Accepted for ICANN2010 (June 20, 2010)

Integrative Learning between Language and Action: A Neuro-Robotics Experiment

Hiroaki Arie[1], Tetsuro Endo[3], Sungmoon Jeong[2], Minho Lee[2], Shigeki Sugano[3], and Jun Tani[1]

[1] Brain Science Institute, RIKEN
[2] School of Electrical Engineering and Computer Science, Kyungpook National University
[3] Department of Modern Mechanical Engineering, Waseda University

Abstract. The current paper introduces a model for associative learning between linguistic modality and behavior modality. The model constitutes of language and behavior modules both of which are implemented with a hierarchical dynamic network model and they interact densely through hub-like neurons, the parametric biases (PB). By implementing this model to a humanoid robot in the task of multiple objects manipulation, the robot was tutored to associate sentences of two different grammatical types to corresponding sensory-motor schemata. One type is a verb followed by an objective noun as like "hold red" or "hit blue" and the other is a verb followed by an objective noun and further followed by an adverb phrase as like "Put red on blue". Our analysis on the results of a learning experiment showed that two clusters corresponding to these two types of grammatical sentences appear in the PB activity space where a specific micro structure is organized for each cluster.

1 Introduction

The compositionality by meaning that whole can be constituted by reusable parts is one of the essential human cognitive characteristics [1]. In the linguistic processing, diversity of meaning can be generated by combining words by following grammatical rules and semantic constraints. Moreover, diversity in spoken words is originated from compositions through multiple levels from segments to syllable and syllable to lexicons. In the action generation, complex actions can be generated diversely by combining behavior primitives [2]. Both of these compositional systems of language and action have been considered to be organized with specific hierarchy in neuronal anatomy.

In the conventional neuroscience, these two types of compositional processing about language and action have been treated as independent processes. Recently, however, some researches those look at these two functions by utilizing various brain imaging techniques including fMRI, PET and EEG, began to suggest that there are certain dependency between the two. Hauk etal [3] showed in their functional MRI experiment that reading action related words with different end effectors, "Lick", "Pick" and "Kick" evoke neural activities in the motor areas

those overlap with the local areas responsible for generating motor movements in face, arm and leg, respectively. This result as well as [4] suggest that understanding action related words or sentences may require specific motor circuits responsible for generating those actions and therefore brain functions for language and actions might be organized as interdependent.

If everyday experiences of speech and its corresponding sensory-motor signal tend to overlap during infant development, synaptic connectivity between the two circuits can be reinforced by the hebbian learning as discussed by Pulvemuller [5]. This suggests a possibility that meaning and concepts of words and sentences can be acquired as associated with the related sensory-motor experiences, as discussed in the usage-based approach [6] in Cognitive linguistics. Sugita and Tani [7] conducted synthetic neuro-robotics study to examine the idea of the usage-based approach. They proposed a connectionist architecture which consists of a linguistic recurrent neural network (RNN) [8] module and an action (RNN) module which are interacted via associative learning of proto-language and actions of robots. The results of the robot learning experiment showed that the robot can acquire a set of action related concepts by self-organizing certain compositional structure related to verbs and object nouns.

The current paper introduces a trial to extend the aforementioned study [7, 9]. The main motivation is to introduce functional hierarchy both in the linguistic and the behavioral modalities by employing a dynamic neural network model so-called the multiple timescale RNN (MTRNN) developed by our group [10–12]. It is expected that the behavioral modality could have a functional hierarchy where behavior primitives are acquired in the lower level with fast dynamics network and the action compositions do in the higher level with slow dynamics network [10]. Also the linguistic modality could be developed with organizing hierarchy consisting of the alphabetical level, the lexical level and the sentence level by utilizing the time scale difference at each level of the network[9]. The processes for these two modalities could be associated by a similar scheme of the PB binding as described in [7]. In the current task setting, a humanoid robot learns a set of multiple object manipulation behaviors as associated to command sentences. The sentences are comprised of two classes of grammars where one type of sentence is organized as verbs followed by object nouns, as like "Hold Object-A". The other is verbs followed by object nouns and further followed by adverb phrases, as like "Put Object-A on Object-B". These actions are more complex than the ones described in [7] because these require adequate visual attentions on the objects as well as their sequential shifts. The current study will examine what sorts of internal representations can be self-organized in the consequences of the associative learning of these classes of sentences and corresponding actions accompanied with visual attention shifts.

