The Neurophonetic Model of Speech Processing ACT: Structure, Knowledge Acquisition, and Function Modes

Bernd J. Kröger[1], Jim Kannampuzha[1], Cornelia Eckers[1],
Stefan Heim[2,3,4], Emily Kaufmann[5], and Christiane Neuschaefer-Rube[1]

[1] Department of Phoniatrics, Pedaudiology, and Communication Disorders,
University Hospital Aachen and RWTH Aachen University, Germany
{bkroeger,jkannampuzha,ceckers,cneuschaefer}@ukaachen.de
[2] Section Functional Brain Mapping, Department of Psychiatry,
Psychotherapy, and Psychosomatics
[3] Section Neurological Cognition Research, Department of Neurology,
University Hospital Aachen and RWTH Aachen University, Germany, and
[4] Institute of Neuroscience and Medicine (INM-1), Research Centre Jülich, Germany
sheim@ukaachen.de
[5] Education and Rehabilitation of the Deaf and Hard of Hearing, University of Cologne,
Germany
emily.kaufmann@uni-koeln.de

Abstract. Speech production and speech perception are important human capabilities comprising cognitive as well as sensorimotor functions. This paper summarizes our work developing a neurophonetic model for speech processing, called ACT, which was carried out over the last seven years. The function modes of the model are production, perception, and acquisition. The name of our model reflects the fact that vocal tract ACTions, which constitute motor plans of speech items, are the central units in this model. Specifically (i) the *structure* of the model, (ii) the acquired *knowledge*, and (iii) the correspondence between the model's structure and specific *brain regions* are discussed.

Keywords: neurophonetic model, speech processing, speech production, speech perception, speech acquisition, vocal tract actions, motor plan.

1 Introduction

Speech production and speech perception are important human capabilities comprising cognitive as well as sensorimotor functions. Realistic modeling of speech processing is an important part of understanding multimodal face-to-face interaction and thus of understanding important parts of social interactions. The neurophonetic model of speech processing presented in this paper comprises three function modes: speech production, speech perception, and speech acquisition. On the one hand, the model is based on a specific *neuroanatomical structure* for motor, sensory, and phonemic representations of speech [1, 2]. On the other hand, the model acquires *linguistic knowledge* as well as *speech skills* for a specific language. This knowledge and skills become integrated into the model based on synaptic weights (i.e. the degree

A. Esposito et al. (Eds.): Cognitive Behavioural Systems, LNCS 7403, pp. 398–404, 2012.

of excitatory and/or inhibitory synaptic connections) between neurons of different neural maps [2]. Purely cognitive linguistic approaches are beyond the scope of this paper but a blueprint for integrating our model into a complete approach to speech processing including lexical representations is outlined in [3].

Basically, our neurophonetic model, in which *vocal tract ACTions* are assumed to constitute the basic units of motor plans (leading to the name ACT), is inspired by the organization of the previously only quantitative sensorimotor speech production model, i.e. the DIVA model [4, 5, 6]. Both approaches comprise sensorimotor feed-forward and feedback loops (Fig. 1). Starting from a phonemic representation, both approaches (DIVA and ACT) are capable of generating proper articulator movement patterns and subsequently proper acoustic speech signals. From the viewpoint of speech perception, it has been demonstrated that ACT is capable of modeling *categorical perception* [2, 7]; in this paper, we report on the model's ability to assign categorical perception to the topology of the *phonetic map*. A phonetic map is assumed to constitute the central supramodal neural map within a (language specific) speaking skills repository (i.e. vocal tract action repository [7]), and it associates motor, sensory, and phonemic states of speech items (Fig. 1).

2 The Structure of the Model and Its Function Modes of Production and Perception

The structure of our neurophonetic model is given in Fig. 1. *Speech production* starts with the phonemic representation of a speech item. This speech pattern (e.g. a word or a short utterance) is processed syllable by syllable. In the case of a *frequent syllable*, for which the motor plan has already been acquired, first the motor plan state is activated via the phonetic map, and subsequently the motor neuron activation pattern (level of the primary motor map) is generated for each vocal tract action occurring within the syllable. The subsequent neuromuscular processing leads to articulator movements and allows the generation of the acoustic speech signal via our articulatory-acoustic model [8]. The previously-acquired sensory state of this syllable is co-activated in parallel via the phonetic map (internal or trained state TS; Fig. 1). This state TS is matched with the state ES (external state ES; Fig. 1), resulting from the current production of that syllable. In the case of noticeable differences, an auditory and somatosensory error signal (Δau and Δss; Fig 1) is propagated via the phonetic map in order to alter the motor plan of that syllable for a new (corrected) production of that syllable. In the case of *infrequent syllables*, a motor plan is generated via the motor planning module by activating the plan of a phonetically and phonotactically similar syllable via the phonetic map [9].

