditory capabilities, but can also use information from other modalities to disambiguate situations that are not separable in the purely auditory domain. Another important factor is that human infants are not only listening, but also experimenting with speech production. Infants are equipped with an articulatory system that gives them access to the constraints and possibilities of speech sound generation, revealing another representation of speech acts that is not linear with respect to the auditory domain.

In general, what kind of sub-lexical and lexical structures can be actually learned from speech with different types of approaches, constraints, and assumptions is well worth investigating. As will be seen in the following two sections, partial success has been achieved with both lexicon-first and subwords-first approaches, but no single model has been so far able to convincingly explain the integral development and interdependence of the two systems.

3. Computational models of phonetic learning

One of the basic hypotheses in the NLM-e theory [24] is that the first stages in LA are dominated by the attunement of the infant to the distributional properties of speech sounds. More

specifically, NLM-e states that the exposure to infant directed speech drives distributional learning of native phonetic categories which then form the basis for phonotactic segmentation of words from continuous speech [24]. This theory is supported by behavioral findings. Although infants are born with equal sensitivity towards all phonetic contrasts in the world's languages [50,51], studies indicate that infants show heightened sensitivity to native phonetic contrasts towards the end of their first year (e.g., [1]), whereas sensitivity to non-significant non-native contrasts decreases [2]. Moreover, studies show that success in native phonetic category perception predicts later proficiency in the language (e.g., [52,53], see also [24] and references therein for a more comprehensive review on the topic). Several computational models have been proposed to demonstrate NLM-e like distributional acquisition of phonetic categories from speech.

A computational study by de Boer and Kuhl [54] concentrated on the acquisition of American English /i/,/a/, and /u/ vowels, comparing the effects of infant-directed speech (IDS) and adult-directed speech (ADS). They used vowel data from single-vowel words recorded in natural conversations between a mother and an infant or a mother and another adult. Formant frequencies of the vowels were extracted automatically from the vowel frames. Then the standard EM algorithm [55] was applied to fit the data with a Gaussian mixture with three components, i.e. the correct number of categories was specified manually. The learning procedure was performed separately for ADS and IDS material. Qualitative analysis of the results showed that the IDS produced more realistic clusters with resemblance to the underlying vowel distributions, whereas ADS runs often led to one unrealistic outlier cluster or only two clusters with significant contribution in explaining the data distribution. This led the authors to conclude that IDS led to more effective learning of vowel categories than ADS, supporting the idea that the exaggerated articulation of IDS facilitates distributional learning in infancy. However, the conclusions are somewhat limited by the fact that the correct number of vowel categories

was provided manually to the system and due to the lack of comprehensive quantitative analysis of the results.

In the work of Vallabha, McLelland, Pons, Werker and Amano [56], the emergence of categories for a subset of Japanese and English vowels was studied. In the study, two model variants were further developed from the original algorithm presented in [57] and were used to learn explicit probability distributions for two first formants and duration for English /I, i, ε, e/ and Japanese /i, i:, e, e:/ vowels extracted from monosyllabic words. Their first variant of the algorithm, Parametric Algorithm for Online Mixture Estimation (OME), assumes that the vowel stimuli are drawn from multivariate Gaussian distributions and that each vowel token is independent of the previous tokens. Given these assumptions, the task of the algorithm was then to estimate the correct number of Gaussians, their means and covariances, and the respective mixing probabilities from a set of vowel tokens. Instead of the algorithm of [57], the estimation of the parameters was performed with an on-line version of the standard EM algorithm. After training the OME algorithm with the vowel data and labeling the distributions in a post-hoc manner for evaluation, the correspondence of the categories was compared to the ground truth. 92.7% of the English and 91.1% of the Japanese vowel tokens were observed to b e classified to correct categories if data from a same talker was used for both training and testing. When the model taught by one talker was used to classify tokens from other talkers of the same language, classification rates of 69% and 77% were obtained for English and Japanese vowels, respectively [56].

The second variant of the algorithm studied by Vallabha et al. [56], Topographic OME (TOME), does not make the assumption that vowels are drawn from a Gaussian distribution, but attempts to discover category distributions by dividing the feature space into a 3-D grid (the first two formants i and j and duration k) with 25x25x25 = 15625 cells and then studying the vowel data density

flowing to these cells, essentially modeling the discretized joint distributions of the features. In TOME, each cell $c_{i,j,k}$ is fully connected to a set of category units $r \in R$ with weights $w^r_{i,j,k}$. For each data point in the training set, weight w of the best matching cell is updated with respect to categories r, leading ultimately to a situation where only a small number of dominant categories r cover the majority of the feature space, whereas the remaining categories become pruned due to diminishing mixing weights. This approach allows the modeling of arbitrarily shaped distributions, only limited by the resolution of the used grid and the sufficiency of the training data to estimate proper weights for each cell. However, when tested for within-talker data, the mean classification rates were 83.0% and 85.2% for English and Japanese, respectively, being notably lower than with the OME variant. The authors hypothesized that this was probably due to less constrained modeling of TOME that leads to worse generalization for slightly deviant tokens, since weights of all grid points that occur in the test set need also to be covered during the training [56].

In a later study, Lake, Vallabha and McLelland [59] applied OME to a variety of categorical learning tasks and compared the evolution of the OME categories to human category learning in the same tasks. More specifically, they first investigated acquired distinctiveness and acquired similarity of tokens from [da]-[ta] continuum endpoints when the training tokens were drawn from either unimodal or bimodal distributions (see the same test performed with infants in [60]. The hypothesis was that the distinctiveness of the endpoints would be higher when the training data originates from bimodal distribution. Their results showed that this was the case for adult listeners, and the performance of the OME replicated these findings with a good accuracy. In addition, their other experiments showed that the distinctiveness of categories learned by OME resembled those of humans in the discrimination of English vowels /e, /E/, and /I/ and in the learning of visual categories for simple one-dimensional bars [59]. In addition, Toscano and McMurray [61] have studied cue integration in categorization of speech

sounds using the gradient descent-based GMM estimation presented in [57]. They show that when the reliability of the cue for category identity is defined to be inversely dependent on the variance of the corresponding Gaussian receptive fields, the model exhibits similar trading relations to human listeners [57]. In a similar vein, Feldman et al. [62] have shown that categorical perception, when modeled as an optimal Bayesian inference using Gaussian distributions, leads directly to the so-called perceptual magnet effect observed in human listeners (see, e.g., [24]). These results suggest that the behavioral effects observed in categorical perception of speech sounds can be largely explained by a perceptual system that adapts to the distributional characteristics of the speech input, although the current models are unable to account for perfect, speaker independent, category learning from continuous speech.

Lately, Kouki, Kikuchi and Mazuka [63] have applied self-organizing maps (SOM; [64]) for unsupervised clustering of audio features (see also [65] which uses a similar technique to explain the perceptual magnet effect for /r/ and /l/ discrimination in the behavioral data of [66]). In their work, Kouki et al. extracted a series of consecutive MFCC spectral features from continuous speech produced by twelve Japanese talkers and used the SOM to learn topological clustering for the features. Then the output nodes of the SOM (which has a fixed pre-defined number of output nodes) were clustered again using a method called data density histograms in order to discover clusters of output nodes that were concurrently activated from the same input vectors. The classification accuracy was measured as the selectivity of each final cluster. For the Japanese vowels, the following classification accuracies were obtained: /i/ = 56.1%, /e/ = 79%, /o/ = 66.1%, /a/ = 71.9%, and /u/ = 41.2%. This indicates that the categorization of speech sounds based on spectral features extracted from randomly extracted portions of speech is not trivial, since the overall performance (63.5%) is relatively low if categorical perception of speech sound is to be expected (see also similar results from [67]). On the other hand, the authors

refer to the study of [68] which showed that human performance in a similar speaker-independent recognition task may be even worse, namely 58% of correct vowel identifications.

3.1 Audiovisual phonetic clustering

Instead of performing category acquisition based purely on acoustic features, support from the visual domain can be utilized. Coen [69] has studied unsupervised acquisition of American English vowel categories in a situation where both formant data and visual lip data were available to the learning algorithm. Coen's algorithm is based on the knowledge that those vowels that are ambiguous in the auditory domain tend to be more easily separable in the visual domain and vise versa. By performing *intersensory disambiguation* the creation of clusters of the data that correspond to the phonetic categories as defined by proficient language users may be possible. The algorithm uses so called Hebbian projections to learn conditional probabilities of the possible states in a modality given a state in another modality, where states are initial unsupervisedly acquired clusters of sensory data. Then the states sharing similar cross-modal conditional distributions are merged to form larger categories (see also [70]).

The data used in the experiments of Coen were recorded using the pronunciation protocol of Peterson & Barney ([71] and spoken by a single female talker. Each vowel was spoken approximately 90–140 times in a CVC structure beginning with [h] and ending with [d] (e.g., "had" for vowel /ae/). In addition to extraction of formant frequencies, visible lip contours of the talker were automatically extracted in synchrony with the audio. When the cross-modal clustering model was trained using the multimodal data, notably enhanced disambiguation of phone classes was seen in comparison to the original formant charts or to the original lip contour data [69]. Overall, the study illustrates how visual information of externally perceivable articulators can significantly aid in disambiguation of

acoustically overlapping phone distributions by increasing cross-categorical distances in multisensory feature space.

3.2 Phonetic learning in an integrative framework

Although important for understanding how the statistical structure of sensory signals can be utilized, the previously mentioned approaches provide only a narrow perspective to the process of phonetic learning. In practical learning situations, the infant is not only faced with a single sensory stream of acoustic speech, but is also embodied in an interactive communicative situation where the infant's own articulatory activity and the caregiver's flexible responsiveness provide further constraints and cues to the learning process. For example, the interaction strategy of the caregiver is known to have a significant effect on the quality of vocalizations produced by the infant [72] and that the caregiver provides positive feedback to communicative attempts of an infant by imitating the infant's utterances in a linguistically corrected form [73]. So far, only two models have attempted simultaneous modeling of speech perception and production without imposing strong a priori linguistic assumptions on the learning process.

In [48], a computational system for phonological learning and speech production is described. The system called HABLAR first learns to classify incoming (but synthesized) automatically segmented speech sounds into a finite number of acoustic categories and then also learns to imitate these sounds with an articulatory synthesizer equipped with reinforcement learning techniques. A method called soft competitive learning (SCL; [74]) is used to infer parameters of a mixture of Gaussian distributions that represents the acoustic feature values of the phonetic categories. In the experiments of Markey, static spectral vectors from stable parts of the spectrum and dynamic spectral representations from changing parts of the spectrum were segmented and trained to separate models. When the static Gaussians were later labeled by hand and evaluated with synthetic CV-syllables

containing one of the three stop consonants, /b/, /d/, or /g/, and one of the ten American English vowel sounds, the selectivity of the Gaussians towards specific vowels was between 32% for /O/ and 100% for /i/, with a mean of 79%. Although a slightly larger number of Gaussian categories was obtained than there are vowel sounds in American English, the experiment demonstrates incremental learning of phone-like categorical representations for synthesized speech data.

As for the speech production, HABLAR utilizes a reinforcement learning scheme based on the learner's imitation of perceived speech. During learning, the model first perceives a spoken utterance and interprets it as a sequence of phone-like units provided by the categorical perception module. Then the model searches for articulatory motor programs that, when activated, lead to an acoustically similar sequence of speech sounds to what was perceived. Successful articulatory gestures are rewarded and reinforced using the Q-learning algorithm, leading to the learning of the correspondence between the acoustic categories and the articulatory gestures needed to produce these sounds.