2 Model

2.1 Brain Model

We propose a brain inspired model in which three cognitive functions of speech comprehension, action generation and visual attention switching are integrated via their mutual interactions. Firstly, the action generation pathway is considered. We [13, 11] hypothesized that the inferior parietal lobe (IPL) may play a role of sensory forward model for given action programs from recent neurophysiological evidences [14]. This means that the IPL may predicts coming visuoproprioceptive sensation flow which is associated with action plans provided from the frontal cortex. For example, when the frontal cortex sends an abstract action plan to the IPL for grasping a mug in front of us, the sensory forward model predicts how our arm and hand postures would change in time, how our hands reaching to the mug would be visually perceived and how tactile stimulus of touching the mug would arise.

Skilled behaviors of acting toward objects as like manual object manipulations require adequate timing of visual attention shifts to the target objects. Here, we consider a functional hierarchy where an abstract plan of visual attention shift is generated in the frontal eye field (FEF) [15] and the exact eye saccadic movement to achieve the attention shift is generated in the intraparietal sulcus (IPS) [16] in the downstream. The current model assumes that the FEF predicts sequences of shifts of visual attention to particular objects for given action programs and the IPS generates eye movements to attended objects by following our prior model [17].

Although it has been considered that speech comprehension is performed in the Wernicke's area in the temporal cortex, recent evidences [18] have shown that the Broca's area in the inferior frontal gyrus, which is considered to be responsible for speech generation, actively participates in the process. Also, Tettamanti etal [4] showed that listening action related sentences evoke activation spreading from the Broca's area to specific premotor and motor cortex regions which is considered to be topographically responsible for generating the corresponding motor activities expressed in the sentences. Here, we could draw a hypothesis from these evidences that listening action related sentences could evoke corresponding activation in the Broca's area which can lead to regeneration of neuronal activities in two ways simultaneously. One is activation in the premotor and motor cortices which generates the corresponding motor imagery and the other is that in the Wernicke's area to generate the auditory imagery.

The core part of our hypothesis for the speech comprehension is that streams of auditory signal might be recognized by inferring the corresponding neural activation patterns in the Broca's area which can regenerate them via forward model assumed in the Broca's area. Furthermore, it is speculated that the forward model is constituted hierarchically by the Broca's area responsible for sentence level, the MTG for lexical level and the STG for phonetic level.

Finally, our basic idea of integrating three neural processing systems of the speech comprehension, the visual attention and the action generation is

overviewed. Because all of these neural processing systems seem to constitute functional hierarchy by connecting different local networks, we propose to model each of them by MTRNN. Then these three neural processes are integrated with the Broca's area as a "hub" of connecting these three neural pathways as shown in Fig. 1. For given speech inputs in STG as targets, the speech comprehension system consisting of the Broca's area, the MTG and the STG attempts to reconstruct them in its forward computation (depicted by a blue arrow) by inferring adequate activation patterns in the Broca's area (depicted by a red dotted arrow). Then, the obtained activation patterns in the Broca's area initiate forward computation in two pathways of the visual attention and action generation. In the visual attention system, the FEF generates predictive sequences of attention shifts from one object to another in the workspace and the IPS generates the corresponding eye saccadic motion while the premotor generates predictive sequences of shifts of behavior primitives and the IPL generates detailed prediction of visuo-proprioceptive flow. The prediction of the posture change in time is utilized to compute necessary motor commands in motor cortex to achieve the change.

2.2 Computational Model

Overview As described in previous section, the architecture is based on our prior proposed model of MTRNN [10]. In the current study, as shown in Fig.1, the whole network consists of behavior module network in the right-hand side, the liguistic module network in the left-hand side and the binding network which may correspond to Broca's area in human brains in the upper part. The binding network contains parametric biases (PB) neurons[7]. The idea is that specific static vector values of the PB maps to generation of a linguistic temporal pattern in the linguistic network and the corresponding behavioral temporal pattern in the behavior network as a generative model.

