Towards Multimodal Perception and Semantic Understanding in a Developmental Model of Speech Acquisition

Anja Kristina Philippsen and Britta Wrede
Applied Informatics Group, Cognitive Interaction Technology Center (CITEC), Bielefeld University
{aphilipp, bwrede}@techfak.uni-bielefeld.de

***Abstract*—Babbling is a crucial process in young infants for acquiring articulatory control. By constantly trying out motor commands and observing the consequences that they cause in the environment, they develop an understanding of their own body and learn to control their articulators in order to produce meaningful speech. In earlier works, we proposed a developmental model of speech acquisition that learns to control a 3-d vocal tract simulation for producing vowel or syllable sounds. The system self-organizes learning according to speech it perceives from its environment, so that ambient speech shapes the learning process. Here, we discuss how the proposed model could be extended to form a bridge between perception, on the one end, and semantics, on the other end. The idea is to connect acoustic perception with other perceptual modalities. As an example, we discuss how the system could integrate visual input in its learning loop. By learning associations between acoustic targets and simultaneous visual perceptions, such an enhanced model could produce speech not only in reaction to acoustic input, but also triggered by visual input. Vision, thus, could help to establish a common ground between learner and tutor for interactive articulatory learning.**

I. A DEVELOPMENTAL MODEL OF SPEECH ACQUISITION

Infants explore in a goal-directed manner from the very beginning. Studies show that even neonates orient themselves towards more interesting targets [1] and prenatal exposure to their native language seems to influence infants' early babbling behavior (e.g. [2]).

We combined these ideas in a developmental model of speech acquisition in which we model the influence of ambient speech on the learning process [3]. We provide a set of speech sounds to the system from which it extracts the important components from a high-dimensional acoustic space representation via linear discriminant analysis. The resulting 2-d *goal space* forms a low-dimensional representation of speech that the system is exposed to in its environment. Through this dimension reduction, full syllables are projected onto a single point in goal space. Fig. 1 depicts such a goal space trained from ambient speech sound sets consisting of the three syllables /a/, /ba/ and /ma/.

This research has been supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG), and is related to the European Project CODEFROR (FP7-PIRSES-2013-612555)

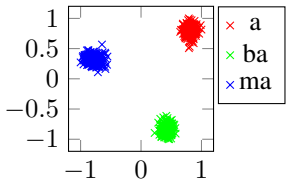

Fig. 1. A goal space for the three syllables /a/, /ba/ and /ma/.

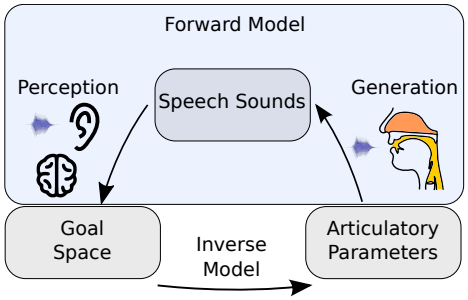

Fig. 2. Forward and inverse model map between articulatory parameters and points in the low-dimensional goal space

Using this goal space, we employ goal babbling, a developmental approach to learning motor coordination [4], [5]. Goal babbling learns by drawing targets from the goal space and trying to achieve them. Collected action–outcome pairs are used to learn an inverse model, mapping from target positions in goal space to motor commands (see Fig. 2). Whereas the inverse model is updated with a radial basis function neural network in an online fashion, the forward model is available to the system before babbling starts. Executing the forward model includes two steps. In the generation step, the articulatory parameters are fed into the speech synthesizer and a speech sound is produced. In the perception step, this speech sound is "perceived", i.e. it is mapped to the previously trained goal space. This perception step models that learning is shaped by ambient language exposure.

II. BABBLING A SET OF SYLLABLES

Motor commands are either static vocal tract shapes (for vowels), or articulatory trajectories modelled with dynamic

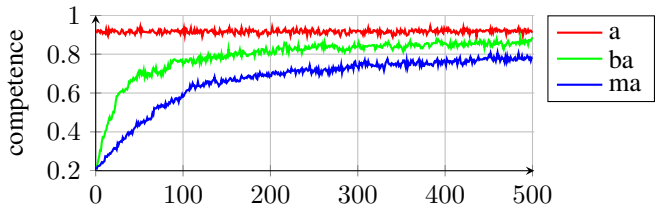

Fig. 3. Competence increase during babbling /a/, /ba/ and /ma/ for 500 iterations (using GBFB features).