Note that the learning in HABLAR is purely sequential and the articulatory learning and interaction with the environment does not affect the organization of perceptual categories. However, the perceptual categories have a notable effect on the articulatory development since the learned speech segments determine the basic phonological units whose production is then learned as articulatory programs. Therefore, in HABLAR, speech perception determines the fundamental units of speech production [48]. Also, the articulatory learning in HABLAR is based on the ecologically problematic assumption that the articulatory gestures are learned by minimizing the distance between the agent's own acoustic output and that of a caregiver. It is well known that the vocal tract of an infant is too short to be able to produce configurations of formant frequencies similar to those of an adult speaker. Also, there is evidence that the infants do not actually imitate their caregivers as much as caregivers imitate their children [75,76].

Recently, Howard & Messum [77] presented an integrative model of phonological development. Their model concentrates more on the acquisition of articulatory gestures for speech production than modeling the acquisition of categorical perception of phone-like units. Still, the work of Howard & Messum is an excellent example of how an integrative framework including the modeling of the learner-caregiver interaction, auditory and motor learning, and the modeling of shared communicative context can produce human-like learning results. Their computational learning agent, Elija, learns to produce native speech sounds and words through interaction with a human caregiver. Initially, Elija explores the space of different articulations and receives internal rewards for acoustically salient or motorically diverse productions. After the initial learning, Elija's vocalizations start to draw the caregiver's attention. The caregiver interprets Elija's output in terms of the native phonetic system and provides feedback for successful articulations via the imitative reformulation of Elija's speech output. This then reinforces Elija's native-like articulatory gestures and causes the speech production system to converge towards the set of native speech sound categories. Moreover, mediated by the shared communicative context, Elija is able to associate his own speech to that of the caregiver, allowing Elija to learn the mapping between the acoustics of adult speech and his own articulatory gestures. When the communicative situation is supplemented with referential objects that are being repeatedly named by the caregiver, Elija gradually learns the correspondence between object identifiers ("visual tags") and their respective auditory and motor representations.

In more technical detail, Elija's articulatory apparatus is modeled with an articulatory synthesizer based on the work of Maeda [78] and the voice source model of Fant et al. [79]. The articulatory system is represented by nine parameters of which seven control the vocal tract area functions indirectly by adjusting the positions of the articulatory organs and the remaining two control the characteristics of the voice source. Each articulatory gesture is defined in terms of target positions

of the articulators, and the trajectories between targets are calculated according to the minimum-jerk principle. Somatosensory feedback from the vocal tract is simulated by detecting articulations that lead to full closure of the vocal tract at some point. Elija's hearing is based on a 21-channel filterbank that is used to convert both Elija's and the caregiver's speech into auditory representations and on dynamic time warping (DTW) that allows the comparison and clustering of temporally varying acoustic patterns. Both motor and auditory patterns are represented in Elija's memory as cluster centroids (motor patterns) or best exemplars in clusters (auditory patterns) in order to simplify the learning process. k-means clustering and a DTW-variant of the k-means are used to construct cluster codebooks, meaning that the total number of motor and perceptual categories have to be defined by hand.

During the first stages of learning, Elija explores his articulatory possibilities, leading to unsupervised discovery of a large number of vowel-like sounds. Articulatory gestures of these initial vowels are then combined with randomly initialized consonantal sounds, leading to canonical babbling of CV structures. At this stage, caregiver feedback is used to reinforce those CV articulations that resemble phonologically relevant syllables, while unreasonable ones are pruned away. In the final integrative stage of phonological learning, the vowels and consonants in the consolidated CV patterns are recombined in new ways to create a full spectrum of possible CV structures. Imitative feedback from the caregiver is again used to consolidate those articulations that have linguistic value. Caregiver reformulations are also used to learn the correspondences between Elija's articulatory gestures and the speech of the caregiver. In the original simulations of Howard & Messum [77], this learning process led to the acquisition of 915 motor patterns for CV structures for which the corresponding auditory patterns of caregiver speech were also learned.

From the perspective of unsupervised language learning, possibly the largest shortcoming of the work is that the Elija's learning process is not fully automatic so that it would be purely based on pre-

defined internal criteria and external feedback. Instead, there are several stages in which human intervention is required to define a proper number of signal categories, or, e.g., in setting rewards in the articulatory exploration so that a phonetically proper set of sounds is learned. Also, technical ambiguity in [77] leaves it unclear whether Elija is actually able to represent longer linguistic units, such as words, in terms of their constituent phonetic units, or whether word learning is based on storing full utterances associated with the simultaneous word referent. Finally, as the authors also note, the automatic evaluation of child-like speech is difficult due to the lack of reliable transcription methods. Therefore, the authors decided to exclude a quantitative analysis of the Elija's perceptual and production skills, but provided a set of exemplar productions and imitations by Elija, making direct comparison to behavioral studies or future models difficult. On the other hand, Elija clearly shows the capability to associate spoken words with external referents, to understand the link between its own vocalizations and adult speech, and the ability to reformulate object names using its own vocal apparatus. This raises the question whether evaluation in terms of correct categorical classifications of incoming speech sounds similarly to earlier work is even necessary.

3.3 Conclusions from phonetic learning simulations

The existing computational models on learning of categorical perception of speech sounds show how the distributional properties of speech sounds can be estimated with properly formulated statistical learning mechanisms. The obtained categories exhibit similar properties to human perception in terms of distinctiveness and similarity of the tokens belonging to these categories. From the computational point of view, the work in [56,57,59,61,62] suggests that the human sensory grouping of variable stimuli shares similarities with incremental competitive learning of (Gaussian) distributions that try to maximize the probability of the perceived data. This is interesting from the point of view of neural mechanisms underlying sensory plasticity, since there is strong support for the idea that the receptive

fields of early sensory cortices also perform competitive self-organization in order to adapt to the

prevailing statistics of the incoming sensory data (e.g., [80]).

It should be noted that none of the above approaches for purely bottom-up phonetic learning,

except for the work of de Boer & Kuhl [54] where the correct number of phonetic categories was

specified in advance, make assumptions that would clearly make the algorithms impossible to be

implemented in a biological brain. As shown by the OME [56], the EM-based estimation of mixtures of

distributions is possible in an incremental manner, and the possibility for incrementality also applies for

the approaches in [48,54,69]. Also, the SOMs are also inherently sequential in nature, making them

feasible for continuous and gradual learning. While a biological brain is certainly not explicitly storing

means and variances of Gaussian distributions, it is not meaningful to make plausibility comparisons

between biological systems and computational models at the implementation level but in terms of

computational principles (cf. [8]). None of the models clearly violate the computational premises or

contain extraneous sources of information that are known to be available to the human brain during the

early language acquisition process.

However, an important limitation in many of the clustering studies is that the speech data do not

represent randomly drawn segments from continuous speech, but carefully chosen maximally stable

portions of context-limited vowel-segments. The only exception is the work of Kouki et al. [63], but

they obtained only limited success in the clustering of features into vowel categories. In order to

increase the ecological plausibility of the other approaches, a mechanism for segmentation of these

vowel segments from continuous speech would be needed, or otherwise the methods should be

evaluated directly on continuous speech. Notably, Markey has already proposed a segmentation method

for speech but the method was evaluated only on simplified synthetic speech [48]. Howard & Messum

also describe the principle of representing incoming speech as a sequence of categorical units, but their

existing work assumes that the perceptual category learning is mainly based on isolated caregiver reformulations of canonical babbling [77]. Although not fully implausible, this behavioral hypothesis is to be confirmed with experimental studies.

Another issue is the generalization across talkers. For example, in the work of Vallabha et al. [56], the categorization of vowel segments is much more accurate for the experiments in which the same talker is used in the training and in the evaluation of the categorization. When additional talkers are used for evaluation, the performance drops significantly. Although this difficulty is expected due to well-known acoustic variability across different talkers, how human infants may solve this challenge remains unknown. From the computational point of view, distributional learning leads to much more ambiguous category boundaries if the learner receives data from several talkers instead of a single caregiver. On the other hand, modeling each talker separately or using data from only one talker, such as the primary caregiver, leads to much sharper category boundaries, but then these categories are not compatible with those produced by other talkers of the language. The generalization problem is also inherent to the lexicon-first models of LA (see next section), leading to incompatibilities between lexical items learned from the speech of different talkers [15].

It should be also noted that, except for the work in [48,63,69], the inventory of phonetic categories to be learned in the purely acoustic experiments is much smaller than that of the normal number of vowel categories in the world's languages. Currently, how the other proposed methods would scale up to a full vowel repertoire of a language and how they can deal with the temporal ambiguity of vowel boundaries and the coarticulatory effects between subsequent vowels still remains unknown.

In general, the current results suggest that the categorization of speech sounds into phonetic categories is far from perfect for unconstrained speech, although distributional properties of speech

sounds clearly follow language- and talker-specific patterns. This means that assigning a unique and correct phonetic label to each acoustic percept is not possible (cf., [56, 63]). This is especially pronounced in generalization across multiple talkers, where distributional clusters of one talker are not compatible to those of another, or where distributions learned from multiple talkers become very ambiguous at the category boundaries. On the other hand, if categorization performance notably below 100% is allowed at the phone perception level (note that human performance is not perfect either for isolated vowel recognition), the question is how the learning of lexical items can be realized with such partial and distorted input. Is there a way to learn a robust lexicon directly upon the layer of unsupervisedly acquired discrete units, and should the representations to be refined later on in order to better accommodate the requirements of the lexical layer or due to additional constraints emerging from the visual [69] and/or articulatory domain?

If caution is used to derive a conclusion, one could say that the behavioral evidence and computational experiments both show how the perceptual system can make use of the distributional properties of dominant auditory patterns in order to allocate processing resources to those parts of the acoustic space that are densely populated (cf., NLM-e, [24]). Still, it is not clear that the adaptive organization would be sufficiently detailed to allow representation of continuous speech as a sequence of abstract independent elements, each element corresponding to a well-specified (phonetic) category upon which lexical items are organized.

As for the de Boer & Kuhl [54] study, it would be interesting to see whether the notable differences in learnability of ID and AD speech could be replicated with a purely unsupervised algorithms. Also, it would be interesting to see how IDS-based vowel training generalizes to ADS, since there is some evidence from ASR research that the mismatch between IDS and ADS may actually hinder the recognition of adult spoken words [81]. This leaves open the question of how small children

compensate for the mismatch between acoustic properties of IDS and ADS in order to make sense of more adult-like conversations.

It should also be noted that the integrative frameworks including both perceptual and articulatory learning [48] and realistic learner-caregiver interaction [77] may provide new insights on the categorical perception and phonological development. Ultimately, the perception of phonetic contrasts serves the organization of lexical contrasts, which then in turn serve the emergence of *semantic contrasts* between different linguistic messages. Also, some phonological contrasts that are difficult to differentiate in acoustic space can be very distinct in the articulatory domain (such as plosives; cf., Motor theory of speech perception, [82]). Finally, given that the current multimodal context modulates the perceptual processing in the auditory cortex [83], it is highly unlikely that the organization of the categorical perception of speech sounds would be purely independent of speech production or the semantics of the language. However, the existing models have not focused on how perceptual development can be constrained by the simultaneous articulatory learning. Instead, perceptual learning has been considered independent of the speech production system and the feedback from interaction with the environment. This is a topic that needs further investigation.