Two control mechanisms are featured in our model. Firstly, *lower level compensatory corrections* occur in real time by correcting articulator position and velocity of articulators with respect to action goals as defined on the motor plan level (module of subcortical/cortical motor programming, execution and control; Fig. 1). On this level, compensation results from previously-acquired knowledge about possible vocal tract action realizations, especially if more than one articulator is involved (e.g. lower jaw and lips in the case of labial closing). This type of compensation has been exemplified in bite block experiments [10] and in experiments introducing unexpected jaw perturbations [11].

Secondly, *sensorimotor adaptation* can be modeled in our approach by comparing internal (or trained) and external sensory states TS and ES (cortical level; Fig. 1). Basically, these states do not show noticeable differences (also called “error signals” Δau and Δss; Fig 1) after speech acquisition. But error signals can occur, for example if the lower-level perceptual processing system is modified artificially (e.g. by permanently shifting the second formant F2 via a specific real-time signal processing procedure [12]). The resulting adaptation effects have been explained in detail in the DIVA approach [4].

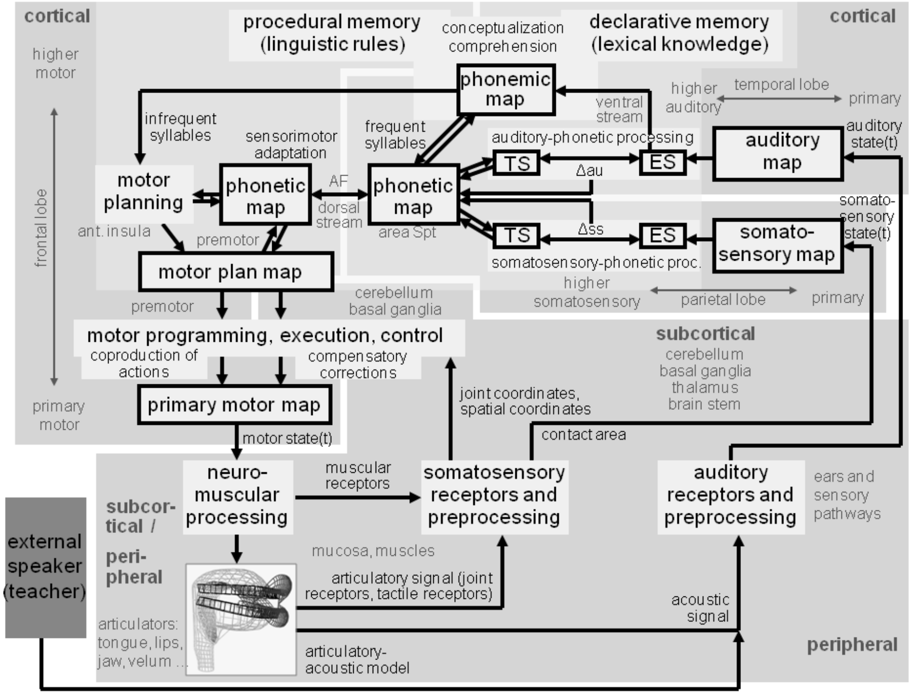

Fig. 1. Structure of the neurophonetic model ACT. Non-framed boxes indicate processing modules, framed boxes indicate neural maps. Single arrows indicate topology-preserving connections or streams (one-to-one mappings; parallel neural connections only); double arrows indicate complex neural mappings (all-to-all mappings; crossing neural connections). Dark grey: external speaker (mainly teacher or caretaker during speech acquisition); medium grey and light grey: the neural model; the medium grey region comprise modules and maps with temporal processing in short time intervals (12.5 msec in the current implementation of the model); the light grey region comprises modules and maps which process syllables as a whole unit (state maps within this region are part of short-term memory; mapping from state maps onto the phonetic map, as well as the phonetic map itself, are part of long-term memory). TS: trained or internal (sensory) states; ES: external (sensory) states; Δau: auditory error signal; Δss: somatosensory error signal; area Spt: area in the Silvian fissure at the parieto-temporal boundary [13]; AF: arcuate fasciculus.

