A self-referential childlike model to acquire phones, syllables and words from acoustic speech

Holger Brandl, Britta Wrede
Research Institute for Cognition and Robotics
Bielefeld University
{*hbrandl, bwrede*}*@techkfak.uni-bielefeld.de*

Frank Joublin, Christian Goerick
Honda Research Institute Europe GmbH
Offenbach am Main
{*Frank.Joublin, Christian.Goerick*}*@honda-ri.de*

***Abstract*—Speech understanding requires the ability to parse spoken utterances into words. But this ability is not innate and needs to be developed by infants within the first years of their life. So far almost all computational speech processing systems neglected this bootstrapping process. Here we propose a model for early infant word learning embedded into a layered architecture comprising phone, phonotactics and syllable learning. Our model uses raw acoustic speech as input and aims to learn the structure of speech unsupervised on different levels of granularity.**

We present first experiments which evaluate our model on speech corpora that have some of the properties of infant-directed speech. To further motivate our approach we outline how the proposed model integrates into an embodied multimodal learning and interaction framework running on Honda's ASIMO robot.

***Index Terms*—Language Acquisition, Multimodal Integration, Robotics, Speech Recognition, Statistical Language Modeling**

I. INTRODUCTION

Most computational models for word acquisition suffer from two major weaknesses. First, they tackle the problem of speech acquisition in the symbolic domain only, although it is not clear how and whether these approaches can be generalized to the acoustic domain. Second, most models rely on some kind of innate representation, which is mostly at the level of syllables. But because syllables depend strongly on the language to be learned, it is not clear how these approaches can be extended to become valid models for language acquisition as observed in infants.

To our best knowledge existing computational models offer only very limited explanation for the marvelous process of speech acquisition observed in infants. One of the most important, but in the literature often neglected ability required for word acquisition, is the segmentation of speech into words (c.f. c.f. [1]). There is evidence, that this ability allows already infants as young as 8 months to bootstrap new words based on the principle of subtraction (c.f. [2]).

The idea of this work is to bootstrap a word representation based on the statistics of raw acoustic input speech only. Following Occam's razor we first evaluated the most straightforward word acquisition approach, which is to use utterances of word length to bootstrap new words and to apply the principle of subtraction to learn also words which do not appear as isolated utterances (c.f. [3]). Unfortunately this approach failed, because length is not a reliable cue for word segmentation. Therefore word learning needs to be modeled using more elaborate methods like metric segmentation strategies, transitional probabilistic models, or the unique stress constraint principle, which are believed to play a role in the early infant lexical word learning (c.f. [4], [5]). All these principles depend on a representation of the speech input in terms of syllables. But because syllables strongly depend on the language they cannot be assumed to be innate. Although there are promising results for syllable acquisition on symbolic corpora as described in [6], it remains a challenging task to bootstrap a syllable representation from raw acoustic speech, because only very few syllables appear as isolated utterances in spoken language. However, there is broad agreement that phonotactics, that are the rules which restrict how phones can be assembled to syllables, play an important role for syllable learning. In contrast to phonemes we consider phones to be acoustically distinguishable units without any relation to meaning.

Therefore, in order to actually build a system which is able to learn words based on these above mentioned principles, we propose a three-layered framework for speech acquisition: First, it learns a phone-representation including a phonotactic model. Second, based on the syllabic constraints implied by these learned phonotactics and input speech obeying some properties of infant-directed speech, a syllable representation becomes learned. Finally, our framework acquires a word lexicon based on the above mentioned word acquisition principles which are believed to play a critical role in early-infant language development. Technically, our system can be defined as a cascade of HMM-based speech unit spotting instances which rely on incomplete speech unit representations on phone, syllable and word level.