4. Computational models of lexical learning

4.1 On lexical representations of speech

When literate adults discuss the concept of *words,* they talk about well-defined entities that have both written and spoken form with a finite number of discrete elements (phonemes) in a specific order. The majority of the words either have significant associations that instantly stimulate multimodal perceptions related to the concept denoted by the word, or they play a significant role in the construction of grammatically correct sentences by disambiguating causal and temporal relationships of

the actors and events involved in the verbal description at hand. Either way, the words show themselves as meaningful symbols and the symbol is perceived as "incorrect" when it is misspelled or mispronounced, requiring additional cognitive resources to recover from the aberration.

From the viewpoint of a young infant, the situation is very different. This is especially true if the nativist views are completely abandoned and the infant is considered as a tabula rasa cognitive agent with efficient innate learning capabilities and bias for social behavior. An infant does not know what a lexical item or symbol is. Moreover, it does not even know what speech is about. Much of the learning effort during the first year of the infant's life is about discovering that the world can be perceived through senses and that it can be also manipulated by motor activity. Through the development of the action-perception loop and maturation of the brain, the infant acquires understanding that the world is a 3-D realm with distinct objects with varying properties, and with actors (living objects) that can have an impact on the state of the other objects in the realm or on the (needs of the) infant itself. Although the development of auditory perception is affected by the exposure to speech associated with social and emotional interaction with the caregiver (cf. NLM-e, [24]), our claim is that *the first real contact with the language faculty occurs when the infant first realizes that sensory patterns originating from other people's mouths have correspondence to the state of the surrounding world*. This is probably already preceded by the realization that the objects and events in the environment are sometimes associated with distinct non-speech auditory patterns (note that the sensory patterns need not to be auditory, but signed language will also do the trick, e.g. [84]; cf. also "*goes with*" vs. "*stands for*" distinction of words in [85]). The core of the language is in the ability to activate representations in other people's minds about things that are not necessarily available in the present sensory domain or at least not in the current focus of attention. The learning of these links is necessarily bootstrapped by associating the sensory patterns to internal active representations of the

concepts describing the world. For very young infants, these associative links are necessarily tied to the surface form (i.e., acoustic or visual realization) of the patterns, since it is the most directly observable and statistically significant structure that has correspondence to the external world. This type of learning can be accomplished in the absence of any kind of linguistically motivated knowledge in the learner [15].

What this all means from the viewpoint of early lexical learning is that the learning does not have an ecological pressure to "find" words from speech nor code these words in any specific format (such as precisely defined sequences of discrete elements like phones or phonemes). Instead, the functional advantage of spoken language emerges from sufficiently detailed but sufficiently general representations of the spectrotemporal acoustic patterns that systematically indicate the co-occurrence of objects and events in the environment, i.e., the useful units of a language are defined by their *semantic content*. These patterns can match individual words, but they can also be part-words, compounds, frequently co-occurring words ("*doyou*"; "*Isee*") or even entire phrases if they are systematically used in specific situations. As long as there is equally good predictive power (or functional consequences) in "wrong" lexical representations of the language that do not match the adult vocabulary, there is really no need for the learner to refine these representations. As the complexity of the interactions with the environment increases and as the number of proto-lexical representations accumulates, the early representations of spoken language become refined in order to answer to the increasing communicative challenges and to reduce the internal contradictions in the previously acquired lexico-conceptual system (cf., principle of conventionality, [24]). For example, the increasing semantic awareness imposes new distinctions to the linguistic representations and gives rise to concept of lexical synonymy. The increasingly structured parsing of speech and increased size of the lexicon

also possibly gives rise to the sub-word/morphological representation of spoken language as it provides a more efficient means of coding (cf. PRIMIR, [25]).

4.2 On the grounding of auditory patterns

In computational models of lexical learning, the word semantics are typically assumed to emerge directly from the established link between an acoustic word form and an internal representation of a word referent, such as a visual or haptic representation of an object or action. Based on this idea, the computational models of lexical learning from continuous speech can be divided into two main categories, based on the principle how words (learned acoustic patterns) are grounded to their referents. These categories will be referred to as models with *indirect* and *direct grounding* of words.

Indirect grounding refers to a learning process in which the learning agent first learns speech patterns such as words from continuous speech independently of other modalities. The criterion for the initial word segmentation can be arbitrary, but assuming the absence of a priori phonetic or linguistic categories, it has to be some type of statistical measure that reacts to the specific organization or recurrence in the acoustic features computed from the auditory signal. Once a pattern is learned and can be recognized from future input, its occurrence can be then studied in the context of other modalities and internal states[1] of the agent in order to find statistical correlations between the pattern and the contextual variables. Once such a correlation is found, it is said that the pattern (or word) is grounded to the contextual variable, providing meaning to the pattern.

On the contrary, direct grounding means that the learning agent perceives speech simultaneously with internal active representations of contextual variables. Due to the immediate co-occurrence, the representations of the objects and events in the active internal state become associated

---

[1] Note that while some words such as nouns typically refer to perceivable physical objects, some others such as "sad" or "hungry" are ultimately grounded to the internal needs or emotions of the learner.

with the heard auditory patterns, allowing instantaneous (but originally vague) meaning to emerge for the spoken utterances. In the simplest case, the internal state may simply reflect the visual objects in the immediate surroundings that the learner is attending to, leading to direct cross-modal associations between acoustic patterns and visual objects. For a more complex cognitive system, the internal state may reflect a combination of the task-modulated short-term memory contents (which may consist partly of immediate sensory consequences and partly of items recalled from the long term memory) and some internal variables of the system such as the emotional state of the agent. The major difference from indirect grounding is that now the contextual variables such as visual objects can directly affect the patterning of the auditory stream. This provides the learning system with additional statistical constraints that can help in the learning process.

For both indirect and direct grounding of word forms, the main problem is that one exposure to the auditory pattern is not sufficient to obtain meaning of the word since there are typically multiple potential word referents available [86]. This is where so-called cross-situational learning mechanism comes into play [87,88] and due to the multiple exposures to situations with several possible referents simultaneously with speech containing the word of interest, the ambiguity in word-referent mappings is gradually resolved (e.g., [89,90]).

The models that discover patterns from continuous speech in the absence of referential information will be discussed first in the next section, and then attention will be turned to the models of direct lexical grounding (section 4.4).

4.3 Models with indirect lexical grounding

So far, no computational model of indirect lexical grounding exists that would combine unsupervised acquisition of phonetic categories with lexical learning. Instead, a number of models have been

proposed that attempt to learn lexical representations directly from continuous speech without relying on an intermediate phonetic layer.

The computational models of LA from continuous real speech based on indirect lexical grounding include the PERUSE algorithm by Oates [91,92], the P&G algorithm of Park & Glass [93,94] and its incremental variant by McInnes & Goldwater [95], and the transitional probability-based algorithm of Räsänen [15].

4.3.1 PERUSE

PERUSE is an algorithm for learning recurring patterns from a multivariate time-series [91,92]. It is based on the assumption that structurally significant patterns occur as sequences of multivariate observations of features, where each temporal spot in the sequence has a unique mean and variance describing the local acoustic properties. Likelihood of a sequence of data for a given pattern model can be directly computed by temporally aligning the model with a pattern using dynamic programming and then summing the log-likelihoods of individual observations across the entire sequence.

In order to discover the first one of such patterns, PERUSE performs a global and exhaustive search over all available speech data in order to find a signal segment of length L_{min} that has the highest likelihood of having at least N_{min} other occurrences in the data (L_{min} and N_{min} are user specified parameters). Once the most likely pattern is discovered, the length of the segment is increased incrementally by taking into account all $N_{min}+1$ realizations of the pattern and a statistical test is performed for the quality of the new length patterns in order to determine the overall length of the recurring pattern. Finally, the total number of occurrences of the pattern is estimated from the data.

Oates has demonstrated the performance of the PERUSE algorithm in a word learning task from English, German and Mandarin speech, where it successfully detected more than 65% of frequent

words used by a single talker. Oates has also represented a framework that allows grounding of the detected word forms to contextual sensory data collected by a robot [92]).

The main drawback of the PERUSE algorithm is that it requires all speech data to be in the memory of the system already at the beginning of the learning. The underlying assumption is that the longest words that have most occurrences in the data have most significance and are therefore learned first. Additionally, the algorithm is computationally complex, as it has to search and evaluate the data set iteratively numerous times in order to converge to the final set of words. In addition, each word has to occur several times in the data before a representation can emerge for it. This makes the approach implausible to for a biological system that needs to deal with the continuous flow of sensory information here and now without access to globally determined statistical significances between different choices of signal patterning. Oates has also acknowledged this limitation, noting that iterative batch processing is an unreasonable requirement for a computational agent that should support continuous long-term LA [92]. Still, PERUSE is statistically well-justified approach for pattern mining from multivariate time-series such as speech and demonstrates nicely how an unsupervised system can converge to a set of word models learned from real speech. Unfortunately, a detailed quantitative analysis of word model properties is missing from the studies, making comparison to other approaches difficult.

4.3.2 P&G algorithm

The word discovery algorithm by Park and Glass [93,94], hence P&G algorithm, is based on a modified dynamic time warping (DTW) of feature representations of auditory patterns. In the standard DTW, speech signals are represented as sequences of spectral vector time series and the aim of the DTW algorithm is to discover the cheapest path across a distance matrix whose elements describe the distances between the spectral frames of the two signals. As an outcome, the obtained shortest path

describes the temporal correspondence between spectrotemporal patterns in both sequences and DTW is therefore especially suitable for temporal alignment of signals that are known to contain identical utterances but spoken at a different tempo. However, the standard DTW as such is not suitable for word discovery from continuous speech since spoken utterances may contain any number of arbitrary words with varying word order, making global pair-wise alignment of two or more blindly chosen utterances an ill-defined problem. Instead of using the standard DTW, Park and Glass [93,94] devised an algorithm that divides the distance matrix of two utterances into several overlapping diagonal segments and then finds the best alignment separately for each segment. Next, the best matching subsequence of each aligned section is extracted, where the best subsequence is defined as the one with the smallest average distance between the two utterances but still exceeds a minimum length. Finally, the overall best alignment between the two utterances is chosen and stored for further processing along its mean distance measure ("*distortion*"). The process is repeated for all utterance pairs perceived by the system, leading to a collection of pairs of aligned spectrotemporal signals. Then the signals are clustered using an agglomerative graph clustering method [96] in order to find categories of similar patterns, or words.

Park and Glass evaluated the performance of the algorithm with the MIT lecture corpus containing recordings from a variety of academic lectures, showing that the algorithm successfully discovers a large portion of the ten most frequent words recurring in the data. They also reported two major types of error made by the algorithm that have relevance to the LA theme: in the first error type, acoustically similar but lexically distinct patterns are accidentally grouped together during the clustering process (e.g., "*this miss*", "*misses*" and "*which is*"). The second type of error category consists of clusters that contain only partial representation of the frequently occurring phrases or words. They give an example of "*square root*", from which only "*square*" is learned, and "*times ten to*", from which the section "*tent*" is acquired. Both error types are understandable considering the architecture of

the learning algorithm. The first error type is a direct consequence of indirectly grounded word learning, since there is no way for the purely bottom-up algorithm to distinguish the situation of two occurrences of the same word spoken in two different speaking styles from a situation with two different but acoustically similar words. The latter error type is at least partially explained by the manner that the matching subsequences of aligned sections are extracted. Since the best matching path between two utterances is the one leading to the smallest mean distance with minimum length L, there is an inherent bias to extract shorter segments close to L due to increased probability of variation in pattern realizations as the length of the patterns increases.