The behavior network learns to generate two types of sequence patterns, one for proprioception in terms of arm posture p_t and the other for so-called the visual attention command a_t. The behavior network outputs the attention command with specific color category to the visual attention module. Then, the visual attention module searches for an object of the specified color in the retina image (see the detail implementation in [17]). Then, the camera head of the robot moves to target the attended object by means of a hand-coded program. The current angle position of the camera head v_t is fed into the input of the behavior network. In summary, the behavior network predicts which color of the object to be attended next and it receives the relative position of the attended object in terms of the camera head angle positions. At the same time the network predicts how the arm posture changes in time with receiving sensation of the relative position of the currently attended object.

The linguistic network learns to generate alphabetic sequences l_t for command sentences. It can generate sequences autonomously with a closed-loop operation in which the input of the current alphabet l_t is fed from its prediction in the previous step instead of the one given externally. The behavior network

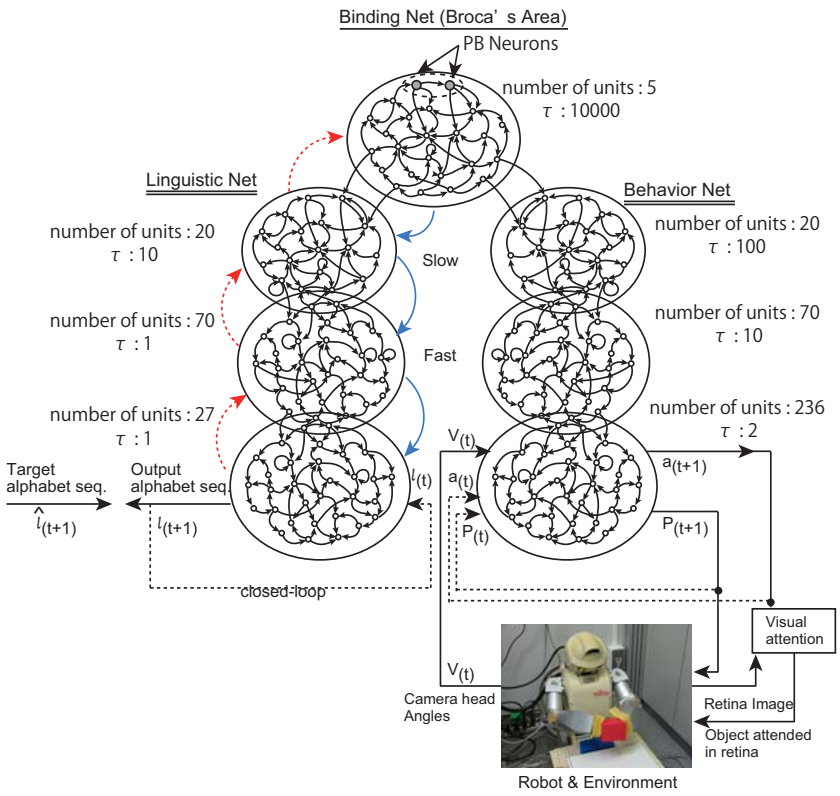

Fig. 1. MTRNN Model with Robot Platform.

can be operated only with the open-loop with receiving the external input v_t of representing relational position of the attended object.

The current model is adopted to a robot task which proceeds with the linguistic phase followed by the behavior phase. In the linguistic phase, the robot receives an alphabetic command sequence without moving and infer the PB values as will be detailed later. Then the robot starts to move with the acquired PB values in the following behavioral phase. The associative training is conducted for each pair of alphabetic sequence (command sentence) and behavior sequence. The behavior training sequence corresponding to each command sentence is generated by guiding the arm posture associated with the visual attention command at each time step. A set of command sentences and their corresponding behavior sequences can be associated by determining specific PB value for each pair. The delta errors generated in both module networks during the association learning were propagated through the both networks to the PB units in the binding network. The PB values responsible for each associative pair in the training

sequences are updated by utilizing this delta error while the optimal synaptic weights for minimizing the learning errors for all the training pairs are searched.

After the learning for all pairs is converged with minimizing the training error, the robot is tested as folllows. A command sentence in terms of alphabetic sequence is shown to the linguistic network as the target to be recognized. This can be done by reconstructing the target sequence by inferring an optimal PB value by back-propagating the error between the target sequence and the regenerated one. Once the PB value is determined, the robot is operated by the behavior network with setting the obtained PB value into the neural units in the binding network.

Mathematical Detail The current model consists of seven groups of neural units as shown in Fig.1, namely linguistic input-output units (IO_l), behavioral input-output units (IO_b), linguistic fast context units (CF_l), behavioral fast context units (CF_b), linguistic slow context units (CS_l), behavioral slow context units (CS_b) and binding units (PB).