The remainder of this work is organized as follows. In section II we give an overview of our framework. Subse-

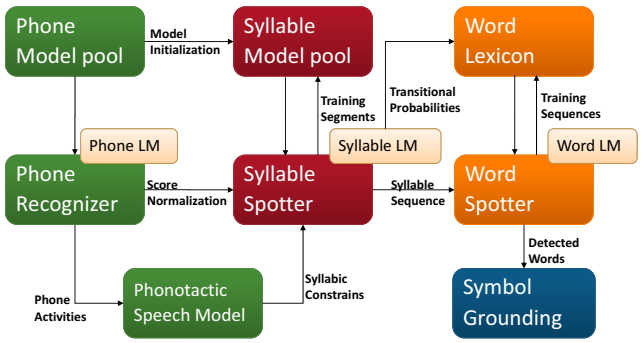


Fig. 1. The proposed three-layered architecture for speech acquisition. As indicated by the visualization all layers have a very similar structure consisting of a pool of unit-models, a statistical grammar (LM), and a recognizer which detects learned units in the incoming feature stream.

quently in section III we describe in more detail the used phone acquisition approach. Syllable learning and a set of regulatory metrics necessary to keep the syllable part of our system in homeostasis are defined in section IV. In Section V we describe the details of our lexical word acquisition method and possible extensions. The used evaluation metrics, the different kinds of evaluation scenarios and first results are subsequently reviewed in section VI. There, we also outline how our approach embeds into an embodied multi-modal learning and interaction framework. Finally, we discuss ideas for further improvements in section VII.

II. System Architecture

The proposed system architecture is shown in figure 1. Three interconnected layers are used to learn the phone-, the syllable- and finally the word representation of the input language. Initially all representations are empty. Processing and learning are organized in a bottom-up manner. The learning of phones and phonotactics completely priors syllable and word acquisition which allows to neglect some stability and plasticity issues. In contrast syllables and words are acquired incrementally in parallel.

Each layer comprises a pool of models, a detector and a statistical speech unit grammar. Because the speech units are modeled as Hidden Markov Models we can use state of the art speech recognition methods for detection. Being learned in one single clustering step, phones can be directly recognized with a phonotactically constrained Viterbi-decoder (c.f. [7]). In contrast, syllable and word representation are learned incrementally. Viterbi-decoding is not directly applicable in such a case, and we use a keyword-spotter to detect already learned syllables/words. Thereby the next lower level representation is used as background (aka. world-, filler-, OOV-) model.

Although not shown in figure 1 for sake of simplicity, input speech is framed by a voice activity detector as described in [8] into segments. These contain utterances of different complexity starting from isolated mono-syllabic words up to complex utterances comprising many poly-syllabic words. Such segments become converted to mel-frequency cepstrum coefficients including energy, and their first and second time derivatives. Resulting features define the input to the phone clustering module, the phone recognizer and to the syllable spotter. Word acquisition and spotting are based solely on the results provided by the syllable spotter, which is a discrete, initially incomplete stream of syllable symbols.

In contrast to other approaches on language acquisition, we make no assumptions on the language to be learned, except the idea that speech is organized in terms of syllables. Therefore, the first step to bootstrap a complex speech representation is to learn some basic units of speech: in our case, phones which we think to be speech segments that possess distinct physical and perceptual properties. Because the set of possible phones in a language is extremely small compared to the number of syllables or words we attempt to find such a phone representation using an unsupervised clustering approach.

III. Phone Learning

A phone representation is crucial to make our proposed acquisition model operative: firstly, it allows the conversion of speech into phone symbol sequences, which is a prerequisite to learn the phonotactics of language. Secondly, syllable models can be created by concatenation of phones HMMs as indicated in fig. 1. As shown in fig. 2 the syllable spotter requires a phone representation as background model. Finally, a phone representation allows us to normalize acoustic scores while recognizing syllables as described in [9].