It should be noted that the original P&G algorithm was not intended to model infant language acquisition, but was designed as an engineering tool for unsupervised pattern discovery from speech signals. Lately, McInnes and Goldwater [95] have modified the P&G algorithm in order to achieve a higher ecological plausibility for LA simulations. Instead of performing word discovery as a batch process as in the original P&G algorithm, their system works incrementally by comparing the current input only to the previously discovered word fragments and to a finite number of previously perceived utterances. The word fragment extraction of the algorithm was also modified in order to allow the discovery of multiple separate word fragments from the same diagonal section of the distance matrix between utterances. The performance of the new algorithm was evaluated using speech from the Brent corpus [97] of the CHILDES database [98] that contains real recordings of interaction of mother-infant dyads. The results showed that the DTW-based detection of recurring words is feasible also in an incremental manner, especially when facilitated by infant-directed speech that contains multiple repetitions of salient words occurring close in time.

4.3.3 Transitional probability based learning

Räsänen [15] has presented a computational model for unsupervised word discovery from speech that is based on transitional probabilities (TPs) between atomic acoustic events. The model is inspired by the finding that eight-month-old infants can already segment recurring words out of speech by analyzing TPs of subsequent syllables [99,100] and may treat the detected segments as lexical items when presented in a proper linguistic context [101]. However, the model of Räsänen does not assume that the learner can recognize phonetic or linguistic units such as phones or syllables from continuous speech, but simply represents the acoustic speech signal as a sequence of discrete elements obtained by unsupervised vector quantization of spectral vectors. Recurring speech patterns are modeled by analyzing the TPs between the discrete acoustic events, i.e. each unique pattern model is characterized by a specific set of TPs between the discrete elements in the signal. However, the modeling is not performed only for transitions between the two subsequent elements, but in parallel for a number of different temporal distances in order to capture long-range statistical dependencies and to enhance robustness of the model against noise variability in the signals. Learning in the algorithm produces a non-predefined number of word models, the number being mainly defined by a novelty threshold parameter. These models can be then used to recognize similar words from novel speech input, inherently segmenting the new input into word-like units [15].

The algorithm was evaluated using the CAREGIVER Y2 UK corpus [102] that contains 50 keywords (1-4 per utterance) plus the surrounding carrier sentences, yielding a total vocabulary size of 80 unique words. The results showed that the algorithm successfully learned a number of ungrounded word models that were selective towards specific words in the material when compared against the word level annotation (many of the models responded only to one word above 80% of the time). The word segmentation accuracy was also notably above chance level. It was also observed that the

learning performance was much higher when the training and recognition was performed with data from a same talker. When data from multiple talkers were used, the generalization across talkers was relatively poor in terms of word recognition accuracy, although the segmentation accuracy generalized better for models learned from one talker to speech from another talker. This replicated the common finding from ASR research that acoustic models learned for one speaker are not easily generalized to other speakers, especially to speakers of different gender. The results also showed that the typical errors in word segmentation and model selectivity were either related to a situation where a model had considered that two frequently co-occurring short words are one word (such as "*doyousee*", "*doyoulikethe*"), or when the words were acoustically similar ("*small*" and "*ball*", or "*cow*" and "*cat*"). Also some oversegmentation similarly to the P&G algorithm [94] was observed, e.g. for the word "*telephone*" that was sometimes represented and recognized as "*phone*".

Although not intended as a neurophysiologically valid model, the TP-based approach has some biological validity due to the property that the incoming speech signal has to be stored in detail only for a few hundreds of milliseconds during which TPs ("weights") of the internal representations are updated. It is also purely incremental and utilizes short-term dependency statistics that the human test subjects are known to be sensitive to (see also [103] for further analysis). Also, the statistics of the model can be easily represented in a sparse, distributed form, analogous to brain-like synaptic weights (cf., [104]). The most significant shortcoming in the model of Räsänen [15] is in the threshold parameter that defines whether the current input is sufficiently familiar to be updated to an existing word model or whether it should be used to create a new model. If set too low, only a small number of unselective lexical models are learned despite the complexity of the data. On the other hand, too high a threshold leads to a situation where a new model is learned from nearly every window of data. The experiments were performed with three different thresholds, showing that the performance of the

algorithm depends on the proper setting of this parameter, but no automated method for proper estimation of this parameter value was presented.

4.3.5 Summary of models with indirect lexical grounding

The studies and models reviewed above demonstrate that learning of words (or more like proto-words) from continuous speech is possible in the absence of contextual support or external feedback and without any a priori linguistic or phonetic knowledge. However, these word models are not always perfectly aligned with the words defined by a proficient language user, but more likely represent statistically significant continuous spectrotemporal structures that systematically recur in the speech data.

It is noteworthy that the successful discovery of word-like units is achieved by three totally different methodological approaches. The P&G approach [93,94] and its modification [95] basically perform exemplar based learning by extracting recurring fragments of speech and then comparing these fragments to novel utterances. The approach in [15] does not store word exemplars per se, but represents each word as a construct that defines probabilities at which specific acoustic events follow each other in the word. Finally, the PERUSE algorithm [91,92] treats each word as a probabilistic construct, but solves the problem in purely continuous time and feature domains. Instead of defining transitional probabilities between feature configurations, a word is defined by an ordered set of observations with a specific mean and variance allowed for each observation. Despite their differences all the algorithms show a similar pattern of results, suggesting that the recurring word structure in speech can be captured using a variety of pattern representations.

4.4 Models with direct lexical grounding

4.4.1 Statistical word discovery algorithm.

The statistical word discovery (SWD) algorithm described by ten Bosch and Cranen [105] utilizes segmental representation of speech by blindly segmenting the input signals into phone-like units based on spectral changes in the signal. The obtained segments are then aligned with DTW and clustered with the k-means algorithm [106] so that each segment category is represented by an integer value from 1 to 25 (the cluster number). In other words, the utterances are converted into discrete sequences, one element spanning approximately one phone-sized unit. During the word learning process, the segmental representation of each utterance is represented in association with a bag of abstract tags that describes which words are present in the utterance but do not reveal the temporal locations or ordering of the words. Each utterance is then compared to all previously perceived utterances and the best matching subsequence of each pair is extracted. If the current utterance shares the same abstract tag with the one in the memory that it is being compared to, the best matching subsequence is appended to a B_{match} list. Otherwise it is added to $B_{no\text{-}match}$. When the learning process is repeated across several utterances, the match and no-match lists grow in number. The lists are sorted so that the most frequently occurring sequences are placed at the top of the lists, revealing the most typical sequential representations of each word. The sequences in the $B_{no\text{-}match}$ are considered as negative examples of a word and therefore the equivalent sequences in the B_{match} list are eliminated in order to facilitate the contrast in the cross-situational learning situation [105].

Ten Bosch and Cranen evaluated the performance of the algorithm using the Aurora 2.0 database [107] that contains continuously spoken English digit sequences with 1–7 digits per utterance and speech from a large number of talkers. The results showed that their algorithm achieved approximately a 90% word recognition rate after perceiving 1000 tokens per each digit when the

hypothesized word tags were compared to the ground truth. The authors also noted that the number of false alarms (words being hypothesized to points in time where there are no corresponding words) was relatively high (above 10%). They hypothesized that it may indicate that the learned word representations were somewhat shorter than the true lengths of the words, since the correct recognitions did not cover the entire timeline of the utterances.

In general, the SWD algorithm is interesting because it is one of the rare attempts to segment speech into phone-like units before further processing (cf. Kuhl's basic cuts, [29,30]). While the idea of making an exhaustive comparison of the current speech token against all previously heard utterances with a shared context seems drastic, it is not totally unreasonable from the perspective of exemplar based theories of human memory. Still, the work of ten Bosch & Cranen focuses mainly on the question whether statistical regularities in unsupervisedly learned phone-like segments can form a basis for a lexicon. Analogues to human-like processing are not given by the authors.

4.4.2 NMF-based word discovery.

In the work of ten Bosch et al. [108] and Van hamme [109], a non-negative matrix factorization (NMF) based framework for word discovery from continuous speech was represented. The NMF is a method for matrix factorization that allows a large matrix V of non-negative components to be decomposed into constituent non-negative matrices $\mathbf{W}$ (basis) and $\mathbf{H}$ (weights) so that $\mathbf{V} \approx \mathbf{WH}$. At a conceptual level, the factorization can be considered as a process in which the data observations presented as columns in $\mathbf{V}$ can be described as a linear combination of typical patterns $\mathbf{W}$ and the strengths of their occurrences $\mathbf{H}$ in each column of $\mathbf{V}$. Because the results of the factorization can be given a probabilistic interpretation (see [110]), the NMF is especially suitable for speech analysis and pattern recognition. The factorization is obtained by iteratively minimizing a divergence metric between $\mathbf{V}$ and $\mathbf{WH}$, as described in [111].

In order to discover words using the NMF, each utterance was represented as a histogram of acoustic co-occurrences (HAC). HAC is a long vector constructed from the frequencies of co-occurrences of discrete acoustic events at different temporal distances. When the HAC vector $\mathbf{v}_{HAC}$ is concatenated with a grounding vector $\mathbf{v}_g$ (e.g., visual context vector) corresponding to the utterance (i.e., $\mathbf{v}_f = [\mathbf{v}_{HAC}\ \mathbf{v}_g]^T$) and then these vectors are accumulated across *n* utterances, matrix $\mathbf{V}$ of size $|\mathbf{v}_f|$ x *n* is obtained. When factorized, $\mathbf{V} \rightarrow \mathbf{WH}$, the matrix $\mathbf{W}$ represents the typical audiovisual patterns that recur in the data, whereas $\mathbf{H}$ describes the activation level of these patterns in each utterance. During word recognition, the activation matrix $\mathbf{H}$ is computed from the HAC formed from the utterance under consideration (without the visual part). In order to obtain the activation values for the visual tags, the obtained $\mathbf{H}$ matrix is then multiplied by the submatrix $\mathbf{W}_g$ of the $\mathbf{W}$ that contains only the vectors related to the grounding information. Since the normal factorization process does not allow recovery of the pattern (e.g., word) order from $\mathbf{H}$, but simply reveals what patterns are present in an utterance, Van

hamme also introduced a time-scaled histogram variant of the system for temporal decoding of utterances [109].

Van hamme [109] demonstrated the feasibility of the NMF approach in a connected digit speech recognition task, showing that the matrix factorization based framework leads to successful learning of grounded word patterns. In [108] and [112], the same algorithm was evaluated more comprehensively from the perspective of LA in a keyword discovery and recognition task. The authors use infant directed speech data from the CAREGIVER corpus [102]. The data contains utterances that each have one of the 10 possible keywords surrounded by a carrier sentence (e.g., "*Do you see the ball*?", "Where is **mommy** now?" keywords emphasized). The visual tags corresponding the utterances denote the identity of the keyword in the utterance. The results showed that the NMF-based system can learn the ten unique keywords with high accuracy when grounded directly with the related visual tags during learning. For further ecological plausibility, the NMF-based LA model was modified in [113] so that the processing is no longer performed in a batch mode, but allows incremental learning. This makes the approach intriguing for the LA studies since the NMF, and especially the HAC representation of the sensory input, share many properties with human-like information processing, including the representation of information in a distributed form and the ability to include multiple sources of information at different granularities into the same computational framework.