The activation value of the i-th neural unit at time step t is calculated as follows.

$$y_{t,i} = \begin{cases} \frac{exp(u_{t,i})}{\sum_{j \in IO} exp(u_{t,j})}, & (i \in IO) \\ \frac{1}{1+exp(-u_{t,i})}, & (i \notin IO) \end{cases} \tag{1}$$

$$u_{t,i} = \begin{cases} 0, & (t = 0 \wedge i \notin PB) \\ PB_i, & (t = 0 \wedge i \in PB) \\ (1 - \frac{1}{\tau_i})u_{t-1,i} + \frac{1}{\tau_i} \sum_{j \in I_{all}} w_{ij} x_{t,j}, & (otherwise) \end{cases} \tag{2}$$

$$x_{t,j} = y_{t-1,j} \tag{3}$$

$u_{t,i}$: internal state of the i-th unit at time step t
PB_i : neural activation of binding units (PB value)
τ_i : time constant of i-th unit
w_{ij} : connection weight from j-th unit to i-th unit
$x_{j,t}$: input from j-th unit at time step t

Number of neural units and time constant are shown in Fig.1.The time constant of the binding network is set with large value so that the neural activation of binding units can be considerd as static vector values like PB.

Connection weights and PB value are udjusted using the Back Propagation Through Time(BPTT) algorithm as follows.

$$w_{ij}^{(n+1)} = w_{ij}^n - \eta \frac{\partial E}{\partial w_{ij}} = w_{ij}^n - \frac{\eta}{\tau_i} \sum_t x_{t,j} \frac{\partial E}{\partial u_{t,i}} \tag{4}$$

$$PB_i^{(n+1)} = PB_i^n - \alpha \frac{\partial E}{\partial PB_i} = PB_i^n - \alpha \frac{\partial E}{\partial u_{0,i}} \tag{5}$$

$$E = \sum_{t} \sum_{i \in IO} y^*_{t,i} log(\frac{y^*_{t,i}}{y_{t,i}}) \tag{6}$$

$$\frac{\partial E}{\partial u_{t,i}} = \begin{cases} y_{t,i} - y^*_{t,i} + (1 - \frac{1}{\tau_i}) \frac{\partial E}{\partial u_{t+1,i}} & (i \in IO) \\ y_{t,i}(1 - y_{t,i}) \sum_{k \in IO} \frac{w_{ki}}{\tau_k} \frac{\partial E}{\partial u_{t+1,k}} + (1 - \frac{1}{\tau_i}) \frac{\partial E}{\partial u_{t+1,i}} & (otherwise) \end{cases} \tag{7}$$

n : iteration number in updating process

E : prediction error

$y^*_{t,i}$: value of current training sequence for i-th neural unit at time step t

η, α : $learningrate$

The PB values are determined for each training sequence independently while the connection weights do for all the sequences. To recognize a linguistic sequence after the associative learning, the PB value corresponding to given alphabetic command sequence is searched using BPTT with fixed connection weights.

3 Experiment

3.1 Task Design

A small humanoid robot as shown in Fig.1 was used as an experimental platform. The robot was fixed to a chair, and a table was set in front of the robot. The robot was suppoed to associate a set of alphabetic sequences of two different grammatical types to corresponding behavioral sequence. The first type (type-1) was a verb followed by an objective noun where the verbs can take two words of "hold" and "up-down" and the objective nouns for "red", "blue" and "green". This generates six sentences. The other type (type-2) was a verb followd by an objective noun and further followd by an adverb phrase. In this type the verb can be just one word of "put" and both objective nouns and adverbs can take three words of "red", "blue" and "green". This generates six sentences. For each action, the robot was tutored three times with changing the initial position of the object as 4cm left from the original position, the original one and 4cm right from original one for the purpose of gaining generalization in object manipulation behaviors. Each action was tutored in every possible combination of object position (left, center and right) and color of the object (red, blue and green). Totally there were 18 behavior sequences for the type-1 and 54 for the type-2.