To bootstrap a phone representation we adopted the approach proposed by [10]: First, a few minutes of input speech are accumulated to give a sufficiently large training sample. Single state HMMs with mixtures of Gaussians including 8 component densities as output probability distribution functions (OPDF) are estimated using k-means. Thereby a transition matrix between the different single-state HMMs is estimated. To obtain the initial-phone models, a Monte-Carlo-sampling governed by these transition probabilities is used. This gives us the most frequent state-sequences. The N most frequent state sequences are concatenated to 3-state phone-models with Bakis-topology. These initial phone models become further refined using the above mentioned Baum-Welch training. Additionally, the number of required phones N is optimized based on the Akaike information criterion as proposed in [11].

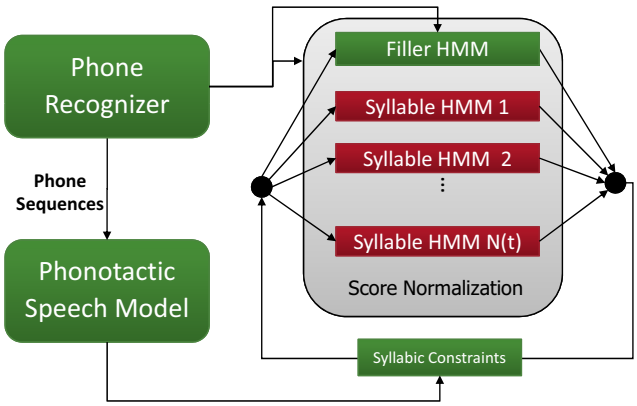


Fig. 2. The syllable spotter implementation. The phone model is used as filler model and to normalize acoustic scores. Learned phonotactics further constrain the Viterbi decoding.

A. Phonotactic modeling

Phonotactics refer to the rules that govern the structure of syllables in a particular language. Given a phone representation, a phonotactic model can only be estimated from the initial and final parts of recognized phone sequences. This is because without an explicit syllable model it is not possible to induce further phonotactically meaningful training segments than the initial phone-symbols of the syllable at utterance start and the coda phone-symbols of the syllable at the utterance end.

The probability for a syllable change is modeled by combining two N-Gram models for the syllable initial P_{SI} and final P_{SF} respectively. Thereby $P_{SI}(X)$ gives the probability that a phone-sequence X starts with a syllable (and vice versa for P_{SF}). Both are trained using the initial/final phones sequences of the training utterances only.

The probability for a syllable change after a phone symbol k in a sequence of phone-symbols $X_{1:N}$ comprising N phones is computed as the product of as P_{SI} and final P_{SF} by splitting the argument phone sequence after phone k:

$$P(k|X_{1:N}) = P_{SF}(X_{1:k}) \cdot P_{SI}(X_{k+1:N}) \tag{1}$$

The context size of the both N-Gram models was chosen to be 3 which we assume to be an appropriate trade-off between discriminative power, trainability and applicability given the task to learn syllable structure statistics. To interpolate unobserved phone-sequence probabilities due to insufficient training-samples a Katz-smoothing was used (c.f. [12]).

IV. Syllable Acquisition

Initially the syllable representation does not contain any models. Incoming speech is analyzed solely by the phone-recognizer and the voice activity detector. Inspired by the properties of infant-directed speech uttered by adults to ease the word model bootstrapping of their children, we assume the input speech to occasionally contain isolated monosyllabic words. These allow to bootstrap a first set of syllable models. In contrast to our previous work described in [3] training segments for syllable learning are now restricted to mono-syllabic segments by performing a simple hypothesis test about the number of syllables contained in a segment based on the previously learned phonotactic model. These segments are used to trigger the syllable acquisition.

As soon as a new syllable model is acquired, it becomes integrated into the syllable spotter depicted in fig. 2. Spotting results can be further employed to give new training segments for bootstrapping: Inspired by the *principle of subtraction* as described in [13] or [2] these results are fused with speech-segments to give additional training segments: Given an already acquired model for the syllable *[si]* and a sequence of syllables *[a] [si] [mo]* as speech input, the syllable spotter will be able to detect *[si]* within this sequence. That allows to *subtract* the spotted syllable-segment from the framing voice activity segment which gives two additional training segments for *[a]* and *[mo]*.