4.4.3 Weakly-supervised transitional probability analysis.

Blind segmentation of speech into phone-like units was also used in the work of Räsänen, Laine and Altosaar [114]. Similarly to ten Bosch and Cranen [105], the segments were vector quantized in order to obtain discrete sequences of phone-like units representing the speech signals. Also, the utterances were paired with abstract tags denoting the "visual objects" perceived simultaneously with the utterance. However, the learning procedure was now based on TPs between subsequent phone-like

units (cf. [99]) in the context of each tag, i.e. the probability of a transition from discrete phone-like segment S_t to segment S_{t+1} was measured separately for the presence of each contextual tag. During recognition, the probability of each tag was computed by following the transitions through the sequential representation of the utterance and retrieving the corresponding tag specific TPs from the memory. When evaluated with the CAREGIVER Y1 FIN corpus [87] with a total of ten unique keywords (the visual tags), one keyword embedded in each utterance in addition to the surrounding carrier sentences, the algorithm obtained a keyword recognition rate of 74.5% for speech from one talker.

The experiment of Räsänen et al. showed that there is some feasibility in the transitional probability approach in word recognition, but the overall word recognition rate was relatively low considering the simplicity of the task. In [37] it was found that the low performance was mainly due to segmental representation of speech that did not capture spectrotemporal details in sufficient accuracy in order to obtain efficient models for words. Insertions and deletions of segments was also a concern. Finally, the analysis of only subsequent segments (bigrams) of [114] did not yield sufficiently strong statistical models for speech. This led to the discarding of the segmentation based approach and a further developed mathematical framework for TP analysis is presented in [115]. The approach makes use of the normal fixed-frame windowing with 10 ms frame shifts and vector quantization of speech signals. In addition, similarly to [15], the TP analysis is performed at a variety of different lags, increasing notably the robustness of the learned models. When evaluated with the CAREGIVER Y2 UK corpus with 50 unique keywords, 1–4 keywords occurring in each utterance, a word recognition rate of above 92% was obtained for data from four different talkers (two male, two female).

4.4.4 DP-ngrams-based weakly-supervised learning.

Lately, Aimetti [116] has proposed a dynamic programming-based system for word learning from continuous speech. However, unlike the P&G algorithm [93,94], the system also utilizes direct grounding of the lexical items to the co-occurring visual objects (visual objects are simulated with abstract and discrete semantic tags). The learning proceeds by first comparing the visual tags of the current utterance to the tags of previously perceived utterances in the short-term memory (STM) of the system. For a tag never perceived before, the entire utterance is stored as an acoustic entry for the new tag and the system proceeds to the next utterance. In case of matching tags, the utterances are aligned with a method called DP-ngrams, which is a modification of standard DTW that allows efficient extraction of best matching temporally contiguous aligned sequences. The part of the novel utterance containing the best matching alignment with existing memory entries is then extracted and appended to the list of exemplars representing the corresponding tag. When the process is repeated over the entire training data, each tag becomes associated with a list of exemplar occurrences of the corresponding word. These exemplars can then be matched with novel utterances in order to determine which tag is most likely given the audio signal. Also, a clustering process can be applied to the list of exemplars in order to obtain a prototypical representation of each word [116].

Evaluation of the algorithm was carried out with the Y1 UK version of the CAREGIVER corpus [102]. When evaluated in terms of word recognition accuracy (correspondence between true and hypothesized visual tag of a novel utterance), convergence to a word recognition rate of approximately 90% was observed with exemplar-based recognition after perceiving approximately 140 utterances. For prototype-based recognition, the accuracy was notably worse (around 70%), suggesting that a word even from a single talker is not very well represented by an "acoustic mean" of its realizations. For experiments with four talkers instead of one, a recognition rate of slightly below 50% was obtained

after observing 200 utterances, again giving a clear indication of the notable acoustic mismatch between different talkers even on material with very limited vocabulary.

4.4.5 Other experiments with direct lexical grounding.

In addition to the methods described above, a number of additional experiments of LA have been reported using the algorithms. In [117], the TP-based learning algorithm presented in [115] was studied in different caregiver-learning agent interactions. Different interaction strategies were simulated by varying the realiability of the visual labels associated with the spoken utterances, revealing the somewhat expected result that the more ambiguous contextual situations lead to slower learning.

In [118], the effect of the number of different caregivers was studied using the TP-based algorithm of Räsänen & Laine [115], DP-ngram algorithm of Aimetti [116], and the NMF-based system presented in [108,109]. In the learning phase, the learner perceived speech from either one or four different caregivers. The word recognition accuracy was then probed for the four main caregivers and for six additional talkers available in the CAREGIVER Y2 corpus. The results from all three learning algorithms revealed that the generalization from one caregiver to the novel talkers was much worse than the word recognition performance evaluated for novel speech from the caregiver used in the training. When four different caregivers were used (two male, two female), the generalization to six additional talkers was slightly better but still far from the matched talker condition. The NMF was also noted to have obtained the best generalization to unseen talkers from the three algorithms [118].

Lately, Versteegh, ten Bosch and Boves [119] have studied how the word learning performance is affected when the learning agent can actively decide whether the visual information (tags) is sufficiently reliable to be used in grounding of the co-occurring speech input. They devised a confidence measure that indicated the reliability of the utterance-tag pair. Based on that confidence measure and a user set threshold, the agent was able to decide whether the content of an utterance was

related to the concurrent visual tag. If the confidence was too low, no learning occurred. Otherwise the model contents were updated according to the standard NMF learning procedure [108,109]. The results of the experiments revealed that the learner was partially able to overcome the uncertainty in the visual domain when actively questioning the reliability of the correspondence between visual and audio domains. If active learning was disabled, the word learning performance was notably hindered [119].

4.4 Conclusions from computational models of word learning

The computational studies reviewed in the previous subsection have successfully demonstrated that the word learning from continuous speech is indeed feasible without explicit top-down information using a variety of techniques. The models with indirect lexical grounding show that, in principle, proto-lexical representations of recurring word forms can be discovered based on the acoustic similarity of word tokens and in the absence of any linguistically motivated expert knowledge in the task. This type of discovery can be based on global matching of similar subsequences (exemplar based view; [91,93,94] or on incremental analysis of TPs between automatically discovered speech sounds ([15]; see also [120]).

When contextual support in the form of abstract tags that correlate with the contents of the concurrent utterances is available, the discovery of auditory patterns that are relevant for each tag can be done efficiently using a variety of approaches ([105,108,109,114-116]) This simulates a learning situation where the learning agent simultaneously hears the speech of the caregiver and shares attention with the caregiver towards some specific objects in the environment.

In theory, direct grounding enables more efficient pattern models due to additional statistical constraints. If the learning is based on the assumption that the visual objects are always present when they are being discussed, the learning algorithm can assume that the auditory patterns occurring in the absence of the visual object are not related to it. This makes it possible to contrast the statistical models

so that those aspects of the models that are relevant only for the given visual tag are given higher priority (cf. [105,115]). However, this also means that auditory patterns that are not systematically represented in the visual domain as possible referents do not obtain their own representations.

Indirect grounding, on the other hand, enables acquisition of word forms independently of the surrounding context, making accumulation of vocabulary faster since any words can be learned without requiring evident referents in the external world. However, the obtained word models are initially weaker since the relevance of the patterns must be inferred solely from the statistical properties of the auditory stream. Because the models are learned in isolation from other domains such as the visual world, there is no guarantee that the learned patterns have optimal correspondence to objects and events in the world (e.g., the agent may learn "*redball*" instead of "*red*" and "*ball*" if the combination occurs sufficiently often, although they are clearly dissociative entities in the world of visuomotoric experience). It is also possible that no distinct model emerges for a word at all, even though the word constantly has a visual referent in a normal learning situation. For example, the word models obtained from the unsupervised TP-based algorithm of [15] can be afterwards grounded to co-occurring visual tags in the 50 keyword recognition task of the CAREGIVER Y2 corpus by simply analyzing the co-occurrence frequencies of the words and tags. If these associative links are then used to recognize the most likely visual objects associated with novel utterances, a word recognition rate of approximately 67% can be obtained [121]. This is notably worse than the result obtained with an algorithm applying the direct grounding approach for which above 92% word recognition rate has been achieved using the same material [115].

In general, indirect and direct grounding, when used in isolation, do not seem to directly correspond to the learning challenge faced by infants in early word learning. Requiring the learner always to have a concrete and attended referent for the spoken language it hears in order to learn

something seems an unreasonable limitation. On the other hand, as language serves the purpose of lighting up associations in the minds of the receivers, learning a language in total isolation from the environment does not make sense either. The need for some kind of contextual support is already indicated by the fact that detection and the precise modeling of the word patterns is not an easy task due to immense acoustic and temporal variability in speech. However, one should note that the models of lexical acquisition presented in this work aim to explain the very first steps of the LA, i.e. they show possible ways to bootstrap the learning system. After being able to segment a novel utterance into word-like units and unfamiliar segments, the demands of the acquisition process change and additional hybrid mechanisms of unsupervised pattern discovery and cross-situational grounding may become feasible. Also, since the human memory is highly based on associative links between perceived patterns and events, and since the processing at early sensory cortices is modulated by the multimodal context [83], it may well be that there never is truly unsupervised unimodal learning, but all sensory processing takes place in the context of other modalities and internal states of the learner, even in the absence of "correct" referents. These contextual cues can be then used to store and retrieve patterns from the memory and to categorize them according to the similarity of the contexts in which they occur (cf., [121]). Over time, the relationships between the patterns and their referential contexts simply become more distinct, enhancing the predictive value of spoken messages.

As for the ecological plausibility of the proposed algorithms, PERUSE is the only one that directly runs into trouble with its batch training. The ecological plausibility of the DTW-based approaches (P&G algorithm [93,94] and the work of McInnes and Goldwater [95]) is greatly enhanced by the work of Unal & Tepedelenlioglu [9] who showed that the DTW computations can be accomplished with artificial neural networks. In a similar manner, the TP-based modeling of the signal structure in terms of short-term acoustic events is possible with recurrent neural networks with

sufficient temporal memory (cf. [22]), and the average temporal dependency structure explicitly modeled by the TPs has even been shown to have a close correspondence to the integration times measured in the human auditory system [122].

It should be also noted that despite the absence of an explicit phonetic layer, the distributional properties of speech sounds are implicitly taken into account in the approaches utilizing vector quantization (VQ) of speech frames (e.g., [15,105,108,109,114,115]). However, these clusters and the corresponding sequence elements are by no means comparable to phones, not least because they are not defined in duration, but have fixed and short (typically 10 ms) length, they are not fully selective to speech sounds from only one phonetic category independently of the talker, and because the typical number of elements is much higher than the number of phones in any language. The basic reason for the conversion is not the belief that human infants would perceive speech as sequences of symbolic elements, but because the computational pattern discovery problem is simplified notably. Since the discretization can be considered as lossy compression, the word modeling results obtained with discrete representations provide a lower bound to the learnability from the data from which semi-continuous or fully continuous methods designed to do the same task should be able to improve, if properly formulated.