3.2 Results

After the training, the robot was tested wheather it can recognize all 12 linguistic command sentences and generate corresponding behavior with different object position situations (left, center and right). The recognition was done by the searching optimal PB values. The search calculation was iterated 2000 epochs

with $\alpha = 0.2$ for each linguistic command sentences. The performance was scored in terms of a success rate across all trials. It was considered that a trial was successful if the robot can generate corresponding behavior sequence with the obtained PB values. As the results, it was confirmed that the robot could generate correct behavior with 82% success rate. It was further confirmed that the robot recognize all the 12 sentences because it can generate correct behaviors at the least one specific object position case for each command sentence.

The Fig.2 shows two examples of time development of sequences. Here it can be seen that the profile of fast context containes more complex patterns as compared to those in the slow context in both linguistic and behavior network.

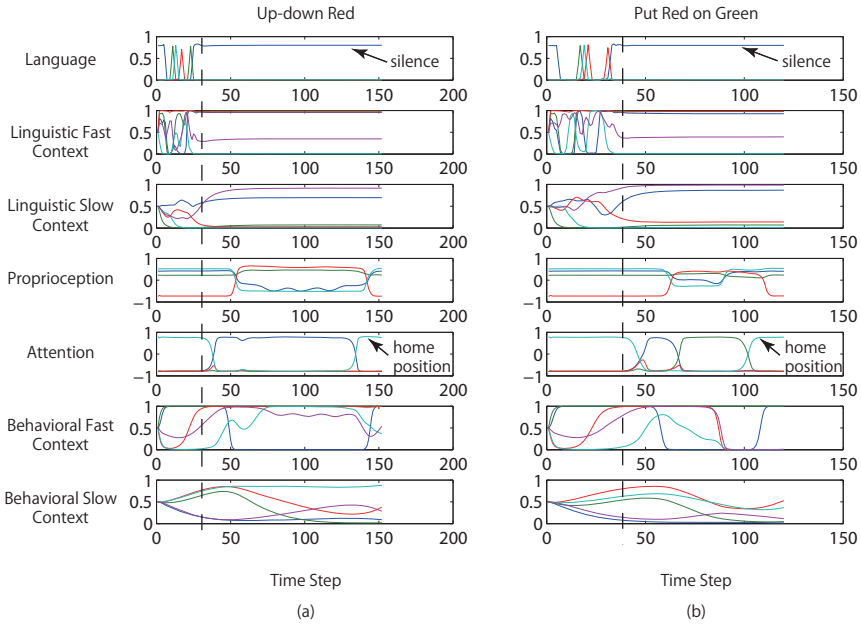

Fig. 2. Examples of sequences generated with obtained PB values. (a) shows "Up Down Red" and (b) shows "Put Red on Green" case. Vertical dashed lines indicate the onset of the behavioral phase. The first row shows alphabetic sequence by representative four characters ("silence", "d", "n" and "o"). The second and the third raw show activations of fast and slow context units in linguistic network. The fourth row shows joint angles of right arm of the robot. The fifth raw shows visual attention command corresponding to "home", "red", "blue" and "green". The sixth and seventh raws show activations of fast and slow context units in behavior network.

4 Analysys

We applied principal component analysys to visualize the structure of the PB space. Fig.4 shows the 1st and 2nd principal components of neural activation of binding units (PB). It can be observed that there are two clusters corresponding to the type-1 sentences and the type-2 sentences. The cluster for type-1 sentences shows that a compositional structure of two verbs multiplied by three objective nouns appears in a two dimensional grid which is similar to the structure observed in our previous study[7]. On the other hand, any systematic structures can not be found for tye type-2 although the mapping from these sentences to the actions was successfully generated but without forming generalized representation.

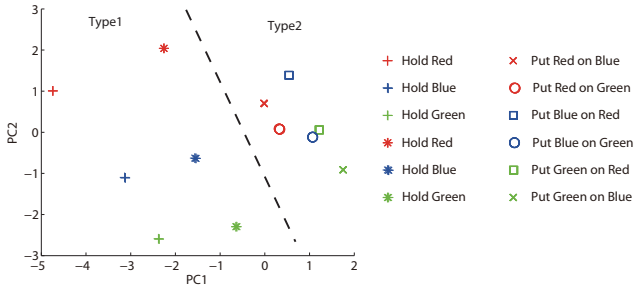

Fig. 3. The PB space. The dimension are reduced from 5 to 2 by PCA. The sentences are clusterd based on their grammatical structure.