Triggered by a syllable training segment the unsupervised clustering method to syllable bootstrapping proceeds as follows: A new training segment X will be processed in a twofold way. First the model λ^* which is most likely to explain the given segment is determined by

$$\lambda^* = \arg\max_{\lambda:\mathcal{M}} P(X|\lambda) \tag{2}$$

Thereby $P(X|\lambda)$ denotes the data likelihood. For the second step we assume the histogram of former training to be approximated by a probability distribution with the density $f_{\lambda^*}(p)$. The corresponding cumulative distribution function F_{λ^*} is than used to map $P(X|\lambda^*)$:

$$\nu(\lambda^*, X) = F_{\lambda^*}(P(X|\lambda^*)) = \int_{-\infty}^{P(X|\lambda^*)} f_{\lambda^*}(p)dp \tag{3}$$

Given decision threshold θ two cases have to be considered:

1) $\nu(\lambda^*, X) \geq \theta$: X seems to be sufficiently new. The best matching sequence of phone-models will give an initialization model for the new syllable model to be created. This case applies also if the the syllable model pool is still empty.
2) $\nu(\lambda^*, X) < \theta$: The model λ^* seems to be appropriate to model the current segment X, which therefore will be used to improve/reestimate λ^* using MAP-training.

After a segment has been processed $f_{\lambda_{update}}(p)$ becomes incrementally updated with $P(X|\lambda_{update})$. Given that a specific amount of training segments was used to estimate

the parameters of a syllable model, it is tagged as *stable*. Using such a bootstrapping approach syllables are modeled incrementally based on their appearance in time.

A. Regulation

Spotted segments are used to update metrics commonly used to score unsupervised learning tasks: completeness Γ, orthogonality η and stability ψ. Applied to the problem of speech acquisition we realized these regularization terms as follows:

Model spotting coverage $\Gamma(t)$ measures the completeness of the representation at a given time. It is defined as the ratio of speech covered by at least one of the detected syllable-segments to the overall amount of speech.

Model coactivity , measures the mutual dependence between all syllable models for in a given time. Optimally syllable models are orthogonal with respect to their discriminative power, i.e. only one model is active at a time. It is measured pairwise in terms of correlated keyword spotting activity. For two models i and j the model coactivity is denoted with $\eta(\lambda_i, \lambda_j, t)$.

Pool stability $\psi(t)$ is defined as the ratio of stable models to the non stable models.

Based on these terms the acquisition problem can be reformulated as an optimization problem to provide a unified framework for speech acquisition:

$$\Gamma + \psi - |\eta| \rightarrow \max! \tag{4}$$

Thereby $|\bullet|$ denotes a common matrix norm. Intuitively this regularization function attempts to establish homeostasis as soon as the syllable representation allows to completely model the input speech. Because it is not possible to find a closed from solution, we propose two heuristics which attempt to maximize this criterion function.

(I) A first approach to limit the pool growth is chosen to be based on pool stability. New models are created only if

$$\psi(t) > \Gamma(t) \tag{5}$$

Otherwise the best pool model is updated. Using this heuristic the creation of new models is eased if speech coverage is low. Vice versa this heuristic prevents the creation of new models if the current AM is already able to model the speech input sufficiently.

(II) Whereas the default acquisition loop assumes $\nu(\lambda^*, X)$ to be greater than a fixed threshold it might be more appropriate to use an adaptive threshold. Such a threshold can be chosen by:

$$\theta = \theta_0 \cdot (1 + \beta \cdot \psi) \tag{6}$$

This heuristic is inspired by the idea to ease the creation of new models if the AM is sufficiently stable. Vice versa low stability prevents the creation of new models. Thereby β defines a weighting factor.

