Performing Identification and Discrimination Experiments for Vowels and Voiced Plosives by Using a Neurocomputational Model of Speech Production and Perception

Bernd J. Kröger, Jim Kannampuzha, Christiane Neuschaefer-Rube

RWTH Aachen University and University Hospital Aachen
E-mail: bkroeger@ukaachen.de, jim.kannampuzha@rwth-aachen.de, cneuschaefer@ukaachen.de

Abstract

A neurocomputational model of speech production and speech perception is introduced. After training, i.e. after mimicking early phases of speech acquisition, the model is capable of producing and perceiving vowels and CV-syllables (C = voiced plosives). Different instances of the model were trained for representing different "virtual subjects" which are then used as listeners in identification and discrimination experiments. First results indicate that a typical feature of speech perception – i.e. categorical perception – occurs in a straight forward way from our neurocomputational production-perception model.

1 Introduction

Our work on this modeling topic started with the development of a neural model of speech production (Kröger et al. 2006a and Kröger et al. 2006b). The work was inspired by the approach introduced by Guenther et al. (2006) and Guenther (2006). A main feature of our model is the separation of the control module into feedforward and feedback control components. First training results indicated that motor and sensory states are ordered with respect to phonetic features like "front-back" and "high-low" in the case of vowels and with respect to phonetic features like "place of articulation" in the case of voiced plosives at the neural level of sensorimotor associations. This phenomenon is called "phonetotopy" (Kröger et al. in press). Since sensory (auditory and somatosensory) feedback is important for getting a phonetotopic ordering of speech items, the feedback loop of the production model was extended in order to integrate perception for speech sounds and syllables. It is described in this paper how the model now can be used in experiments focusing on speech sound identification and discrimination.

2 The Model

The organization of the model is given in Fig. 1. The model comprises *neural maps* for motor planning and motor programming/execution, for auditory and somatosensory representations of speech items (sounds, syllables, or words), for phonetic features, and for linguistic information (i.e. phonemic representations of speech items). The *neural mappings* connecting these maps are trained using pre-linguistic proto-vocalic and proto-consonantal training data (babbling training data) and later on language-specific vocalic and consonantal training data (imitation training data, Kröger et al. in press). Thus the model separates neural *structure* (i.e. maps) and *knowledge* (i.e. synaptic link weight adjustment for the mappings during training). For *speech production* feed-forward and feedback control loops (or control streams) are implemented. For speech perception the dual-stream approach (Hickok and Poeppel 2007) is implemented. The *ventral stream* can be interpreted in terms of passive non-motor assisted perception using a direct link from auditory maps to the mental lexicon while the *dorsal stream* – activating frontal motor areas – uses existing neural networks of the feed-forward speech production part of the model. It is assumed that the dorsal stream of speech perception is mainly active during lower level perception tasks like

recognition of phonetic-phonological sound features, while the ventral stream directly activates lexical items as a whole. For the perception experiments described in this paper, exclusively the dorsal stream is used.