Speech perception starts with an external acoustic signal. If phonemic identification is intended, the signal must be a realization of a frequent syllable. For this purpose the signal is preprocessed at peripheral and subcortical levels and loaded to the short-term memory as an external auditory state (ES; Fig 1). Then its neural activation pattern is passed to the trained state map (TS; Fig 1), firstly leading to the co-activation of a neuron region on the level of the phonetic map and secondly to the co-activation of a specific neuron on the level of the phonemic map; the first representing that syllable phonetically, the second phonologically. This neural pathway via the phonetic map, also referred to as the *dorsal stream* of speech perception [13], also co-activates a motor plan pattern for this frequent syllable. A second stream in speech perception, i.e. the *ventral stream*, directly links the auditory activation pattern with the phonological processing module [ibid.]. The dorsal stream is assumed to be important during speech acquisition, while the ventral stream is dominant later on during adult speech perception. The ventral stream is indicated in Fig. 1, but it has not yet been integrated into ACT.

3 Training Experiments for Acquiring Speech Knowledge and Speaking Skills

Our training experiments always comprise a *babbling phase* and an *imitation phase* [2] (see also DIVA model [4]). During babbling training, the model associates motor plan states with auditory states. On this basis, the model is capable of generating motor plan states during imitation training. We started with experiments on a phonotactically simple "*model language*" comprising V- and CV-syllables, with five vowels (V = /i/, /e/, /u/, /o/, /u/) and three consonants (C = /b/, /d/, /g/). All combinations of vowels and consonants occurred within the syllables with equal frequency [2]. We were able to show that on the level of the phonetic map, a strict phonetic ordering of realizations (exemplars) of these syllables occurred during babbling training (*phonetotopy* [14]). During imitation training, *phoneme regions* appeared on the level of the phonetic map [2, 7]. After these initial experiments we proceeded with a more complex model language which comprised V-, CV-, and CVV-syllables and which is based on a larger set of consonants (/b/, /d/, /g/, /p/, /t/, /k/, /m/, /n/, /l/). The training once again resulted in a strict ordering of the phonetic map with respect to phonetic features (e.g. place and manner of consonant articulation) and phonotactic features such as syllable type (C, CV, or CCV) and types of consonants within the CC-cluster [9].

In the present experiments, we trained sets of the *200 most frequent syllables of Standard German* [15]. One of the most interesting results is that phonetic and phonotactic ordering in a real language is less strict than in the "model languages" we trained. This may be due to the fact that a real language is not constructed in a strictly regular way with respect to phonotactics, meaning not all CV or CCV combinations occur in a real language, or at least they do not occur with equal frequency. Furthermore, syllable exemplar regions are of different size in the case of a real language. Here, size is corresponding to the frequency of occurrence of a syllable in that particular language (Fig. 2). Thus in the case of this experiment up to 10 model neurons represent different realizations of a syllable.

Production and perception quality of the model was checked for the 50 most frequent syllables. Perception quality was quantified by the percentage of syllables, produced by a natural speaker and correctly identified by the model. This rate was 92% in the case that test items and training items were produced by the same speaker. The identification rate dropped to 84% if syllables were produced by a different male speaker. In the case of the same speaker, test and training items were different, i.e. chosen from different subsets of syllable realizations. Production quality was quantified by the percentage of syllables, correctly identified by natural subjects (listeners). The 50 most frequent syllables were produced by the model. Perception tests were performed by 5 persons between 25 and 28 years old with no known speech perception deficits. The mean rate of correct identifications was 96%.