V. Word learning

In contrast to syllables and phones, words are modeled as simple sequences of syllable-symbols. Accordingly the word-models are discrete HMMs. We propose a lexical acquisition mechanism which allows to bootstrap a lexicon L. Its input are sequences of syllable symbols as highlighted in fig. 1. To exclude noisy models, only speech-segments which are completely covered by syllable segments are used to learn new words. Starting with an empty lexicon we combine algebraic learning (c.f. [4]) with co-occurrence based word acquisition (c.f. [5]). Given a sequence $S = S_1 S_2 ... S_N$ of syllables, lexical learning works as follows:

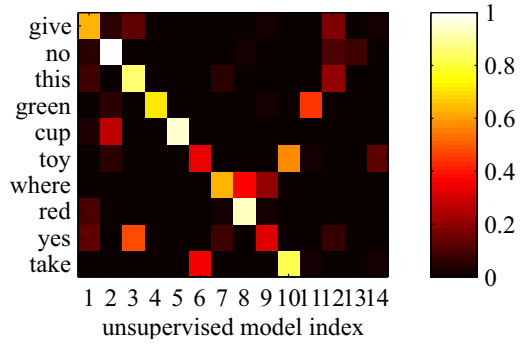
- **if**($S \in L$) return, because words are modeled as discrete symbol sequences, and a syllable sequence which is already in L does not need to be re-added.
- **if**($N == 1$) add S to L because every isolated syllable is a word.
- **else if**($P_{SLM}(S_) > \Theta$) add S to L because the syllables in S co-occur with such a high probability that it is reasonable to assume that S is a word.
- **else** Find the best matching word sequence and apply the principle of subtraction to compute the residual segments of S which are not in the lexicon yet. Apply this algorithm to all residual sequences.

It is clear to us that such an approach is far less powerful compared to symbolic lexical learning approaches like the model of Gambell and Yang proposed in [4]. The main difference lies in the fact that many word segmentation approaches rely on stress. Here, this is not applicable because of missing stress information. Unfortunately stress is highly language dependent, and we're not aware of signal processing techniques for language independent stress detection. However, we think this approach to be a first valid step in the direction of unsupervised word acquisition.

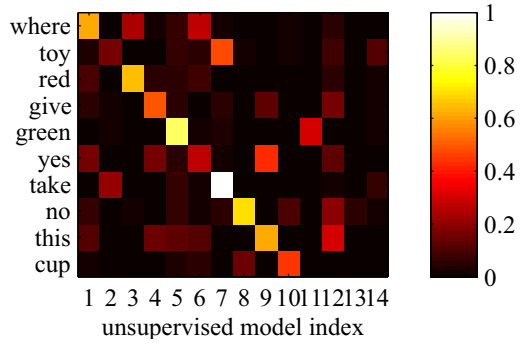
VI. Results

As depicted in fig. 3 we evaluated the system using a speech corpus comprising 40 minutes of speech containing 10 different isolated monosyllabic words in arbitrary order. The regularization module of the syllable bootstrapping was parameterized with $\theta = 0.05$. Although not shown phone-learning preceded syllable and word learning as described above. Phonotactic learning was active but had no significant influence due to simplified structure of the input data in this first experiment.

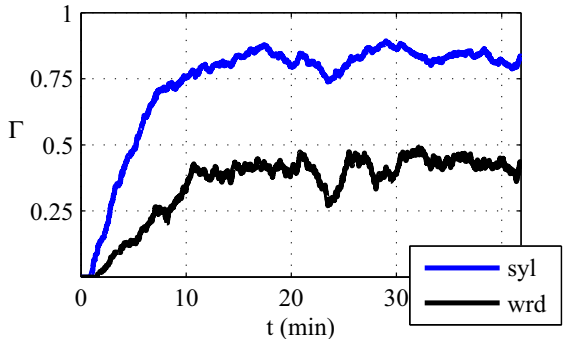
As shown in fig. 3(a) about 15 syllable models were learned which exceeds the number words in the training