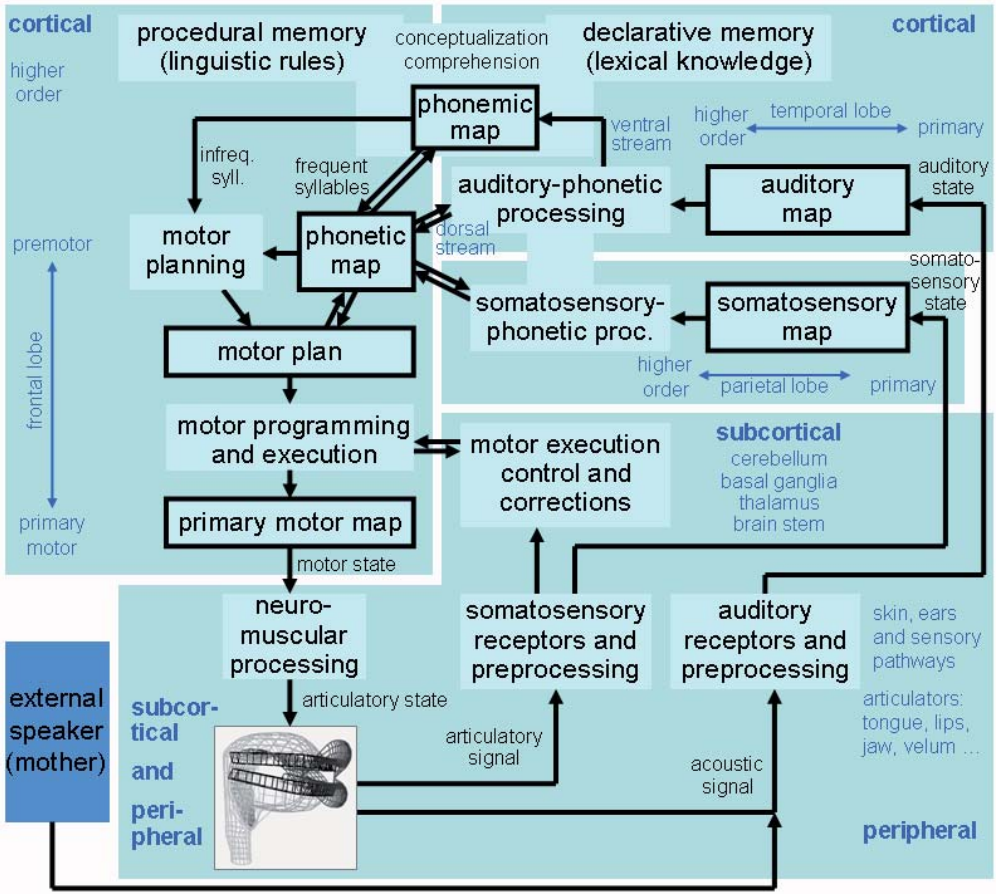

Figure 1: *Neurocomputational model of speech production and speech perception. Boxes with black outline represent neural maps; arrows indicate processing paths or neural mappings. Boxes without black outline indicate processing modules (not specified in detail currently).*

Feed-forward control starts with a linguistic representation of a speech item, activated on the level of the phonemic map. Each frequent syllable activates its prestored sensory states (auditory and somatosensory; somatosensory comprises tactile and proprioceptive states) and its prestored motor plan state (a detailed description of motor plan states is given in Kröger et al. in press). Prestored motor and sensory states are trained or learned during speech acquisition training stages. This results in a self-organization of the phonetic map and its phonetic to sensory and phonetic to motor plan mappings (Kröger et al. in press). The motor plan of an infrequent syllable is generated by the motor planning module. The motor plan of frequent as well as of infrequent speech items is executed via the motor execution module and leads to a time course of position and velocities for all model articulators of the vocal tract model. A description of the vocal tract model is given by Birkholz et al. (2006), Birkholz and Kröger (2006 and 2007), and by Kröger and Birkholz (2007). The vocal tract model is capable of generating articulatory movement patterns (i.e. location, velocity, and shape of each model articulator for each time instant) and the acoustic speech signal. These output signals of the vocal tract model serve as input information for the feedback control component. Following preprocessing, the current sensory (auditory and somatosensory) signals activate neurons within the primary sensory maps. Then a comparison of the current sensory state with the prestored sensory state for the speech item under production is done within the sensory-phonetic processing modules. As a result an error signal is calculated for correcting the current motor execution, if the current sensory state deviates from the stored sensory state for the current speech item.

3 Experiments using the trained model

Identification and discrimination experiments are based on a vocalic or on a consonantal stimulus continuum (e.g. [i]-[e]-[a]-continuum or [ba]-[da]-[ga]-continuum). Each stimulus continuum comprises 13 stimuli in the case of our experiments. Classic identification and discrimination experiments are performed by human subjects. In the case of identification experiment subjects are asked to assign each stimulus to a phoneme category (forced choice test). In the case of discrimination experiments subjects are asked to determine whether two stimuli are identical or not. The discrimination experiment is always designed in a way that both stimuli are never identical but exhibit a consonant (small) distance on the physical scale defining the stimulus continuum.