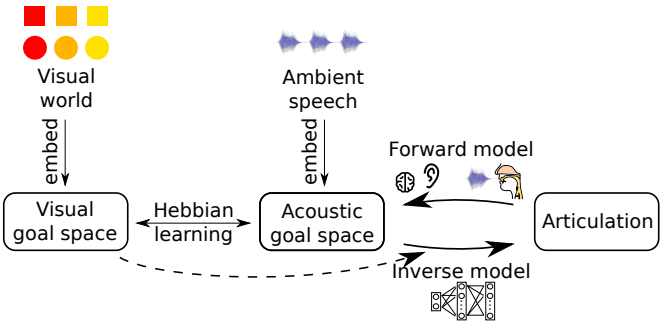

Fig. 4. Associations between visual and acoustic perceptions learned via Hebbian learning could extend the model towards multimodel perception and, eventually, towards a better representation of semantics.

movement primitives (for syllables). Because the dimension reduction step extracts a low-dimensional representation, arbitrary high-dimensional acoustic features can be applied, e.g. MFCCs or Gabor Filterbank (GBFB) Features [6].

In the beginning, the inverse model only knows a single speech sound (in the example in Fig. 3 this is /a/). Then babbling proceeds: the system draws targets randomly from the distribution of ambient speech sounds, estimates the required motor command by executing the inverse model, adds exploratory noise and observes the actual outcomes by executing the forward model. Then, the inverse model gets updated with the new action–outcome pairs (see [3]).

In this way, the system gradually increases its competence. We measure the system's competence by testing how well it can produce the three target sounds that are present in ambient speech. Competence is computed as the exponential of the negative distance between the desired and the actually achieved point in goal space.

Fig. 3 shows how the competence increases during 500 babbling iterations for learning /a/, /ba/ and /ma/.

III. Linking Acoustic and Visual Goal Space

After learning, our model is capable to produce speech sounds in two ways, namely via imitation or via self-production. *Imitation* means that the system perceives an (external) speech sound, maps it to its goal space and tries to achieve this point in goal space by executing its inverse and forward model in a row. *Self-production* means that the system selects its own target in goal space. Computationally, this target could be drawn randomly from the distribution of ambient speech. Such spontaneous productions might occur in young infants from time to time, however, human speech production is rarely purely random.

The reason is that we learn how to speak in order to interact with and learn about our environment. Speech sounds are associated with visual impressions, tactile feedback, emotional states and various other sensations. Semantic emerges from an association of speech with multimodal experiences [7].

How could multimodal perception be integrated into our speech acquisition model? Analogously to how the acoustic goal space forms a low-dimensional representation of ambient speech, a dimension reduction on the visual perceptions could generate a *visual goal space*. As visual and acoustic perception occur in parallel, points in both goal spaces are activated simultaneously, so Hebbian learning, which strengthens connections between co-occurring events, could be applied to learn associations between these two modalities (similar to applications in e.g. [8], [9]).

In consequence, the resulting system could produce speech sounds not only imitatively or randomly, but also in reaction to a visual percept. Either the visual percept activates associated acoustics which then can serve as an imitation target. Or the visual percept is directly inputted into the inverse model.

Furthermore, such multimodal information could alleviate the correspondence problem, i.e. the problem of mapping between speech sounds produced by different individuals. Currently, this problem is avoided by generating ambient speech via the same vocal tract model that the learner uses for babbling. Goal spaces in additional modalities such as vision are not affected by speaker variations and could help to build a common ground between learner and tutor. Multimodal associations, thus, could extend our system towards the integration of semantic information, eventually connecting articulatory learning to interactive language learning.

References

[1] C. Von Hofsten, "An action perspective on motor development," *Trends in cognitive sciences*, vol. 8, no. 6, pp. 266–272, 2004.

[2] A. J. DeCasper and M. J. Spence, "Prenatal maternal speech influences newborns' perception of speech sounds," *Infant behavior and Development*, vol. 9, no. 2, pp. 133–150, 1986.

[3] A. K. Philippsen, R. F. Reinhart, and B. Wrede, "Goal babbling of acoustic-articulatory models with adaptive exploration noise," in *IEEE Intern. Conf. on Development and Learning*, 2016.

[4] M. Rolf, J. J. Steil, and M. Gienger, "Goal babbling permits direct learning of inverse kinematics," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 3, pp. 216–229, 2010.

[5] A. Baranes and P.-Y. Oudeyer, "Active learning of inverse models with intrinsically motivated goal exploration in robots," *Robotics and Autonomous Systems*, vol. 61, no. 1, pp. 49–73, 2013.

[6] M. R. Schädler, B. T. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 131, no. 5, pp. 4134–4151, 2012.

[7] L. Fernandino, J. R. Binder, R. H. Desai, S. L. Pendl, C. J. Humphries, W. L. Gross, L. L. Conant, and M. S. Seidenberg, "Concept representation reflects multimodal abstraction: A framework for embodied semantics," *Cerebral Cortex*, vol. 26, no. 5, pp. 2018–2034, 2015.

[8] V. R. De Sa and D. H. Ballard, "Category learning through multimodality sensing," *Neural Computation*, vol. 10, no. 5, pp. 1097–1117, 1998.

[9] B. Golosio, A. Cangelosi, O. Gamotina, and G. L. Masala, "A cognitive neural architecture able to learn and communicate through natural language," *PloS one*, vol. 10, no. 11, p. e0140866, 2015.