5. Future work and open questions

It is evident from experimental psychology that the language exposure shapes the perception of speech sounds, and there are now multiple computational models that are able to replicate these findings to a large degree. However, whether this type of development towards categorical perception of speech sounds is indicative of emerging representation of language as a sequence of categorical subword elements is not clear. The other possibility is that, although there are statistical biases in the perception of speech sounds towards typical distributional patterns of native speech, the output of this perception

layer to the lexicon is not discrete and sequential, but more like a continuous probabilistic stream of speech patterns. The recognition of sub-word units as unique objects with well-defined boundaries may not explicitly precede lexical access, but may follow from lexical processing and experience with the written language [27].

Current computational methods reveal that, in the absence of any innate linguistic or phonetic knowledge, learning words is possible if the speech signal is represented as sequence of fine-grained spectral features. On the other hand, there is only limited success in approaches that first learn to describe speech as a sequence of spectrally and temporally defined phone-like units and then to discover words from these phone sequences [105,114]. The current evidence, however, is not conclusive, proving only what is possible instead of proving what is impossible. For example, none of the existing models have been able to address the hypothesis of PRIMIR that the phonemic representation of language emerges later in the development through accumulation of lexicon and discovery of similarities across different lexical items.

If a learning agent would somehow obtain sufficiently systematical and invariant sequential representation of speech sounds comparable to the phonetic or syllabic transcriptions made by expert phoneticians, the various word segmentation methods described in the literature show how word learning can be accomplished using this type of representation (e.g., [123-129], and how syntactic categories could be also inferred from phonological representations using distributional information [130].

Before that, more research is required in order to understand how the interplay between lexical and sub-lexical representations drives the development of language proficiency and communicative capability in early development.

5.1 Some open issues in unsupervised word learning

The current computational models face a number of issues that are not fully addressed by any of the existing models. One of the biggest problems is the generalization across tokens with variable acoustic properties such as different talkers. Distributional categories of speech sounds learned from acoustic signals are talker specific, and speaker-independent representations overlap so largely that discrimination of context-independent vowels is far from perfect. The same is true for learned lexical items, where generalization towards new talkers is poor [15,118]. How infants overcome the generalization problem is currently not understood.

Another open issue is the role of prosody in early LA. Although behavioral studies indicate that the infants are sensitive to prosodial aspects such as intonation and stress [131-134]), no computational model dealing with continuous speech has so far been able to utilize these features efficiently in its functionality. One should however note that the prosodial features are inherently included in all standard features that encode the wide-band spectrum of the speech signal such as the FFT and MFCCs. The question then remains whether young infants treat prosody or other suprasegmental cues as a separate source of information and process it in isolation from the systems dealing with phonetic and lexical identity. If so, inclusion of a separate mechanism for prosodial processing should show some value in computational simulations. If not, a mechanism explaining the development of the ability to separate linguistic and paralinguistic information from the one and same signal is required.

Temporal representation of speech signals is also an open question. Current computational models typically describe speech as a sequence of feature frames extracted at fixed intervals. While the segmentation of the signal into phone-like units before lexical access has been studied [48,105,114], bottom-up discovery of phone-like units is evidently difficult (see section 3). Interestingly, no computational approach has truly utilized syllabification of speech signals, although, e.g., the

WRAPSA model of LA [26] directly states that the syllables serve as the basic temporal slices of speech input in perception. This is justified by the argument that seeing any other units than syllables that would enable automatic temporal normalization of speech is difficult [135]. Also Werker and Curtin [25] argue that there is an innate preference for syllabification of speech input. The unsupervised segmentation of speech signals into syllabic units is known to be much more systematic and accurate with computational algorithms than the blind segmentation into phone-like units [136]. Finally, EEG studies suggest that human speech comprehension performance correlates with the synchronization of the auditory cortex to the energy envelope of speech, and this synchronization shows deficits in dyslexic patients [137]. Therefore the role of syllables in early LA should be further investigated with computational models.

The question of the role of grounding in the learning of internal representations for acoustic patterns corresponding to words also needs more attention. There is clear evidence that highly accurate models for words can be learned in a cross-situational learning situation where the learning mechanism receives utterances paired with a set of possible word referents and that this scales up to at least vocabularies of 50 words and multiple target words in each utterance [108,109,114,115,116]. However, the statistical linkage between the acoustics and the referents is direct and the algorithms cannot learn models for word patterns without clear referential information. Moreover, no lexical learning occurs at all if no referents are available. This also means that the models cannot learn words that do not have distinctive contextual referents. Also, the simulations have very strong assumptions (but not necessarily unreasonable; see the section 2) regarding the coherence between the contents of the spoken utterances and the attention of the learner. On the other hand, unsupervised acquisition of words in the absence of referential information is demonstrated in the works of Oates [91,92], Park and Glass [93,94] and Räsänen [15], but the generalization performance of these word models is worse due to their inability

to overcome significant acoustic differences between word tokens in the absence of any contextual constraints, possibly leading to multiple parallel models for each word spoken in different acoustic conditions.

The general problem is that none of the existing models provide a systematic strategy to exploit both bottom-up statistical cues and cross-situational cues in concert in order to find the best possible representations for the incoming speech. In language learning literature, the word forms are often assumed to be first segmented from the speech stream before meaning can be attached to them (e.g., [25]), being in line with the idea of indirect lexical grounding. On the other hand, the contents of the early receptive vocabulary of infants mainly consists of nouns with very distinctive external referents (e.g., [138] and MacArthur–Bates communicative development inventories [139]), suggesting that the bootstrapping of early lexicon could be also explained by acoustic patterning based on direct cross-situational learning. Given the current evidence, it is difficult to say whether word forms come first based solely on their acoustic properties or whether contextual constraints play a role in the lexical learning all the way from the beginning.

Joint attention and intentionality in language learning are also a topics of future research. The social-pragmatic theory of word learning states that a linguistic symbol itself is a tool to share attention between multiple persons and that learning children already know that the function of language is to direct the attention of others [140]. The theory also assumes that the linguistic competence arises from the learner's ability to infer the intentions of the speaker and then relate these to the spoken messages instead of just associating superficially perceived objects and actions to concurring words. Intentionality is also claimed to drive attention so that the task of the perceptual system is to search for goal-relevant features in the environment [141]. In other words, the entire language use is about modulating attention toward internal and external concepts. However, the social-pragmatic theory does

not state how the intentions are extracted or represented by the learner. This poses a difficult task for computational models of LA, since modeling of intentionality calls for agents with a highly developed ability to understand and reason the state of matters in the surrounding world, not only for their own percepts and actions, but also for actions and states of other agents. The question how intentionality could be modeled in simulated environments in a plausible manner is way beyond the scope of this paper, but the reader is recommended to see [142] for a related review.

Finally, there are numerous other important phenomena that are barely touched in the existing research on computational models of LA. For example, syntax has been largely ignored in the existing work, assuming that syntactic learning follows from lexical knowledge. However, syntax may actually provide additional cues to the word learning problem: instead of assuming that the spoken words are purely independent of each other, the statistical dependencies across hypothesized word-like patterns may be used as an additional criterion in the learning process. In a similar vein, articulatory learning and the role of feedback in perceptual learning are not currently understood, although the hypothetical role of the articulatory domain in speech perception has been disputed since the introduction of the motor theory of speech perception [82]. There is also a plethora of knowledge from experimental studies related to bilingualism, language-related developmental disorders, and, e.g., distributional learning in non-speech domains that could be used to inspire and to evaluate future models of language acquisition.

5.2 Concluding remarks

This review has mainly focused on off-line computational experiments performed with pre-recorded or simulated data. However, the advances in technology have enabled study of language learning cognitive systems by building autonomous robots capable of sensing and acting in the real world environment (e.g. [143-145]). These systems are inherently embodied and the realistic learning data is

there to be explored. This overcomes many challenges with the necessary assumptions and simplifications of computational models discussed previously.

However, direct integration to the real world does not come without drawbacks. The first obvious limitation is the time-scale of the experiments. Since interaction with the real world and real people takes place in real-time, each experiment takes a long time to run. Controlling the effects of all environmental variables is also difficult in all but the simplest learning situations. Replication of the same experiments with different parameter settings is also difficult when humans are involved in the interaction. Finally, mechanical robots come with plenty of design, maintenance, and programming work that is not directly related to the research questions at hand, requiring notable extra resources.

Together with the huge number of open issues in LA research, this suggests that there still is room for computational simulations that can efficiently model specific aspects of the learning problem in repeatable and a well-controlled manner. Embodied robots will likely be the platform for the ultimate evaluation of unifying computational models and algorithms that claim to explain how LA takes place. Before that, many virtual steps remain to be taken.

Acknowledgements

References

[1] Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P., 2006. Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, 9, F13–F21.

[2] Werker, J., & Tees, R., 1984. Cross-language speech perception evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7, 49–63.

[3] Best, C., & McRoberts, G. W., 2003. Infant perception of non-native consonant contrasts that adults assimilate in different way. *Language and Speech*, 46, 183–216.

[4] Caselli, M. C., Bates, E., Casadio, P., Fenson, J., Fenson, L., Sanderl, L., & Weir, J., 1995. A cross-linguistic study of early lexical development. *Cognitive Development*, 10, 159–199.

[5] Hamilton, A., Plunkett, K., & Schafer, G., 2000. Infant vocabulary development assessed with a British communicative development inventory. *Journal of Child Language*, 27, 689–705.

[6] Levitt, A., & Utman, J., 1992. From babbling towards the sound systems of English and French: a longitudinal two-case study. *Journal of Child Language*, 19, 19–49.

[7] Meltzoff, A., Kuhl, P., Movellan, J., & Sejnowski, T., 2009. Foundations for a New Science of Learning. *Science*, 323, 284–288.

[8] Marr, D., 1982. *Vision: A Computational Approach*. San Francisco, Freeman & Co.

[9] Unal, F. A., & Tepedelenlioglu, N., 1992. Dynamic time warping using an artificial neural network. *Proceedings of International Joint Conference on Neural Networks* (*IJCNN*), pp. 715-721.

[10] Kanerva, P., Kristoferson, J., & Holst, A., 2000. Random Indexing of Text Samples for Latent Semantic Analysis. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, pp. 103-106, 2000.

[11] Beal, J., & Roberts, J., 2009. Enhancing Methodological Rigor for Computational Cognitive Science: Complexity Analysis. *Proceedings of the 31th Annual Conference of the Cognitive Science Society*, pp. 99-104, 2009.

[12] Feldman, N., Griffiths, T., & Morgan, J., 2009. Learning phonetic categories by learning a lexicon. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, Austin, Texas, pp. 2208–2213.

[13] Smith, K., Smith, A. D., Blythe, R. A., & Vogt, P., 2006. Cross-situational learning: a mathematical approach. *Proceedings of the Third International Workshop on the Emergence and Evolution of Linguistic Communication*, Rome, Italy, pp. 31–44.

[14] McMurray, B., & Hollich, G., 2009. Core computational principles of language acquisition: can statistical learning do the job? Introduction to Special Section. *Developmental Science*, 12, 365-368.

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

[16] Witner, S., 2010. Computational Models of Language Acquisition. *Proceedings of the 11th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing'2010)*, pp. 86–99.