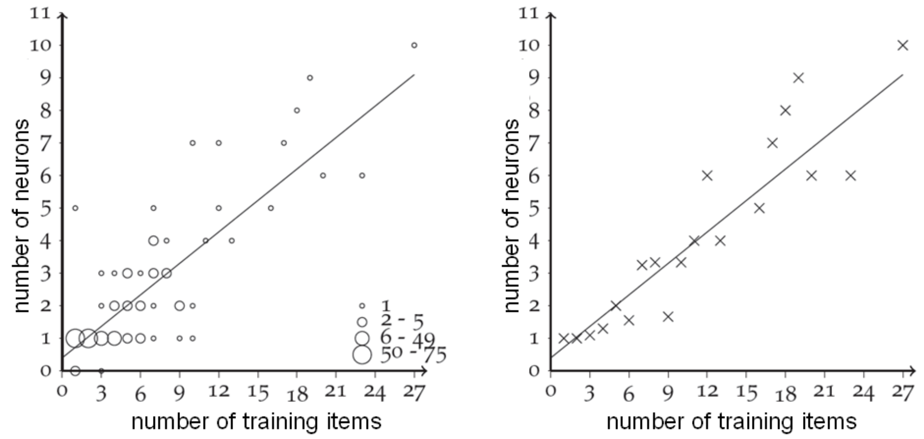

Fig. 2. Number of neurons representing a certain syllable as functions of number of training items which are used for training of that certain syllable. Left side: the size of the circle represents the number of different syllables exhibiting the same number of training items and leading to the same number of neurons representing that specific syllable; Right side: arithmetic mean of neurons representing all syllables with an identical number of training items. The training was performed for 200 most frequent syllables of Standard German uttered in sentence context uttered by a 33 year old male speaker without any known speaking or hearing abnormalities [15].

4 Neuroanatomical Correlates

In order to specify the neuroanatomical correlated neural maps, mappings and pathways needed to be differentiated. *Neural maps* comprise all state maps (i.e. phonemic, auditory, somatosensory, motor plan, primary motor maps; Fig. 1) as well as the self-organizing map (i.e. the phonetic map; see Fig. 1). *Neural mappings* occur between the self-organizing map and all state maps (double arrows in Fig. 1). *Neural pathways* occur between maps or between a map and a processing module (single arrows in Fig. 1). In contrast to neural mappings neural pathways are not capable of *generating new* neural activation pattern, but rather *forwards* an already existing neural activation pattern from one map to another. Thus, neural streams are represented as bundles of *parallel neural fibers* and are capable of connecting non-adjacent brain regions (e.g. arcuate fasciculus [16]), while neural mappings are realized by complex all-to-all cross-connection networks. Mappings mainly connect neural maps which are in close range of each other.

The short-term memory state maps as defined in our model (state maps within the light grey area in Fig. 1) comprise a temporal storage over the time interval of at least one syllable [2, 3]. These higher-level state maps are assumed to be located near the brain regions of the sensory error maps postulated in [4, 5]. In contrast, the lower-level state maps (state maps within the medium grey area in Fig. 1), process the stream of motor data for articulation and the stream of sensory data, already preprocessed by peripheral modules. These state maps are located in primary sensory and motor areas (Fig. 2).

A further comparison of the structure of our model to the structure of the speech processing model proposed by Hickok and Poeppel [13] leads to the conclusion that the supramodal phonetic map cannot be located in *one* particular brain region. Moreover, it is assumed that the phonetic map is copied from the area in the Silvian fissure at the parieto-temporal boundary (labeled as area Spt [13]) onto the premotor area by a neural stream (i.e. by the arcuate fasciculus) in order to allow close-range neural mappings between the phonetic map and the motor plan state map, on the one hand, and between the phonetic map and the phonemic, auditory, and somatosensory short-term memory state maps on the other hand (Fig. 1). Thus the phonetic map could be interpreted as a mirror neuron system at the phonetic level in contrast to the well-known semantic level mirror neurons [17].

5 Discussion and Further Work

Although both quantitative sensorimotor models of speech production and speech processing in principle are compatible, what makes our neurophonetic model ACT [2, 7, 9, 14, 15] different from the DIVA model [4, 5, 6] is that it allows an alternative view on the neurophonetics of speech production, perception, and acquisition. While the DIVA model mainly focuses on exemplifying sensorimotor adaptation [4, 5], our model focuses on exemplifying the development of a vocal tract action repository (i.e. phonetic map) as the central repository for sensorimotor speech knowledge and speaking skills on the basis of principles of neural self-organization [2, 14]. Brain imaging experiments are planned in order to verify or falsify our hypotheses, especially on the mirror-neuron-character of the phonetic map. Further modeling experiments are planned in order to gain more insight into the development of phonetic knowledge and phonological structure of a specific language within a complete neurolinguistic approach. This approach would also include the development of the mental lexicon along with the development of the vocal tract action repository [3].