In this study human subjects are replaced by "virtual subjects" or "virtual listeners". Each virtual listener is realized by an instance of our neurocomputational model. All instances of the model indicate the same organization as is given in Fig. 1 but they are trained using (i) a different (random) initialization of link weights and (ii) different (random) ordering of the training items during learning. After training or learning these different random orderings lead to qualitatively the same organization of the phonetic map and the phonetic to sensory, phonetic to motor plan, and phonetic to phonemic mapping. But it can be seen that the phonetic map and the related mappings indicate individual differences from model instance to model instance. Examples for phonetic maps of different model instances are given in Fig. 2 (for training details see Kröger et al. in press and Kröger et al. accepted).

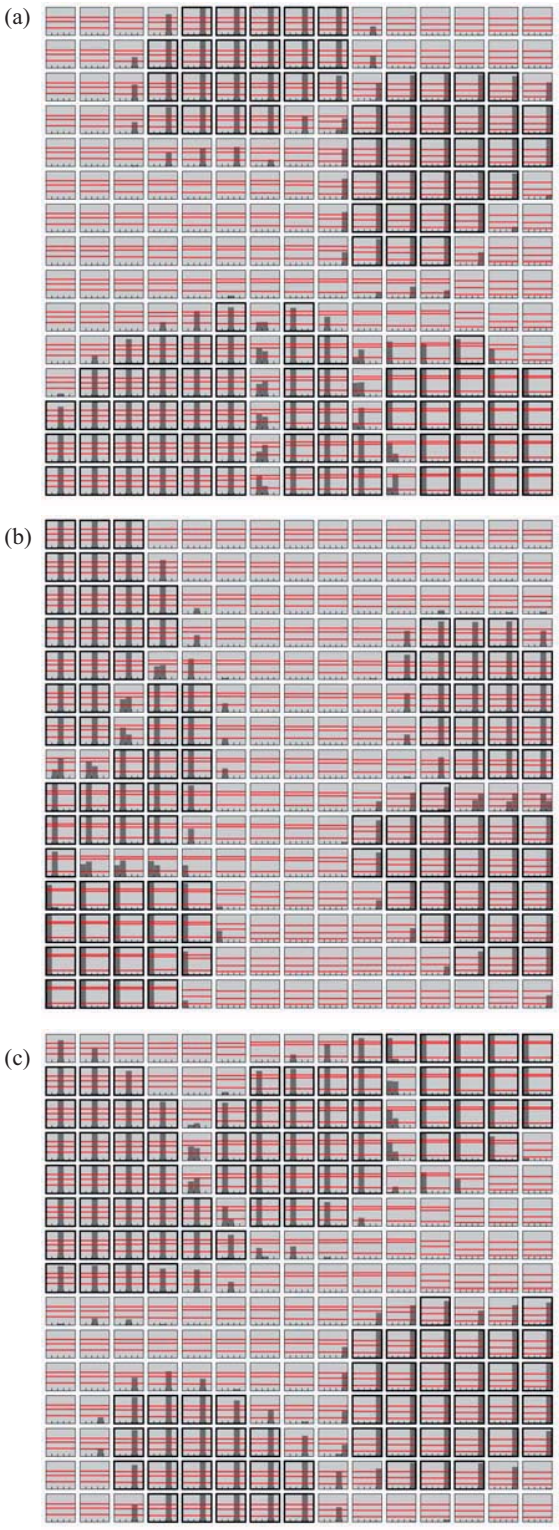

Figure 2: *Self-organizing phonetic map (15x15 neurons) for vowels for three different instances of the neurocomputational model. Bars from left to right within each neuron square: phonemic link weight values for /i/, /e/, /a/, /o/, and /u/. Horizontal lines: bark scaled values of the first three formants, i.e. auditory link weight values. Outlined boxes indicate phoneme link weight values above 80% for a phoneme and thus indicate neurons, representing different realizations of a phoneme. Neurons, realizing the same phoneme, form clusters within the phonetic map.*