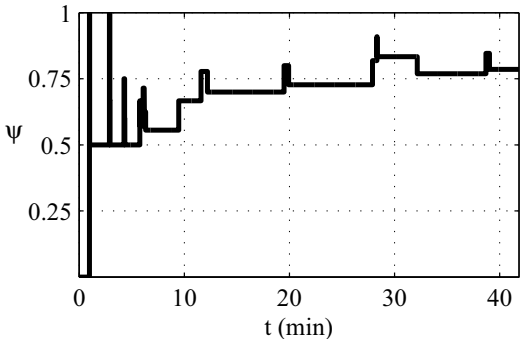
(a) *Training confusion* computed after 40 minutes of speech. The matrix was permuted to maximize its trace. Because of the applied trace maximization the relation between models and labels becomes evident.



(b) Detection confusion D_{conf} computed after 40 minutes of speech. The clear matrix trace indicates that the acquired syllable models are sufficiently discriminative to classify syllables into the correct categories.



(c) Speech Coverage Γ of syllables and words. Syllable coverage quickly approaches 90% whereas word coverage increases slower due to the more strict assumptions on lexical learning process.



(d) Syllable pool stability ψ as defined in section IV-A. High fluctuations are due to active homeostasis which triggers learning only for stable representations.

Fig. 3. Results for isolated word acquisition

sample. The syllable representation therefore over-represents the syllable structure of the input language. Compared with the training confusion matrix in 3(a) the detection confusion matrix in figure 3(b) lacks of the low orthogonality. Syllable pool stability and speech coverage converge during the acquisition process against almost full coverage and stability as shown in fig. 3(c) and 3(d). Word coverage underperforms compared with the syllable results, which is probably due to the restrictive training data pruning as described in section V.

A. Embodied language acquisition

The focus of this work is on developmentally inspired models for acoustic language acquisition. However, language acquisition requires embodiment in order to ground acquired words. Therefore we've extended our work presented in [15] towards a new system for autonomous learning and interaction (ALIS2) running on Honda's ASIMO robot (c.f. [16]). As a first step we've integrated the above mentioned syllable-learning layer into this system, allowing us to teach new auditory labels online to ASIMO. Because neither phonotactically constrained syllable-segmentation nor the word-acquisition algorithm presented in section V are part of this integration yet, only mono-syllabic words can be learned until now. Additionally, we biased learning with a set of predefined semantic classes.

Learning of new words works as follows: Given an object, the user restricts the system attention to the an object property which should be labeled (e.g. object height). Independently of a concrete appearance the system is able to detect object motion, height, planarity, and object location relative to the robot's upper body. In our current implementation, new words are then taught by providing a few (2-5) isolated samples for each word. The temporal grouping these speech segments was given to the system as an additional cue to ease learning in our first experiments.

Based on the novelty detection mechanism presented in IV the system is able to distinguish automatically between already known synonyms and new synonyms. The same mechanism allows to retrain already learned synonyms to improve the recognition performance. The system is language-independent and was successfully used to acquire mixed-language representations. In our experiments up to 20 words could be learned online solely through contingent

Fig. 4. Interaction between tutor and robot. During evaluation, the tutor pronounces a previously learned word, which is converted into an expectation towards the object property associated with it. A presented an object will lead to a positive feedback (head nodding) if the word is associated to the referred object property. Otherwise the robot will give a negative feedback (head shaking), but will keep its expectation until it becomes fulfilled.

verbal and gestural interactions based alone. No offline computation was necessary and no words were confused (c.f. [16]).

VII. Discussion

We've proposed a model for unsupervised acoustic speech acquisition inspired by principles which are believed to play a role in early infant speech acquisition. Although far from being a complete functional model for this process, our system seems to be capable to model some of its aspects: The plasticity of the phone representation vanishes after some time of habituation to a certain language. As indicated for infants in [2] the principle of subtraction plays an important role when learning new words and syllables. After some time of habituation to the phonotactics to a particular language, our system is able to bootstrap a stable syllable representation.