Acknowledgments. This work was supported in part by German Research Council (DFG) grants Kr 1439/13-1 and Kr 1439/15-1 and in part by COST-action 2102.

References

1. Kröger, B.J., Kopp, S., Lowit, A.: A model for production, perception, and acquisition of actions in face-to-face communication. Cognitive Processing 11, 187–205 (2010)
2. Kröger, B.J., Kannampuzha, J., Neuschaefer-Rube, C.: Towards a neurocomputational model of speech production and perception. Speech Communication 51, 793–809 (2009)
3. Kröger, B.J., Birkholz, P., Neuschaefer-Rube, C.: Towards an articulation-based developmental robotics approach for word processing in face-to-face communication. PALADYN Journal of Behavioral Robotics (in press)
4. Guenther, F.H.: Cortical interactions underlying the production of speech sounds. Journal of Communication Disorders 39, 350–365 (2006)
5. Guenther, F.H., Ghosh, S.S., Tourville, J.A.: Neural modeling and imaging of the cortical interactions underlying syllable production. Brain and Language 96, 280–301 (2006)
6. Guenther, F.H., Vladusich, T.: A neural theory of speech acquisition and production. Journal of Neurolinguistics (in press)
7. Kröger, B.J., Birkholz, P., Kannampuzha, J., Neuschaefer-Rube, C.: Categorical Perception of Consonants and Vowels: Evidence from a Neurophonetic Model of Speech Production and Perception. In: Esposito, A., Esposito, A.M., Martone, R., Müller, V.C., Scarpetta, G. (eds.) COST 2102 Int. Training School 2010. LNCS, vol. 6456, pp. 354–361. Springer, Heidelberg (2011)
8. Kröger, B.J., Birkholz, P.: A Gesture-Based Concept for Speech Movement Control in Articulatory Speech Synthesis. In: Esposito, A., Faundez-Zanuy, M., Keller, E., Marinaro, M. (eds.) Verbal and Nonverbal Commun. Behaviours. LNCS (LNAI), vol. 4775, pp. 174–189. Springer, Heidelberg (2007)
9. Kröger, B.J., Miller, N., Lowit, A.: Defective neural motor speech mappings as a source for apraxia of speech: Evidence from a quantitative neural model of speech processing. In: Lowit, A., Kent, R. (eds.) Assessment of Motor Speech Disorders, pp. 325–346. Plural Publishing, San Diego (2011)
10. Fowler, C.A., Turvey, M.T.: Immediate compensation in bite-block speech. Phonetica 37, 306–326
11. Kelso, J.S., Tuller, B., Vatikiotis-Bateson, E., Fowler, C.A.: Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for cooperative structures. Journal of Experimental Psychology: Human Perception and Performance 10, 812–832 (1984)
12. Houde, J.F., Jordan, M.I.: Sensorimotor adaptation of speech I : Compensation and adaptation. Journal of Speech, Language, and Hearing Research 45, 295–310 (2002)
13. Hickok, G., Poeppel, D.: Towards a functional neuroanatomy of speech perception. Trends in Cognitive Sciences 4, 131–138 (2007)
14. Kröger, B.J., Kannampuzha, J., Lowit, A., Neuschaefer-Rube, C.: Phonetotopy within a neurocomputational model of speech production and speech acquisition. In: Fuchs, S., Loevenbruck, H., Pape, D., Perrier, P. (eds.) Some Aspects of Speech and the Brain, pp. 59–90. Peter Lang, Berlin (2009)
15. Kröger, B.J., Birkholz, P., Kannampuzha, J., Kaufmann, E., Neuschaefer-Rube, C.: Towards the Acquisition of a Sensorimotor Vocal Tract Action Repository within a Neural Model of Speech Processing. In: Esposito, A., Vinciarelli, A., Vicsi, K., Pelachaud, C., Nijholt, A. (eds.) Communication and Enactment 2010. LNCS, vol. 6800, pp. 287–293. Springer, Heidelberg (2011)
16. Bernal, B., Ardila, A.: The role of the arcuate fasciculus in conduction aphasia. Brain 132, 2309–2316 (2009)
17. Rizzolatti, G., Craighero, L.: The mirror-neuron system. Annual Review of Neuroscience 27, 169–192 (2004)