4 Modeling Identification and Discrimination

While a speech sound (e.g. a vowel) is produced by activating the appropriate neuron in the phonemic map and subsequently a neuron of the phonetic map (see Fig 2 and Kröger et al. accepted) speech sound *identification* is done by looking for the most activated neuron within the phonemic map activated by the neural auditory pathway. The sensory state of the speech sound under identification activates a neuron within the sensory map and subsequently in the phonetic and in the phonemic map (dorsal stream of speech perception; the ventral stream deals with more complex speech items like words). It is assumed in our model that the ability for sound *discrimination* is proportional to the distance of activation for both stimuli on the level of the phonetic map. Thus discrimination of two speech sounds is simple if the distance between the neurons mainly activated by both sound is large. Discrimination becomes more and more difficult the more the distance between both speech items decreases on the level of the phonetic map (Kröger et al. accepted).

5 Results

The results of our preliminary perception experiments performed by using our neurocomputational model are in agreement with the results of the classical identification and discrimination experiments performed by humans. Identification scores exhibit typical phoneme regions and phoneme boundaries. Measured discrimination (i.e. discrimination scores resulting from discrimination experiments) is higher than calculated discrimination (i.e. calculated discrimination on the basis of individual identification scores) in the case of vowels and comparable to measured discrimination in the case of consonants indicating that consonant perception is more categorical than vowel perception (Kröger et al. accepted).

6 Acknowledgments

This work was supported in part by the German Research Council Grand Nr. KR 1439/13-1.

References

Birkholz P, Jackel D, Kröger BJ (2006) Construction and control of a three-dimensional vocal tract model. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2006, Toulouse, France) pp. 873-876

Birkholz P, Kröger BJ (2006) Vocal tract model adaptation using magnetic resonance imaging. Proceedings of the 7th International Seminar on Speech Production (Belo Horizonte, Brazil) pp. 493-500

Birkholz P, Kröger BJ (2007) Simulation of vocal tract growth for articulatory speech synthesis. Proceedings of the 16th International Congress of Phonetic Sciences (Saarbrücken, Germany) pp. 377-380

Guenther FH (2006) Cortical interaction underlying the production of speech sounds. *Journal of Communication Disorders* 39: 350-365

Guenther FH, Ghosh SS, Tourville JA (2006) Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language* 96: 280-301

Hickok G, Poeppel D (2007) Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences* 4: 131-138

Kröger BJ, Birkholz P, Kannampuzha J, Neuschaefer-Rube C (2006a) Learning to associate speech-like sensory and motor states during babbling. Proceedings of the 7th International Seminar on Speech Production. Belo Horizonte, Brazil, pp. 67-74

Kröger BJ, Birkholz P, Kannampuzha J, Neuschaefer-Rube C (2006b) Modeling sensory-to-motor mappings using neural nets and a 3D articulatory speech synthesizer. Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech 2006 – ICSLP, Pittsburgh, Pennsylvania) pp. 565-568

Kröger BJ, Birkholz P (2007) 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 Communication Behaviours, LNAI 4775 (Springer Verlag, Berlin, Heidelberg) pp. 174-189, http://dx.doi.org/10.1007/978-3-540-76442-7_16

Kröger BJ, Kannampuzha J, Neuschaefer-Rube C (accepted) Towards a neurocomputational model of speech production and perception. Speech Communication, http://dx.doi.org/10.1016/j.specom.2008.08.002

Kröger BJ, Kannampuzha J, Lowit A, Neuschaefer-Rube C (in press) 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. (Peter Lang, Berlin) pp. 59-90