5 Conclusion

In this paper, we reported on the integrative learning of linguistic and behavioral sequences by the MTRNN. We trained the model with a set of linguistic and behavioral sequences. As a result of our experiment, we found that the model can aquire the capability to recognize linguistic sentences and generate corresponding behavioral sequence patterns. Our analysys showed that compositional structure can be self-organized for the type-1 sentences but not for the type-2 ones.

Two possibilities might be suggested to account for this result. One possibility is that the PB space cannot embed two distinct compositional structures, i.e. type-1 and type-2 simultaneously. The other possibility is that the number of examples in learning of type-2 structures are too small for achieving the generalization as the type-2 was trained only with one verb case of “put” in the current study. Our future study will examine these accounts and also will pursue scaling of the system in terms of number of words, diversity of grammar types and behavior complexity.

References

1. Evans, G.: Semantic theory and tacit knowledge. In Holzman, S., Leich, C., eds.: Wittgenstein: To Follow a Rule. Routledge and Kegan Paul, London (1981) 118–137
2. Arbib, M.: Perceptual structures and distributed motor control. In: Handbook of Physiology: The Nervous System, II. Motor Control. MIT Press, cambridge, MA (1981) 1448–1480
3. Hauk, O., Johnsrude, I., Pulvermuller, F.: Somatotopic representation of action words in human motor and premotor cortex. Neuron **41**(2) (2004) 301–307
4. Tettamanti, M., Buccino, G., Saccuman, M.C., Gallese, V., Danna, M., Scifo, P., Fazio, F., Rizzolatti, G., Cappa, S.F., Perani, D.: Listening to action-related sentences activates fronto-parietal motor circuits. Journal of Cognitive Neuroscience **17**(2) (2005) 273–281
5. Pulvermuller, F.: Brain mechanisms linking language and action. Nature Reviews Neuroscience **6** (2005) 576–582
6. Tomasello, M.: Constructing a language: A usage-based theory of language acquisition. Harvard University Press, Cambridge, MA (2003)
7. Sugita, Y., Tani, J.: Learning semantic combinatoriality from the interaction between linguistic and behavioral processes. Adaptive Behavior **13**(1) (2005) 33–52
8. Elman, J.: Finding structure in time. Cognitive Science **14** (1990) 179–211
9. Hinoshita, W., Arie, H., Tani, J., Ogata, T., Okuno, H.G.: Emergence of hierarchical structure mirroring linguistic compositionality in recurrent neural network. Neural Networks (submitted)
10. Yamashita, Y., Tani, J.: Emergence of functional hierarchy in a multiple timescale neural network model: a humanoid robot experiment. PLoS Computational Biology **4**(11) (2008)
11. Nishimoto, R., Namikawa, J., Tani, J.: Learning multiple goal-directed actions through self-organization of a dynamic neural network model: A humanoid robot experiment. Adaptive Behavior **16**(2-3) (2008) 166–181
12. Arie, H., Endo, T., Arakaki, T., Sugano, S., Tani, J.: Creating novel goal-directed actions at criticality: A neuro-robotic experiment. New Mathematics and Natural Computation **5**(1) (2009) 307–334
13. Tani, J., Nishimoto, R., Paine, R.: Achieving "organic compositionality" through self-organization: Reviews on brain-inspired robotics experiments. Neural Networks (2008) 584–603
14. Ehrsson, H., Fagergren, A., Johansson, R., Forssberg, H.: Evidence for the involvement of the posterior parietal cortex in coordination of fingertip forces for grasp stability in manipulation. Journal of Neurophysiology **90** (2003) 2978–2986
15. Bruce, C.J., Goldberg, M.E.: Primate frontal eye fields. i. single neurons discharging before saccades. Journal of Neurophysiology **53**(3) (1985) 603–635
16. Colby, C.L., Goldberg, M.E.: Space and attention in parietal cortex. Annual Review of Neuroscience **22** (1999) 319–349
17. Jeong, S., Lee, M., Arie, H., Tani, J.: Developmental learning of integrating visual attention shifts and bimanual behavior in object manipulation tasks. In: Proc. of 2010 IEEE 9th International Conference on Development and Learning. (Submitted)
18. Wilson, S., Saygin, A., Sereno, M., Iacoboni, M.: Listening to speech activates motor areas involved in speech production. Nature Neuroscience **7**(7) (2004) 701–702