[17] Daland, R., & Pierrehumbert, J., 2011. Learning Diphone-Based Segmentation. *Cognitive Science, 35*, 119–155.

[18] Buttery, P., 2006. *Computational Models for First Language Acquisition*. Technical Report No. 675, University of Cambridge, Computer Laboratory, UK.

[19] McClelland, J., & Elman, J., 1986. The TRACE model of speech perception. *Cognitive Psychology*, 18, 1–86.

[20] Norris, D., 1994. Shortlist: a connectionist model of continuous speech recognition. *Cognition*, 52, 189–234.

[21] Scharenborg, O., & Boves, L., 2010. Computational modelling of spoken-word recognition processes. *Pragmatics & Cognition*, 18, 136–164.

[22] Elman, J., (1990). Finding Structure in Time. *Cognitive Science*, 14, 179-211.

[23] ten Bosch, L., & Kirchhoff, K. (Eds.), 2007. Bridging the Gap between Human and Automatic Speech Recognition [Special Issue]. *Speech Communication*, 49, 331–436.

[24] Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson T., 2008. Phonetic learning as a pathway to language: new data and native language magnet theory expanded, NLM-e). *Phil. Trans. Royal Society B*, 363, 979–1000.

[25] Werker, J. F., & Curtin, S., 2005. PRIMIR: A Developmental Framework of Infant Speech Processing. *Language Learning and Development*, 1, 197–234.

[26] Jusczyk, P. W., 1993. From general to language-specific capacities: the WRAPSA model of how speech perception develops. *Journal of Phonetics*, 21, 3–28.

[27] Port, R., 2007. How are words stored in memory? Beyond phones and phonemes. *New Ideas in Psychology,* 25, 143–170.

[28] Norris, D., McQueen, J., & Cutler, A., 2000. Merging information in speech: Feedback is never necessary. *Behavioral and Brain Sciences*, 23, 299–370.

[29] Kuhl, P., 1986. Theoretical contributions of tests on animals to the special mechanisms debate in speech. *Experimental Biology*, 45, 233–265.

[30] Kuhl, P., 2004. Early Language Acquisition: Cracking the Speech Code. *Nature Reviews Neuroscience*, 5, 831–843.

[31] Scharenborg, O., Ernestus, M., & Wan, V., 2007. Segmentation of speech: Child's play?

Proceedings of the Interspeech'07, Antwerp, Belgium, pp. 1953–1956.

[32] Esposito, A., & Aversano, G., 2005. Text Independent Methods for Speech Segmentation. In G. Chollet et al. (Eds.), *Lecture Notes in Computer Science: Nonlinear Speech Modeling*. Springer Verlag, Berlin Heidelberg, pp. 261-290.

[33] Estevan, Y.P., Wan, V., & Scharenborg, O., 2007. Finding Maximum Margin Segments in Speech. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '07)*, Honolulu, Hawaii, USA, pp. IV–937–940.

[34] Aversano, G., Esposito, A., Esposito, A., & Marinaro, M., 2001. A New Text-Independent Method for Phoneme Segmentation. *Proceedings of the IEEE international Workshop on Circuits and Systems*, Dayton, Ohio, USA, pp. 516–519.

[35] Almpanidis, G., & Kotropoulos, C., 2008. Phonemic segmentation using the generalized Gamma distribution and small sample Bayesian information criterion. *Speech Communication*, 50, 38–55.

[36] Räsänen, O. J., Laine, U. K., & Altosaar, T., 2011. Blind segmentation of speech using non-linear filtering methods. In I. Ipsic (Ed.), *Speech technologies*. InTech, Praha, Czech Republic, pp. 105-124.

[37] Räsänen O., & Driesen J., 2009. A comparison and combination of segmental and fixed-frame signal representations in NMF-based word recognition. *Proceedings of 17th Nordic Conference on Computational Linguistics*, Odense, Denmark, pp. 255–262.

[38] Demuynck, K., & Laureys, T., 2002. A Comparison of Different Approaches to Automatic Speech Segmentation. *Proceedings of the 5th International Conference on Text, Speech and Dialogue,* pp. 277–284.

[39] Toledano, D., Hernández Gómez, L., & Villarubia Grande, L., 2003. Automatic Phonetic

Segmentation. *IEEE Transactions in Speech and Audio Processing*, 11, 617–625.

[40] Keshet, J., Shalev-Shwartz, S., Singer, Y., & Chazan D., 2005. Phoneme Alignment Based on Discriminative Learning. *Proceedings of the Interspeech'05*, pp. 2961– 2964.

[41] Pisoni, D. B., 1997. Some thoughts on ''Normalization'' in speech perception. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing*. San Diego: Academic Press, pp. 9-32.

[42] Houston, D. M., & Jusczyk, P. W., 2003. Infants' long-term memory for the sound patterns of words and voices. *Journal of Experimental Psychology: Human Perception and Performance*, 29, 1143–1154.

[43] Curtin, S., Mintz, T. H., & Byrd, D., 2001. Coarticulatory cues enhance infants' recognition of syllable sequences in speech. *Proceedings of the 25th Annual Boston University Conference on Language Development*, Somerville, MA: Cascadilla, pp. 190–201.

[44] Curtin, S., Mintz, T. H., & Christiansen, M. H., 2005. Stress changes representational landscape: Evidence from word segmentation. *Cognition*, 96, 233–262.

[45] Stager, C. L., & Werker, J. F., 1997. Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, 388, 381–382.

[46] Warren, R. M., 2000. Phonemic organization does not occur: Hence no feedback. Commentary to Norris, D., McQueen, J. M., & Cutler, A., 2000. Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, 23, 350–351.

[47] Waterson, N., 1971. Child phonology: a prosodic view. *Journal of Linguistics*, 7, 179–211.

[48] Markey, K. L., 1994. *The Sensorimotor Foundations of Phonology: A Computational Model of Early Childhood Articulatory and Phonetic Development*. Doctoral Thesis, University of Colorado, Dept. Computer Science, Colorado, USA.

[49] Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K., 1995. Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, 97, 3099–3111.

[50] Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J., 1971. Speech perception in infants. *Science* 171, 303–306.

[51] Trehub, S. E., 1976. The discrimination of foreign speech contrasts by infants and adults. *Child Devevelopment*, 47, 466–472.

[52] Tsao, F.-M., Liu, H.-M. & Kuhl, P. K., 2004. Speech perception in infancy predicts language development in the second year of life: a longitudinal study. *Child Development*, 75, 1067–1084.

[53] Kuhl, P. K., Conboy, B. T., Padden, D., Nelson, T., & Pruitt, J., 2005. Early speech perception and later language development: implications for the "critical period". *Language Learning and Development*, 1, 237–264.

[54] de Boer, B., & Kuhl, P., 2003. Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, 4, 129–134.

[55] Dempster, A.P., Laird, N.M., & Rubin, D.B., 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society - Series B (Methodological)*, 39, 1–38.

[56] Vallabha, G. K., McLelland, J. L., Pons, F., Werker, J. F., & Amano, S., 2007. Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of National Academy of Sciences*, 104, 13273–13278.

[57] McMurray, B., Aslin, R., & Toscano, J., 2009. Statistical learning of phonetic categories: insights from a computational approach. *Developmental Science*, 12, 369–378.

[58] Duda, R., Hart, P., & Stork, D., 2001. *Pattern Classification* (2nd Edition). Wiley-Interscience, New York.

[59] Lake, B., Vallabha, G., & McClelland, J., 2009. Modeling Unsupervised Perceptual Category Learning. *IEEE Transactions on Autonomous Mental Development*, 1, 35–43.

[60] Maye, J., Werker, J., & Gerken, L-A., 2002. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82, B101–B111.

[61] Toscano, J. C., & McMurray, B., 2010. Cue Integration With Categories: Weighting Acoustic Cues in Speech Using Unsupervised Learning and Distributional Statistics. *Cognitive Science*, 34, 434–464.

[62] Feldman, N., Griffiths, T. L., & Morgan, J. L., 2009. The Influence of Categories on Perception: Explaining Perceptual Magnet Effect as Optimal Statistical Inference. *Psychological Review*, 116, 752-782.

[63] Kouki, M., Kikuchi, H., & Mazuka, R., 2010. Unsupervised Learning of Vowels from Continuous Speech Based on Self-Organized Phoneme Acquisition Model. *Proceedings of the Interspeech'2010*, pp. 2914–2917.

[64] Kohonen, T., 1990. The Self-organizing Map. *Proceedings of the IEEE*, 78, 1464–1480.

[65] Guenther, F. H., & Gjaja, M. N., 1996. The Perceptual Magnet Effect as an Emergent Property of Neural Map Formation. *Journal of the Acoustical Society of America*, 100, 1111–1121.

[66] Iverson, P., & Kuhl, P. K., 1994. Tests of the perceptual magnet effect for American English /r/ and /l/. *Journal of the Acoustical Society of America*, 95, 2976.

[67] Duran, D., Schütze, H., Möbius, B., & Walsh, M., in press. A Computational Model of Unsupervised Speech Segmentation for Correspondence Learning. *Research on Language and Computation*.

[68] Kuwahara, H., & Sasaki, H., 1972. Perception of Vowels and C-V Syllables Segmented from Connected Speech. *The Acoustical Society of Japan*, 28, 225–234.

[69] Coen, M. H., 2006. Self-supervised acquisition of vowels in American English. *Proceedings of the 21st national conference on Artificial intelligence*, Boston, USA, 2, pp. 1451–1456.

[70] Coen, M.H., 2005. Cross-modal Clustering. *Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI'05)*, Pittsburg, PA, pp. 932–937.

[71] Peterson, G. E., & Barney, H. L., 1952. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24, 175-184.

[72] Goldstein, M. H., & Schwade, J. A., 2008. Social Feedback to Infants' Babbling Facilitates Rapid Phonological Learning, *Psychological Science*, 19, 515-523.

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

[74] Nowlan, S. J., 1991. Maximum likelihood competitive learning. In D. S. Touretsky (Ed.), *Advances in Neural Information Processing Systems*, *Vol. 2*. San Mateo, CA: Morgan Kauffan Publishers, pp. 574-582.

[75] Kokkinaki, T., &Kugiumutzakis, G., 2000. Basic aspects of vocal imitation in infant–parent interaction during the first 6 months. *Journal of Reproductive and Infant Psychology*, 18, 173– 187.

[76] Jones, S. S., 2007. Imitation in infancy: The development of mimicry. *Psychological Science*, 18, 593–599.

[77] Howard, I. S, & Messum, P., 2011. Modeling the Development of Pronunciation in Infant Speech Acquisition. *Motor Control*, 15, 85-117.

[78] Maeda, S., 1990. Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. In W. J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modeling*, Boston: Kluwer Academic Publishers, pp. 131-149.

[79] Fant, G., Liljencrants, J., & Lin, Q., 1985. A four-parameter model of global flow. *Speech Transmission Laboratory. Quarterly Progress and Status Reports* (*STL-QPSR*), 4, 1-13.

[80] Miller, K., 1992. Development of orientation columns via competition between ON- and OFF-center inputs. *Neuroreport*, 3, 73–76.

[81] Kirchhoff, K., & Schimmel, S., 2005. Statistical properties of infant-directed versus adult-directed speech: Insights from speech recognition. *Journal of Acoustical Society of America*, 117, 2238–2246.