We could show that our current system is able to learn a stable set of syllable and word models independently of the complexity of the test language. The key concepts of our approach include a regulation scheme which ensures asymptotic homeostasis, the combination of unsupervised and supervised speech processing, and the extension of recent speech processing techniques which allows to use raw acoustic input as input. Our next steps will include a more elaborate evaluation using semi-synthetic acoustic speech corpora with defined statistic regularities as input.

For our first experiments we've concentrated on a solely perception driven processing and learning architecture. However, it is reasonable to believe some kind of top-down expectation to be beneficial to the performance of the emerging speech representation. E.g. syllable segments detected with high confidence could be converted into further training samples for the learning of phonotactics. Another possibility would be to use the learned word lexicon to bias the syllable recognition. Such ideas are straightforward to realize. However, whether and how incomplete and partially unstable representations on each level can be used to bias more basic bootstrapping processes will be subject of further research.

Here we restricted our embodied research platform ALIS2 to learn mono-syllabic words only. However, a complete integration of the proposed language acquisition architecture as well as a tighter coupling with other perceptual processes are subject of our ongoing research. Whereas word acquisition is currently driven by speech coverage as sole criterion function to be maximized, we're especially interested in more sophisticated task-models as driving force behind emerging language abilities.

References

[1] P. W. Jusczyk, "How infants begin to extract words from speech," *Trends in Cognitive Sciences*, vol. 3, no. 9, pp. 323–328, September 1999.

[2] H. Bortfeld, J. L. Morgan, R. M. Golinkoff, and K. Rathbun, "Mommy and me," *Psychological Science*, vol. 16, no. 4, pp. 298–304, 2005.

[3] H. Brandl, F. Joublin, and C. Goerick, "Towards unsupervised online word clustering," in *Proc. ICASSP*. IEEE, 2008, pp. 5073–76.

[4] T. Gambell and C. Yang, "Mechanisms and constraints in word segmentation," June 2005, yale University.

[5] R. N. Aslin, J. R. Saffran, and E. L. Newport, "Computation of conditional probability statistics by 8-month-old infants," *Psychological Science*, vol. 9, no. 4, pp. 321–324, July 1998.

[6] S. Goldwater and M. Johnson, "Representational bias in unsupervised learning of syllable structure," in *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL)*. Ann Arbor: Association for Computational Linguistics, June 2005, pp. 112–119.

[7] I. Bazzi and J. Glass, "Modeling out-of-vocabulary words for robust speech recognition," in *Proc. ICSLP*, Beijing, 2000.

[8] W. Walker, P. Lamere, and P. Kwok, "Sphinx-4: A flexible open source framework for speech recognition," 2004.

[9] S. O. Kamppari and T. J. Hazen, "Word and phone level acoustic confidence scoring," in *Proc. ICASSP*, vol. 3, Istanbul, 2000, pp. 1799–1802.

[10] N. Iwahashi, "Robots that learn language: Developmental approach to human-machine conversations," in *Symbol Grounding and Beyond - EELC*, P. Vogt, Y. Sugita, E. Tuci, and C. Nehaniv, Eds., 2006, pp. 143–167.

[11] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[12] S. M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *Acoustics, Speech and Audio Processing, IEEE Transactions on*, vol. 3, pp. 400–401, 1987.

[13] C. D. Marcken, "Acquiring a lexicon from unsegmented speech," in *Meeting of the Association for Computational Linguistics*, 1995, pp. 311–313.

[14] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions SAP*, vol. 2, pp. 291–298, 1994.

[15] C. Goerick, B. Bolder, H. J. en, and M. Gienger, "Towards incremental hierarchical behavior generation for humanoids," *IEEE-RAS International Conference on Humanoids*, 2007.

[16] I. Mikhailova, M. Heracles, B. Bolder, H. Janssen, H. Brandl, J. Schmüdderich, and C. Goerick, "Coupling of mental concepts to a reactive system: incremental approach in system design," in *Submitted to the Eighth International Conference on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, 2008.