[82] Liberman, A., & Mattingly I., 1985. The motor theory of speech perception revised, *Cognition*, 21, 1-36.

[83] Brosch, M., & Scheich, H., 2005. Non-acoustic influence on neural activity in auditory cortex. In R. König, P. Heil, E. Budinger, & H. Scheich (Eds.), *The auditory cortex: A synthesis of human and animal research* (pp. 127–144). New Jersey: Lawrence Erlbaum Associates.

[84] Emmorey, K., 2006. The signer as an embodied mirror neuron system: neural mechanisms underlying sign language and action. In M. Arbib (Ed.), *From Action to Language via the Mirror Neuron System*. Cambridge University Press, New York, pp. 110-135.

[85] Golinkoff, R. M., Mervis, C. B., & Hirsh-Pasek, K., 1994. Early object labels: the case for a developmental principles framework. *Journal of Child Language*, 21,125–155.

[86] Quine, W. V. O., 1960. *Word and object*. Cambridge, MA: MIT Press.

[87] Pinker, S., 1989. *Learnability and cognition: The acquisition of argument structure*. Cambridge,

MA: MIT Press.

[88] Gleitman, L. R., 1990. The structural sources of verb meanings. *Language Acquisition*, 1, 3–55.

[89] Smith, L. B., & Yu, C., 2008. Infants rapidly learn word-referent mappings via cross-situational statistics. Cognition, 106, 1558–1568.

[90] Smith, K., Smith, A. D., & Blythe, R. A., 2011. Cross-Situational Learning: An Experimental Study of Word-Learning Mechanisms. *Cognitive Science*, 35, 480–498.

[91] Oates, T., 2002. PERUSE: An unsupervised algorithm for finding recurrent patterns in time-series. *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, Maebashi City, Japan, pp. 330–337.

[92] Oates, T., 2001. *Grounding Knowledge in Sensors: Unsupervised Learning for Language and Planning*. Doctoral Thesis, University of Massachusetts Amherst, MA, USA.

[93] Park, A., & Glass, J. R., 2005. Towards Unsupervised Pattern Discovery in Speech. *Proceedings of 2005 IEEE Workshop Automatic Speech Recognition and Understanding (ASRU'05)*, Cancún, Mexico, pp. 53–58.

[94] Park, A., & Glass, J. R., 2006. Unsupervised word acquisition from speech using pattern discovery. *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06)*, Toulouse, France, pp. 409–412.

[95] McInnes, F., & Goldwater, S., 2011. Unsupervised extraction of recurring words from infant-directed speech. *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, Boston, MA, pp. 2006-2012.

[96] Newman, M. E. J., 2004. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69, 0066133–1–5.

[97] Brent, M. R., & Siskind, J., 2001. The role of exposure to isolated words in early vocabulary development. *Cognition*, 81, B33–44.

[98] MacWhinney, B., & Snow, C., 1985. The child language data exchange system. *Journal of Child Language*, 12, 271–296.

[99] Saffran, J. R., Aslin, R. N., & Newport, E. L., 1996. Statistical Learning by 8-Month-Old Infants. *Science*, 274, 1926–1928, 1996.

[100] Saffran, J. R., Newport, E. L., & Aslin R. N., 1996. Word Segmentation: The Role of Distributional Cues. *Journal of Memory and Language*, 35, 606–621.

[101] Saffran, J. R., 2001. Words in the sea of sounds: the output of infant statistical learning. *Cognition,* 81, 149–169.

[102] Altosaar, T., ten Bosch, L., Aimetti, G., Koniaris, C., Demuynck, K., & van den Heuvel, H., 2010. A Speech Corpus for Modeling Language Acquisition: CAREGIVER. *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Malta, pp. 1062–1068.

[103] Räsänen, O., & Rasilo, H., 2012. Acoustic analysis supports the existence of a single distributional learning mechanism in structural rule learning from an artificial language. Accepted to *Proceedings of the 34th Annual Conference of the Cognitive Science Society (CogSci2012)*, Sapporo, Japan.

[104] Kanerva, P., 2009. Hyperdimensional Computing: An introduction to computing in distributed representation with high-dimensional random vectors. *Cognitive Computation*, 1, 139-159.

[105] ten Bosch, L. & Cranen, B., 2007. A computational model for unsupervised word discovery. *Proceedings of Interspeech'07*, Antwerp, Belgium, pp. 1481–1484.

[106] MacQueen, J. B., 1967. Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 281–297.

[107] Hirsch, H. G., & Pearce, D., 2000. The AURORA Experimental framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions. *Proceedings of the ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millennium*, Paris, France, pp. 29–32.

[108] ten Bosch, L., Van hamme, H., & Boves, L., 2008. Discovery of words: Towards a computational model of language acquisition. In F. Mihelič and J. Žibert (Eds.), *Speech Recognition: Technologies and Applications*. Vienna: I-Tech Education and Publishing KG, pp. 205-224.

[109] Van hamme, H., 2008. HAC-models: a Novel Approach to Continuous Speech Recognition. *Proceedings of the Interspeech'08*, Brisbane, Australia, pp. 2554–2557.

[110] Stouten, V., Demuynck, K., & Van hamme, H., 2007. Discovering Phone Patterns in Spoken Utterances by Non-negative Matrix Factorization. *IEEE Signal Processing Letters*, 15, 131–132.

[111] Lee, D. D., & Seung, H. S., 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401, 788–791.

[112] ten Bosch, L., Van hamme, H., Boves, L., & Moore, R.K., 2009. A computational model of language acquisition: the emergence of words. *Fundamenta Informaticae*, 90, 229–249.

[113] Driesen, J., ten Bosch, L., & Van hamme, H., 2009. Adaptive Non-negative Matrix Factorization in a Computational Model of Language Acquisition. *Proceedings of the Interspeech'09*, Brighton, England, pp. 1731–1734.

[114] Räsänen, O., Laine, U. K., & Altosaar, T., 2008. Computational language acquisition by statistical bottom-up processing. *Proceedings of the Interspeech'08,* Brisbane, Australia, pp. 1980–1983.

[115] Räsänen, O., & Laine, U. K., 2012. A method for noise-robust context-aware pattern discovery and recognition from categorical sequences. *Pattern Recognition*, 45, 606–616.

[116] Aimetti, G., 2009. Modelling early language acquisition skills: Towards a general statistical learning mechanism. *Proceedings of EACL-2009-SRWS*, Athens, Greece, pp. 1–9.

[117] ten Bosch L., Boves L. & Räsänen O., 2009. Learning meaningful units from multimodal input - the effect of interaction strategies. *Proceedings of Workshop on Child, Computer and Interaction 2009 (WOCCI),* Boston, MA, United States.

[118] ten Bosch L., Räsänen O., Driesen J., Aimetti G., Altosaar T., & Boves, L., 2009. Do Multiple Caregivers Speed up Language Acquisition? *Proceedings of Interspeech'09,* Brighton, England, pp. 704–707.

[119] Versteegh, M., ten Bosch, L., & Boves, L., 2010. Active word learning under uncertain input conditions. *Proceedings of the Interspeech'10*, Chiba, Japan, pp. 2930–2933.

[120] Miller, M., & Stoytchev, A., 2009. An Unsupervised Model of Infant Acoustic Speech Segmentation. *Proceedings of 9th International Conference on Epigenetic Robotics*, Venice, Italy, 12–14.

[121] Räsänen, O., 2012. Context induced merging of synonymous word models in computational modeling of early language acquisition. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2012)*, Kyoto, Japan, pp. 5037-5040, 2012.

[122] Räsänen, O., 2012. Structure of continuous speech matches with temporal processing in auditory perception. Submitted for publication.

[123] de Marcken, C., 1995. *The unsupervised acquisition of a lexicon from continuous speech*. AI Memo No. 1558, Massachusetts Institute of Technology, MA.

[124] Christiansen, M. H., Allen, J. A., & Seidenberg, M. S, 1998. Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, 13, 221–268.

[125] Brent, M. R., & Cartwright, T. A., 1996. Distributional regularity and phonotactics are useful for segmentation. *Cognition*, 61, 93–125.

[126] Brent, M. R., 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34, 71–105.

[127] Venkataraman, A., 2001. A statistical model for word discovery in transcribed speech. *Computational Linguistics*, 27, 351–372.

[128] Swingley, D., 2005. Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, 50, 86–132.

[129] Blanchard, D., Heinz, J., & Golinkoff, R., 2010. Modeling the contribution of phonotactic cues to the problem of word segmentation. *Journal of Child Language*, 37, 487–511.

[130] Christiansen, M. H., Onnis, L., & Hockema, S. A., 2009. The secret is in the sound: from unsegmented speech to lexical categories. *Developmental Science*, 12, 388-395.

[131] Thiessen, E., & Saffran, J. R., 2003. When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology*, 39, 706–716.

[132] Thiessen, E., & Saffran, J. R., 2004. Spectral tilt as a cue to word segmentation in infancy and adulthood. *Perception and Psychophysics*, 65, 779–791.

[133] Cutler, A., 1994. Segmentation problems, rhythmic solutions. *Lingua*, 92, 81–104.

[134] Jusczyk, P. W., 1993. Discovering sound patterns in the native language. *Proceedings of the 15th Annual Meeting of the Cognitive Science Society*, Colorado, Boulder, pp. 49–60.

[135] Mehler, J., Dupoux, E., & Segui, J., 1990. Constraining models of lexical access: the onset of word recognition. In G. T. Altmann (Ed.), *Cognitive models of speech processing*. Hillsdale, NJ:Erlbaum.

[136] Villing, R., Ward, T., & Timoney, J., 2006. Performance Limits for Envelope based Automatic Syllable Segmentation. *Proceedings of the Irish Signals and Systems Conference: ISSC2006*, pp. 521–526.

[137] Ahissar, E., & Ahissar, M, 2005. Processing of the temporal envelope of speech. In R. König, P. Heil, E. Budinger, & H. Scheich (Eds.), *The auditory cortex: A synthesis of human and animal research* (pp. 295–314). New Jersey: Lawrence Erlbaum Associates.

[138] Gentner, D., 1983. Why nouns are learned before verbs: linguistic relativity versus natural partitioning. In S. Kuczaj (Ed.), *Language Development*, *Vol. 2, Language, cognition and culture*. Hillsdale, NJ: Erlbaum.

[139] Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., & Bates, E., 2003. *MacArthur–Bates communicative development inventories (CDIs)* (2nd ed.). Baltimore, MD: Brooks Publishing.

[140] Tomasello, M., 2000. The Social-Pragmatic Theory of Word Learning. *Pragmatics*, 10, 401–413.

[141] Tomasello, M., 1995. Joint attention as social cognition. In C. Moore & P. Dunham (Eds.), *Joint attention: Its origins and role in development* (pp. 103–130). Hillsdale, NJ: Erlbaum.

[142] Kaplan, F., & Hafner, V. V., 2006. The challenges of joint attention. *Interaction Studies*, 7 135–169.

[143] Steels, L., & Kaplan, F., 2000. Aibo's first words: The social learning of language and meaning. *Evolution of Communication*, 4, 3–32.

[144] Steels, L., 2003. Evolving grounded communication for robots. *Trends in Cognitive Science*, 7, 308–312.

[145] Roy, D., 2003. Grounded Spoken Language Acquisition: Experiments in Word Learning. *IEEE Transactions on Multimedia*, 5, 197–209.